DUAL-CHANNEL ROBUST GROUP-RELATIVE POLICY OPTIMIZATION VIA ADVANTAGE AND SEQUENCE-WEIGHT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Group-relative policy optimization relies on reward-derived advantages and sequence-level likelihood weights, both of which can be sensitive to localized outliers. Extreme rewards can collapse the contrast among clean responses after group normalization, while token-level log-ratio perturbations can alter sequence weights and clipping decisions. We introduce RoVR-GSPO, a dual-channel robust optimizer that addresses these failure modes separately. Its reward channel combines robust reference estimation with bounded residual credit, while its ratio channel uses differentiable SoftRoVR aggregation to construct robust sequence weights. We provide stability and efficiency analyses for both channels. Experiments on mathematical reasoning, long-context summarization, and tool-call annotation show consistent improvements over GSPO, while controlled perturbation studies demonstrate stronger robustness to reward contamination and token-ratio anomalies.

1 INTRODUCTION

Reinforcement learning is central to language-model post-training (Ouyang et al., 2022; Lightman et al., 2023; Xu et al., 2025). PPO uses a learned value function (Schulman et al., 2017), whereas GRPO removes the critic by comparing several responses sampled for the same prompt (Shao et al., 2024). DAPO and GSPO retain this group-relative structure while changing training and likelihood-correction choices (Yu et al., 2025; Zheng et al., 2025). These methods reduce the need for a learned critic, but they still depend on small-sample statistics: one reward reference determines all response advantages, and a sequence-level likelihood weight summarizes many token log-ratios.

The reward-side failure is more specific than the statement that rewards are heavy-tailed. If one reward is replaced by $R_j + \Delta$, the mean moves by Δ/G while the RMS grows with $|\Delta|$. The standardized advantages stay bounded, but as $|\Delta| \to \infty$ the anomalous response approaches $\operatorname{sign}(\Delta)\sqrt{G-1}$ and all clean responses approach $-\operatorname{sign}(\Delta)/\sqrt{G-1}$. Their pairwise advantage differences therefore vanish. This separates two roles that are often conflated: a robust reference limits shared displacement, whereas a bounded residual map protects focal leverage and the scale used to normalize clean responses.

We use this separation as the design principle for RoVR-GSPO. On the reward channel, RoVR computes a robust reference and then applies a bounded odd map and mapped RMS, yielding $\widehat{A}^{\mathrm{credit}}$. On the ratio channel, SoftRoVR applies differentiable robust aggregation to token log-ratios and supplies sequence weights to the clipped GSPO objective. The method therefore exposes the reference, credit, and sequence-weight operations as distinct objects that can be analyzed and evaluated separately. Our contributions are:

- We identify two distinct update sensitivities: reward perturbations can erase clean-response advantage contrast, while localized token log-ratio perturbations can change sequence weights and clipping branches.
- We introduce RoVR-GSPO, which combines robust reference estimation with bounded credit on the reward channel and a differentiable robust sequence surrogate on the ratio channel. This separation makes the role of each operation explicit.

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

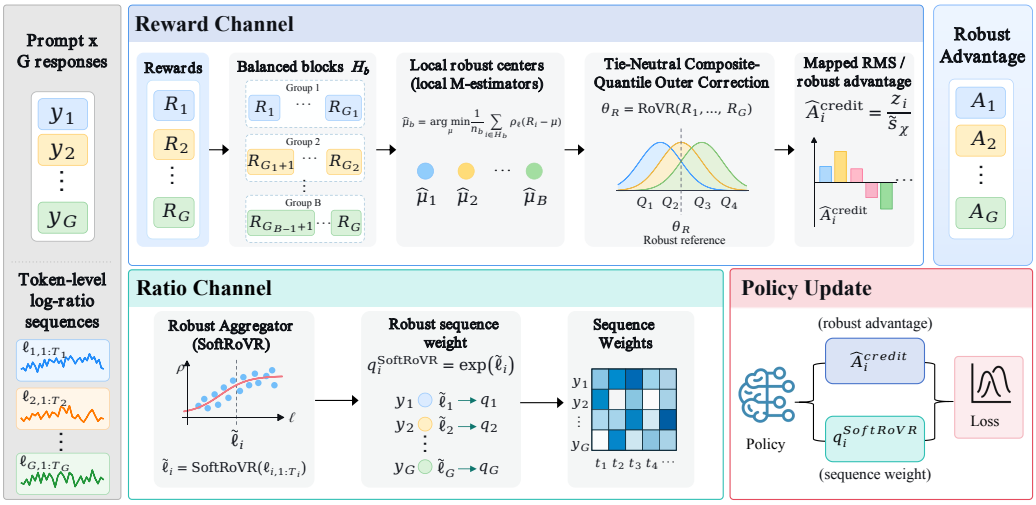

Figure 1: Overview of RoVR-GSPO. The reward channel first estimates a robust reference, then applies the bounded residual map and mapped RMS to form $\widehat{A}^{\mathrm{credit}}$. The ratio channel converts token-level log-ratio sequences into differentiable robust sequence weights via SoftRoVR. These two signals jointly determine the policy update.

- We derive conditional reference, credit-stability, outer-efficiency, and clipping-stability results, and connect them to downstream comparisons, channel placement, and paired perturbation studies.

Here "variance reduction" refers to the outer reference-estimation factor relative to median aggregation. SoftRoVR is a sequence-weight surrogate whose robustness properties are studied directly; it is not treated as an exact importance ratio.

2 Background: From Group Statistics to Policy Updates

RoVR-GSPO is motivated by two distinct sensitivities: reward perturbations can collapse clean-response advantage contrast, while token log-ratio bursts can alter sequence weights and clipping branches. We analyze these effects below.

2.1 Mean–RMS group-relative advantages

For a prompt x, let the old policy sample responses $\{y_i\}_{i=1}^G$ with rewards $\{R_i\}_{i=1}^G$. GRPO constructs

$$\bar{R} = \frac{1}{G}\sum_{j=1}^{G} R_j, \qquad s_R^2 = \frac{1}{G}\sum_{j=1}^{G}(R_j - \bar{R})^2 + \varepsilon_s, \qquad \widehat{A}_i^{\mathrm{GRPO}} = \frac{R_i - \bar{R}}{s_R}, \tag{1}$$

where $\varepsilon_s \geq 0$. The next proposition shows that self-normalization masks anomaly magnitude but broadcasts its effect across the group.

Proposition 2.1. *Let $G \geq 2$. Fix $R_1, \ldots, R_G$, replace R_j by $R_j + \Delta$, and consider values of Δ for which $s_\Delta > 0$ (in particular, all Δ when $\varepsilon_s > 0$). Let $\bar{R}_\Delta, s_\Delta, A_{i,\Delta}$ denote the resulting quantities. Then*

$$\bar{R}_\Delta = \bar{R} + \frac{\Delta}{G}, \qquad s_\Delta^2 = s_R^2 + \frac{2\Delta(R_j - \bar{R})}{G} + \frac{G-1}{G^2}\Delta^2. \tag{2}$$

For every finite group, $\sum_i A_{i,\Delta} = 0$, $\sum_i A_{i,\Delta}^2 \leq G$, and $\max_i |A_{i,\Delta}| \leq \sqrt{G-1}$. Moreover,

$$A_{j,\Delta} \longrightarrow \mathrm{sign}(\Delta)\sqrt{G-1}, \qquad A_{i,\Delta} \longrightarrow -\frac{\mathrm{sign}(\Delta)}{\sqrt{G-1}} \quad (i \neq j) \tag{3}$$

as $|\Delta| \to \infty$. Consequently, when $G \geq 3$, any two clean responses $i, k \neq j$ satisfy $A_{i,\Delta} - A_{k,\Delta} = (R_i - R_k)/s_\Delta \to 0$.

For a clean index set H with $|H| \geq 2$, define the retained pairwise advantage contrast $\mathcal{C}_H(A) = \{2\sum_{i<k,\, i,k\in H}(A_i - A_k)^2/|H|(|H|-1)\}^{1/2}$. Equation (3) gives $\mathcal{C}_H(A_\Delta) \to 0$ when H contains the fixed clean responses. This quantity is measurable in saved batches and is therefore a primary diagnostic rather than an informal interpretation.

Proposition 2.2. *Let a location estimate m_Δ remain bounded when one reward is $R_j + \Delta$ and all other rewards are fixed. Form raw-residual credits $C_{i,\Delta} = (R_i^{(\Delta)} - m_\Delta)/d_\Delta$ with $d_\Delta > 0$. If $d_\Delta = O(1)$, then $|C_{j,\Delta}| \to \infty$. If $C_{j,\Delta} = O(1)$, then $d_\Delta = \Omega(|\Delta|)$ and $C_{i,\Delta} \to 0$ for every clean $i \neq j$. In particular, an RMS about m_Δ gives $d_\Delta \sim |\Delta|/\sqrt{G}$, $C_{j,\Delta} \to \operatorname{sign}(\Delta)\sqrt{G}$, and clean-advantage erasure.*

For raw-residual credits with a shared scale, protecting both focal leverage and clean-response contrast therefore requires an additional residual transformation. We use the bounded map in Section 3.2, which also makes its clean-signal fidelity explicit.

2.2 Sequence weights and clipping asymmetry

For response length T_i, define token log ratios $\ell_{i,t}(\theta) = \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}) - \log \pi_{\text{old}}(y_{i,t} \mid x, y_{i,<t})$. GSPO uses

$$q_i^{\text{GSPO}}(\theta) = \exp\left(\frac{1}{T_i}\sum_{t=1}^{T_i} \ell_{i,t}(\theta)\right). \tag{4}$$

With $l = 1 - \epsilon_{\text{lo}} > 0$ and $u = 1 + \epsilon_{\text{hi}}$, the per-sequence clipped term is

$$g(q, A) = \min\{qA, \text{clip}(q, l, u)A\}, \qquad \mathcal{L}_{\text{GSPO}}(\theta) = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} g(q_i(\theta), \widehat{A}_i)\right]. \tag{5}$$

Writing $q = e^m$, away from $q \in \{l, u\}$,

$$\frac{\partial g(e^m, A)}{\partial m} = \begin{cases} e^m A, & A > 0,\ e^m < u, \\ 0, & A > 0,\ e^m > u, \\ 0, & A < 0,\ e^m < l, \\ e^m A, & A < 0,\ e^m > l. \end{cases} \tag{6}$$

Therefore an upper-ratio excursion is clipped for positive advantage but remains on the qA branch for negative credit. If a contiguous token set $\mathcal{C}$ receives perturbations δ_t, then

$$\frac{q'}{q} = \exp\left(\frac{1}{T}\sum_{t\in\mathcal{C}} \delta_t\right), \tag{7}$$

If $q < u$ initially, the perturbation crosses the upper boundary whenever $T^{-1}\sum_{t\in\mathcal{C}} \delta_t > \log(u/q)$. If $q \geq u$ already, the same upward perturbation magnifies the active negative-advantage branch. Thus localized token perturbations can affect both the sequence weight and the active optimization branch.

3 Dual-Channel Robust Estimation for Group-Relative Policy Optimization

RoVR-GSPO robustifies the two compact statistics that drive a GSPO update. For prompt-level rewards $R_{1:G}$, the reward channel separates reference estimation from credit shaping: it computes a hard RoVR reference and converts centered residuals into the bounded normalized advantage $\widehat{A}^{\text{credit}}$. For the token log-ratios $\ell_{i,1:T_i}$ of each response, the ratio channel applies SoftRoVR to obtain a differentiable sequence-weight surrogate. The complete update is therefore $(R_{1:G},\, \ell_{i,1:T_i}) \longrightarrow (\widehat{A}_i^{\text{credit}},\, q_i^{\text{SoftRoVR}}) \longrightarrow \mathcal{L}_{\text{GSPO}}$. Building on the composite-quantile aggregation principle used in VRMOM (Tu et al., 2021), RoVR uses bounded-score local centers and an explicit credit map so that reference protection and clean-contrast protection remain distinguishable. We reserve ***RoVR*** for the hard reward-reference estimator, ***SoftRoVR*** for the differentiable ratio-side aggregator, and ***RoVR-GSPO*** for the complete dual-channel optimizer. The single-channel variants serve as placement ablations in Section 5.3.

3.1 RoVR ROBUST REFERENCE ESTIMATION

The main method fixes Barron's shape to $\alpha = 1$ (Barron, 2019). Its loss and score are

$$\rho_c(u) = \sqrt{1 + (u/c)^2} - 1, \qquad \psi_c(u) = \frac{u}{c^2}\left(1 + \frac{u^2}{c^2}\right)^{-1/2}, \qquad \psi_c'(u) = \frac{1}{c^2}\left(1 + \frac{u^2}{c^2}\right)^{-3/2}. \tag{8}$$

The empirical objective is strictly convex, its score equation has a unique root, and $M_c := \sup_u |\psi_c(u)| = 1/c$. We fix the convex shape $\alpha = 1$ throughout the main experiments and partition N scalar observations into $B \geq 1$ balanced blocks $\{\mathcal{H}_b\}_{b=1}^{B}$ of sizes n_b, where $\sum_b n_b = N$ and $|n_b - n_{b'}| \leq 1$. The local robust center is

$$\widetilde{\mu}_b = \arg\min_{u \in \mathbb{R}} \frac{1}{n_b} \sum_{i \in \mathcal{H}_b} \rho_c(x_i - u). \tag{9}$$

Its population target is the M-functional

$$\theta_\rho(P): \qquad \mathbb{E}_P[\psi_c(X - \theta_\rho)] = 0. \tag{10}$$

For a symmetric distribution, θ_ρ equals its center and, when finite, its mean. Under skewness it is an explicit robust reference rather than an undisclosed mean estimator. For each block, define its empirical sandwich scale by

$$\widehat{a}_b = \frac{1}{n_b} \sum_{i \in \mathcal{H}_b} \psi_c'(x_i - \widetilde{\mu}_b), \qquad \widehat{b}_b = \frac{1}{n_b} \sum_{i \in \mathcal{H}_b} \psi_c^2(x_i - \widetilde{\mu}_b), \qquad \widehat{\nu}_b = \frac{\sqrt{\widehat{b}_b}}{\widehat{a}_b \vee a_{\min}}.$$

We pool and cap these local sandwich scales,

$$\widehat{\nu} = \mathrm{clip}(\mathrm{median}_b\, \widehat{\nu}_b, \nu_{\min}, \nu_{\max}). \tag{11}$$

For blocks of only two or three observations, a blockwise sandwich estimate can be unstable; the implementation must then use an independent pilot or a lagged running scale and report that choice. The caps are numerical safeguards, not quantities known from theory.

The outer correction is a tie-neutral composite-quantile update. For even B, the initializer is the midpoint of the two central local centers; for odd B, it is the central local center. Let $\widehat{\mu}_0 = \mathrm{median}_b\, \widetilde{\mu}_b$ and define

$$\tau_k = \frac{k}{K+1}, \qquad \Delta_k = \Phi^{-1}(\tau_k), \qquad D_K = \sum_{k=1}^{K} \phi(\Delta_k), \qquad W_B = \sum_{b=1}^{B} \sqrt{n_b}.$$

Discrete rewards require a tie convention. We use the mid-indicator

$$J_0(u) = \mathbf{1}\{u < 0\} + \frac{1}{2}\mathbf{1}\{u = 0\}. \tag{12}$$

The RoVR robust reference is

$$\widehat{\theta}_{\mathrm{RoVR}} = \widehat{\mu}_0 - \frac{\widehat{\nu}}{D_K W_B} \sum_{b=1}^{B} \sum_{k=1}^{K} \left[J_0\left(\widetilde{\mu}_b - \widehat{\mu}_0 - \frac{\widehat{\nu}\Delta_k}{\sqrt{n_b}}\right) - \tau_k \right]. \tag{13}$$

For equal $n_b = n$, (13) reduces to the usual prefactor $\widehat{\nu}/(B\sqrt{n}D_K)$. The denominator W_B is the Newton derivative implied by heteroscedastic normal block centers. Its associated effective sample size is

$$N_{\mathrm{eff}} = \frac{W_B^2}{B} \leq N, \tag{14}$$

which equals N for equal blocks. For balanced blocks, $N_{\mathrm{eff}}/N \to 1$ when $n_{\min} \to \infty$; (14) records the exact finite-block loss otherwise.

Proposition 3.1. *For every $a \in \mathbb{R}$, the hard RoVR map is translation equivariant and constant preserving:*

$$\mathrm{RoVR}(x_1 + a, \ldots, x_N + a) = \mathrm{RoVR}(x_1, \ldots, x_N) + a, \qquad \mathrm{RoVR}(a, \ldots, a) = a.$$

When $B = 1$, it equals the single convex M-center in (9)*. These statements hold for even or odd K and for discrete inputs.*

The half weight at a tie is essential. With the conventional "$\le$" indicator, odd K gives a spurious correction of $-\widehat{\nu}B/(2D_KW_B)$ on constant data. The logistic relaxation below takes value $1/2$ at zero, so (12) also aligns the hard and smooth estimators. After the local roots are computed, direct evaluation costs $O(BK)$; the equal-block version admits the same rank-count simplification as VRMOM (Tu et al., 2021).

Table 6 summarizes the reference estimators. Relative to MOM or Robust MOM, RoVR replaces the outer median by a symmetric composite-quantile correction. Under the clean local normal-score model, this changes the outer variance factor from $\pi/2$ to V_K, with $V_K \to \pi/3$ as the quantile grid is refined. The complete variance combines this factor with the local sandwich variance. RoVR shares the outer-efficiency principle of VRMOM and adds bounded-score local centers, tie-neutral handling of discrete inputs, and balanced unequal blocks.

3.2 Reward-channel advantage estimation

The robust reference controls the shared reward baseline, while the credit map controls the focal residual and the scale used for normalization. RoVR therefore uses the following bounded odd residual map when forming reward advantages:

$$\chi_\kappa(u) = \frac{u}{\sqrt{1+(u/\kappa)^2}} = \kappa^2\psi_\kappa(u), \qquad |\chi_\kappa(u)| \le \kappa, \qquad 0 < \chi'_\kappa(u) \le 1. \tag{15}$$

It preserves sign and local slope while saturating smoothly. With $\widehat{\theta}_R = \mathrm{RoVR}(R_1, \dots, R_G)$, define

$$\widehat{s}_\chi = \{s_{\min}^2 + \frac{1}{G}\sum_{j=1}^{G} \chi_\kappa^2(R_j - \widehat{\theta}_R)\}^{1/2}, \quad \widehat{A}_i^{\mathrm{credit}} = \frac{\chi_\kappa(R_i - \widehat{\theta}_R)}{\widehat{s}_\chi}. \tag{16}$$

The main bounded advantage is deterministically bounded,

$$|\widehat{A}_i^{\mathrm{credit}}| \le \min\left\{\frac{\kappa}{s_{\min}}, \sqrt{G}\right\}. \tag{17}$$

The map is close to the raw residual on clean moderate rewards:

$$|\chi_\kappa(u) - u| \le \frac{|u|^3}{2\kappa^2}. \tag{18}$$

Thus κ controls the fidelity–robustness trade-off. For bounded rule rewards, choosing κ above the reward range makes the distortion small. All reward-side statistics are detached before they enter the policy loss. The center-only comparison below retains the raw residual and raw RMS while replacing only the arithmetic mean by $\widehat{\theta}_R$: $\widehat{A}_i^{\mathrm{center}} = \frac{R_i - \widehat{\theta}_R}{\{\varepsilon_s + G^{-1}\sum_j (R_j - \widehat{\theta}_R)^2\}^{1/2}}$. This isolates the effect of the RoVR reference from the bounded residual map and mapped scale. For an action-independent control-variate reference, use a linear leave-one-out numerator and a positive leave-one-out scale S_{-i}, both computed without R_i:

$$\widehat{A}_i^{\mathrm{LOO}} = \frac{R_i - \widehat{\theta}_{R,-i}}{S_{-i}}, \qquad \widehat{\theta}_{R,-i} = \mathrm{RoVR}(R_1, \dots, R_{i-1}, R_{i+1}, \dots, R_G). \tag{19}$$

A bounded residual leave-one-out version is a shaped robust surrogate and does not inherit the linear baseline cancellation. Its deviation is quantified in Theorem F.17.

Table 7 separates the roles of the three advantage constructions: $\widehat{A}^{\mathrm{center}}$ isolates reference substitution, $\widehat{A}^{\mathrm{LOO}}$ provides a linear action-independent baseline for analysis, and $\widehat{A}^{\mathrm{credit}}$ is the normalized bounded advantage used by RoVR-GSPO.

3.3 SoftRoVR sequence-weight estimation for GSPO

The reward statistic uses the hard RoVR operator in (13). The ratio-side differentiable surrogate is called SoftRoVR. For token log-ratios, define $H_\gamma(u) = \frac{1}{1+\exp(u/\gamma)}, \gamma > 0$, which converges pointwise to J_0, including at ties. A differentiable outer initializer can be obtained from

$$\mathrm{smed}_\eta(v_1, \dots, v_B) = \arg\min_u \frac{1}{B}\sum_{b=1}^{B} \rho_\eta(v_b - u), \qquad \eta > 0. \tag{20}$$

Let $\mathcal{S}_\gamma$ retain the hard block centers, midpoint-median initializer, and pooled scale, replacing only J_0 by H_γ. This auxiliary map isolates indicator-smoothing error. The implemented surrogate $\mathrm{SoftRoVR}_{\gamma,\eta,S}$ additionally uses (20) for both median operations and unrolls S safeguarded solver iterations. At fixed $\eta > 0$, the smooth objective specifies a unique initializer, including for even B. We analyze this finite-temperature surrogate directly; comparison with hard RoVR separates indicator, initializer, scale, and solver errors. The zero-temperature indicator limit agrees with J_0 at ties.

For response i, the robust sequence weight is

$$q_i^{\mathrm{SoftRoVR}}(\theta) = \exp\big[\mathrm{SoftRoVR}_{\gamma,\eta,S}\big(\ell_{i,1}(\theta), \ldots, \ell_{i,T_i}(\theta)\big)\big]. \tag{21}$$

RoVR-GSPO combines $\widehat{A}^{\mathrm{credit}}$ from (16) with (21). Disabling one channel gives a placement ablation. Robust aggregation defines a sequence-weight surrogate with a different likelihood functional; its calibration and local stability are analyzed in Section 4.

3.4 Resolution, failure geometry, and fallback

The block design determines the contamination budget. To tolerate q arbitrary blocks while retaining a strict majority of usable blocks, one needs $B \geq 2q + 1$. To leave a strict clean majority after at most s replacements inside each usable block, one needs $n_{\min} \geq 2s + 1$. Hence a non-degenerate worst-case two-level design requires

$$N \geq (2q + 1)(2s + 1). \tag{22}$$

This condition describes the resolution needed for simultaneous worst-case protection at both levels. In particular, $G < 9$ permits budgets with $q = 0$ or $s = 0$, but cannot support both $q \geq 1$ and $s \geq 1$. RoVR itself is defined for every balanced partition; the budget determines the applicable guarantee. The $B = 1$ case provides a global-M boundary case. Contiguous token blocks encode a burst geometry, while random partitions support the independently assigned contamination model.

4 Theoretical Guarantees

We analyze the three quantities modified by RoVR-GSPO: the reward reference, bounded response credit, and sequence weight. The results cover finite-group stability, clean outer efficiency, calibration, and clipping margins; auxiliary statements and proofs appear in Sections F to I.

Consider B declared blocks, with a fixed set of $q < r_B := \lceil B/2 \rceil$ arbitrary blocks and honest complement $\mathcal{H}$. Each $b \in \mathcal{H}$ contains n_b independent clean draws followed by at most s_b adaptive replacements. If $g(t) = \mathbb{E}\psi_c(X - t)$ satisfies $g(\theta + u) \leq -au$ and $g(\theta - u) \geq au$ for $0 \leq u \leq r_0$, define, for $0 < t \leq r_0$,

$$H = B - q, \quad \eta_q = (r_B - q)/(B - q), \quad p_b(t) = \exp\left[-\frac{n_b}{2M_c^2}\left(at - \frac{2s_b M_c}{n_b}\right)_+^2\right],$$
$$\bar{p}_H(t) = H^{-1}\sum_{b\in\mathcal{H}} p_b(t), \quad \beta_H(t) = \min\{1, 2e^{-2H(\eta_q - \bar{p}_H(t))^2}\},$$
$$\widehat{C}_B = |\widehat{\theta}_{\mathrm{RoVR}} - \widehat{\mu}_0|, \qquad C_B = \nu_{\max} BK/(2D_K W_B).$$

Only the population score separation is distributional; replacement values may be unbounded.

Theorem 4.1 (Finite-group reference and credit stability). *If $\bar{p}_H(t) < \eta_q$, then $\widehat{C}_B \leq C_B$ and*

$$\mathbb{P}\Big(|\widehat{\theta}_{\mathrm{RoVR}} - \theta| > t + \widehat{C}_B\Big) \leq \beta_H(t), \qquad \mathbb{P}\Big(|\widehat{\theta}_{\mathrm{RoVR}} - \theta| > t + C_B\Big) \leq \beta_H(t). \tag{23}$$

Suppose additionally that the observed reward group and its clean precursor differ in at most r coordinates. Let $A_i^{\theta,\star}$ be the bounded credit of the precursor computed with population reference θ, and set $e_\theta(t) = t + C_B$, $e_{R,i} = |R_i - R_i^\star|$, and $\bar{e}_s(t) = \{r\kappa^2/G + 2\kappa e_\theta(t)\}/(2s_{\min})$. With probability at least $1 - \beta_H(t)$, simultaneously for all responses,

$$|\widehat{A}_i^{\mathrm{credit}} - A_i^{\theta,\star}| \leq \frac{\min\{e_{R,i} + e_\theta(t), 2\kappa\}}{s_{\min}} + \frac{\kappa \bar{e}_s(t)}{s_{\min}^2}. \tag{24}$$

The credit bound includes the perturbed focal response. Its dependence on $e_{R,i}$ is capped by 2κ, yielding magnitude-independent control under the stated geometry and population separation. Smaller κ limits leverage more strongly, while larger κ preserves moderate clean residuals more closely. Equation (17) gives the corresponding uniform advantage bound.

To characterize clean reference-estimation efficiency, write $\nu^2 = \mathbb{E}\psi_c^2(X-\theta_\rho)/\{\mathbb{E}\psi_c'(X-\theta_\rho)\}^2$. For equal blocks of size n, assume independent honest block centers, $q_B = o(\sqrt{B})$ arbitrary blocks, a $B^{-1/2}$-accurate initializer, an $o_{\mathbb{P}}(B^{-1/4})$ scale error, and the uniform local normal-score and equicontinuity conditions in Theorem F.11.

Theorem 4.2 (Clean outer efficiency). *For fixed K, under the preceding transfer conditions,*

$$\sqrt{Bn}(\widehat{\theta}_{\mathrm{RoVR}} - \theta_\rho) \xrightarrow{d} \mathcal{N}(0, \nu^2 V_K), \qquad V_K := \frac{\sum_{k,\ell}[\min(\tau_k,\tau_\ell) - \tau_k\tau_\ell]}{\{\sum_k \phi(\Delta_k)\}^2} \longrightarrow \frac{\pi}{3}. \tag{25}$$

The limit first sends $(B, n) \to \infty$ at fixed K and then refines the quantile grid.

The outer factor improves on the median value $\pi/2$ under the stated local normal-score model. Total reference-estimation variance combines this factor with the local sandwich factor and, for unequal blocks, the effective sample size in (14).

The ratio channel instead needs calibration and a clipping-margin statement. Let $|\widehat{m} - m^\star| \le e_m$ and $(\widehat{q}, q^\star) = (e^{\widehat{m}}, e^{m^\star})$.

Proposition 4.3 (Calibration and clipping-margin control). *At differentiable inputs, the converged SoftRoVR map, and any fixed number of equivariantly initialized solver steps, obey*

$$\mathrm{SoftRoVR}(\ell + a\mathbf{1}) = \mathrm{SoftRoVR}(\ell) + a, \quad \sum_t \partial\,\mathrm{SoftRoVR}(\ell)/\partial\ell_t = 1, \quad e^{-e_m} \le \widehat{q}/q^\star \le e^{e_m}.$$

A clipping branch can change only if $\mathrm{dist}(m^\star, \{\log l, \log u\}) \le e_m$.

The proposition links log-weight error to the distance from a clipping boundary. Theorem F.15 extends this relation to same-branch objective and gradient perturbations under explicit Jacobian conditions.

5 Experiments

We evaluate RoVR-GSPO on mathematical reasoning, long-context summarization, and tool-call annotation, followed by channel ablations and controlled perturbation analyses. Detailed protocols and supplementary diagnostics appear in Sections B and C.

5.1 Experimental setup

We evaluate three downstream tasks. For mathematical reasoning, Qwen3-4B-Base and OLMo-3-7B-Instruct are trained on the 7,500-example MATH split (Hendrycks et al., 2021) and evaluated on MATH500, AIME2025, AMC23, Gaokao2023-Math-En (Zhang et al., 2024), and Minerva-Math (Lewkowycz et al., 2022), using either a bounded answer verifier or InternLM2-1.8B-Reward. For GovReport summarization (Huang et al., 2021), training uses LongDocFACTScore (Bishop et al., 2024) while evaluation uses ROUGE-1/2/L (Lin, 2004); the evaluation metrics are therefore distinct from the training reward. For offline tool-call annotation, Qwen3-4B-Base is trained on the RLLA-4K split from ToolRL (**?**), and the held-out metric is the training scorer's format-plus-tool-match reward. No tools are executed in this third task.

GSPO (Zheng et al., 2025) is the parent baseline. Initialization, response and optimizer budgets, decoding, and evaluation code are matched within each comparison. RoVR-GSPO combines the full RoVR reward reference, bounded credit, and SoftRoVR sequence weights; channel ablations disable the corresponding component. The mathematical-reasoning main comparison and channel ablation report mean $\pm$ standard deviation over three independent runs; summarization and tool-call annotation are single-run estimates. Section E separately evaluates the reference estimator.

Table 1: Mathematical-reasoning accuracy (%), mean ± standard deviation over three runs. Bold marks the higher mean for each setting.

Model	Dataset	Rule reward		InternLM2 reward	
		GSPO	RoVR-GSPO	GSPO	RoVR-GSPO
Qwen3-4B-Base	MATH500	77.00 ± 0.43	$\mathbf{77.47 \pm 0.57}$	74.53 ± 0.41	$\mathbf{75.80 \pm 0.28}$
	Gaokao2023en	39.74 ± 0.21	$\mathbf{45.37 \pm 0.95}$	38.53 ± 0.32	$\mathbf{42.08 \pm 0.37}$
	MinervaMath	19.36 ± 0.46	$\mathbf{21.69 \pm 1.04}$	15.81 ± 0.30	$\mathbf{16.67 \pm 0.34}$
	AIME2025	14.44 ± 1.57	$\mathbf{15.56 \pm 1.57}$	8.89 ± 1.57	$\mathbf{15.56 \pm 1.57}$
	AMC23	62.50 ± 2.04	$\mathbf{63.33 \pm 1.18}$	56.67 ± 1.18	$\mathbf{64.17 \pm 3.12}$
OLMo-3-7B-Instruct	MATH500	89.80 ± 0.16	$\mathbf{90.73 \pm 0.25}$	88.93 ± 0.74	$\mathbf{89.53 \pm 0.41}$
	Gaokao2023en	73.77 ± 0.21	$\mathbf{74.64 \pm 0.74}$	71.94 ± 0.36	$\mathbf{72.90 \pm 0.25}$
	MinervaMath	28.80 ± 0.76	$\mathbf{29.17 \pm 0.34}$	28.56 ± 0.17	$\mathbf{29.29 \pm 0.63}$
	AIME2025	37.78 ± 1.57	$\mathbf{41.11 \pm 1.57}$	37.78 ± 1.57	$\mathbf{41.11 \pm 3.14}$
	AMC23	86.67 ± 2.36	$\mathbf{87.50 \pm 2.04}$	86.67 ± 2.36	$\mathbf{90.83 \pm 1.18}$

5.2 Downstream results

Mathematical reasoning. RoVR-GSPO achieves a higher mean than matched GSPO in all 20 settings. The five-benchmark average improvements are $2.08\%/3.97\%$ for Qwen3-4B-Base and $1.27\%/1.96\%$ for OLMo-3-7B-Instruct under rule/InternLM2 rewards.

Long-context summarization. For GovReport, both backbones are fine-tuned with LongDoc-FACTScore rewards using eight sampled summaries per input and evaluated with ROUGE-1/2/L.

Table 2: GovReport ROUGE point estimates on a 0–1 scale; bold marks the higher score.

Model	Method	ROUGE-1	ROUGE-2	ROUGE-L
Qwen3-4B-Base	GSPO	0.309	0.142	0.166
	RoVR-GSPO	**0.333**	**0.164**	**0.175**
OLMo-3-7B-Instruct	GSPO	0.552	0.189	0.221
	RoVR-GSPO	**0.553**	**0.195**	**0.229**

RoVR-GSPO improves ROUGE-1/2/L for both backbones: $0.024/0.022/0.009$ for Qwen3-4B-Base and $0.001/0.006/0.008$ for OLMo-3-7B-Instruct.

The validation-reward trajectories in Figure 2 are consistent with the held-out ROUGE results.

Tool-call annotation. On RLLA-4K, using 80 held-out prompts and one greedy response per prompt, the highest observed mean Total reward is 1.924 for GSPO and 2.009 for RoVR-GSPO. Each value is selected from four logged validations. The score measures format and tool matching without executing tools, and the available logs cover only the early portion of the configured training budget.

5.3 Channel placement ablation

We compare GSPO, reward-only RoVR-Adv, ratio-only SoftRoVR, and the full RoVR-GSPO model (Table 3 and Fig. 3).

Across the five benchmarks, the macro-average gains over GSPO are 1.00, 2.20, and 3.97 percentage points for RoVR-Adv, SoftRoVR, and RoVR-GSPO, respectively, based on the three-run means. RoVR-GSPO has the highest mean on all five benchmarks and exceeds SoftRoVR by 1.77 points on average.

The combined variant records the highest late-stage reward, while the robust variants show lower late-stage entropy (Figure 3).

Table 3: Channel-ablation accuracy (%) on mathematical reasoning. Values are reported as mean ± standard deviation over three independent runs; bold marks the highest mean in each column.

Method	MATH500	Gaokao	Minerva	AIME25	AMC23
GSPO	74.53 ± 0.50	38.53 ± 0.40	15.81 ± 0.37	8.89 ± 1.92	56.67 ± 1.44
RoVR-Adv	74.73 ± 0.81	40.61 ± 0.40	15.18 ± 0.23	12.22 ± 1.92	56.67 ± 1.44
SoftRoVR	74.53 ± 0.23	40.87 ± 1.28	16.42 ± 0.34	14.44 ± 1.93	59.17 ± 1.44
RoVR-GSPO	$\mathbf{75.80 \pm 0.35}$	$\mathbf{42.08 \pm 0.45}$	$\mathbf{16.67 \pm 0.42}$	$\mathbf{15.56 \pm 1.93}$	$\mathbf{64.17 \pm 3.82}$

Mechanism checks. Offline perturbations support both mechanisms. Robust centering reduces reference displacement from 1.000σ to 0.060σ, while bounded credit limits scale inflation from $28.194\times$ to $1.911\times$ and raises clean-contrast retention from 0.035 to 0.520. SoftRoVR reduces log-weight displacement by about 26% for strong token spikes and 87% for 20% bursts, with fewer clipping flips under strong spikes.

6 Conclusion

RoVR-GSPO treats reward-derived advantages and sequence-level weights as separate statistical interfaces to group-relative policy optimization. Blockwise M-estimation and tie-neutral composite-quantile aggregation construct the reward reference; bounded credit protects the normalized response contrast, and SoftRoVR aggregates token log-ratios before clipping. The analysis links these design choices to reference robustness, clean outer efficiency, and local update stability.

Across mathematical reasoning and GovReport, RoVR-GSPO consistently improves over matched GSPO baselines, with additional gains on the offline tool-call annotation score. Controlled reward and token perturbations further show that the two channels address complementary sources of update sensitivity. These results support robust advantage and sequence-weight estimation as a practical statistical interface for group-relative language-model post-training.

AI use statement

OpenAI Codex was used to assist with manuscript organization, language editing, LaTeX formatting, and stress-testing the exposition and logical consistency of the motivation and theoretical arguments. The authors are responsible for independently verifying every mathematical claim, proof, citation, experimental datum, and conclusion before submission, and they take responsibility for the final content of the paper.

Ethics statement

This work studies robust advantage and sequence-weight estimation for language-model post-training and introduces no new human-subject study or personally identifiable dataset. Robust aggregation can reduce sensitivity to isolated scoring errors, but it does not guarantee reward-model validity, factual correctness, fairness, or safe deployment. Any deployment should therefore retain independent reward audits, distribution-shift evaluation, and the data-governance requirements of the underlying models and datasets.

Reproducibility statement

The estimator, credit map, token extension, and optimization procedure are specified in Sections 3 and D. The assumptions and complete proofs of the theoretical claims appear in Sections G to I. The experimental protocol, placement ablation, and reporting conventions are specified in Sections C and 5.

References

Jonathan T. Barron. A general and adaptive robust loss function. *CVPR*, 2019.

Jennifer A Bishop, Sophia Ananiadou, and Qianqian Xie. Longdocfactscore: Evaluating the factuality of long document abstractive summarisation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10777–10789, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Thomas Hitchcox and James Richard Forbes. Mind the gap: Norm-aware adaptive robust loss for multivariate least-squares problems. *IEEE Robotics and Automation Letters*, 7(3):7116–7123, 2022.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. In *Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: Human language technologies*, pp. 1419–1436, 2021.

Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pp. 1248–1251. Springer, 2011.

Pierre Humbert, Batiste Le Bars, and Ludovic Minvielle. Robust kernel density estimation with median-of-means principle. In *International Conference on Machine Learning*, pp. 9444–9465. PMLR, 2022.

Kyungmin Jung, Thomas Hitchcox, and James Richard Forbes. An adaptive graduated nonconvexity loss function for robust nonlinear least-squares solutions. *IEEE Transactions on Robotics*, 2024.

Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. Catastrophic goodhart: regularizing rlhf with kl divergence does not mitigate heavy-tailed reward misspecification. *arXiv preprint arXiv:2407.14503*, 2024. doi: 10.48550/arXiv.2407.14503.

Guillaume Lecué, Matthieu Lerasle, and Timlothée Mathieu. Robust classification via mom minimization. *Machine learning*, 109(8):1635–1665, 2020.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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, pp. 27730–27744, 2022.

Tao Ren, Jinyang Jiang, Hui Yang, Wan Tian, and Yijie Peng. RiskPO: Risk-based policy optimization with verifiable reward for LLM post-training. In *NeurIPS 2025 Workshop MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making*, 2025. URL https://openreview.net/forum?id=8hxqmh25ZH.

Helmut Rieder. *Robust Asymptotic Statistics: Volume I*. Springer Science & Business Media, 2012.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jiyuan Tu, Weidong Liu, Xiaojun Mao, and Xi Chen. Variance reduced median-of-means estimator for byzantine-robust distributed inference. *Journal of Machine Learning Research*, 22(84):1–67, 2021.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, et al. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark, 2024. URL `https://arxiv.org/abs/2305.12474`.

Chuheng Zheng, Siyuan Liu, Yibo Zhu, Zhijie Deng, Tao Ji, Kai Yang, Zhenyu Li, Ming Liu, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Related Work

RoVR-GSPO connects two strands of work: robust construction of reward-derived advantages and robust aggregation of the likelihood statistics used by clipped policy objectives. PPO clips likelihood ratios and is commonly paired with a learned value function in RLHF (Schulman et al., 2017; Ouyang et al., 2022); GRPO replaces the critic with within-group reward normalization (Shao et al., 2024); DAPO studies large-scale choices such as asymmetric clipping and dynamic sampling (Yu et al., 2025); and GSPO uses a length-normalized sequence correction (Zheng et al., 2025). RoVR-GSPO keeps the parent clipped objective and contributes a statistical interface that separates reference estimation, bounded credit, and sequence-weight aggregation.

Robust reward and advantage estimation. Learned reward models can exhibit label errors, calibration shifts, and exploitable tails, while outcome verifiers are bounded and often tie-rich. Process supervision supplies denser feedback (Lightman et al., 2023), and heavy-tailed reward misspecification remains a concern even with KL regularization (Kwa et al., 2024). Reward winsorization and post-normalization advantage clipping limit extremes; RoVR separates robust reference estimation from bounded residual normalization so their effects can be analyzed independently. Risk-sensitive objectives intentionally change the optimized reward functional (Ren et al., 2025), whereas our reward channel exposes the robustness–fidelity trade-off through the explicit map χ_κ.

Group-relative normalization also induces prompt-dependent difficulty weighting (Yang et al., 2026). This statistical bias concerns the target of advantage estimation, whereas RoVR characterizes sensitivity to localized reward perturbations and the fidelity of bounded credit. The leave-one-out analysis in Theorem F.16 makes the prompt-reweighting effect explicit.

Robust reference estimation. Medians, trimming, and bounded-score M-estimators provide classical contamination protection (Huber, 2011; Rieder, 2012). A global Huber or pseudo-Huber estimator is the $B = 1$ boundary of our reference construction. MOM methods add an explicit block interface (Lecué et al., 2020; Humbert et al., 2022), while VRMOM reduces the outer median variance through a composite-quantile one-step correction but retains arithmetic local means (Tu et al., 2021). RoVR combines bounded-score local centers with tie-neutral composite-quantile aggregation and then adds a policy-specific bounded credit map, so reference protection and clean-response contrast can be examined separately.

Robust sequence-weight estimation. GSPO summarizes token log-ratios into a sequence-level weight before clipping. SoftRoVR targets this aggregation step directly: it smooths the RoVR construction so gradients can pass through the sequence statistic and then exponentiates the resulting log-weight. Because this robust aggregation changes the likelihood functional, we analyze it as an optimization surrogate and study its calibration and clipping-branch stability rather than treating it as an exact importance ratio.

Barron's family connects several classical robust penalties (Barron, 2019) and supports adaptive and graduated robust estimation (Jung et al., 2024; Hitchcox & Forbes, 2022). RoVR fixes $\alpha = 1$ to obtain a convex pseudo-Huber/Charbonnier-type local loss, then combines its bounded score with tie-neutral composite-quantile aggregation and policy-specific bounded residual normalization. This choice gives a unique local root and the score properties used in our analysis.

B Additional Experimental Results

This appendix reports training dynamics and the complete token-ratio and reward-contamination diagnostics using the sampling units specified in each protocol.

B.1 TRAINING DYNAMICS

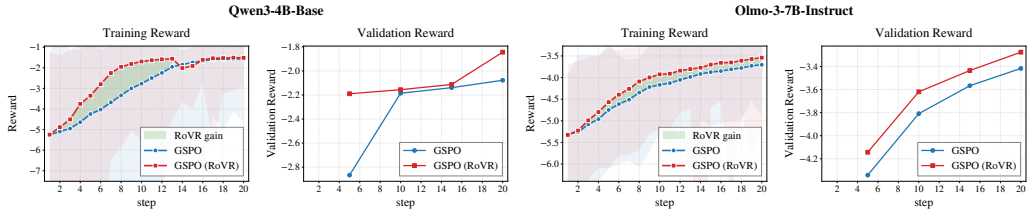

Figure 2: GovReport training dynamics for Qwen3-4B-Base (left pair) and OLMo-3-7B-Instruct (right pair). Each pair shows training reward followed by validation reward. Blue circles denote GSPO, and red squares denote RoVR-GSPO, labeled GSPO (RoVR) in the figure. The green fill visualizes the reward gap.

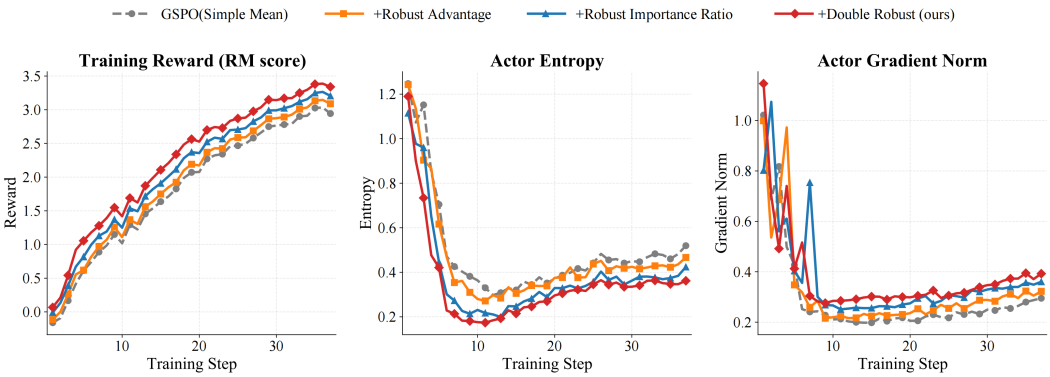

Figure 3: Training dynamics for the channel ablation: reward, actor entropy, and actor gradient norm. Orange, blue, and red denote RoVR-Adv, SoftRoVR, and RoVR-GSPO; dashed gray denotes GSPO.

B.2 LOCALIZED TOKEN-RATIO STRESS TEST

We evaluate aggregation sensitivity on saved Qwen3-4B-Base records from a dual-channel run using $\widehat{A}^{\text{credit}}$ and SoftRoVR. No policy is retrained: GSPO applies arithmetic-mean aggregation to the recorded token log-ratios, while SoftRoVR uses the run's differentiable aggregation. We then measure how localized token-ratio perturbations change sequence log-weights and clipping branches.

The saved snapshot contains 35 complete training steps, each with 192 prompt groups and 16 responses (107,520 responses in total). The partially recorded step 36 contains only 1,408 responses and is excluded from the stress analysis. We select step 18 using a fixed upper-middle rule on the ordered complete steps and use all 3,072 responses, including all eight within-step optimizer-update positions. For response i, let $\boldsymbol{\ell}_i$ be the valid token log-ratios and let σ_i be their population standard deviation. A single-token spike adds $\Delta = \pm c\sigma_i$ at a uniformly sampled valid position, with $c \in \{0, 2, 4, 8, 16\}$. A contiguous burst adds $\pm 8\sigma_i$ to $\lceil fT_i \rceil$ tokens, with $f \in \{0, 0.01, 0.05, 0.10, 0.20\}$. Each response uses five paired random positions across methods and severities. We average positions within a response, then responses within a prompt, and finally give every prompt equal weight. Pointwise intervals are 5,000 prompt-cluster percentile bootstrap replicates.

We measure the absolute log-weight displacement

$$D_q = |m_{\text{corrupt}} - m_{\text{clean}}|$$

and the fraction of responses whose clipping branch changes relative to its own clean branch. Branches are evaluated with the recorded advantage and the actual GSPO clipping interval $[0.9997, 1.0004]$. The zero-scale responses (12.5% of responses in the complete steps) remain in the primary analysis; their perturbation is exactly zero. The paired differences in Table 4 are GSPO minus SoftRoVR, so positive values favor SoftRoVR.

At the strongest single-token spike, SoftRoVR reduces D_q by 26.30% for positive perturbations and 26.09% for negative perturbations; clipping-flip rates decrease from 22.526% to 19.362% and from

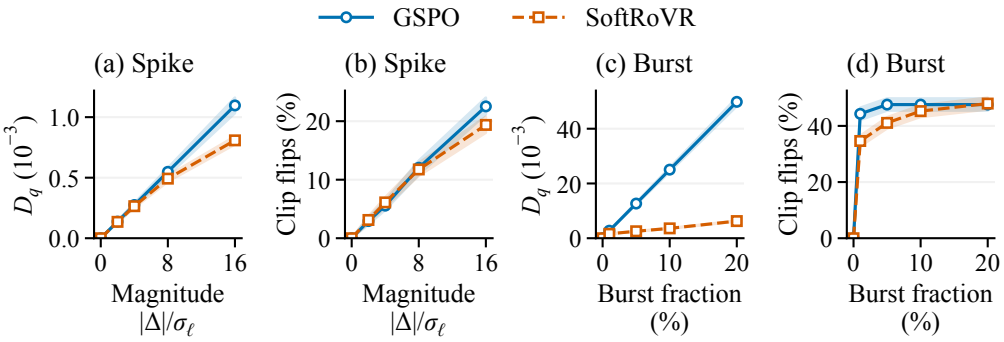

Figure 4: Localized token-ratio stress test at recorded step 18. Curves show means over 192 prompt groups (3,072 responses); shaded regions show pointwise 95% prompt-cluster bootstrap intervals. SoftRoVR attenuates log-weight displacement for both isolated spikes and contiguous bursts. For strong spikes, the paired flip-rate difference favors SoftRoVR; for the positive 20% burst, its interval includes zero.

Table 4: Paired offline token-ratio stress test at step 18 using 192 prompt groups and 3,072 responses. D_q is reported in 10^{-3} units; for Flips, differences are reported in %. Intervals are pointwise 95% prompt-cluster bootstrap intervals.

Perturbation	Sign	Metric	GSPO	SoftRoVR	Difference [95% CI]
Spike $16\sigma_\ell$	+	D_q	1.097	0.808	0.288 [0.256, 0.324]
Spike $16\sigma_\ell$	+	Flips (%)	22.526	19.362	3.164 [2.246, 4.082]
Burst 20%	+	D_q	49.812	6.258	43.554 [42.112, 44.963]
Burst 20%	+	Flips (%)	47.591	47.956	-0.365 [-0.944, 0.221]
Spike $16\sigma_\ell$	−	D_q	1.097	0.811	0.286 [0.254, 0.321]
Spike $16\sigma_\ell$	−	Flips (%)	21.061	17.370	3.691 [2.949, 4.453]
Burst 20%	−	D_q	49.812	6.283	43.529 [42.088, 44.940]
Burst 20%	−	Flips (%)	39.909	39.102	0.807 [0.202, 1.413]

21.061% to 17.370%. For a 20% burst, D_q decreases by 87.44% and 87.39% in the positive and negative directions. The positive-burst flip rate changes from 47.591% to 47.956% with interval $[-0.944\%, 0.221\%]$, while the negative-burst rate changes from 39.909% to 39.102% with interval $[0.202\%, 1.413\%]$. Overall, SoftRoVR reduces aggregate displacement and strong-spike clipping sensitivity, while long-burst clipping effects depend on perturbation direction and training stage.

Section C.3 provides the negative-direction and across-step analyses.

B.3 Robustness to reward contamination

We perturb saved InternLM2 rewards from Qwen3-4B-Base rollouts while keeping responses fixed. From 16,704 complete groups, we select 5,000 evenly spaced groups with $G = 16$ and replace one reward at a time by

$$R'_t = R_t \pm \alpha\sigma, \qquad \alpha \in \{0.5, 1, 2, 4, 8, 16\}, \tag{26}$$

where $\sigma = 1.165$ is the population standard deviation of the saved reward stream. We compare GRPO/GSPO normalization, the RoVR-reference variant $\widehat{A}^{\text{center}}$, and the bounded-credit variant $\widehat{A}^{\text{credit}}$. Both robust variants use the full RoVR reference, including blockwise M-centers with $c = 1$ and the outer composite-quantile correction. The credit map uses $\kappa = 1$ and $s_{\min} = 10^{-3}$. Keeping the reference and block assignment matched isolates the contribution of the bounded residual map.

The baseline divides mean-centered rewards by the sample standard deviation (denominator $G - 1$) plus 10^{-6}. The RoVR-reference and bounded-credit variants use the RMS of their respective residuals with 10^{-6} inside the square root. At $G = 16$ and negligible regularization, sample standard deviation is $\sqrt{16/15}$ times mean-centered RMS. The center-to-credit comparison holds both the RoVR reference and the RMS convention fixed.

We measure reference displacement, scale inflation, clean-response contrast retention $\mathcal{C}_H$, and clean-advantage RMS deviation relative to each estimator's clean outputs. Results use within-group medians, averaged over perturbation signs, followed by medians across prompt groups. Pointwise 95% intervals use 500 group-bootstrap replicates.

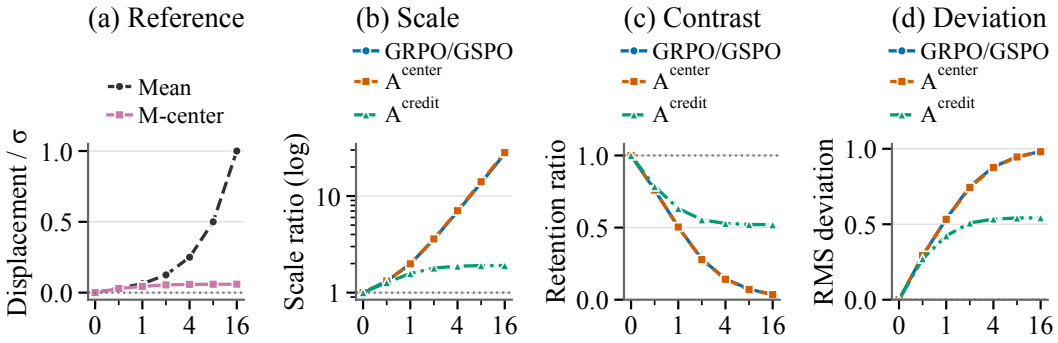

Figure 5: Offline sensitivity to additive reward contamination. One of $G = 16$ rewards is perturbed by $\pm\alpha\sigma$; curves summarize 5,000 prompt groups and shaded regions denote bootstrap 95% intervals. The RoVR reference (labeled M-center) limits location displacement, while bounded-credit normalization additionally limits scale inflation and preserves more clean-response advantage contrast.

Table 5: Estimator sensitivity at the strongest offline stress level ($\alpha = 16$). Scale inflation and clean contrast retention are ideal at 1; advantage RMS deviation is ideal at 0. Values are medians across the same 5,000 complete prompt groups used in Figure 5.

Estimator	Scale inflation	Clean contrast retention	Advantage RMS deviation
GRPO/GSPO	27.400	0.036	0.982
Robust center	28.194	0.035	0.980
A^{credit}	**1.911**	**0.520**	**0.543**

Figure 5 and Table 5 separate protection of the reference from protection of the complete advantage. At 16σ, the arithmetic reference shifts by 1.000σ, whereas the robust reference shifts by only 0.060σ. With the reference held fixed, the bounded-credit map further limits scale inflation to $1.911\times$, retains 0.520 of the clean contrast, and reduces clean-advantage RMS deviation to 0.543; the center-only variant retains only 0.035 of the clean contrast and inflates the scale by $28.194\times$, close to standard normalization (0.036 and $27.400\times$).

At the tested 4σ shift, bounded-credit normalization yields $1.881\times$ scale inflation and 0.529 contrast retention, compared with $6.918\times$ and 0.145 for standard normalization. These fixed-batch results match the mechanism in Theorem 2.2: robust centering controls shared reference displacement, while the bounded residual map additionally controls the influence of the perturbed response on scale and advantage. Complementary simulations in Section E isolate the clean outer-efficiency result and compare reference estimators under heavy tails and matched contamination geometries.

C Experimental Protocol and Hyperparameters

This appendix specifies the training configurations, RoVR components, and statistical reporting conventions for Section 5. The channel definitions follow Section 5.3; the offline studies evaluate the corresponding update statistics on saved responses.

C.1 Training configuration

The mathematical-reasoning experiments use Qwen3-4B-Base and OLMo-3-7B-Instruct on a 7,500-example MATH split, with group size $G = 16$, actor learning rate 10^{-6}, prompt/response limits 2,048/8,192, global and mini-batches of 192/24 prompts, asynchronous vLLM rollout with tensor parallelism 2, GPU memory utilization 0.8, and at most 500 steps over 10 epochs. The GovReport runs use the same backbones, sample eight summaries per input with temperature and top-p equal to 1.0, and train with LongDocFACTScore.

The tool-call comparison uses the identical 3,920/80 train/test parquet files and rule scorer in both runs, with Qwen3-4B-Base, $G = 16$, actor learning rate 10^{-6}, prompt/response limits 3,072/1,024, global and mini-batches of 64/16 prompts, and asynchronous vLLM rollout with tensor parallelism 2. Validation uses one greedy response per prompt every 20 steps. The only recorded configuration differences are experiment names and output directories; the RoVR-GSPO launcher selects block-credit advantages and SoftRoVR sequence weights, while the GSPO launcher selects ordinary group mean/standard-deviation advantages and arithmetic sequence weights. Both configurations request 500 steps, but the supplied logs contain training only through the step-100 checkpoint and validation through step 80. The run identifiers are `sqy23z50` (RoVR-GSPO) and `y1c3z2ii` (GSPO).

The dual-channel Qwen3-4B-Base run used in Section B.2 records reward parameters $\kappa = 1$ and $s_{\min} = 10^{-3}$. Its ratio channel uses $B = 8$ balanced contiguous blocks, a minimum block size of 4, $K = 9$ quantiles, $c = 1$, $\gamma = \eta = 0.01$, $S = 32$ safeguarded solver iterations, $a_{\min} = \nu_{\min} = 10^{-6}$, $\nu_{\max} = 10$, and denominator regularization 10^{-8}. The recorded clipping interval is $[0.9997, 1.0004]$. These settings describe the trajectory used by the paired offline comparison.

The reward-contamination analysis uses groups of $G = 16$, the full RoVR reference with local scale $c = 1$, credit parameters $\kappa = 1$ and $s_{\min} = 10^{-3}$, bootstrap seed 20250904, and 500 bootstrap replicates. The global reward standard deviation is 1.165, and the median within-group population standard deviation is approximately 0.165. A 4σ global shift therefore corresponds to about 28.28 times this typical within-group scale.

C.2 Resolution and reporting conventions

The RoVR configurations in the experiments use blockwise M-centers and the composite-quantile outer correction. The reward-reference ablation retains this construction and changes only the residual normalization; the channel-placement ablations enable the reward channel, the ratio channel, or both. The theoretical conditions $B \geq 2q + 1$ and $n_{\min} \geq 2s + 1$ specify the contamination budgets covered by the two-level guarantee. Section E studies six scalar reference estimators under matched partitions.

Downstream comparisons match initialization, sampled-response budget, decoding, optimizer budget, and evaluation code; prompt order is not verified across the two tool-call runs. Tables 1 and 3 report means and cross-run standard deviations over three independent runs. GovReport reports single-run point estimates; the tool-call comparison reports one run per method and selects the highest Total reward from each run's four logged validations. The token-ratio analysis uses 5,000 prompt-cluster bootstrap replicates from a fixed training trajectory, the reward analysis uses 500 group-bootstrap replicates, and the scalar simulations quantify Monte Carlo uncertainty. The GovReport figure displays reward trajectories and their visual separation.

C.3 Additional token-ratio stress diagnostics

The primary token-ratio stress test in Section B.2 uses positive perturbations in the main figure; here we apply the same protocol to negative perturbations. At step 18, the negative $16\sigma_\ell$ spike reduces D_q by 26.09%, while the flip rate decreases from 21.061% to 17.370% with difference 3.691% and interval $[2.949\%, 4.453\%]$. For a negative 20% burst, D_q decreases by 87.39% and the flip-rate difference is 0.807% with interval $[0.202\%, 1.413\%]$. The corresponding positive-burst interval includes zero.

For a stage-sensitivity check, every complete recorded step is evaluated at spike magnitude $16\sigma_\ell$ and burst fraction 20% using a fixed seeded sample of 64 complete prompt groups (1,024 responses) and five paired positions per response. Across all 35 complete steps and both directions, the paired differences for D_q and strong-spike flips are positive, with pointwise intervals above zero. For 20% bursts, the reduction in mean D_q ranges from 85.29% to 90.12%. The burst flip difference changes sign across steps: its point estimate favors SoftRoVR in $28/35$ positive-direction checks and $20/35$ negative-direction checks. These results support lower aggregation sensitivity across recorded stages, while clipping effects remain direction- and stage-dependent. These steps are observations from one training run, not independent seeds, and the pointwise intervals are not adjusted for multiple comparisons.

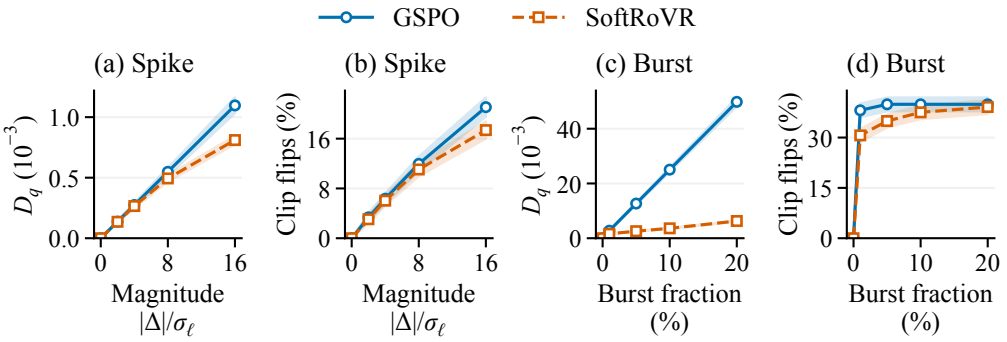

Figure 6: Negative-direction sensitivity analysis at step 18 (192 prompt groups, 3,072 responses). Bands are pointwise 95% prompt-cluster bootstrap intervals. The perturbations are $-c\sigma_\ell$ for spikes and $-8\sigma_\ell$ for bursts. SoftRoVR reduces D_q throughout the tested nonzero settings. Its clipping benefit is larger for strong spikes and short bursts than for the longest burst.

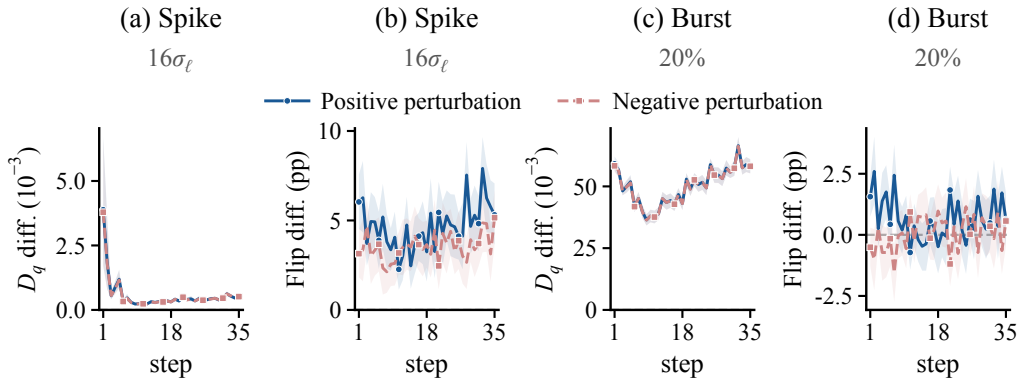

Figure 7: Paired differences across 35 complete steps for spike $16\sigma_\ell$ and burst 20% stress, using 64 prompt groups per step. Lines connect the recorded estimates without smoothing; shaded bands are pointwise 95% prompt-cluster bootstrap intervals. Positive values favor SoftRoVR. Effects on log-weight displacement are consistent across the recorded stages; long-burst clipping effects are stage-dependent.

Table 6: Comparison of representative location/reference estimators. RoVR combines a bounded local M-estimator with tie-neutral composite-quantile aggregation while retaining explicit block protection.

Estimator	Local statistic	Outer aggregation	Bounded local score	Block protection	Clean outer factor
Arithmetic mean	Mean	None	–	–	1
Global M-center	Bounded-score M-center	None	✓	–	–
MOM	Arithmetic block mean	Median	–	✓	$\pi/2$
VRMOM	Arithmetic block mean	Composite quantile	–	✓	$V_K \to \pi/3$
Robust MOM	Robust local M-center	Median	✓	✓	$\pi/2$
RoVR	**Robust local M-center**	**Tie-neutral composite quantile**	✓	✓	$\boldsymbol{V_K \to \pi/3}$

D Implementation Details

This appendix specifies the numerical object corresponding to the definitions in the main text. It records the full Barron family for completeness, then gives the convex solver, tie convention, balanced-block handling, SoftRoVR path, and complexity. Main results use only $\alpha = 1$.

Table 7: RoVR-based advantage constructions used in our analysis and algorithm.

Variant	RoVR reference	Numerator	Scale	Role
$\widehat{A}_i^{\mathrm{center}}$	$\mathrm{RoVR}(R_{1:G})$	$R_i - \widehat{\theta}_R$	Raw RMS	Center-only ablation
$\widehat{A}_i^{\mathrm{LOO}}$	$\mathrm{RoVR}(R_{-i})$	$R_i - \widehat{\theta}_{R,-i}$	S_{-i}	Leave-one-out theoretical reference
$\widehat{A}_i^{\mathrm{credit}}$	$\mathrm{RoVR}(R_{1:G})$	$\boldsymbol{\chi_\kappa(R_i - \widehat{\theta}_R)}$	$\widehat{s}_\chi$	**Default robust advantage**; bounded leverage and fidelity control

Only $\widehat{A}_i^{\mathrm{credit}}$ is used by default in RoVR-GSPO. $\widehat{A}_i^{\mathrm{center}}$ is a center-only ablation, whereas $\widehat{A}_i^{\mathrm{LOO}}$ provides a leave-one-out reference for the theoretical analysis.

D.1 Estimator and update summaries

Algorithm 1 RoVR-GSPO: dual-channel robust GSPO update

Require: rewards $R_{1:G}$; proposed B and balanced sizes $\{n_b\}$; declared failure unit and budgets (q, s)

Require: $(K, c, \kappa, s_{\min}, a_{\min}, \nu_{\min}, \nu_{\max})$; SoftRoVR switch z_{ratio}

$z_{\mathrm{block}} \leftarrow$ failure unit supported and budgets (q, s) declared
$z_{\mathrm{block}} \leftarrow z_{\mathrm{block}} \wedge (B > 1) \wedge (B \geq 2q + 1)$
$z_{\mathrm{block}} \leftarrow z_{\mathrm{block}} \wedge (\min_b n_b \geq 2s + 1) \wedge (\sum_b n_b = G)$
$z_{\mathrm{block}} \leftarrow z_{\mathrm{block}} \wedge (\max_b n_b - \min_b n_b \leq 1)$
if $z_{\mathrm{block}} = 0$ or the realized assignment violates the declared sizes **then**
    Set $B \leftarrow 1$ and $\mathcal{H}_1 \leftarrow \{1, \ldots, G\}$ ▷ global-M fallback
else
    Use the predeclared balanced assignment sampled independently of corruption labels used by the guarantee
end if
Compute every convex local center by (9)
if $B = 1$ **then**
    $\widehat{\theta}_R \leftarrow \widetilde{\mu}_1$
else
    Compute the capped pooled scale (11) and RoVR reference (13)
end if
$z_i \leftarrow \chi_\kappa(R_i - \widehat{\theta}_R)$ and $\widehat{s}_\chi \leftarrow \{s_{\min}^2 + G^{-1} \sum_j z_j^2\}^{1/2}$
$\widehat{A}_i^{\mathrm{credit}} \leftarrow z_i / \widehat{s}_\chi$ for $i = 1, \ldots, G$; detach reward statistics
if $z_{\mathrm{ratio}} = 1$ **then**
    Require a GSPO parent; compute the SoftRoVR sequence weight in (21)
else
    Keep the canonical GSPO sequence aggregate
end if
return $\widehat{A}_{1:G}^{\mathrm{credit}}$ and the selected sequence weights

D.2 Barron family and convex root solver

For residual u, shape α, and scale $c > 0$, Barron's loss is (Barron, 2019)

$$\rho(u; \alpha, c) = \begin{cases} \frac{1}{2}(u/c)^2, & \alpha = 2, \\ \log(1 + \frac{1}{2}(u/c)^2), & \alpha = 0, \\ 1 - \exp[-\frac{1}{2}(u/c)^2], & \alpha = -\infty, \\ \frac{|\alpha-2|}{\alpha} \left[\left(1 + \frac{(u/c)^2}{|\alpha-2|} \right)^{\alpha/2} - 1 \right], & \text{otherwise.} \end{cases}$$

For $\alpha = 1$, safeguarded Newton or bisection solves the strictly decreasing empirical score. A bracket is given by the sample minimum and maximum. Newton steps leaving the bracket are replaced by bisection, and termination uses both score and interval tolerances.

Shapes $\alpha < 1$ are redescending and may have multiple stationary points. If reported, they use a fixed median initialization, a declared graduated schedule, and a separate label; no convex-root or finite-group theorem is transferred to them.

D.3 Blocks, ties, scales, and differentiation

The entire balanced random assignment or permutation, not merely the block sizes, is sampled independently of corruption labels whenever Theorem F.2 is invoked. The exact n_b, partition seed, and whether corruption could adapt to the partition are logged. For $B = 1$, the composite correction is identically zero by Theorem 3.1. For even B, the midpoint median is used consistently in hard reward code; the token map uses the unique finite-temperature initializer in (20).

For exact discrete rewards, J_0 assigns half weight to equality. A numerical implementation may use a fixed tolerance ϵ_{tie} to assign $1/2$ when $|u| \le \epsilon_{\mathrm{tie}}$; this convention is part of the estimator specification. The logistic map returns $1/2$ at zero.

For small local blocks, an independent pilot or a lagged running statistic can supply the scale. Gradients are stopped through reward-side centering, scale estimation, and partitioning. On the ratio side, gradients pass through the unrolled safeguarded solvers and the finite-temperature correction. The pooled smooth scale uses capped local sandwich scales; derivative checks should evaluate the same finite-temperature map.

With S root iterations, local centers cost $O(SN)$ scalar operations. Hard median selection costs $O(B)$, while the general outer correction costs $O(BK)$. SoftRoVR's two smooth median solves add $O(SB)$ operations; bounded reward-credit construction costs $O(G)$. Forward aggregation stores local statistics and correction terms, while differentiating through the unrolled solves additionally retains token-level intermediate states. Actual memory usage depends on solver unrolling, token packing, and the autograd implementation.

E Supplementary Experiment: Reference-Estimator Variance and Robustness

This appendix studies the scalar reference estimators under controlled sampling and contamination. Each method receives the same observations and balanced partition, with uncontaminated center 0 as the target. The design isolates reference-estimation variance and error, complementing the downstream policy comparisons.

E.1 Design and estimands

We use $N = 128$ observations partitioned into $B = 8$ balanced blocks of size 16, with the RoVR parameters $K = 9$ and $c = 1$. The clean distributions are standard Gaussian and $t_3/\sqrt{3}$, so both have center and variance equal to zero and one, respectively. For the contamination experiments, 16 observations (12.5%) are shifted by $+8$. In the dispersed condition these observations are allocated as evenly as possible across the eight blocks; in the coherent condition one entire block is shifted. Thus the two geometries have the same contamination budget, while only their placement relative to the block interface differs. The estimator comparison uses 3,000 independent trials, the geometry sweep uses 2,000 trials per shift, and the oracle outer-factor experiment uses 20,000 trials. All random streams are seeded with seed 0.

We compare the arithmetic mean, the global convex M-center, MOM, a matched *VRMOM-style* estimator, Robust MOM, and RoVR. The VRMOM-style comparator uses arithmetic local means, composite-quantile correction, and the median block standard-deviation scale. We report sampling variance and RMSE to the uncontaminated center. Error bars are 95% bootstrap intervals over Monte Carlo trials.

E.2 Oracle outer-factor validation

Table 8 and Figure 8 validate the outer aggregation in isolation. The theoretical factor V_K decreases from $\pi/2$ at $K = 1$ toward $\pi/3$ as the quantile grid is refined. At the paper's setting $K = 9$, $V_K = 1.0691$ and the empirical oracle factor is 1.0747; at $K = 31$, they are 1.0498 and 1.0540. The empirical median factor is 1.5649, close to $\pi/2 = 1.5708$. The finite-sample agreement supports the implementation of the composite-quantile correction and its stated outer-factor interpretation. The complete estimator additionally incorporates the local sandwich scale, finite block sizes, and estimated initialization.

Table 8: Oracle outer-factor validation. The empirical factors use 20,000 trials with $B = 101$ and block size $n = 32$. The median factor is shown for reference; $\pi/2$ and $\pi/3$ are the clean asymptotic reference values.

K	V_K (theory)	Oracle correction (emp.)	Outer median (emp.)
1	1.5708	1.5812	1.5649
3	1.1680	1.1802	1.5649
5	1.1034	1.1102	1.5649
9	1.0691	1.0747	1.5649
15	1.0564	1.0614	1.5649
31	1.0498	1.0540	1.5649
	$\pi/2$		1.5708
	$\pi/3$		1.0472

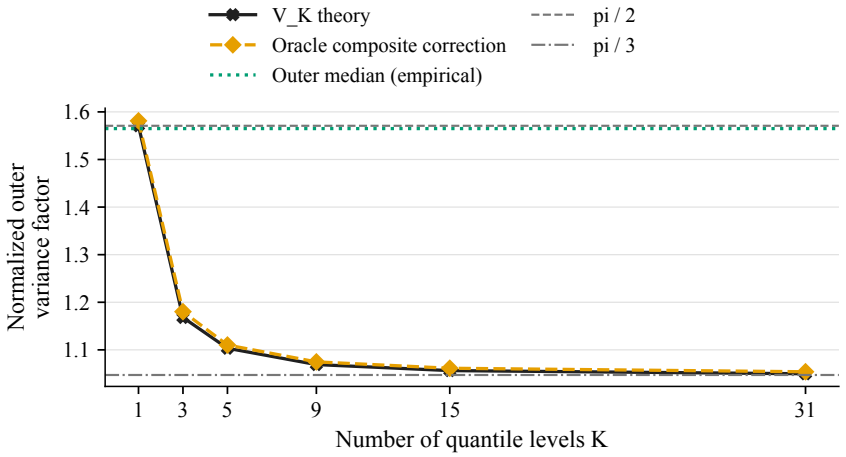

Figure 8: Oracle outer-factor validation. The oracle composite-quantile estimator follows V_K closely, while the outer median remains near $\pi/2$.

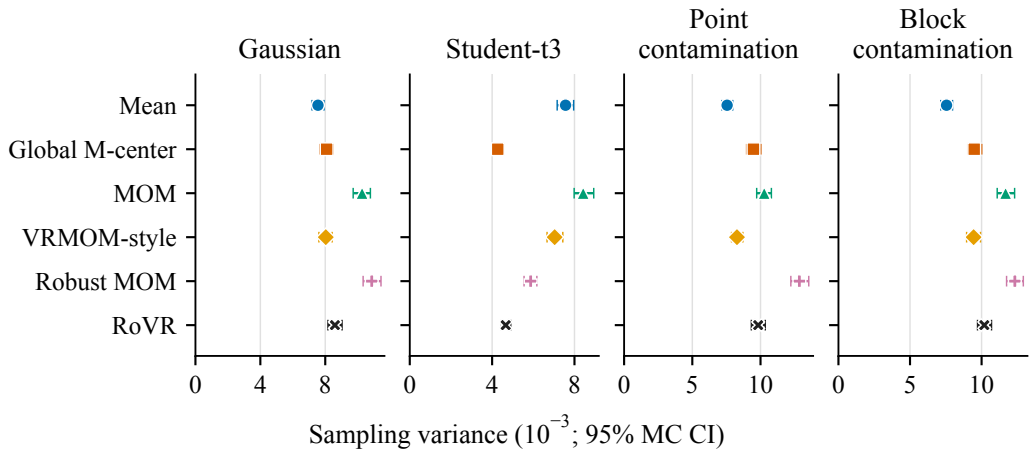

Figure 9: Sampling variance across clean and contaminated scenarios. Error bars are 95% bootstrap Monte Carlo intervals over 3,000 estimator trials. Under contamination, variance alone is not a sufficient robustness metric because it does not include systematic displacement.

E.3 FINITE-SAMPLE COMPARISON

The sampling variances are reported in Table 9 and Figure 9. Under clean Gaussian data, the arithmetic mean is the most efficient estimator in this comparison (0.007552), while RoVR has variance 0.008596. This is expected: bounded local scores and block aggregation trade some Gaussian efficiency for robustness. RoVR nevertheless has lower variance than MOM (0.010271) and Robust MOM (0.010865) in this finite configuration. Under the unit-variance t_3 distribution, robust score estimators benefit from suppressing tail observations; RoVR has variance 0.004660, close to the global M-center (0.004274) and below the arithmetic mean (0.007567) and MOM (0.008424).

Table 9: Sampling variance of the six reference estimators. Each entry is computed from 3,000 trials; the full CSV also reports bootstrap Monte Carlo intervals.

Scenario	Mean	Global M	MOM	VRMOM-style	Robust MOM	RoVR
Gaussian	0.007552	0.008088	0.010271	0.008033	0.010865	0.008596
Student-t_3	0.007567	0.004274	0.008424	0.007038	0.005871	0.004660
Point contamination	0.007552	0.009483	0.010271	0.008273	0.012848	0.009828
Block contamination	0.007552	0.009486	0.011677	0.009443	0.012323	0.010190

E.4 CONTAMINATION GEOMETRY AND INTERPRETATION

Table 10: RMSE to the uncontaminated center 0 under the fixed $+8$ contamination. Both contamination geometries modify 16 of 128 observations.

Scenario	Mean	Global M	MOM	VRMOM-style	Robust MOM	RoVR
Point contamination	1.0059	0.2719	1.0067	1.0064	0.2746	0.2695
Block contamination	1.0059	0.2718	0.1171	0.1129	0.1206	0.1170

Table 10 and Figure 10 show why the failure geometry must be declared. With dispersed point contamination, the mean, MOM, and VRMOM-style estimator have RMSE approximately 1.006, whereas RoVR has RMSE 0.270, close to the global M-center (0.272) and Robust MOM (0.275). The bounded local score prevents the shifted observations from directly controlling every block mean. With one fully corrupted block, block-based estimators are substantially more effective: MOM, VRMOM-style, and RoVR have RMSE 0.117, 0.113, and 0.117, respectively, while the global M-center has RMSE 0.272. Both RoVR and the VRMOM-style comparator benefit from the block interface in this setting.

The shift sweep in Figures 11 and 12 gives the same qualitative picture over contamination magnitudes from 0 to 16. Under dispersed corruption, the mean-like block estimators grow approximately linearly with the shift, while the global and block-robust estimators saturate. Under coherent block corruption, the explicit block interface allows the median and composite-quantile outer aggregators to reject the bad block for the declared one-block failure geometry. RoVR inherits this behavior through its robust local centers and composite-quantile correction. The sweep evaluates the declared, fixed block-failure geometry across perturbation magnitudes.

The simulations reproduce the predicted outer-efficiency factor and show how local and block robustness contribute under different contamination geometries. RoVR improves over median-based aggregation in the clean settings and achieves low error under both dispersed and coherent contamination. The downstream effects of the complete optimizer are evaluated in Tables 1 to 3; the exploratory tool-call proxy is reported in Section 5.2.

F DETAILED THEORETICAL STATEMENTS

This section connects estimator robustness to the group-relative update in four steps. We first give deterministic and fixed-group guarantees under mixed block and observation contamination. We then control the complete bounded credit and its target-fidelity cost. A clean transfer result separates the local sandwich variance from the outer factor and exploits scale orthogonality. The final results cover

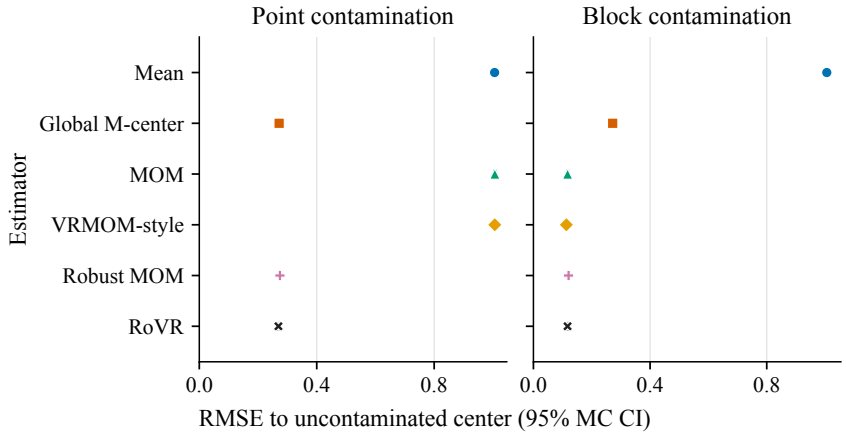

Figure 10: RMSE to the uncontaminated center under the two matched contamination geometries. Error bars are 95% bootstrap Monte Carlo intervals over 3,000 trials.

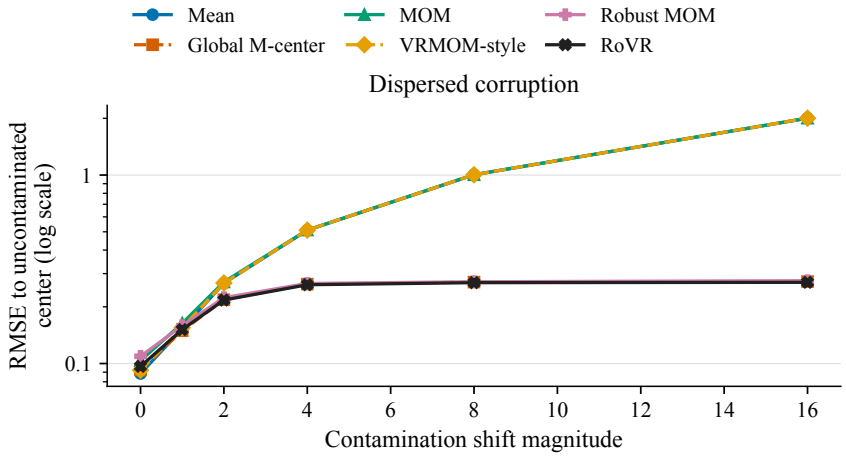

Figure 11: RMSE sweep for dispersed point corruption. The y-axis is logarithmic to show both mean-like and robust estimators over the same range.

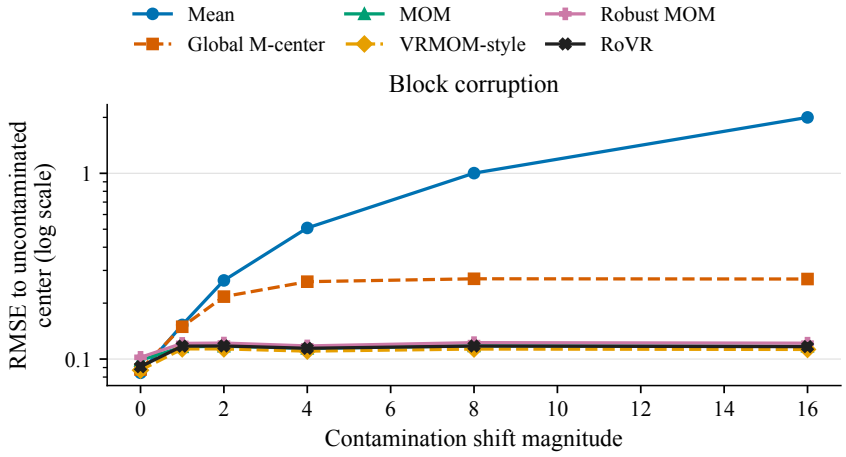

Figure 12: RMSE sweep for one-block coherent corruption. The favorable behavior of block estimators is conditional on the declared block failure geometry.

smoothing, clipping branches, linear leave-one-out baselines, and the bias introduced by robust credit shaping. Throughout, $\mathcal{I}(a,b) = [\min\{a,b\}, \max\{a,b\}]$ denotes the closed interval between two real numbers.

F.1 Mixed contamination at fixed group size

Bounded score alone does not bound the displacement of a convex M-root because ψ_c' vanishes in the tails. A deterministic comparison of two realized roots therefore requires curvature on the interval between them. This result is useful for token replacements, where no independence model is imposed.

Proposition F.1. *Let datasets $D = (x_1, \dots, x_n)$ and $D' = (x_1', \dots, x_n')$ differ in at most s coordinates, with convex M-centers T, T'. If*

$$\inf_{u \in \mathcal{I}(T,T')} \frac{1}{n} \sum_{i=1}^{n} \psi_c'(x_i - u) \geq a_0 > 0,$$

then

$$|T' - T| \leq \frac{2sM_c}{na_0}. \tag{27}$$

The same conclusion holds when the curvature condition is imposed on D'.

For reward groups, the main stochastic result avoids conditioning on this unknown root path. Let $g(t) = \mathbb{E}\psi_c(X - t)$ and suppose, for $0 \leq u \leq r_0$,

$$g(\theta + u) \leq -au, \qquad g(\theta - u) \geq au$$

for $a > 0$. In honest block b, let at most s_b observations be replaced after drawing n_b independent clean observations. Define

$$p_b(t) = \exp\left[-\frac{n_b}{2M_c^2} \left(at - \frac{2s_b M_c}{n_b} \right)_+^2 \right]. \tag{28}$$

This bound follows by evaluating the contaminated score at the fixed threshold $\theta \pm t$; it does not require the empirical root to remain in a postulated curvature basin.

Random reward blocks only justify a within-block load statement if the partition is drawn independently of the corrupted set. The next proposition makes that scope explicit.

Proposition F.2. *Fix a set of m corrupted indices among N observations, and draw a balanced partition independently and uniformly at random. If S_b is the number of corrupted indices in a block of size n_b, then for every $u > 0$,*

$$\mathbb{P}\left(\max_{1 \leq b \leq B} \left\{ S_b - \frac{n_b m}{N} \right\} \geq u \right) \leq \sum_{b=1}^{B} \exp\left(-\frac{2u^2}{n_b} \right) \leq B \exp\left(-\frac{2u^2}{n_{\max}} \right). \tag{29}$$

The statement does not apply to corruption selected after observing the partition.

We now combine non-identical honest-block tails with two distinct contamination units. Let $\mathcal{A}$ contain q arbitrary blocks and let $\mathcal{H}$ be its complement, with $H = B - q$, $r_B = \lceil B/2 \rceil$, and $q < r_B$. The set $\mathcal{A}$ must be fixed before the clean draws or sampled independently of them; values in these blocks may then be arbitrary and may depend on all clean observations. Latent clean observations in the honest blocks are mutually independent within and across blocks. After those observations are drawn, an adversary may choose the coordinates and values of at most s_b replacements inside each honest block b. Thus the replacement mechanism may be adaptive to all clean values, but neither the honest/arbitrary block labels nor a partition invoked through Theorem F.2 may be selected from those values. Define

$$\overline{p}_H(t) = \frac{1}{H} \sum_{b \in \mathcal{H}} p_b(t), \qquad \eta_q = \frac{r_B - q}{B - q},$$

and define the observable outer-correction magnitude and its deterministic envelope

$$\widehat{C}_B = |\widehat{\theta}_{\mathrm{RoVR}} - \widehat{\mu}_0|, \qquad C_B = \frac{\nu_{\max} BK}{2D_K W_B}, \qquad \widehat{C}_B \leq C_B. \tag{30}$$

The value $\widehat{C}_B$ is computable from a realized sample, but it is not by itself a confidence radius. The probability statement below additionally fixes t and depends on the population-separation quantities (a, r_0) through $p_b(t)$; these quantities are not estimated post hoc from the same batch.

Theorem F.3. *Suppose the contamination timing and independence conditions above hold and every honest block satisfies the population score separation used in* (28)*. For $0 < t \le r_0$, if $\overline{p}_H(t) < \eta_q$, set*

$$\beta_H(t) = \min\{1,\, 2\exp[-2H\{\eta_q - \overline{p}_H(t)\}^2]\}.$$

Then the tie-neutral estimator in (13) *satisfies both*

$$\mathbb{P}\Big(|\widehat{\theta}_{\mathrm{RoVR}} - \theta| > t + \widehat{C}_B\Big) \le \beta_H(t), \qquad \mathbb{P}\Big(|\widehat{\theta}_{\mathrm{RoVR}} - \theta| > t + C_B\Big) \le \beta_H(t). \tag{31}$$

Corollary F.4. *Under Theorem F.3 with $q = 0$, suppose the complete set of m replacement indices is fixed before an independently sampled uniform balanced partition. For $0 < \delta_{\mathrm{part}} < 1$, use the block-specific deviation*

$$u_b(\delta_{\mathrm{part}}) = \sqrt{\frac{n_b}{2}\log\frac{B}{\delta_{\mathrm{part}}}}, \qquad s_b(\delta_{\mathrm{part}}) = \min\Big\{n_b, \Big\lceil \frac{n_b m}{N} + u_b(\delta_{\mathrm{part}})\Big\rceil\Big\}.$$

Thus each block uses its own n_b rather than a common deviation based on $n_{\max}$. Form $p_{b,\delta}(t)$ from (28)*, and let $\overline{p}_{B,\delta}(t) = B^{-1}\sum_b p_{b,\delta}(t)$ and $\eta_0 = \lceil B/2\rceil/B$. If $\overline{p}_{B,\delta}(t) < \eta_0$, define*

$$\beta_{B,\delta}(t) = \min\{1,\, \delta_{\mathrm{part}} + \min\{1,\, 2\exp[-2B\{\eta_0 - \overline{p}_{B,\delta}(t)\}^2]\}\}.$$

Then, jointly over the clean sample and partition,

$$\mathbb{P}\Big(|\widehat{\theta}_{\mathrm{RoVR}} - \theta| > t + \widehat{C}_B\Big) \le \beta_{B,\delta}(t), \qquad \mathbb{P}\Big(|\widehat{\theta}_{\mathrm{RoVR}} - \theta| > t + C_B\Big) \le \beta_{B,\delta}(t). \tag{32}$$

The conclusion does not cover indices selected after the partition, a block designation selected by inspecting realized block values, or corruption that can observe the partition RNG before choosing its support. As in Theorem F.3, the conditional error statement still depends on the declared population-separation model.

The theorem covers fixed B and balanced unequal block sizes under the declared population-separation model. The quantity $\widehat{C}_B$ measures the realized outer correction; an error radius $t + \widehat{C}_B$ additionally uses the ex ante threshold t and separation parameters (a, r_0). At $B = 1$, the estimator equals the global M-center. Clean outer-efficiency rates are characterized separately in Section F.3.

F.2 Target, bounded response advantage, and retained contrast

Robust location changes its population target under asymmetry, while bounded credit changes the reward utility. This subsection separates both effects from sampling and contamination error.

Proposition F.5. *Let $\mu = \mathbb{E}X$ be finite, $g(t) = \mathbb{E}\psi_c(X - t)$, and θ_ρ be a root of g. If*

$$\inf_{u \in \mathcal{I}(\mu, \theta_\rho)} \mathbb{E}\psi_c'(X - u) \ge a_0 > 0,$$

then

$$|\theta_\rho - \mu| \le \frac{|\mathbb{E}\psi_c(X - \mu)|}{a_0}. \tag{33}$$

If $\mathbb{E}|X - \mu|^3 < \infty$ and

$$\inf_{u \in \mathcal{I}(\mu, \theta_\rho)} \mathbb{E}\left[\left\{1 + \frac{(X-u)^2}{c^2}\right\}^{-3/2}\right] \ge \bar{a}_0 > 0,$$

then

$$|\theta_\rho - \mu| \le \frac{\mathbb{E}|X - \mu|^3}{2\bar{a}_0 c^2}. \tag{34}$$

Proposition F.6. *Let $P_\epsilon = (1 - \epsilon)P_0 + \epsilon Q$, where P_0 is symmetric about μ_0. If the clean score obeys $|\mathbb{E}_{P_0}\psi_c(X - t)| \ge a_0|t - \mu_0|$ between μ_0 and $\theta_\rho(P_\epsilon)$, then*

$$|\theta_\rho(P_\epsilon) - \mu_0| \le \frac{\epsilon M_c}{(1 - \epsilon)a_0}. \tag{35}$$

The next result controls every response, including the replaced focal response omitted by a center-only analysis.

Proposition F.7. *Let reward groups $R, R^\star \in \mathbb{R}^G$ differ in at most r coordinates, and suppose their RoVR centers differ by at most e_θ. Construct scales and credits with the same χ_κ as in* (16)*. Then*

$$e_s := |\widehat{s}_\chi - s^\star_\chi| \le \frac{r\kappa^2/G + 2\kappa e_\theta}{2s_{\min}}. \tag{36}$$

For any index i, let $e_{R,i} = |R_i - R^\star_i|$ and $d_{\chi,i} = \min\{e_{R,i} + e_\theta, 2\kappa\}$. Then

$$|\widehat{A}^{\mathrm{credit}}_i - A^{\mathrm{credit},\star}_i| \le \frac{d_{\chi,i}}{s_{\min}} + \frac{\kappa e_s}{s^2_{\min}}, \qquad |\widehat{A}^{\mathrm{credit}}_i| \le \frac{\kappa}{s_{\min}}. \tag{37}$$

The bound is independent of the magnitude of a replaced reward. For an unchanged response, $d_{\chi,i} \le e_\theta$ and its sign is preserved whenever $|R_i - \theta^\star_R| > e_\theta$.

The center and credit bounds can be composed without hiding the focal replacement. Let $R^\star$ denote the clean precursor group and construct $A^{\theta,\star}_i$ from $R^\star$ using population reference θ, the same χ_κ, and the same $s_{\min}$.

Corollary F.8. *Under Theorem F.3, suppose the observed and precursor reward groups differ in at most r coordinates. Set*

$$e_\theta(t) = t + C_B, \qquad \overline{e}_s(t) = \frac{r\kappa^2/G + 2\kappa e_\theta(t)}{2s_{\min}},$$

and $e_{R,i} = |R_i - R^\star_i|$. With probability at least $1 - \beta_H(t)$, simultaneously for all responses,

$$|\widehat{A}^{\mathrm{credit}}_i - A^{\theta,\star}_i| \le \frac{\min\{e_{R,i} + e_\theta(t), 2\kappa\}}{s_{\min}} + \frac{\kappa \overline{e}_s(t)}{s^2_{\min}}. \tag{38}$$

The same population-separation-dependent probability statement holds with the observable correction magnitude $e_\theta(t) = t + \widehat{C}_B$ and the corresponding $\overline{e}_s(t)$. Hence the complete credit perturbation is independent of replacement magnitude, although it can still be numerically loose when $s_{\min}$ is small. A magnitude-independent policy-gradient statement additionally requires a score-Jacobian bound, a stable clipping branch, or explicit gradient clipping; Theorem F.15 states only the corresponding local conditions.

Boundedness need not erase moderate clean contrasts. If two clean residuals lie in $[-M, M]$, the mean-value theorem and $\chi'_\kappa(u) \ge m_{\kappa,M} := (1 + M^2/\kappa^2)^{-3/2}$ give

$$|\widehat{A}^{\mathrm{credit}}_i - \widehat{A}^{\mathrm{credit}}_j| \ge \frac{m_{\kappa,M}}{\sqrt{s^2_{\min} + \kappa^2}} |R_i - R_j|. \tag{39}$$

Together, (18) and (39) state the trade-off: larger κ more closely preserves raw credit, while smaller κ more strongly limits focal leverage.

F.3 Clean efficiency and scale orthogonality

The finite-group bound is conditional on the declared population score separation and uses a deterministic envelope for the outer correction. Clean efficiency instead follows from a separate local block-center model. We first state an exact oracle calculation for unequal blocks, then give the transfer result for the implemented estimator.

Let

$$a_\rho = \mathbb{E}\psi'_c(X - \theta_\rho), \qquad b_\rho = \mathbb{E}\psi^2_c(X - \theta_\rho), \qquad \nu^2 = \frac{b_\rho}{a^2_\rho}.$$

The clean transfer assumes $0 < a_\rho < \infty$, $0 < b_\rho < \infty$, $a_\rho > a_{\min}$, and $\nu \in (\nu_{\min}, \nu_{\max})$, so the implemented floor and cap are asymptotically inactive. Let

$$A_K = \sum_{k=1}^{K}\sum_{\ell=1}^{K} [\min(\tau_k, \tau_\ell) - \tau_k\tau_\ell], \qquad V_K = \frac{A_K}{D^2_K}.$$

Proposition F.9. *Suppose independent block centers satisfy $\widetilde{\mu}_b \sim \mathcal{N}(\theta, \nu^2/n_b)$ exactly. If the one-step correction in* (13) *is evaluated at the oracle initializer θ and scale ν, then it is unbiased and*

$$\mathrm{Var}(\widehat{\theta}_{\mathrm{oracle}}) = \frac{\nu^2 V_K}{N_{\mathrm{eff}}}, \qquad N_{\mathrm{eff}} = \frac{W_B^2}{B}. \tag{40}$$

The symmetric quantile grid makes the one-step score locally insensitive to scale. Define

$$M_\Phi(t,s) = \sum_{k=1}^{K} \{\Phi(t + s\Delta_k) - \tau_k\}.$$

Proposition F.10. *At $(t,s) = (0,1)$,*

$$M_\Phi(0,1) = 0, \qquad \partial_t M_\Phi(0,1) = D_K, \qquad \partial_s M_\Phi(0,1) = \sum_{k=1}^{K} \Delta_k \phi(\Delta_k) = 0. \tag{41}$$

For fixed K and (t,s) in a neighborhood of $(0,1)$,

$$M_\Phi(t,s) = D_K t + O\big(t^2 + (s-1)^2\big). \tag{42}$$

Scale orthogonality removes the first-order scale term under the local normal-score model. The transfer below requires only $\widehat{\nu}/\nu - 1 = o_{\mathbb{P}}(B^{-1/4})$; a root-$B$ pilot is sufficient but not necessary. This remains a conditional asymptotic statement: the local normal-score approximation, stochastic equicontinuity, and pilot-scale rate are explicit assumptions rather than consequences of a pointwise block-center CLT. For a concise statement, take equal blocks of size n and define

$$Z_{b,n} = \frac{\sqrt{n}(\widetilde{\mu}_b - \theta_\rho)}{\nu}, \qquad F_n(z) = \mathbb{E}[J_0(Z_{b,n} - z)],$$

$$M_n(t,s) = \sum_{k=1}^{K} \{F_n(t + s\Delta_k) - \tau_k\},$$

For the realized row, let $\mathcal{H}_B$ be the honest-block set and $H_B = |\mathcal{H}_B| = B - q_B$. Define the all-block composite score and its honest centered fluctuation by

$$\widehat{M}_{B,n}(t,s) = \frac{1}{B} \sum_{b=1}^{B} \sum_{k=1}^{K} \{J_0(Z_{b,n} - t - s\Delta_k) - \tau_k\},$$

$$\mathbb{U}_{B,n}(t,s) = \frac{1}{B} \sum_{b \in \mathcal{H}_B} \sum_{k=1}^{K} \{J_0(Z_{b,n} - t - s\Delta_k) - F_n(t + s\Delta_k)\}.$$

Assumption F.11. For fixed K, honest block centers are independent and identically distributed within each row, while $q_B = o(\sqrt{B})$ blocks may be arbitrary. The population quantities satisfy $0 < a_\rho < \infty$, $0 < b_\rho < \infty$, $a_\rho > a_{\min}$, and $\nu \in (\nu_{\min}, \nu_{\max})$. With $t_0 = \sqrt{n}(\widehat{\mu}_0 - \theta_\rho)/\nu$ and $\widehat{s} = \widehat{\nu}/\nu$:

(T1) $t_0 = O_{\mathbb{P}}(B^{-1/2})$.

(T2) $\widehat{s} - 1 = o_{\mathbb{P}}(B^{-1/4})$.

(T3) For every fixed C and a deterministic $\varrho_B = o(B^{-1/4})$ containing $|\widehat{s} - 1|$ with probability tending to one,

$$\sup_{|t| \le C/\sqrt{B},\, |s-1| \le \varrho_B} |M_n(t,s) - M_\Phi(t,s)| = o(B^{-1/2}).$$

(T4) On the same neighborhood,

$$\sup |\mathbb{U}_{B,n}(t,s) - \mathbb{U}_{B,n}(0,1)| = o_{\mathbb{P}}(B^{-1/2}).$$

(T5) The vector $\{J_0(Z_{b,n} - \Delta_k)\}_{k=1}^K$ has covariance converging to $\min(\tau_k, \tau_\ell) - \tau_k\tau_\ell$; in particular $F_n(\Delta_k) \to \tau_k$ and atoms at the thresholds vanish.

Conditions (T3)–(T4) are local normal-score approximation and stochastic equicontinuity. They are not implied by a pointwise CLT, especially for tie-rich rule rewards. Smooth M-root linearization, a local density approximation, and a root-B pilot scale are sufficient routes to these conditions; the fixed-group theorem remains the relevant result when they fail.

Theorem F.12. *Under Theorem F.11,*

$$\sqrt{Bn}(\widehat{\theta}_{\mathrm{RoVR}} - \theta_\rho) = -\frac{\nu}{D_K\sqrt{B}} \sum_{b\in\mathcal{H}_B} \sum_{k=1}^{K} [J_0(Z_{b,n} - \Delta_k) - F_n(\Delta_k)] + o_{\mathbb{P}}(1), \tag{43}$$

and therefore

$$\sqrt{Bn}(\widehat{\theta}_{\mathrm{RoVR}} - \theta_\rho) \xrightarrow{d} \mathcal{N}(0, \nu^2 V_K).$$

For the sequential limit in which $(B, n) \to \infty$ first at fixed K and then $K \to \infty$, $V_K \to \pi/3$.

The theorem isolates the outer asymptotic factor. Combining it with the local sandwich factor and the unequal-block effective sample size yields the complete asymptotic variance decomposition.

When $\theta_\rho = \mu$ and $0 < \sigma^2 < \infty$, clean efficiency relative to the sample mean is

$$\mathrm{ARE}_{\mathrm{mean}} = \frac{\sigma^2}{\nu^2 V_K}, \tag{44}$$

not $1/V_K$. Under asymmetry the estimands differ, so this is only a variance ratio. For quadratic local centers under a Gaussian model, $\nu^2 = \sigma^2$ and the outer limit is $3/\pi \approx 0.955$; for the robust local center, the sandwich factor must be included.

F.4 Smoothing, clipping branches, and policy scope

For fixed block centers, initializer, and scale, let

$$u_{bk} = \widetilde{\mu}_b - \widehat{\mu}_0 - \frac{\widehat{\nu}\Delta_k}{\sqrt{n_b}}, \qquad m_\star = \min_{b,k:\, u_{bk}\neq 0} |u_{bk}|.$$

At a tie $u_{bk} = 0$, both H_γ and J_0 equal $1/2$, so that indicator term contributes zero discrepancy. Initializer, scale, and solver residuals of the full smooth map remain separate.

Proposition F.13. *If at least one u_{bk} is nonzero, then*

$$|\mathcal{S}_\gamma - \mathrm{RoVR}\,| \le \frac{\widehat{\nu}BK}{D_K W_B} \exp\left(-\frac{m_\star}{\gamma}\right). \tag{45}$$

If each margin has no nonzero atoms and its continuous density is bounded by L_z, while $\widehat{\nu} \le \nu_{\max}$, then $\mathbb{E}|\mathcal{S}_\gamma - \mathrm{RoVR}\,| \le 2\nu_{\max}BKL_z\gamma \log 2/(D_K W_B)$. Atoms at zero contribute exactly zero under the mid-rank convention.

For block-center vectors T, T' on equal blocks of size n and pooled scales $0 < \lambda, \lambda' \le \lambda_{\max}$, let $d_T = \|T - T'\|_\infty$, $d_\lambda = |\lambda - \lambda'|$, and $\Delta_{\max} = \max_k |\Delta_k|$. With the raw-residual logistic parameterization $H_\gamma(u)$ defined above, the deterministic Lipschitz calculation in the appendix gives

$$\begin{aligned} |\mathcal{S}_\gamma(T, \lambda) - \mathcal{S}_\gamma(T', \lambda')| \le d_T + \frac{Kd_\lambda}{2\sqrt{n}D_K} \\ + \frac{\lambda_{\max}K}{4\gamma\sqrt{n}D_K}\left(2d_T + \frac{\Delta_{\max}d_\lambda}{\sqrt{n}}\right). \end{aligned} \tag{46}$$

Together with Theorem F.1, this is a distribution-free token replacement statement. The $1/\gamma$ term is the explicit robustness–smoothness trade-off.

The converged SoftRoVR map is translation equivariant. At differentiable inputs,

$$\mathrm{SoftRoVR}_{\gamma,\eta,S}(\ell + a\mathbf{1}) = \mathrm{SoftRoVR}_{\gamma,\eta,S}(\ell) + a, \qquad \sum_{t=1}^{T} \frac{\partial\, \mathrm{SoftRoVR}_{\gamma,\eta,S}(\ell)}{\partial \ell_t} = 1. \tag{47}$$

The same holds for a fixed number of equivariantly initialized solver steps. This common-shift calibration does not make the robust aggregate an exact likelihood ratio; asymmetry and position drift can still change its target.

Let ideal and observed log aggregates satisfy $|\widehat{m} - m^\star| \le e_m$. Then

$$e^{-e_m} \le \frac{\widehat{q}}{q^\star} \le e^{e_m}, \qquad \left|\frac{\widehat{q}}{q^\star} - 1\right| \le e^{e_m} - 1. \tag{48}$$

Branch stability can be checked rather than assumed globally.

Proposition F.14. *For sequence i, let $m_i^\star = \log q_i^\star$ and $|\widehat{m}_i - m_i^\star| \le e_{m,i}$. A ratio clipping boundary can change only if*

$$\operatorname{dist}(m_i^\star, \{\log l, \log u\}) \le e_{m,i}.$$

Let $e_{R,i} = |R_i - R_i^\star|$. An advantage sign can change only if $|R_i^\star - \theta_R^\star| \le e_{R,i} + e_\theta$; for an unchanged reward, $e_{R,i} = 0$. Hence

$$N_{\text{ratio-flip}} \le \sum_i \mathbf{1}\{\operatorname{dist}(m_i^\star, \{\log l, \log u\}) \le e_{m,i}\}, \quad N_{\text{sign-flip}} \le \sum_i \mathbf{1}\{|R_i^\star - \theta_R^\star| \le e_{R,i} + e_\theta\}. \tag{49}$$

Let $e_A = |\widehat{A} - A^\star|$, using (37) for $\widehat{A}^{\text{credit}}$. The matched clean reference here uses the same robust functionals; bias relative to the arithmetic-mean objective is a separate target/fidelity term.

Proposition F.15. *Suppose $|A^\star| \le A_{\max}$, $q^\star, \widehat{q} \le Q_{\max}$, and $|m^\star|, |\widehat{m}| \le L$. Set $\overline{Q} = \max\{Q_{\max}, u\}$. Then*

$$|g(\widehat{q}, \widehat{A}) - g(q^\star, A^\star)| \le \overline{Q} e_A + A_{\max} e^L e_m. \tag{50}$$

If both pairs have the same advantage sign, remain in the same differentiable clipping branch, $\|\nabla m^\star\| \le J_{\max}$, and $\|\nabla \widehat{m} - \nabla m^\star\| \le e_J$, then

$$\|\nabla g(\widehat{q}, \widehat{A}) - \nabla g(q^\star, A^\star)\| \le e^L \{e_A(J_{\max} + e_J) + A_{\max}(e_J + J_{\max} e_m)\}. \tag{51}$$

The same-branch premise is guaranteed whenever the margins in Theorem F.14 exceed their errors. No global policy-improvement claim follows at clipping kinks.

Finally, linear leave-one-out centering and bounded credit have different semantics.

Theorem F.16. *Fix a prompt x and let $Y_1, \dots, Y_G$ be independent samples from $\pi_\theta(\cdot \mid x)$. Suppose support is parameter independent and differentiation may pass through integration. If B_i and $S_i > 0$ are measurable with respect to (x, Y_{-i}, U) for independent auxiliary randomness U, assume $\mathbb{E}[S_i^{-1} \mid x] < \infty$ and*

$$\mathbb{E}\left[\|\nabla_\theta \log \pi_\theta(Y_i \mid x)\| \frac{|R(x, Y_i)| + |B_i|}{S_i} \middle| x\right] < \infty.$$

Then

$$\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^G \nabla_\theta \log \pi_\theta(Y_i \mid x) \frac{R(x, Y_i) - B_i}{S_i} \middle| x\right] = w(x) \nabla_\theta \mathbb{E}_{\pi_\theta}[R(x, Y) \mid x], \tag{52}$$

where $w(x) = G^{-1} \sum_i \mathbb{E}[S_i^{-1} \mid x] \in (0, \infty)$. If $S_i \equiv 1$, then $w(x) = 1$; a deterministic $S_i \equiv s$ instead gives $w(x) = 1/s$. Exact off-policy likelihood ratios give the analogous identity under old-policy sampling, with the expectation defining w taken under the old policy.

Equation (52) preserves each prompt's gradient direction, but prompt-dependent $w(x)$ reweights the global objective. Clipping, the robust sequence surrogate, and bounded credit introduce additional changes.

Proposition F.17. *Under the conditions of Theorem F.16, set $U_i = R(x, Y_i) - B_i$ and suppose $S_i \ge s_{\min}$. At a common leave-one-out scale, the numerator-shaping difference between bounded-credit and linear score directions satisfies*

$$\left\| \mathbb{E}\left[\nabla_\theta \log \pi_\theta(Y_i \mid x) \frac{\chi_\kappa(U_i) - U_i}{S_i} \middle| x\right] \right\| \le \frac{1}{2\kappa^2 s_{\min}} \mathbb{E}\left[\|\nabla_\theta \log \pi_\theta(Y_i \mid x)\| \, |U_i|^3 \middle| x\right]. \tag{53}$$

The preceding bound isolates numerator shaping at a common scale. The full $\widehat{A}^{\mathrm{credit}}$ construction also replaces a raw scale by the mapped scale; algebraically,

$$\left|\frac{\chi_\kappa(U)}{S_\chi}-\frac{U}{S_{\mathrm{lin}}}\right| \le \frac{|\chi_\kappa(U)-U|}{S_\chi}+\frac{|U|\,|S_\chi-S_{\mathrm{lin}}|}{S_\chi S_{\mathrm{lin}}},$$

so the scale-construction term must be bounded under additional moments or reported empirically. Thus the error to the original arithmetic objective decomposes into contamination error to the matched robust clean reference, robust-location target bias, numerator-shaping distortion, scale-construction distortion, and smoothing or solver error. Only the first component is addressed by a stability theorem; the others are deliberate surrogate choices.

G Proofs for Failure Propagation and Finite-Group Robustness

This appendix proves the exact GRPO phase diagram, the translation, constant-preservation, and tie-neutral properties of RoVR, and the mixed-contamination theorem. The fixed-group proof evaluates bounded empirical scores at deterministic thresholds, thereby avoiding a curvature event along an unknown contaminated root path.

G.1 Single-reward propagation and the center-only impossibility

Proof of Theorem 2.1. Write $R_i^{(\Delta)} = R_i + \Delta\mathbf{1}\{i=j\}$. The mean identity is immediate. Expanding the centered sum of squares, or applying the one-point variance update, gives

$$\frac{1}{G}\sum_i (R_i^{(\Delta)}-\bar R_\Delta)^2 = \frac{1}{G}\sum_i (R_i-\bar R)^2 + \frac{2\Delta(R_j-\bar R)}{G} + \frac{G-1}{G^2}\Delta^2,$$

which proves (2), including the unchanged floor ε_s.

The centered residuals sum to zero, so $\sum_i A_{i,\Delta}=0$. Moreover,

$$\sum_i A_{i,\Delta}^2 = \frac{\sum_i (R_i^{(\Delta)}-\bar R_\Delta)^2}{G^{-1}\sum_i (R_i^{(\Delta)}-\bar R_\Delta)^2+\varepsilon_s} \le G.$$

If $M=\max_i |A_{i,\Delta}|$, the other $G-1$ values sum to a number of magnitude M. Cauchy–Schwarz gives $\sum_{k\neq i} A_{k,\Delta}^2 \ge M^2/(G-1)$, hence $G \ge M^2G/(G-1)$ and $M \le \sqrt{G-1}$.

As $|\Delta|\to\infty$, (2) gives $s_\Delta \sim |\Delta|\sqrt{G-1}/G$. The focal centered residual is $\Delta(G-1)/G+O(1)$ and every other centered residual is $-\Delta/G+O(1)$. Division proves (3). Finally, $A_{i,\Delta}-A_{k,\Delta} = (R_i-R_k)/s_\Delta \to 0$ for clean i,k. □

Proof of Theorem 2.2. Because $m_\Delta = O(1)$, the focal raw residual is $R_j+\Delta-m_\Delta = \Delta+O(1)$, while every clean residual is $O(1)$. If $d_\Delta = O(1)$, the focal quotient diverges. Conversely, bounded focal credit requires $d_\Delta=\Omega(|\Delta|)$, and division sends every fixed clean residual to zero. For an RMS around m_Δ, the single squared focal residual dominates its average, so $d_\Delta \sim |\Delta|/\sqrt{G}$ and the stated limits follow. □

G.2 Invariance and deterministic root sensitivity

Proof of Theorem 3.1. Adding a to every observation shifts each convex block objective by the change of variable $u \mapsto u+a$. Hence every $\widetilde\mu_b$ and their median shift by a, while all residuals, local sandwich scales, and the arguments of J_0 remain unchanged. Equation (13) therefore shifts by exactly a.

If every observation equals a, each block center and the median equal a. Symmetry gives $\Delta_{K+1-k} = -\Delta_k$. For every noncentral pair,

$$J_0(-\widehat\nu\Delta_k/\sqrt{n_b}) + J_0(-\widehat\nu\Delta_{K+1-k}/\sqrt{n_b}) = 1.$$

When K is odd, the central term is $J_0(0)=1/2$. Thus the sum over k is $K/2=\sum_k \tau_k$ for every block, and the correction is zero. The same pairing applies when $B=1$ because its only block center equals the initializer, so RoVR reduces to that local M-center. □

Proof of Theorem F.1. Replacing at most s observations changes the empirical score uniformly by at most $2sM_c/n$. Since the score derivative is the negative empirical curvature, the mean-value theorem between the two roots gives

$$a_0|T' - T| \le \sup_u |S_D(u) - S_{D'}(u)| \le \frac{2sM_c}{n},$$

which proves (27). Interchanging D, D' gives the alternative condition. □

G.3 Randomized load and the fixed-threshold block tail

Proof of Theorem F.2. For a fixed block of size n_b, S_b is hypergeometric with mean $n_b m/N$. Hoeffding's inequality for sampling without replacement yields

$$\mathbb{P}\Big(S_b - \frac{n_b m}{N} \ge u\Big) \le \exp(-2u^2/n_b).$$

A union bound over blocks proves the first inequality in (29). Since $n_b \le n_{\max}$, each summand is at most $\exp(-2u^2/n_{\max})$, proving the second. □

Proof of Theorem F.4. Apply Theorem F.2 using the block-specific threshold for each block b, $u_b = \sqrt{n_b \log(B/\delta_{\text{part}})/2}$, and then take a union bound. With probability at least $1 - \delta_{\text{part}}$, every realized load is at most $s_b(\delta_{\text{part}})$. Conditional on the sampled partition and this load event, the clean block samples remain independent, the replacement support obeys the deterministic budgets, and Theorem F.3 applies with $q = 0$. Its observable-correction-magnitude and deterministic-envelope statements give the two conditional bounds. A final union bound over the load event proves (32). If the replacement indices were chosen after seeing the partition, this conditioning step would not supply the prescribed independent load event. □

Let $S_b^0(v) = n_b^{-1} \sum_i \psi_c(X_{bi} - v)$ be the clean score and $S_b(v)$ the score after at most s_b replacements. Uniformly in v,

$$|S_b(v) - S_b^0(v)| \le \frac{2s_b M_c}{n_b}.$$

Because S_b is decreasing, $\{T_b > \theta + t\} \subseteq \{S_b(\theta + t) > 0\}$. Therefore

$$\{T_b > \theta + t\} \subseteq \left\{ S_b^0(\theta + t) - \mathbb{E}S_b^0(\theta + t) > at - \frac{2s_b M_c}{n_b} \right\}.$$

Each score is in $[-M_c, M_c]$, so Hoeffding's inequality gives the upper-tail probability $p_b(t)$ in (28); the lower tail is identical.

Proof of Theorem F.3. For every honest block define the clean upper-tail event

$$E_b^+(t) = \left\{ S_b^0(\theta + t) - \mathbb{E}S_b^0(\theta + t) > at - \frac{2s_b M_c}{n_b} \right\}.$$

The fixed-threshold inclusion gives $\{T_b > \theta + t\} \subseteq E_b^+(t)$ regardless of how the allowed replacement coordinates and values are selected after the clean sample is observed. The events $\{E_b^+(t) : b \in \mathcal{H}\}$ depend only on mutually independent latent clean blocks, so they remain independent and have probabilities at most $p_b(t)$. This is why adaptivity of replacement values is allowed but data-dependent selection of $\mathcal{H}$ is not.

If the midpoint median for even B, or the central median for odd B, exceeds $\theta + t$, at least $r_B = \lceil B/2 \rceil$ block centers lie above the threshold. Even if all q arbitrary blocks do so, at least $r_B - q$ clean dominating events must occur. Hoeffding's inequality for independent, non-identically distributed Bernoulli variables gives

$$\mathbb{P}(\widehat{\mu}_0 > \theta + t) \le \exp[-2H\{\eta_q - \overline{p}_H(t)\}^2].$$

The lower-tail argument is identical, so a union bound yields $\mathbb{P}(|\widehat{\mu}_0 - \theta| > t) \le \beta_H(t)$.

On the complementary event, the triangle inequality combines the population-dependent threshold with the observable correction magnitude

$$|\widehat{\theta}_{\mathrm{RoVR}} - \theta| \le t + |\widehat{\theta}_{\mathrm{RoVR}} - \widehat{\mu}_0| = t + \widehat{C}_B.$$

For each block, $0 \le \sum_k J_0(\cdot) \le K$ and $\sum_k \tau_k = K/2$, so its centered composite score has magnitude at most $K/2$. Together with $\widehat{\nu} \le \nu_{\max}$, this gives

$$\widehat{C}_B \le \frac{\nu_{\max}}{D_K W_B} \frac{BK}{2} = C_B.$$

The two probability statements in (31) follow. □

H Proofs for Target, Credit, and Clean Efficiency

This appendix proves the target, bounded-credit, unequal-block oracle variance, scale-orthogonality, and clean transfer results. The arguments keep functional bias, contamination bias, finite-sample error, and credit shaping separate.

H.1 Asymmetry and population contamination

Proof of Theorem F.5. Since $g(\theta_\rho) = 0$, the mean-value theorem and the curvature lower bound give $|g(\mu)| \ge a_0|\mu - \theta_\rho|$, proving (33). For the third-moment statement, let $U = X - \mu$. The inequality $1 - (1+x)^{-1/2} \le x/2$ gives

$$\left| \frac{U}{\sqrt{1 + U^2/c^2}} - U \right| \le \frac{|U|^3}{2c^2}.$$

Since $\mathbb{E}U = 0$, $|\mathbb{E}\psi_c(U)| \le \mathbb{E}|U|^3/(2c^4)$. The dimensionless curvature condition corresponds to $a_0 = \bar{a}_0/c^2$, which proves (34). □

Proof of Theorem F.6. At $\theta_\epsilon = \theta_\rho(P_\epsilon)$,

$$(1-\epsilon)\mathbb{E}_{P_0}\psi_c(X - \theta_\epsilon) + \epsilon \mathbb{E}_Q \psi_c(X - \theta_\epsilon) = 0.$$

The contamination term has magnitude at most M_c, while the clean term has magnitude at least $a_0|\theta_\epsilon - \mu_0|$. Rearranging proves (35). □

H.2 Bounded-advantage stability and retained contrast

Proof of Theorem F.7. The map χ_κ is 1-Lipschitz and lies in $[-\kappa, \kappa]$. Its square takes values in $[0, \kappa^2]$ and is 2κ-Lipschitz. Each of the r replaced rewards can change a squared-credit summand by at most κ^2; each unchanged reward changes by at most $2\kappa e_\theta$ through the center. The two quantities inside the scale square roots therefore differ by at most $r\kappa^2/G + 2\kappa e_\theta$. Since the square root is $1/(2s_{\min})$-Lipschitz on $[s_{\min}^2, \infty)$, (36) follows.

For response i,

$$|\chi_\kappa(R_i - \widehat{\theta}_R) - \chi_\kappa(R_i^\star - \theta_R^\star)| \le \min\{e_{R,i} + e_\theta, 2\kappa\} = d_{\chi,i}.$$

Adding and subtracting the first numerator divided by the oracle scale gives

$$|\widehat{A}_i - A_i^\star| \le \frac{d_{\chi,i}}{s_{\min}} + \kappa \frac{|\widehat{s}_\chi - s_\chi^\star|}{s_{\min}^2},$$

which is (37). The global magnitude bound follows from $|\chi_\kappa| \le \kappa$. Sign preservation uses the sign-preserving map and the fact that an unchanged residual cannot cross zero when its margin exceeds e_θ.

For (39), apply the mean-value theorem to the two residuals. On $[-M, M]$, the derivative is at least $m_{\kappa,M}$. The shared center cancels from their raw difference, and $\widehat{s}_\chi \le \sqrt{s_{\min}^2 + \kappa^2}$. □

Proof of Theorem F.8. On the event in Theorem F.3, the observed center is within $e_\theta(t)$ of the population reference θ. Apply Theorem F.7 to the observed group and the clean precursor centered at θ, then substitute $e_\theta(t)$ into (36) and (37). The theorem's tail probability gives the simultaneous statement. □

Equation (18) follows from

$$|\chi_\kappa(u) - u| = |u| \left\{1 - (1 + u^2/\kappa^2)^{-1/2}\right\} \le \frac{|u|^3}{2\kappa^2}.$$

H.3 Unequal-block oracle variance and scale orthogonality

Proof of Theorem F.9. For block b, standardization makes $Z_b = \sqrt{n_b}(\widetilde{\mu}_b - \theta)/\nu$ standard normal. The block composite score

$$\xi_b = \sum_{k=1}^{K} [J_0(Z_b - \Delta_k) - \tau_k]$$

has mean zero and variance A_K; ties have probability zero. Independence across blocks gives

$$\mathrm{Var}\left(\frac{\nu}{D_K W_B} \sum_b \xi_b\right) = \frac{\nu^2 B A_K}{D_K^2 W_B^2} = \frac{\nu^2 V_K}{N_{\text{eff}}}.$$

□

Proof of Theorem F.10. Because $\Phi(\Delta_k) = \tau_k$, $M_\Phi(0,1) = 0$. Differentiation gives

$$\partial_t M_\Phi(0,1) = \sum_k \phi(\Delta_k) = D_K, \qquad \partial_s M_\Phi(0,1) = \sum_k \Delta_k \phi(\Delta_k).$$

The quantile grid is symmetric, so terms in the last sum cancel pairwise and the central term, if present, is zero. A second-order Taylor expansion on a fixed neighborhood proves (42). □

H.4 Clean efficiency transfer

Proof of Theorem F.12. Let

$$R_{B,n}(t,s) = \frac{1}{B} \sum_{b \notin \mathcal{H}_B} \sum_{k=1}^{K} \{J_0(Z_{b,n} - t - s\Delta_k) - \tau_k\}.$$

Each arbitrary-block composite score has magnitude at most $K/2$, so $|R_{B,n}(t,s)| \le q_B K/(2B) = o(B^{-1/2})$. The exact decomposition is

$$\widehat{M}_{B,n}(t,s) = \frac{H_B}{B} M_n(t,s) + \mathbb{U}_{B,n}(t,s) + R_{B,n}(t,s).$$

For honest blocks, (T3), Theorem F.10, and (T1)–(T2) give

$$M_n(t_0, \widehat{s}) = D_K t_0 + o_{\mathbb{P}}(B^{-1/2}),$$

because $t_0^2 = O_{\mathbb{P}}(B^{-1})$ and $(\widehat{s} - 1)^2 = o_{\mathbb{P}}(B^{-1/2})$. Moreover, $(H_B/B - 1)M_n = -(q_B/B)O_{\mathbb{P}}(B^{-1/2}) = o_{\mathbb{P}}(B^{-1/2})$. Condition (T4) therefore yields

$$\widehat{M}_{B,n}(t_0, \widehat{s}) = D_K t_0 + \mathbb{U}_{B,n}(0,1) + o_{\mathbb{P}}(B^{-1/2}).$$

Substitution into (13) gives

$$\begin{aligned}\widehat{\theta}_{\mathrm{RoVR}} - \theta_\rho = & -\frac{\nu}{\sqrt{n} D_K} \mathbb{U}_{B,n}(0,1) + \frac{\nu(1-\widehat{s})t_0}{\sqrt{n}} \\ & - \frac{\nu(\widehat{s}-1)}{\sqrt{n} D_K} \mathbb{U}_{B,n}(0,1) + o_{\mathbb{P}}((Bn)^{-1/2}).\end{aligned}$$

The second residual is $o_{\mathbb{P}}(B^{-1/2}/\sqrt{n})$ because $(\widehat{s}-1)t_0 = o_{\mathbb{P}}(B^{-3/4})$, and the third has the same order because $\mathbb{U}_{B,n}(0,1) = O_{\mathbb{P}}(B^{-1/2})$. This proves (43).

The honest block score is bounded by K. Under (T5), its covariance converges to the Brownian-bridge covariance and its variance to A_K; also $H_B/B \to 1$ because $q_B = o(\sqrt{B})$. The triangular-array Lindeberg condition is automatic, so the central limit theorem gives variance $\nu^2 A_K/D_K^2$. Finally,

$$D_K/K \to \int_{\mathbb{R}} \phi^2(z)\,dz = \frac{1}{2\sqrt{\pi}}, \qquad A_K/K^2 \to \int_0^1\int_0^1 [\min(u,v)-uv]\,du\,dv = \frac{1}{12},$$

and their ratio tends to $\pi/3$. □

I Proofs for Smoothing and Policy Scope

This appendix proves the tie-aware smoothing, translation calibration, branch-count, policy-stability, leave-one-out, and credit-distortion results. It does not turn the robust token surrogate into an exact likelihood ratio or the bounded advantage into an unbiased raw-reward gradient.

I.1 Smoothing and translation calibration

Proof of Theorem F.13. For $u \neq 0$, $|H_\gamma(u) - J_0(u)| \le e^{-|u|/\gamma}$; at $u = 0$, both equal $1/2$. Summing at most BK nonzero errors and multiplying by the prefactor $\widehat{\nu}/(D_K W_B)$ proves (45). Integrating the logistic-step discrepancy over a continuous margin gives $2\gamma \log 2$, hence the stated $O(\gamma)$ expected error under a bounded continuous density. Atomic mass at zero has zero discrepancy. □

To verify (46), write $M(T) = \mathrm{median}_b\, T_b$ and $u_{bk}(T,\lambda) = T_b - M(T) - \lambda\Delta_k/\sqrt{n}$. The sample median is 1-Lipschitz in input sup norm, so the raw residuals satisfy

$$\max_{b,k} |u_{bk}(T,\lambda) - u_{bk}(T',\lambda')| \le 2d_T + \frac{\Delta_{\max} d_\lambda}{\sqrt{n}}.$$

For $F_\gamma(T,\lambda) = B^{-1}\sum_{b,k}[H_\gamma(u_{bk}(T,\lambda)) - \tau_k]$, symmetry of the quantile grid gives $\sum_k \tau_k = K/2$ and hence $|F_\gamma| \le K/2$. Also, $|H_\gamma'| \le 1/(4\gamma)$ implies

$$|F_\gamma(T,\lambda) - F_\gamma(T',\lambda')| \le \frac{K}{4\gamma}\left(2d_T + \frac{\Delta_{\max} d_\lambda}{\sqrt{n}}\right).$$

Finally, $\mathcal{S}_\gamma(T,\lambda) = M(T) - \lambda F_\gamma(T,\lambda)/(\sqrt{n} D_K)$. Changing the initializer, scale prefactor, and logistic arguments in turn, and using $\lambda, \lambda' \le \lambda_{\max}$, gives (46) for the raw-residual parameterization used by the algorithm.

For (47), translating all token values translates each local convex root and every equivariantly initialized smooth median by the same amount. Residuals, scales, and correction scores are unchanged. Differentiating the translation identity with respect to the common shift gives the gradient-sum identity.

I.2 Branch counts and local policy stability

Proof of Theorem F.14. If $|\widehat{m}_i - m_i^\star| \le e_{m,i}$ and the oracle log ratio is farther than $e_{m,i}$ from both log boundaries, the observed and oracle log ratios lie in the same interval cut out by those boundaries. A union count gives the first inequality. The center sign statement follows because an unchanged residual can cross zero only if its oracle magnitude is at most the center displacement. Summing gives (49). □

Proof of Theorem F.15. For nonnegative ratios, $g(q,A)$ is $\overline{Q}$-Lipschitz in A over the stated domain and $|A|$-Lipschitz in q. The exponential mean-value bound gives $|\widehat{q} - q^\star| \le e^L e_m$, proving (50).

Away from clipping kinks and under the same active branch, the gradient is either zero or $Aq\nabla m$. The log-aggregate conditions imply

$$\|\widehat{q}\nabla\widehat{m} - q^\star\nabla m^\star\| \le e^L(e_J + J_{\max} e_m), \qquad \|\widehat{q}\nabla\widehat{m}\| \le e^L(J_{\max} + e_J).$$

Adding and subtracting $A^\star\widehat{q}\nabla\widehat{m}$ proves (51). □

I.3 Linear leave-one-out direction and bounded-advantage distortion

Proof of Theorem F.16. Condition on (x, Y_{-i}, U). Then B_i, S_i are constant with respect to Y_i and

$$\mathbb{E}[\nabla_\theta \log \pi_\theta(Y_i \mid x) B_i / S_i \mid x, Y_{-i}, U] = 0.$$

The reward term equals $S_i^{-1} \nabla_\theta \mathbb{E}[R(x, Y) \mid x]$. Averaging over Y_{-i}, U and then over i gives (52). Exact likelihood weighting changes old-policy sampling to the same on-policy expectation. □

Proof of Theorem F.17. Equation (18) gives $|\chi_\kappa(U_i) - U_i| \le |U_i|^3/(2\kappa^2)$. Multiply by the score norm, use $S_i^{-1} \le s_{\min}^{-1}$, and apply the triangle inequality inside expectation to prove (53). □