RESIDUAL AGGREGATION AS NONLINEAR SCALARIZATION: EXACT GRADIENT REALLOCATION AND SHARED-ENCODER TRANSMISSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Block-structured objectives can encode the same residual evidence yet allocate gradient mass differently depending on how block losses are scalarized. Let block residuals be r_v and losses $\ell_v = r_v^q$. We compare linear scalarization $\sum_v \alpha_v \ell_v$ with weighted power scalarization $\kappa(\sum_v \beta_v \ell_v^p)^{1/p}$, where $p = m/q$. We derive an exact pairwise log-odds law showing that native-to-composed allocation distortion separates into relative weight mismatch and $(q - m)$ times the log-residual ratio. The law yields matching conditions, exact worst-case distortion and inverse rigidity, exponent identification, and state sensitivity. Power-mean and Jacobian-conditioned Gram analyses connect the same scalarization change to candidate rankings and shared-parameter updates. Common-state audits on Energy, Parkinsons, and CelebA-40 quantify the predicted reallocation and its transmission through shared encoders. In paired 200-epoch regression runs, native $q = 1$ consistently lowers task-1 error, whereas composed $q = 1, m = 2$ lowers task-2 and worst-task error on both datasets. Macro-average ordering differs across datasets, and the Parkinsons contrast largely attenuates under validation-selected stopping. Together, theory and experiments provide a quantitative account of when scalarization geometry produces persistent, task-selective optimization effects.

1 INTRODUCTION

Learning objectives often decompose evidence into blocks such as sensors, modalities, views, tasks, network branches, or residual channels. Once each block produces a nonnegative residual magnitude, the objective must specify how those magnitudes are organized. A native construction evaluates each block separately and sums the losses; a direct-sum construction first combines the residuals through a joint geometry and then applies an outer loss. The two constructions can agree, but they need not. For example,

$$r_a = (1, 1), \qquad r_b = (0, 1.5). \tag{1}$$

An additive $q = 1$ rule prefers r_b, whereas a Euclidean direct-sum rule prefers r_a. The underlying residual evidence is identical; only its cross-block organization changes.

Aggregation can affect a sequence of increasingly downstream objects: objective values, candidate rankings, residual-head gradient allocation, and ultimately shared-parameter updates. We analyze these levels separately and identify when agreement at one level lifts to the next. Our central object is the **gradient-allocation map** $r \mapsto s(r)$, which records the relative residual-head gradient mass assigned to each block at a fixed residual state. The fixed-state comparison is the mechanism-isolating intervention; trajectory consequences are tested separately through paired shared-encoder training. Complete learning trajectories add the downstream update machinery and are treated through an operator-level lifting result in Appendix B.10.

Our main specialization is

$$\mathcal{N}_q(r; \alpha) = \sum_{v \in \mathcal{A}} \alpha_v r_v^q, \qquad \mathcal{C}_{m,q}(r; \beta, \kappa) = \kappa \left(\sum_{v \in \mathcal{A}} \beta_v r_v^m \right)^{q/m}, \tag{2}$$

where $q > 0$ is the loss power, $m > 0$ is the cross-block composition exponent, $\kappa > 0$, and $|\mathcal{A}| \geq 2$.

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.

These are two scalarizations of the same block losses. Setting $\ell_v = r_v^q$ and $p = m/q$ gives

$$\mathcal{N}_q = \sum_v \alpha_v \ell_v, \qquad \mathcal{C}_{m,q} = \kappa \left(\sum_v \beta_v \ell_v^p \right)^{1/p}. \tag{3}$$

Thus $q = m$ is exactly the linear-scalarization boundary $p = 1$; $q \neq m$ compares linear scalarization with a weighted power scalarization. General loss-compositional and shared-norm formulations already contain these forms (Mehta et al., 2012; Dekel et al., 2006; 2007). Our focus is the exact forward and inverse geometry of their induced residual-head allocation maps and its transmission through a shared Jacobian.

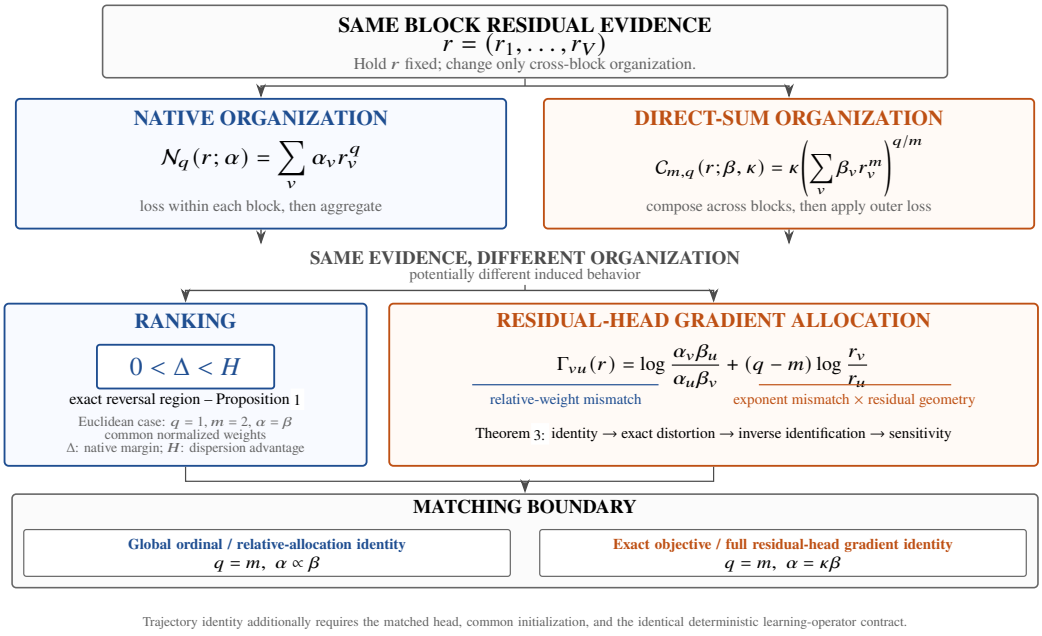


Figure 1: Residual-aggregation mechanism map. Holding the block residual evidence fixed, native and direct-sum organizations can alter ranking and residual-head allocation; the lower strip summarizes the exact matching boundary.

We first identify the representation-level invariance boundary and then characterize the optimization change quantitatively. Define the normalized residual-head gradient shares

$$s_v^N = \frac{|\partial \mathcal{N}_q / \partial r_v|}{\sum_j |\partial \mathcal{N}_q / \partial r_j|}, \qquad s_v^C = \frac{|\partial \mathcal{C}_{m,q} / \partial r_v|}{\sum_j |\partial \mathcal{C}_{m,q} / \partial r_j|}.$$

For positive residuals and active blocks u, v, define their native-to-composed pairwise log-odds distortion by

$$\Gamma_{vu}(r) = \log \frac{s_v^N / s_u^N}{s_v^C / s_u^C} = \log \frac{\alpha_v \beta_u}{\alpha_u \beta_v} + (q - m) \log \frac{r_v}{r_u}$$

so aggregation-induced relative reallocation separates exactly into a weight mismatch and an exponent mismatch modulated by residual geometry. On positive product boxes, this gives exact worst-case distortion and joint inverse rigidity; on an arbitrary attainable set, pairwise relative-residual excitation identifies exponent mismatch. Equivalently, defining $\eta_{\text{state}}(r) := \max_{u,v} |\Gamma_{vu}(r)|$, whenever $\eta_{\text{state}}(r) \leq \tau$, every pairwise block-share ratio changes by a factor in $[e^{-\tau}, e^{\tau}]$. Here $\tau \geq 0$ is a user-chosen tolerance, not an estimated universal threshold. Under proportional weights, the same law becomes an exponential tilt governed by log-residual heterogeneity.

Ranking is the next consequence of the same organization change. For arbitrary positive q, m, native and direct-sum rankings reduce to two generalized power means; in the Euclidean $q = 1, m = 2$

specialization, reversal occurs exactly when a challenger's dispersion advantage exceeds its positive native margin. At shared parameters, the chain rule adds a second geometry: residual-head coefficient changes are projected through block Jacobians, yielding an exact Gram-form characterization of amplification, cancellation, and rotation. The empirical section follows this chain through allocation sensitivity, reversal-region occupancy, Jacobian-conditioned transmission, and paired task-level endpoints.

Contributions. Existing work already permits general loss composition and shared norm-based losses (Mehta et al., 2012; Dekel et al., 2006; 2007). Against that background, our contributions are: (i) an exact forward-and-inverse comparison of the residual-head allocation maps induced by linear and weighted-power scalarization, including a log-odds identity, matching conditions, product-box bounds, inverse identification, and local sensitivity; (ii) an exact connection from the same scalarization change to fixed-candidate ranking, including a Euclidean margin–dispersion criterion; and (iii) a Jacobian-conditioned Gram identity that separates residual-head reallocation from its state-dependent transmission to shared parameters. Common-state and paired endpoint studies then test the chain from allocation to shared updates and task-level trade-offs without claiming universal superiority.

2 Related Work and Setup

Additive representations and rigidity. Additive conjoint measurement and transformed additively separable models study the existence and identification of additive numerical representations (Luce & Tukey, 1964; Fishburn & Roberts, 1988; Jacho-Chávez et al., 2010). The exact equation in our representation layer is a restricted-domain Pexider/additive equation with fixed coordinate maps. Broader exact and Hyers-Ulam stability results are classical (Aczél, 2005; Chudziak & Tabor, 2008; Nikodem, 1991; Kominek, 2004; Chung, 2010). Here the learning architecture fixes the residual transforms and coefficients, and the attainable residual range determines the identifiable portion of the outer link. This gives the representation-level invariance boundary used by the subsequent optimization analysis.

Loss composition and shared-norm objectives. Mehta et al. (2012) introduce a general loss-compositional MTL paradigm that applies an outer function to a vector of task losses and analyze its minimax endpoint through a learning-to-learn bound. Shared-loss formulations based on task-loss norms likewise yield task-update coefficients proportional to powers of the task losses (Dekel et al., 2006; 2007). Equation (3) places our power family inside this established class. These precedents establish composition and loss-dependent coefficients as known ingredients. We complement them with an exact residual-coordinate comparison of linear and weighted-power scalarizations, covering forward distortion, inverse rigidity and identification, ranking consequences, and Jacobian-conditioned transmission.

Scalarization, task balancing, and downstream gradient methods. Mixed-norm learning separates within-group from across-group geometry (Yuan & Lin, 2006), while multi-view and multimodal learning distinguish source-specific from fused organization (Baltrušaitis et al., 2019; Chao et al., 2021). Recent work studies Pareto coverage and large-scale training dynamics under linear weighting (Hu et al., 2023; Royer et al., 2023), and smooth nonlinear scalarizers target Pareto trade-off coverage (Lin et al., 2024). Adaptive loss-weighting methods alter task weights during training (Kendall et al., 2018; Chen et al., 2018; Liu et al., 2023), while multi-objective methods modify shared update directions (Sener & Koltun, 2018; Yu et al., 2020; Navon et al., 2022). Empirical analyses also show that more elaborate multi-task optimizers need not outperform well-tuned scalarization or even random weighting (Kurin et al., 2022; Lin et al., 2022). These findings reinforce the distinction central to our paper: a fixed-state gradient difference is a mechanistic fact, whereas downstream consequences require paired training and task-level evaluation.

Setup and equivalence hierarchy. Let $r = (r_v)_{v\in\mathcal{A}} \in \mathbb{R}_+^{|\mathcal{A}|}$ collect block residual magnitudes, with positive active weights α_v, β_v. The two formulations use the same residual maps and differ only in scalarization. The normalized quantities s_v^N and s_v^C measure relative residual-head gradient mass, so they intentionally remove any common gradient scale; Γ_{vu} compares their pairwise odds. We distinguish objective identity, global ordinal identity, fixed-candidate ranking, relative residual-head allocation, full residual-head gradients, shared-parameter gradients, and complete learning trajectories. Equality at one level does not automatically imply equality downstream. The theory characterizes the

first five levels; the Gram analysis characterizes transmission through a shared Jacobian; Appendix B.10 states sufficient conditions for deterministic trajectory identity.

3 Rigidity and Allocation Theory

Theorems 1–2 close the exact equivalence boundary. Theorem 3 is the main quantitative result: once the organizations are non-equivalent, it characterizes the induced allocation distortion in both forward and inverse directions and identifies its state sensitivity.

3.1 Representation invariance boundary

For continuous strictly increasing block transforms $\phi_v : [0,\infty) \to [0,\infty)$ with $\phi_v(0) = 0$, write $\mathcal{R}_v = \phi_v([0,\infty))$ and let $g : [0,\infty) \to [0,\infty)$ be continuous with $g(0) = 0$. Consider

$$\mathcal{F}_\phi(r;\alpha) = \sum_v \alpha_v \phi_v(r_v), \qquad \mathcal{L}_{\phi,g}(r;\beta) = g\left(\sum_v \beta_v \phi_v(r_v)\right), \tag{4}$$

and let $J = \sum_v \beta_v \mathcal{R}_v$ be the attainable aggregate range.

Theorem 1 (attainable-domain rigidity, compact form). If $|\mathcal{A}| \geq 2$ and active coefficients are positive, then

$$\mathcal{F}_\phi(r;\alpha) = \mathcal{L}_{\phi,g}(r;\beta) \ \forall r \quad \Longleftrightarrow \quad \exists c > 0 : \ g(s) = cs \ \forall s \in J, \ \alpha_v = c\beta_v \ \forall v. \tag{5}$$

Thus exact equality identifies a common linear slope on the attainable aggregate range. A finite-box approximate version gives explicit coefficient and link bounds in terms of the uniform error and attainable interval widths; Supplementary Theorem S1 gives the full statement.

For the power family in Eq. (2), exact and ordinal boundaries become especially simple.

Theorem 2 (power-family boundary, compact form). For $q, m > 0$,

$$\boxed{\begin{array}{c} \mathcal{N}_q \equiv \mathcal{C}_{m,q} \Longleftrightarrow q = m, \ \alpha_v = \kappa\beta_v \ \forall v, \\ \text{global ordinal identity} \Longleftrightarrow q = m, \ \alpha_v = c\beta_v \ \forall v \text{ for some } c > 0. \end{array}} \tag{6}$$

The first line fixes scale; the second preserves only order and therefore allows a common positive rescaling.

3.2 Exact gradient-allocation law and inverse rigidity

Let $e = (e_v)_{v\in\mathcal{A}}$, $e_v = z^{(v)} - c^{(v)}$, $r_v = \|e_v\|_2 > 0$, and set $L_N = \mathcal{N}_q$, $L_C = \mathcal{C}_{m,q}$. Normalizing block-gradient norms gives

$$s_v^N = \frac{\alpha_v r_v^{q-1}}{\sum_u \alpha_u r_u^{q-1}}, \qquad s_v^C = \frac{\beta_v r_v^{m-1}}{\sum_u \beta_u r_u^{m-1}}. \tag{7}$$

The direct-sum outer power contributes a common scale and therefore cancels from relative allocation; its cross-block emphasis is governed by m, while native emphasis is governed by q.

Theorem 3 (gradient-allocation rigidity, identification, and sensitivity). Assume $|\mathcal{A}| \geq 2$, $q, m > 0$, and positive active weights and scale.

(a) Exact log-odds law. For active blocks u, v, define $\Gamma_{vu}(r) = \log[(s_v^N/s_u^N)/(s_v^C/s_u^C)]$. Then for every positive residual vector,

$$\boxed{\Gamma_{vu}(r) = \log\frac{\alpha_v\beta_u}{\alpha_u\beta_v} + (q-m)\log\frac{r_v}{r_u}.} \tag{8}$$

This identity separates block-weight mismatch from exponent mismatch modulated by cross-block residual geometry. Define the statewise distortion $\eta_{\text{state}}(r) := \max_{u,v} |\Gamma_{vu}(r)|$.

(b) Universal identity conditions on the positive residual domain. In particular,

$$\boxed{s^N \equiv s^C \Longleftrightarrow q = m, \ \alpha = c\beta; \qquad \nabla_e L_N \equiv \nabla_e L_C \Longleftrightarrow q = m, \ \alpha = \kappa\beta.} \tag{9}$$

Relative allocation therefore requires proportional weights, whereas equality of the complete residual-head gradient field on this domain additionally fixes scale.

(c) Exact compact-box distortion and inverse rigidity. On $\mathcal{B} = [r_-, r_+]^{|\mathcal{A}|}$, $0 < r_- < r_+$, define $D = \log(r_+/r_-)$ and $\eta_{\mathrm{grad}} = \sup_{r\in\mathcal{B}} \eta_{\mathrm{state}}(r)$. Then

$$\boxed{\eta_{\mathrm{grad}} = |q-m|D + \max_{u,v} \left|\log \frac{\alpha_v \beta_u}{\alpha_u \beta_v}\right|.} \tag{10}$$

Consequently, $\eta_{\mathrm{grad}} \le \varepsilon$ implies

$$|q-m| \le \frac{\varepsilon}{\log(r_+/r_-)}, \tag{11}$$

and all ratios α_v/β_v lie within a multiplicative $e^{\pm\varepsilon}$ factor of a common constant. Thus uniform allocation agreement over a domain with nontrivial relative-residual range identifies both exponent proximity and near-proportional weights.

(d) Near-boundary sensitivity. Under proportional weights $\alpha = c\beta$, let $\delta = q - m$, $p_v = s_v^C(r)$, $x_v = \log r_v$, and $\mu_p = \sum_u p_u x_u$. Native allocation is the exact exponential tilt

$$\boxed{s_v^N = \frac{p_v e^{\delta x_v}}{\sum_u p_u e^{\delta x_u}} = p_v + \delta p_v (x_v - \mu_p) + O(\delta^2).} \tag{12}$$

Writing $R(r) = \max_v x_v - \min_v x_v$,

$$\|s^N - s^C\|_1 \le \frac{|\delta|}{2} R(r), \qquad \left\|s^N - s^C - \delta(p_v(x_v-\mu_p))_v\right\|_1 \le \frac{\delta^2}{4} R(r)^2. \tag{13}$$

The same exponent mismatch can therefore induce a small or large allocation change depending on log-residual heterogeneity.

Corollary 1 (attainable-set identification from relative-residual excitation). Let $\mathcal{S} \subset (0,\infty)^{|\mathcal{A}|}$ be any attainable residual set and define

$$\Delta_{vu}(\mathcal{S}) = \sup_{r\in\mathcal{S}} \log \frac{r_v}{r_u} - \inf_{r\in\mathcal{S}} \log \frac{r_v}{r_u}.$$

If $|\Gamma_{vu}(r)| \le \varepsilon$ for every $r \in \mathcal{S}$ and $\Delta_{vu}(\mathcal{S}) > 0$, then

$$\boxed{|q-m| \le \frac{2\varepsilon}{\Delta_{vu}(\mathcal{S})}.} \tag{14}$$

Exact agreement therefore identifies $q = m$ as soon as one pairwise log-residual ratio varies on the attainable set. Identification is governed by relative-residual excitation, not by the ambient dimension of $\mathcal{S}$; Appendix B.9A gives the proof and centered-log formulation.

Equation (8) makes the allocation geometry explicit: product boxes provide exact worst-case distortion, attainable-set excitation provides inverse exponent identification, and proportional weights yield exponential-tilt state sensitivity. Having characterized how fixed residual evidence reallocates gradient mass, we next ask how the same organization changes candidate order.

3.3 Ranking consequences of aggregation geometry

For common normalized weights $\alpha_v = \beta_v > 0$, $\sum_v \alpha_v = 1$, specialize first to $q = 1, m = 2$ and define

$$A_j = \sum_v \alpha_v r_{jv}, \qquad Q_j = \left(\sum_v \alpha_v r_{jv}^2\right)^{1/2}, \qquad g_j = Q_j - A_j. \tag{15}$$

If $\sigma_j^2 = \sum_v \alpha_v (r_{jv} - A_j)^2$, then $Q_j^2 = A_j^2 + \sigma_j^2$ and, whenever $Q_j + A_j > 0$,

$$g_j = \frac{\sigma_j^2}{Q_j + A_j}, \tag{16}$$

Table 1: Universal matching hierarchy from objective organization to deterministic trajectories.

Level	Universal condition / guarantee	Additional structure for the next level
Exact objective identity	$q = m,\ \alpha = \kappa\beta$	downstream state-transition rules
Global ordinal identity	$q = m,\ \alpha = c\beta$	objective scale and update rule
Relative gradient-allocation identity	$q = m,\ \alpha = c\beta$	common residual-head gradient scale
Full residual-head gradient identity	$q = m,\ \alpha = \kappa\beta$	shared downstream update contract
Complete deterministic trajectory	matched head, common initialization, and identical deterministic state-transition contract are sufficient	–

with $g_j = 0$ in the all-zero case. Thus g_j is exactly the RMS-over-mean correction induced by cross-block dispersion. For a native winner a and challenger b, let $\Delta_{ab} = A_b - A_a > 0$ and $H_{ab} = (g_a - g_b)_+$.

Proposition 1 (exact reversal criterion). The direct-sum rule reverses the native order between a and b if and only if

$$\boxed{0 < \Delta_{ab} < H_{ab}.} \tag{17}$$

Hence reversal requires a sufficiently small native margin together with a positive dispersion advantage large enough to overcome it.

The mechanism is not restricted to $q = 1, m = 2$. For common normalized weights define the power mean $M_p(j) = (\sum_v \alpha_v r_{jv}^p)^{1/p}$, $d_j^{(m,q)} = M_m(j) - M_q(j)$, and $\Delta_{ab}^{(q)} = M_q(b) - M_q(a) > 0$. Native and direct-sum rankings are respectively the rankings of M_q and M_m, so reversal occurs exactly when

$$0 < \Delta_{ab}^{(q)} < d_a^{(m,q)} - d_b^{(m,q)}.$$

The Euclidean specialization remains especially interpretable because $d_j^{(2,1)} = \sigma_j^2/(Q_j + A_j)$, giving the explicit margin–dispersion phase diagram used below. Appendix B.6 gives the full derivation and a distribution-dependent top-1 disagreement bound.

3.4 Matching hierarchy and downstream lifting

The allocation and ranking results occupy different levels of agreement. The hierarchy below records when equality lifts further, from objective or allocation identity to full residual-head gradients and deterministic trajectories.

At $q = m$, direct-sum prototype objectives also separate across blocks (Supplementary Theorem S1); the nonmatched Euclidean stationarity contrast is given by Supplementary Proposition S2, and Appendix B.10 gives the complete deterministic operator-lifting principle. With the equality hierarchy closed, we turn to the nonmatched case and ask how the predicted allocation change appears in learned states and shared encoders.

4 Mechanism Audits and Shared-Encoder Transmission

4.1 Design

Following the theory order, the empirical study uses four controlled stages. **First**, matched $q = m = 2$ two-task models provide 230,400 checkpoint–sample residual observations on Energy and Parkinsons; these are repeated states, not independent learning systems. Reusing each state counterfactually over $q \in [1, 3]$ isolates allocation from representation drift. **Second**, a fixed-center audit measures occupancy of the ranking-reversal region on six multi-view configurations. **Third**, paired supervised systems and CelebA-40/ResNet-18 measure transmission through shared encoders at common states. **Fourth**, 80 frozen training runs (two datasets, ten paired seeds, and four objective configurations)

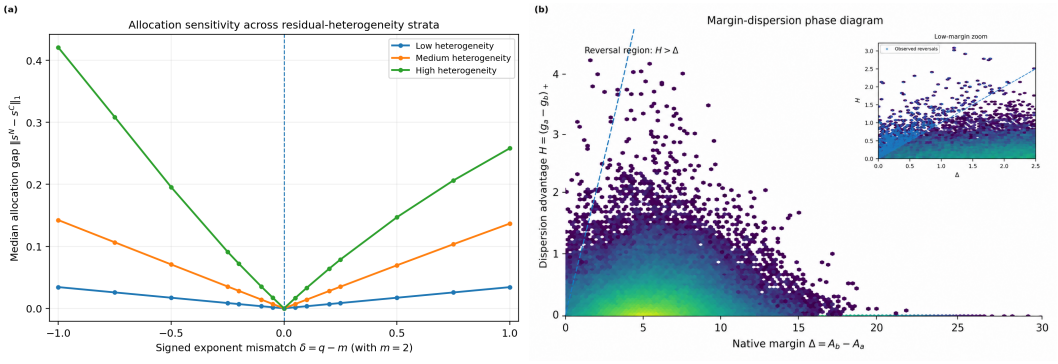


Figure 2: Quantitative mechanisms. (a) Median allocation gap versus signed exponent mismatch $\delta = q - m$ at $m = 2$, stratified by pooled tertiles of the state-level log-residual range $R(r)$. Retaining the sign of δ reveals the asymmetric exponential tilt. (b) Native margin Δ versus dispersion advantage H in the deterministic phase-diagram sample. The dashed line $H = \Delta$ marks the reversal boundary, and the inset enlarges the low-margin region containing the observed reversals. Allocation summaries use the full recovered residual population; the displayed phase density uses a deterministic visualization sample.

provide task-level endpoint and checkpoint-selection comparisons. Appendix C gives full protocols and uncertainty definitions.

4.2 State-dependent allocation audit

The common-state sweep instantiates the state-sensitivity identity on observed residuals: exponent mismatch alone does not determine allocation-change magnitude. We measure state-level residual heterogeneity by the log-residual range $R(r) = \max_v \log r_v - \min_v \log r_v$. At $q = 1, m = 2$, the low-, medium-, and high-heterogeneity tertiles of the pooled recovered-observation population have median L_1 allocation gaps of **0.0343**, **0.1423**, and **0.4207**, respectively. In the two-block equal-weight audit used here, the statewise relation is exact:

$$\|s^N - s^C\|_1 = \tanh\left(\frac{R(r)}{2}\right), \qquad q = 1,\ m = 2,\ \alpha = \beta = (1/2, 1/2).$$

The tertile summaries are numerical effect sizes for the observed, state-weighted residual distribution; because the displayed relationship is an identity, they are not an independent statistical validation of it. Across the evaluated range, the pooled gap increases on both sides of $q = 2$, with the sign dependence predicted by the asymmetric exponential tilt. Near the matched exponent, the normalized per-state remainder over $\delta \in \{\pm 0.05, \pm 0.10, \pm 0.20\}$ has a median log-log slope of **1.9983** (IQR **[1.9978, 1.9985]**). This is a numerical precision and implementation check of the second-order expansion in Eq. (13), rather than a separate empirical law.

4.3 Observed occupancy of the reversal region

In the Euclidean $q = 1, m = 2$ fixed-center controls, top-1 assignment disagreement remains low (approximately **0.3%–3.4%** across configurations), although deeper candidate rankings differ more often. Partitioning the exhaustive reconstructed winner–challenger population by Proposition 1 measures how often observed pairs enter its analytically defined reversal region. Across configurations, **65.3%–95.4%** of native-winner/challenger pairs have no positive dispersion advantage $(H = 0)$. A further **4.5%–34.7%** satisfy $0 < H \leq \Delta$, and only **0.01%–0.54%** fall in $0 < \Delta < H$. Under prespecified within-run margin bins, every observed reversal lies in the lowest native-margin decile. The empirical result is the occupancy of these regions; the equivalence between the pointwise criterion and the objective comparison is algebraic and is treated as an implementation check.

4.4 PAIRED SHARED-ENCODER TRANSMISSION

The supervised bridge tests whether the residual-head difference survives the shared-encoder Jacobian at identical parameter states. Matched controls reproduce the predicted identities across fixed-feature, frozen-embedding, trainable, and supervised systems; detailed counts are reported in Appendix C. For the nonmatched $q = 1, m = 2$ supervised pair, the median common-state shared-encoder relative L_2 gap D_g is 0.099 on Energy and 0.121 on Parkinsons. Per-epoch logs begin with zero parameter-path gap but nonzero encoder-gradient gap, and the median epoch-200 path gap D_θ reaches 0.079 and 0.198, respectively.

The transmission geometry is exact at a common state. Let $c_v^N = \partial L_N/\partial r_v$, $c_v^C = \partial L_C/\partial r_v$, and $\phi_v = \nabla_\theta r_v$ for the shared parameters θ. Then $g_N = \sum_v c_v^N \phi_v$ and $g_C = \sum_v c_v^C \phi_v$. With $\Delta c = c^N - c^C$ and $K_{uv} = \langle \phi_u, \phi_v \rangle$,

$$\boxed{\begin{aligned} &g_N - g_C = \sum_v \Delta c_v \phi_v, \qquad \|g_N - g_C\|_2^2 = \Delta c^\top K \Delta c, \\ &\langle g_N, g_C \rangle = (c^N)^\top K c^C. \end{aligned}}$$

Thus aggregation-induced coefficient changes are amplified, cancelled, or rotated by a state-dependent Jacobian Gram geometry. The paired systems show this transmission at common learned states and its accumulation into positive path separation; Appendix C.4–C.5 gives the repeated-state analysis.

4.5 40-TASK VISUAL SHARED-ENCODER TRANSMISSION

On CelebA-40 (Liu et al., 2015), an ImageNet-pretrained ResNet-18 (He et al., 2016) supplies a higher-dimensional shared representation. Across 2,400 common-state diagnostics, all 480 matched $q = m = 2$ states satisfy $D_s = D_g = 0$; nonmatched states have median residual-head share gap $D_s = 0.149$ and encoder-gradient gap $D_g = 0.317$. At fixed q and checkpoint, $\rho(R, D_g) > 0$ in 56/60 strata. The median native–composed encoder-gradient angle is $11.9°$, indicating directional change together with scale redistribution. Appendix C.10 reports the full protocol, table, and figure.

4.6 PAIRED TASK-LEVEL ENDPOINTS

We re-evaluate the paired Energy and Parkinsons runs at the prespecified 200-epoch budget. Each dataset has ten initialization seeds; native $q = 1$ and composed $q = 1, m = 2$ runs share initialization, data split, and full-batch optimization settings. Errors are test NRMSEs in training-target standard-deviation units. Confidence intervals use 10,000 paired bootstrap resamples of the ten seeds and therefore describe initialization variability on one fixed split, not data-split or population uncertainty.

Table 2: Paired test endpoints at epoch 200. $\Delta = N - C$; negative values favor native and positive values favor composed. Task 1/task 2 are Heating/Cooling for Energy and motor/total UPDRS for Parkinsons.

Dataset	Metric	Native	Composed	Δ [paired 95% CI]
Energy	Task 1	0.3239	0.3292	-0.0053 [$-0.0065, -0.0040$]
	Task 2	0.4434	0.4356	$+0.0077$ [$+0.0067, +0.0087$]
	Macro	0.3837	0.3824	$+0.0012$ [$+0.0007, +0.0019$]
Parkinsons	Task 1	1.1568	1.1851	-0.0283 [$-0.0321, -0.0243$]
	Task 2	1.3995	1.3895	$+0.0100$ [$+0.0048, +0.0150$]
	Macro	1.2782	1.2873	-0.0091 [$-0.0123, -0.0060$]

The fixed-budget result is a consistent task trade-off, not a common winner: native is lower on task 1 and composed is lower on task 2 in both datasets. Because task 2 is also the worse task, composed reduces worst-task NRMSE and the absolute task-error gap in both datasets, whereas the macro-average winner reverses (composed on Energy, native on Parkinsons). The matched $q = m = 2$ control has bitwise-identical trajectories and zero endpoint differences for all 20 dataset–seed pairs.

Checkpoint selection qualifies the fixed-budget result. Energy validation selection remains near epoch 200 and preserves the ordering. On Parkinsons, validation macro-NRMSE worsens during continued optimization, selection chooses epoch 0 in almost every run, and the selected macro difference

shrinks to 0.0006 with a paired interval touching zero. The Parkinsons epoch-200 differences should therefore be read as fixed-budget trajectory effects, not robust model-selected generalization gains. Exploratory mechanism–endpoint correlations are also inconclusive: every bootstrap interval crosses zero (Appendix C.11).

5 Discussion and Limitations

The studied "residual aggregation" is a structured instance of nonlinear scalarization, not a separate objective family. Its residual-coordinate factorization is nevertheless useful because it exposes an exact allocation law: relative reallocation is a weight offset plus $(q - m)$ times a pairwise log-residual ratio. Under proportional weights, $\eta_{\mathrm{state}}(r) = |q - m|R(r)$ makes residual heterogeneity the state-dependent amplifier of exponent mismatch. The power-mean ranking rule and Jacobian Gram identity then show, respectively, when this change can reverse fixed-candidate orderings and how it can be amplified, cancelled, or rotated at shared parameters. The common-state audits establish the mechanism and its transmission; the paired endpoints show that its downstream signature is task reallocation rather than uniform improvement. In both regression datasets the composed objective improves the worse second task at epoch 200, but the macro-average winner changes with the dataset.

Scope and limitations. The exact results concern fixed power-family objectives with positive active residuals and weights; adaptive weighting, nonsmooth boundaries, and learned routing define natural extensions. The 230,400 checkpoint–sample records estimate observed-state effects and are treated as repeated states rather than independent trials. Endpoint uncertainty is quantified over ten paired initialization seeds on one fixed split and training budget, so evaluation across additional splits, architectures, and budgets remains necessary for broader generalization. The Parkinsons results directly illustrate budget dependence: the epoch-200 task trade-off is pronounced, whereas validation-selected differences are small. CelebA complements the endpoint study by isolating common-state transmission in a 40-task encoder. A scale-matched retraining control would further separate relative allocation from overall gradient magnitude. Within this scope, the evidence establishes the allocation mechanism, its shared-encoder transmission, and dataset-dependent task trade-offs.

The next step is to repeat the paired protocol across splits, architectures, and scale-matched controls with model selection fixed in advance, and to extend the theory to constrained attainable sets, stochastic dynamics, and learned aggregation.

6 Conclusion

We give an exact residual-coordinate account of how linear and weighted-power scalarizations allocate gradient mass differently. The log-odds law yields matching conditions, forward bounds, inverse identification, and state sensitivity; the power-mean and Gram analyses connect the same change to fixed-candidate rankings and shared-parameter gradients. Common-state audits quantify the size and transmission of these differences, and paired training reveals a reproducible fixed-budget trade-off between tasks. Checkpoint sensitivity and reversed macro ordering rule out a universal performance claim. The framework therefore identifies when scalarization changes the optimization signal and how that change can appear—selectively rather than uniformly—in task-level outcomes.

AI Use Statement

Generative AI tools were used solely for language polishing, grammar correction, and formatting adjustments during the preparation of this work.

Reproducibility Statement

Appendices A–B provide complete proofs. Appendices C–D document protocols, metrics, deterministic controls, endpoint estimands, and uncertainty. Fixed seeds and paired settings are used where stated, and retained logs and analysis artifacts record the reported quantities.

REFERENCES

János Aczél. Extension of a generalized Pexider equation. *Proceedings of the American Mathematical Society*, 133(11):3227–3233, 2005.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019.

Guoqing Chao, Shiliang Sun, and Jinbo Bi. A survey on multi-view clustering. *IEEE Transactions on Artificial Intelligence*, 2(2):146–168, 2021.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018.

Jacek Chudziak and Józef Tabor. Generalized Pexider equation on a restricted domain. *Journal of Mathematical Psychology*, 52(6):389–392, 2008.

Jae-Young Chung. Stability of functional equations on restricted domains in a group and their asymptotic behaviors. *Computers & Mathematics with Applications*, 60(9):2653–2665, 2010.

Ofer Dekel, Philip M. Long, and Yoram Singer. Online multitask learning. In *Learning Theory*, volume 4005 of *Lecture Notes in Computer Science*, pp. 453–467. Springer, 2006.

Ofer Dekel, Philip M. Long, and Yoram Singer. Online learning of multiple tasks with a shared loss. *Journal of Machine Learning Research*, 8:2233–2264, 2007.

Peter C. Fishburn and Fred S. Roberts. Unique finite conjoint measurement. *Mathematical Social Sciences*, 16(2):107–143, 1988.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yuzheng Hu, Ruicheng Xian, Qilong Wu, Qiuling Fan, Lang Yin, and Han Zhao. Revisiting scalarization in multi-task learning: A theoretical perspective. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

David T. Jacho-Chávez, Arthur Lewbel, and Oliver B. Linton. Identification and nonparametric estimation of a transformed additively separable model. *Journal of Econometrics*, 156(2):392–407, 2010.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Zygfryd Kominek. On Hyers–Ulam stability of the Pexider equation. *Demonstratio Mathematica*, 37 (2):373–376, 2004.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Baijiong Lin, Feiyang Ye, Yu Zhang, and Ivor W. Tsang. Reasonable effectiveness of random weighting: A litmus test for multi-task learning. *Transactions on Machine Learning Research*, 2022.

Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth tchebycheff scalarization for multi-objective optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30479–30509, 2024.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. FAMO: Fast adaptive multitask optimization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

R. Duncan Luce and John W. Tukey. Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1(1):1–27, 1964.

Nishant A. Mehta, Dongryeol Lee, and Alexander G. Gray. Minimax multi-task learning and a generalized loss-compositional paradigm for MTL. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446, 2022.

Kazimierz Nikodem. The stability of the Pexider equation. *Annales Mathematicae Silesianae*, 5: 91–93, 1991.

Amelie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. Scalarization for multi-task and multi-domain learning at scale. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 525–536, 2018.

Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, 2010.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

SUPPLEMENTARY MATERIAL

APPENDIX A. UNIFIED NOTATION AND RESULT MAP

Symbol	Meaning	Scope
$\mathcal{A}$	active block set	$\lvert\mathcal{A}\rvert \geq 2$ in nondegenerate theorems
r_v	nonnegative block residual magnitude	fixed residual map shared by compared formulations
$\mathcal{R}_v$	attainable range of ϕ_v	$\mathcal{R}_v = \phi_v([0,\infty))$ in Theorem 1
ϕ_v	monotone block transform	continuous, strictly increasing in Theorem 1
g	outer link	continuous, $g(0) = 0$ in Theorem 1
J	attainable aggregate range	only this range is identified by exact rigidity
q	loss power	$q > 0$
m	cross-block composition exponent	$m > 0$
α_v	native block weight	positive on active blocks
β_v	direct-sum inner weight	positive on active blocks
κ	positive direct-sum scale	$\kappa > 0$
$\ell_v, \ell_\star, L$	widths of compact attainable box	Theorem 1(b)
ε	uniform approximation error	nonnegative
A_j, Q_j, g_j	native mean residual, RMS residual, dispersion correction	Euclidean $q = 1, m = 2$ with common normalized weights
W_N	native top-1 winner	Proposition 1
Δ_{ab}	native candidate margin	Proposition 1
H_{ab}	positive dispersion advantage	Proposition 1
s_v^N, s_v^C	relative block-gradient shares	Theorem 3
$\Gamma_{vu}(r)$	pairwise gradient-allocation log-odds distortion	Theorem 3(a)
$\eta_{\mathrm{state}}(r)$	statewise maximum pairwise allocation distortion	Theorem 3(a)
η_{grad}	worst-case pairwise allocation distortion on a residual box	Theorem 3(c)
$R(r)$	log-residual dynamic range	Theorem 3(d)
s	complete algorithmic state	Appendix B.10
T	complete deterministic state-transition map	Appendix B.10

Interpretation conventions. "Exact equivalence" means equality for all residual vectors under fixed parameters, and "ordinal equivalence" includes preservation of both strict orders and ties. A fixed-center empirical top-1 match is a dataset-level observation rather than a global equivalence statement. Theorem 1(b) is local to its selected attainable box, and Supplementary Theorem S2 concerns deterministic state-transition maps.

SUPPLEMENTARY THEORETICAL STATEMENTS

These supplementary statements collect supporting results referenced by the main theory and are followed by their proofs in Appendix B.

S.1 THEOREM 1(B): QUANTITATIVE LOCAL RIGIDITY

For each active block choose $\ell_v > 0$ with $[0, \ell_v] \subseteq \beta_v \mathcal{R}_v$, and define

$$a_v = \alpha_v/\beta_v, \qquad \ell_\star = \min_v \ell_v, \qquad L = \sum_v \ell_v. \tag{S.1}$$

If, for some $\varepsilon \geq 0$,

$$\left| \sum_v a_v y_v - g\left(\sum_v y_v \right) \right| \leq \varepsilon \quad \forall (y_v)_v \in \prod_v [0, \ell_v], \tag{S.2}$$

then there exists $c > 0$ such that

$$|a_v - c| \leq \frac{2\varepsilon}{\ell_\star}, \qquad |\alpha_v - c\beta_v| \leq \frac{2\beta_v \varepsilon}{\ell_\star}, \tag{S.3}$$

and

$$\sup_{s \in [0,L]} |g(s) - cs| \leq \varepsilon \left(1 + \frac{2L}{\ell_\star} \right). \tag{S.4}$$

S.2 SUPPLEMENTARY PROPOSITION S1: DIRECT-SUM FIXED-CANDIDATE INVARIANCE

For fixed candidates define

$$S_j = \sum_v \beta_v r_{jv}^m. \tag{S.5}$$

For every $q > 0$,

$$\arg\min_j \mathcal{C}_{m,q}(r_j; \beta, \kappa) = \arg\min_j S_j. \tag{S.6}$$

S.3 PROPOSITION 1: DISPERSION IDENTITY AND DISAGREEMENT BOUND

Under the common normalized-weight assumption in the main text, if $\sigma_j^2 = \sum_v \alpha_v (r_{jv} - A_j)^2$, then for $Q_j + A_j > 0$,

$$g_j = Q_j - A_j = \frac{\sigma_j^2}{Q_j + A_j}; \tag{S.7}$$

when all residuals for candidate j are zero, $A_j = Q_j = \sigma_j = g_j = 0$.

Condition on candidate a being the unique native winner $W_N = a$, and define

$$\Delta_{ab} = A_b - A_a > 0, \qquad H_{ab} = (g_a - g_b)_+. \tag{S.8}$$

Let $\mathcal{D}$ denote top-1 disagreement between the direct-sum and native rules. For arbitrary $\tau_{ab} > 0$ and $p > 0$,

$$\Pr(\mathcal{D} \mid W_N = a) \leq \sum_{b \neq a} \left[\Pr(0 < \Delta_{ab} \leq \tau_{ab} \mid W_N = a) + \frac{\mathbb{E}[H_{ab}^p \mid W_N = a]}{\tau_{ab}^p} \right]. \tag{S.9}$$

Moreover $0 \leq g_j \leq \sigma_j$.

S.4 SUPPLEMENTARY THEOREM S1: UNIVERSAL PROTOTYPE SEPARABILITY

For a finite sample set $\mathcal{I}$, define

$$J_{m,q}(c_1, \ldots, c_V) = \sum_{i \in \mathcal{I}} \left(\sum_v \beta_v \|x_i^{(v)} - c_v\|_2^m \right)^{q/m}. \tag{S.10}$$

Assume $q, m > 0$, at least two active blocks, and $\beta_v > 0$. Call the objective universally block-separable if, for every finite sample set and block-data configuration, there exist functions J_v such that $J_{m,q} = \sum_v J_v(c_v)$. This property holds if and only if $q = m$. At the boundary,

$$J_{m,m}(c_1, \ldots, c_V) = \sum_v \beta_v \sum_{i \in \mathcal{I}} \|x_i^{(v)} - c_v\|_2^m. \tag{S.11}$$

For necessity, the two-block one-sample construction used in Appendix B.7 has mixed derivative

$$\frac{\partial^2}{\partial a \, \partial b} (\beta_a a^m + \beta_b b^m)^{q/m} = q(q-m) \beta_a \beta_b a^{m-1} b^{m-1} (\beta_a a^m + \beta_b b^m)^{q/m-2}, \tag{S.12}$$

which is nonzero for $q \neq m$ at positive residuals.

S.5 Supplementary Proposition S2: block-specific versus shared prototype reweighting

In the Euclidean $q = 1, m = 2$ case with fixed labels, the native objective is

$$J_N = \sum_{i\in\mathcal{I}} \sum_{v} \alpha_v \|x_i^{(v)} - c_v\|_2, \tag{S.13}$$

whose nonzero-residual stationarity contribution in block v is

$$\alpha_v \frac{c_v - x_i^{(v)}}{r_{iv}}, \qquad r_{iv} = \|x_i^{(v)} - c_v\|_2. \tag{S.14}$$

The corresponding joint objective is

$$J_C = \sum_{i\in\mathcal{I}} \left(\sum_{v} \beta_v r_{iv}^2 \right)^{1/2}, \qquad R_i = \left(\sum_{v} \beta_v r_{iv}^2 \right)^{1/2}, \tag{S.15}$$

with stationarity contribution

$$\beta_v \frac{c_v - x_i^{(v)}}{R_i}. \tag{S.16}$$

Thus native fixed-point coefficients are block specific, while the joint denominator is shared across blocks; off-diagonal effects may vanish in degenerate configurations.

S.6 Supplementary Corollary S1: matched-head training-path identity

Consider two deterministic end-to-end systems with identical encoders, initialization, data order, optimizer/update rule and optimizer state, update schedule, candidate and tie handling, prototype-update rules, repair, stopping, and finalization rules. Suppose their only difference is a native/direct-sum residual head satisfying $q = m$ and $\alpha_v = \kappa\beta_v$. Then their losses and gradients agree at every common state, and their parameter, embedding, prototype, label, optimizer, auxiliary, and prediction trajectories are identical.

Appendix B. Proofs

B.1 Proof of Theorem 1(a): exact outer-link rigidity

Let $I_v = \beta_v \mathcal{R}_v$. For a fixed block v, set all residual coordinates to zero except r_v. The equality in Eq. (5) gives

$$g(x) = \frac{\alpha_v}{\beta_v} x \qquad \forall x \in I_v. \tag{B.1}$$

Write $c_v = \alpha_v/\beta_v$. Every I_v is a nondegenerate interval containing zero, so any two active intervals overlap on a nondegenerate interval. On that overlap, the same function g equals both $c_v x$ and $c_u x$, hence $c_v = c_u = c > 0$ for all active blocks.

Now choose attainable $x \in I_a$ and $y \in I_b$ in two blocks and set all remaining coordinates to zero. Using $\alpha_v = c\beta_v$, exact equality yields

$$g(x + y) = cx + cy = c(x + y). \tag{B.2}$$

Repeating the argument over finite Minkowski sums extends $g(s) = cs$ to the full attainable aggregate interval $J = \sum_v I_v$. Conversely, if $g(s) = cs$ on J and $\alpha_v = c\beta_v$, substitution into Eq. (4) gives equality for every residual vector. Values of g outside J never occur in the linked objective and are therefore unconstrained.

B.2 Proof of Theorem 1(b): quantitative local rigidity

Set all coordinates except $y_v = x$ to zero in Eq. (S.2). Then

$$|a_v x - g(x)| \le \varepsilon \quad \forall x \in [0, \ell_v]. \tag{B.3}$$

For two blocks u, v, evaluate at $x = \ell_\star$. By the triangle inequality,

$$|a_v - a_u|\ell_\star \le |a_v \ell_\star - g(\ell_\star)| + |g(\ell_\star) - a_u \ell_\star| \le 2\varepsilon. \tag{B.4}$$

Choose a reference block v_0 and let $c = a_{v_0} > 0$. Then Eq. (S.3) follows.

The Minkowski sum of $[0, \ell_v]$ is $[0, L]$. Thus every $s \in [0, L]$ has a decomposition $s = \sum_v y_v$ with $0 \le y_v \le \ell_v$. Hence

$$\begin{aligned} |g(s) - cs| &\le \left| g\left(\sum_v y_v\right) - \sum_v a_v y_v \right| + \sum_v |a_v - c| y_v \\ &\le \varepsilon + \frac{2\varepsilon}{\ell_\star} \sum_v y_v \le \varepsilon \left(1 + \frac{2L}{\ell_\star}\right), \end{aligned} \tag{B.5}$$

which proves Eq. (S.4).

B.3 Proof of Theorem 2(a): exact power identity

Sufficiency follows immediately from $q = m$ and $\alpha_v = \kappa \beta_v$. For necessity, take $r = te_v$, $t > 0$. Equality gives

$$\alpha_v t^q = \kappa(\beta_v t^m)^{q/m} = \kappa \beta_v^{q/m} t^q,$$

so $\alpha_v = \kappa \beta_v^{q/m}$. Activate two coordinates a, b. After substitution,

$$(\beta_a r_a^m)^{q/m} + (\beta_b r_b^m)^{q/m} = (\beta_a r_a^m + \beta_b r_b^m)^{q/m}. \tag{B.6}$$

Set $x = \beta_a r_a^m$, $y = \beta_b r_b^m$. Free positive variation of the residuals makes x, y range independently over $(0, \infty)$, so

$$x^{q/m} + y^{q/m} = (x + y)^{q/m} \quad \forall x, y > 0.$$

Strict convexity when $q/m > 1$ and strict concavity when $q/m < 1$ rule out both cases. Thus $q = m$, after which $\alpha_v = \kappa \beta_v$.

B.4 Proof of Theorem 2(b): ordinal identity

Assume strict orders and ties coincide globally. Equal $\mathcal{C}_{m,q}$ values therefore imply equal $\mathcal{N}_q$ values, so there exists a strictly increasing φ on the range of $\mathcal{C}_{m,q}$ with

$$\mathcal{N}_q = \varphi \circ \mathcal{C}_{m,q}.$$

Both costs are positively homogeneous of degree q. For any $t > 0$, choose a residual scaling $\lambda = t^{1/q}$. Then

$$\varphi(tx) = t\varphi(x).$$

Hence $\varphi(x) = cx$ on the positive range for some $c > 0$. Ordinal equivalence reduces to proportional exact equivalence, and Theorem 2(a) yields $q = m$ with proportional block weights. Sufficiency is immediate.

A strict reversal exists whenever $q \ne m$. With two equal-weight blocks, take $r = (1, 1)$, $t = (0, c)$. If $q < m$, choose $2^{1/m} < c < 2^{1/q}$; if $q > m$, choose $2^{1/q} < c < 2^{1/m}$. The two constructions then order the pair oppositely.

B.5 Proof of Supplementary Proposition S1

The map $x \mapsto \kappa x^{q/m}$ is strictly increasing on $[0, \infty)$, so it preserves the ordering and ties of the inner statistic S_j. This proves Eq. (S.6).

B.6 Proof of Proposition 1

Under the normalized-weight assumption $\sum_v \alpha_v = 1$, the weighted variance identity gives

$$Q_j^2 = A_j^2 + \sigma_j^2.$$

When $Q_j + A_j > 0$, difference of squares gives

$$g_j = Q_j - A_j = \frac{\sigma_j^2}{Q_j + A_j}.$$

If $Q_j + A_j = 0$, nonnegativity implies that all residuals vanish and hence $A_j = Q_j = \sigma_j = g_j = 0$. If native prefers a, then $A_a < A_b$. The joint rule prefers b exactly when $A_b + g_b < A_a + g_a$, which is equivalent to Eq. (17).

For the probabilistic bound, condition on $W_N = a$, where the native winner is unique. A disagreement requires at least one competitor with

$$0 < \Delta_{ab} \le H_{ab}.$$

For any $\tau_{ab} > 0$,

$$\{0 < \Delta_{ab} \le H_{ab}\} \subseteq \{0 < \Delta_{ab} \le \tau_{ab}\} \cup \{H_{ab} > \tau_{ab}\}.$$

Take a union bound over competitors and apply Markov's inequality to the second event to obtain Eq. (S.9). Finally, when $Q_j + A_j > 0$,

$$g_j = \frac{\sigma_j^2}{Q_j + A_j} \le \sigma_j,$$

because $Q_j \ge \sigma_j$; the same inequality is immediate in the all-zero case. Hence $H_{ab} \le g_a \le \sigma_a$.

General power-mean reversal. The preceding Euclidean criterion is an interpretable specialization of the same algebra for arbitrary positive exponents. For common normalized weights $w_v > 0$, $\sum_v w_v = 1$, define the weighted power mean

$$M_p(j) = \left(\sum_v w_v r_{jv}^p \right)^{1/p}, \qquad p > 0.$$

Native ranking under $\mathcal{N}_q$ is identical to ranking by M_q, while direct-sum ranking under $\mathcal{C}_{m,q}$ is identical to ranking by M_m, because the remaining powers and scales are strictly increasing. If a is a strict native winner over b, let

$$\Delta_{ab}^{(q)} = M_q(b) - M_q(a) > 0, \qquad d_j^{(m,q)} = M_m(j) - M_q(j).$$

Then the direct-sum rule reverses this pair if and only if

$$\boxed{0 < \Delta_{ab}^{(q)} < d_a^{(m,q)} - d_b^{(m,q)}.}$$

Indeed, direct-sum prefers b exactly when $M_q(b) + d_b^{(m,q)} < M_q(a) + d_a^{(m,q)}$, which is equivalent to the displayed criterion. For $q = 1, m = 2$, $M_1 = A$, $M_2 = Q$, and $d_j^{(2,1)} = Q_j - A_j = \sigma_j^2/(Q_j + A_j)$, recovering Proposition 1. The Euclidean specialization is retained in the main text because this correction has the direct ordinary-variance interpretation used by the empirical phase diagram.

B.7 Proof of Supplementary Theorem S1

At $q = m$, Eq. (S.10) expands into Eq. (S.11), proving separability. For necessity, consider two one-dimensional blocks and one sample at $(0, 0)$. Restrict prototypes to $c_a = a > 0$, $c_b = b > 0$, so the two residual magnitudes vary independently and equal a, b. If a universal block-separable representation existed, the mixed derivative with respect to a, b would vanish. Direct differentiation gives Eq. (S.12), which is nonzero for $q \neq m$ and positive residuals. Hence universal prototype-block separability fails outside the matching boundary.

B.8 Proof of Supplementary Proposition S2

Differentiating one native term $\alpha_v \|x_i^{(v)} - c_v\|_2$ at a nonzero residual gives Eq. (S.14). The coefficient multiplying the sample vector in the corresponding fixed-point equation is α_v / r_{iv}, which depends only on block v. Differentiating one joint term gives Eq. (S.16), where the denominator R_i contains residuals from every block. A perturbation to block a therefore changes no native stationarity equation for $b \neq a$, while it generally changes R_i and hence the same sample's coefficient in every joint block.

B.9 Proof of Theorem 3 and Supplementary Corollary S1

For every active block v, let $e_v \neq 0$ and $r_v = \|e_v\|_2$. Differentiating the native term gives

$$\nabla_{e_v} L_N = q\alpha_v r_v^{q-2} e_v,$$

so

$$\|\nabla_{e_v} L_N\|_2 = q\alpha_v r_v^{q-1},$$

and normalization across blocks gives the native share in Eq. (7).

For the direct-sum loss write $S = \sum_w \beta_w r_w^m$. Then

$$\nabla_{e_v} L_C = \kappa q \beta_v S^{q/m-1} r_v^{m-2} e_v,$$

and

$$\|\nabla_{e_v} L_C\|_2 = \kappa q S^{q/m-1} \beta_v r_v^{m-1}.$$

The prefactor is common to every block, so normalization gives the direct-sum share in Eq. (7).

Part (a). For active blocks u, v,

$$\frac{s_v^N}{s_u^N} = \frac{\alpha_v}{\alpha_u} \left(\frac{r_v}{r_u}\right)^{q-1}, \qquad \frac{s_v^C}{s_u^C} = \frac{\beta_v}{\beta_u} \left(\frac{r_v}{r_u}\right)^{m-1}.$$

Dividing these ratios and taking logarithms gives Eq. (8).

Part (b). If $q = m$ and $\alpha_v = c\beta_v$, the common factor c cancels from the normalized native shares, so $s^N = s^C$ for every positive residual vector.

Conversely, assume $s^N(r) = s^C(r)$ for every positive residual vector. Then $\Gamma_{vu}(r) = 0$ for every pair and every such vector. Choosing a positive vector with $r_v = r_u$ in Eq. (8) gives

$$\frac{\alpha_v}{\beta_v} = \frac{\alpha_u}{\beta_u}$$

for every pair, hence $\alpha_v = c\beta_v$ for some $c > 0$. Equation (8) then reduces to

$$0 = (q - m) \log \frac{r_v}{r_u}.$$

Choosing any positive vector with $r_v / r_u \neq 1$ forces $q = m$. This proves the first equivalence in Eq. (9).

At this boundary, $S^{q/m-1} = 1$ and

$$\nabla_{e_v} L_N = qc\beta_v r_v^{q-2} e_v, \qquad \nabla_{e_v} L_C = \kappa q \beta_v r_v^{q-2} e_v,$$

which gives the proportional full-gradient relation. Equality of the complete residual-head gradient fields therefore holds exactly when $c = \kappa$, yielding the second equivalence in Eq. (9). Conversely, universal full-gradient equality implies universal equality of the normalized block-gradient shares, so the allocation converse gives $q = m$ and $\alpha = c\beta$; substitution into the full gradients then forces $c = \kappa$.

Part (c). Let

$$b_{vu} = \log \frac{\alpha_v \beta_u}{\alpha_u \beta_v}, \qquad \delta = q - m, \qquad D = \log \frac{r_+}{r_-}.$$

For a fixed pair, Eq. (8) becomes

$$\Gamma_{vu}(r) = b_{vu} + \delta t, \qquad t = \log \frac{r_v}{r_u}.$$

Because $\mathcal{B}$ is a product box, t ranges over the full interval $[-D, D]$. Hence

$$\sup_{r\in\mathcal{B}} |\Gamma_{vu}(r)| = \sup_{|t|\le D} |b_{vu} + \delta t| = |b_{vu}| + |\delta| D.$$

Taking the maximum over block pairs proves Eq. (10). If $\eta_{\mathrm{grad}} \le \varepsilon$, both nonnegative terms on the right side of Eq. (10) are at most ε. This yields Eq. (11) and also

$$\max_{u,v} \left| \log \frac{\alpha_v \beta_u}{\alpha_u \beta_v} \right| \le \varepsilon.$$

Fixing a reference block v_0 and defining $c = \alpha_{v_0}/\beta_{v_0}$ then gives

$$-\varepsilon \le \log \frac{\alpha_v}{c\beta_v} \le \varepsilon.$$

Exponentiation gives the near-proportionality claim stated after Eq. (11).

Part (d). Suppose $\alpha_v = c\beta_v$, let $\delta = q - m$, and set

$$p_v = s_v^C, \qquad x_v = \log r_v.$$

Then

$$s_v^N = \frac{\beta_v r_v^{m-1} r_v^{\delta}}{\sum_w \beta_w r_w^{m-1} r_w^{\delta}} = \frac{p_v e^{\delta x_v}}{\sum_w p_w e^{\delta x_w}},$$

which proves the exact-tilt identity in Eq. (12); the proportional-weight log-odds specialization follows immediately from Eq. (8).

For the L_1 bound, define the exponential-tilt path

$$p_v(t) = \frac{p_v e^{t x_v}}{\sum_w p_w e^{t x_w}}.$$

Then $p(0) = s^C$, $p(\delta) = s^N$, and with $\mu_t = \sum_w p_w(t) x_w$,

$$p_v'(t) = p_v(t)(x_v - \mu_t).$$

Therefore

$$\|p'(t)\|_1 = \sum_v p_v(t)|x_v - \mu_t| \le \sqrt{\mathrm{Var}_t(X)} \le \frac{R(r)}{2},$$

where the last inequality is Popoviciu's variance bound for a random variable supported on an interval of width $R(r)$. Integrating from 0 to δ proves the L_1 inequality in Eq. (13).

At $t = 0$,

$$p_v'(0) = p_v(x_v - \mu_p), \qquad \mu_p = \sum_w p_w x_w,$$

which gives the first-order term in Eq. (12). Differentiating again,

$$p_v''(t) = p_v(t)\left[(x_v - \mu_t)^2 - \mathrm{Var}_t(X)\right].$$

Thus

$$\|p''(t)\|_1 \le 2\,\mathrm{Var}_t(X) \le \frac{R(r)^2}{2}.$$

Taylor's theorem with integral remainder gives

$$\|p(\delta) - p(0) - \delta p'(0)\|_1 \le \frac{\delta^2}{2} \sup_t \|p''(t)\|_1 \le \frac{\delta^2}{4} R(r)^2,$$

proving Eq. (13). Because $R(r) = 0$ when all block residuals are equal, proportional-weight native and direct-sum allocations can coincide at such a particular state even for $q \ne m$; this is a local coincidence, not universal equivalence.

For Supplementary Corollary S1, under $q = m$ and $\alpha_v = \kappa\beta_v$, Theorem 3 gives equality of loss values and full gradients at every common residual state. If the two deterministic systems also begin from the same complete algorithmic state and share the remaining deterministic state-transition contract, their next states are identical. Induction over update steps gives equality of all subsequent parameter, embedding, prototype, label, optimizer, auxiliary, and prediction states.

B.9A CONSTRAINED ATTAINABLE-SET IDENTIFICATION

The exact distortion formula in Theorem 3(c) uses full relative log-residual variation on a product box. Exponent identification admits a weaker attainable-set condition. Let $\mathcal{S} \subset (0,\infty)^{|\mathcal{A}|}$ be any attainable residual set and fix active blocks u, v. Define

$$\Delta_{vu}(\mathcal{S}) = \sup_{r\in\mathcal{S}} \log \frac{r_v}{r_u} - \inf_{r\in\mathcal{S}} \log \frac{r_v}{r_u}.$$

Attainable-set excitation bound. If $|\Gamma_{vu}(r)| \le \varepsilon$ for every $r \in \mathcal{S}$ and $\Delta_{vu}(\mathcal{S}) > 0$, then

$$\boxed{|q-m| \le \frac{2\varepsilon}{\Delta_{vu}(\mathcal{S})}.}$$

In particular, exact allocation agreement on $\mathcal{S}$, together with variation of a single pairwise log-residual ratio, forces $q = m$. Exact agreement at any state then forces $\alpha \propto \beta$.

To prove the bound, write Eq. (8) as $\Gamma_{vu}(r) = b_{vu} + \delta z_{vu}(r)$, where $\delta = q - m$ and $z_{vu}(r) = \log(r_v/r_u)$. For sequences of attainable states approaching the supremum and infimum of z_{vu}, subtraction cancels b_{vu}, while the two log-odds terms are each bounded in magnitude by ε. Taking the limit gives $|\delta|\Delta_{vu}(\mathcal{S}) \le 2\varepsilon$, proving the bound. When $\varepsilon = 0$, nonzero excitation forces $\delta = 0$; Eq. (8) then reduces to the relative-weight offsets, and exact share agreement gives proportional weights.

An equivalent centered-log formulation makes the low-dimensional case explicit. Let $x(r) = (\log r_v)_v$, let $b_v = \log(\alpha_v/\beta_v)$, and let $P = I - \mathbf{1}\mathbf{1}^\top/|\mathcal{A}|$. Exact share agreement on $\mathcal{S}$ is equivalent to

$$P\{b + (q-m)x(r)\} = 0 \qquad \text{for every } r \in \mathcal{S}.$$

Thus two attainable states with $P[x(r) - x(r')] \neq 0$ already force $q = m$ and then $Pb = 0$, i.e., $\alpha \propto \beta$. Conversely, if every centered log-residual profile is identical on $\mathcal{S}$, exponent mismatch can be absorbed into a fixed relative-weight offset on that set. This is the intrinsic non-identifiable case; the ambient dimension of $\mathcal{S}$ is not itself the obstruction.

B.10 COMPLETE-OPERATOR LIFTING PRINCIPLE (SUPPLEMENTARY THEOREM S2)

Let a complete deterministic state transition be written as

$$T = R \circ U \circ A, \tag{B.7}$$

where A is a decision/assignment map, U contains parameter, prototype, optimizer, and auxiliary-state updates, and R contains deterministic tie resolution, repair, stopping, and finalization.

Supplementary Theorem S2 (complete-operator lifting and assignment converse). Consider two deterministic algorithms T_N and T_C on the same state space and reachable set $\mathcal{R}$. Assume that, at every reachable state, (i) candidate costs are pointwise identical or differ by the same strictly increasing transform whenever only rankings are consumed, and (ii) candidate sets and tie rules agree. Also assume that (iii) parameter/prototype/auxiliary subproblems and deterministic selections agree, and (iv) update order, numerical tolerances, repair, adaptive-state, stopping, and finalization rules agree. Then

$$T_N(s) = T_C(s) \qquad \forall s \in \mathcal{R}, \tag{B.8}$$

and common initialization yields identical finite trajectories. Conversely, if deterministic first-minimizer assignment maps agree for every finite ordered candidate list, every residual configuration, and every permutation of that list, then the underlying costs preserve every strict order and tie. Universal equality of the set-valued argmin correspondence yields the same conclusion without fixing a tie rule.

Under the sufficient hypotheses, the assignment outputs coincide at every reachable state because candidate rankings, candidate sets, and tie handling coincide. The update outputs then coincide because the optimization subproblems and deterministic selections from their solution sets coincide. Repair, adaptive-state, stopping, and finalization maps also coincide. Composition gives Eq. (B.8), and induction gives identical trajectories from a common initial state.

For the assignment converse, suppose two first-minimizer assignment maps agree for every ordered candidate list and every permutation. If two costs strictly order some pair r, t oppositely, a two-element list exposes different minimizers. If one cost ties r, t while the other strictly prefers r, use the ordered list (t, r): the tied cost selects t by the first-minimizer rule, while the strict cost selects r. Thus permutation-complete deterministic equality preserves strict orders and ties. Universal set-valued argmin equality gives the same result without specifying a tie rule.

B.11 Fenchel-dual separability (Supplementary Proposition S3)

For $q, m > 1$ with at least two active blocks, define

$$F_N(x) = \sum_v \alpha_v \|x_v\|_v^q, \qquad F_C(x) = \kappa \left(\sum_v \beta_v \|x_v\|_v^m \right)^{q/m}. \tag{B.9}$$

Writing $q' = q/(q-1)$, $m' = m/(m-1)$, and $c_q = (q-1)q^{-q'}$, the conjugates are

$$F_N^*(s) = c_q \sum_v \alpha_v^{-1/(q-1)} \|s_v\|_{v,*}^{q'}, \tag{B.10}$$

and

$$F_C^*(s) = c_q \kappa^{-1/(q-1)} \left(\sum_v \beta_v^{-1/(m-1)} \|s_v\|_{v,*}^{m'} \right)^{q'/m'}. \tag{B.11}$$

Supplementary Proposition S3 (Fenchel-dual separability). The native conjugate is block-separable for every $q > 1$. The direct-sum conjugate is universally block-separable if and only if $q = m$. This dual observation is supplementary to the main ranking/allocation story.

For $a > 0$, $q > 1$, and any norm $\|\cdot\|$,

$$(a\|\cdot\|^q)^*(s) = (q-1)q^{-q'} a^{-1/(q-1)} \|s\|_*^{q'}.$$

Applying this identity independently to the native blocks gives Eq. (B.10). For the direct-sum term, define

$$\|x\|_{m,\beta} = \left(\sum_v \beta_v \|x_v\|_v^m \right)^{1/m}.$$

Its dual norm is

$$\|s\|_{m',\beta,*} = \left(\sum_v \beta_v^{-1/(m-1)} \|s_v\|_{v,*}^{m'} \right)^{1/m'}.$$

Applying scalar norm-power conjugacy to $\kappa \|x\|_{m,\beta}^q$ yields Eq. (B.11). A nondegenerate function of the form $(\sum_v h_v(s_v))^a$ is universally additive across independent blocks only when $a = 1$. Therefore direct-sum dual separability requires $q'/m' = 1$, which is equivalent to $q = m$.

Appendix C. Experimental and Reproducibility Details

C.1 Fixed-feature matched-control configurations

Configuration	Samples	Views	Clusters	View dimensions
MSRC-v1	210	5	7	24 / 576 / 512 / 256 / 254
ORL	400	3	40	4096 / 3304 / 6750
UCI-Digit	2,000	3	10	216 / 76 / 64
Handwritten	2,000	6	10	216 / 76 / 64 / 6 / 240 / 47
100Leaves	1,600	3	100	64 / 64 / 64
Caltech-5V	1,400	5	7	40 / 254 / 928 / 512 / 1984

Configuration	Samples	Views	Clusters	View dimensions
MNIST-USPS	5,000	2	10	784 / 256
ALOI	10,800	4	100	77 / 13 / 64 / 125
BBCSport	544	2	5	3183 / 3203

UCI-Digit and Handwritten are distinct view subsets of the Multiple Features source and are treated as configurations rather than independent data sources.

C.2 Determinism and exact controls

Every view is standardized independently by one fixed preprocessing routine. All paired operators receive the same transformed data. Initial partitions are hash-checked. Candidate ordering and tie behavior are deterministic. The matched $q = 2$ comparison also uses the deterministic prototype and update contract required by the complete-operator lifting principle in Appendix B.10. The corresponding theorem and implementation controls are regression-tested in the study code.

C.3 Mechanism diagnostics

All fixed-center ADR calculations use one common prototype set and residual tensor. Direct comparison of cluster IDs from independently fitted partitions is avoided because labels are permutation invariant. Zero controls pass at matched powers and in the single-block reduction. The nonmatched fixed-center top-1 ADR is approximately 0.3%–3.4%; candidate Kendall τ is 0.93–0.95; top-3 disagreement is 11%–17%.

Under fixed labels, normalized native-versus-joint prototype displacement is approximately 3%–6%. Native off-diagonal perturbation spillover is zero to numerical precision. Joint spillover is positive in nondegenerate tests because a one-block perturbation changes the shared sample coefficient. The held-out outlier-alignment confirmation contains 160 runs and reproduces a WAI-top contrast of 0.325 after a five-seed pilot value of 0.329.

Native/native (NN) and coupled/coupled (CC) trajectories begin from the same deterministic labels. First-step label divergence is approximately 2%–5%. NN stabilizes in about three iterations and CC in one to three. The experiment supports an interacting trajectory mechanism spanning multiple update components.

C.4 Supervised multi-task transmission protocol and diagnostics

This paired bridge tracks whether the pre-Jacobian aggregation mismatch remains visible in post-Jacobian shared-encoder gradients and paired parameter paths. The supervised bridge uses Energy Efficiency (Tsanas & Xifara, 2012) and Parkinsons Telemonitoring (Tsanas et al., 2010), with two regression targets per sample. A shared MLP encoder feeds two scalar heads. Residual magnitudes are computed per sample and per task before block aggregation, with a smooth magnitude $r = \sqrt{e^2 + 10^{-12}}$. The four aggregation variants correspond to the same power-family constructions used in the main experiments. Both datasets use ten paired seeds, a fixed split seed of 2027, and common initial model states. Training runs for 200 epochs with deterministic CPU float64 execution. Parkinsons is split by subject rather than record. Across the two datasets, all 80 planned primary training runs completed.

Let g_N, g_C denote paired flattened shared-encoder gradient vectors and let θ_N, θ_C denote the corresponding full flattened parameter vectors, with the native branch used as the reference norm. The paired diagnostics are

$$D_g = \frac{\|g_N - g_C\|_2}{\|g_N\|_2 + 10^{-30}}, \qquad D_\theta = \frac{\|\theta_N - \theta_C\|_2}{\|\theta_N\|_2 + 10^{-30}}.$$

For two block-gradient shares s_1, s_2, the paired share difference is $D_s = \frac{1}{2}(|s_{1,N} - s_{1,C}| + |s_{2,N} - s_{2,C}|)$. Common-state diagnostics are evaluated at epochs 0, 20, 50, 100, and 200. For D_s, sample-level share differences are averaged within each checkpoint. The reported dataset summary is the median of those 50 seed–checkpoint means. The gradient diagnostics D_g and encoder-gradient cosine

are each computed once per common-state checkpoint and then summarized by the median across the same 50 checkpoints. The trajectory logs evaluate D_θ for each paired seed at every epoch, and the reported endpoint is the epoch-200 median across the ten paired seeds.

Under the matched $q = m = 2$ exponent-and-scale conditions, stored loss, gradient, parameter-path, and prediction gaps are zero for both datasets. Cross-process checks on Energy seed 0 reproduce identical initialization, final-state, trajectory, and loss hashes for the two matched heads. For the nonmatched pair, the median common-state encoder-gradient cosines are 0.9953 on Energy and 0.9932 on Parkinsons, with minima of 0.9767 and 0.9350. The corresponding median D_g values across the 50 checkpoints are 0.099 and 0.121. The medians of the within-checkpoint mean absolute block-share differences are 0.169 and 0.106. After scalar-value matching at a common state, the median encoder-gradient cosines are 0.932 and 0.963. In the per-epoch trajectory logs, the first logged common parameter state has $D_\theta = 0$ while the mean encoder-gradient relative gap is 0.0896 on Energy and 0.3059 on Parkinsons. The epoch-200 median D_θ values are 0.07884 and 0.19757, corresponding to the rounded endpoint summaries 0.079 and 0.198.

C.5 Common-state transmission analysis

Using the 100 stored supervised common-state checkpoints, we examine how the residual-head allocation discrepancy is transmitted to the shared-encoder gradient field. The audit requires no retraining and leaves the dataset, architecture, split, seeds, and hyperparameters unchanged. Each dataset contributes ten seeds, each observed at epochs 0, 20, 50, 100, and 200. For each checkpoint, let $\bar{D}_s$ denote the stored mean absolute block-share difference and let D_g be the shared-encoder relative gradient gap defined in C.4. Because each seed contributes five repeated checkpoints, common-state point estimates use all stored states. Uncertainty estimates and resampling preserve the repeated-measures structure by treating seeds as clusters. We therefore separate raw, within-epoch, within-seed, and seed-cluster summaries.

Dataset	raw ρ	epoch-adjusted	median within-epoch ρ	median within-seed ρ
Energy	0.596	0.466	0.636	0.700
Parkinsons	0.098	0.460	0.612	0.100

Here ρ is Spearman association. The epoch-adjusted value is obtained by ranking $\bar{D}_s$ and D_g over the 50 checkpoints, residualizing both rank vectors on epoch indicators, and correlating the residuals. A seed-cluster percentile bootstrap resamples the ten seeds with replacement while retaining all five checkpoints for each selected seed. Across 6,000 replicates, the 95% intervals are [0.315, 0.773] (raw) and [0.254, 0.627] (epoch-adjusted) on Energy. On Parkinsons, they are [-0.104, 0.279] (raw) and [0.176, 0.638] (epoch-adjusted). Leave-one-seed-out ranges are 0.534–0.650 and 0.406–0.514 on Energy, and 0.025–0.137 and 0.387–0.535 on Parkinsons, respectively. Thus the Energy association remains visible after conditioning on training stage, while the Parkinsons epoch-adjusted association is stable under leave-one-seed-out variation.

The within-epoch pattern clarifies the Parkinsons result. Its epoch-specific correlations are 0.612, 0.733, 0.891, -0.236, and -0.018 at epochs 0, 20, 50, 100, and 200, respectively. Across training epochs, D_g decreases in all 10 Parkinsons seeds (median Spearman correlation with epoch -0.900), while $\bar{D}_s$ remains near 0.10–0.11 (median correlation with epoch 0.400). The median $\bar{D}_s$ sequence is 0.109, 0.101, 0.102, 0.106, 0.112, whereas the corresponding median D_g sequence is 0.260, 0.128, 0.118, 0.116, 0.078. Hence a similar residual-head allocation discrepancy can be transmitted with different magnitude at different learned states.

The qualitative pattern persists across alternative share and gradient diagnostics. Replacing $\bar{D}_s$ by the stored 90th-percentile share discrepancy gives raw/epoch-adjusted associations of 0.523/0.276 on Energy and 0.094/0.402 on Parkinsons. Replacing D_g by the directional discrepancy $1 - \cos(g_N, g_C)$ gives 0.511/0.368 and 0.091/0.434, respectively. Together, these diagnostics support a learned-state-dependent transmission layer between residual-head allocation and shared-parameter gradients, with its magnitude shaped by the evolving shared-parameter geometry.

C.6 Implementation independence checks

We verify the matched controls with distinct native and direct-sum functions and independently instantiated models and optimizers. Under an equal-weight exponent mismatch, the implementations separate as required. For $r = (1, 3)$, $q = 3$, and $m = 2$, the equal-weight direct-sum branch returns $5^{3/2} \approx 11.1803$, whereas the native branch returns 14. On random positive residual tensors at $q = m = 2$ with matched scaling, the independent generic branches coincide to stored numerical precision.

C.7 Quantitative $q - m$ sweep and residual recovery audit

The quantitative sweep uses the matched `mtl_native_q2` checkpoints at epochs $0, 20, 50, 100, 200$ for both supervised datasets and all ten seeds. Raw per-sample residuals were not stored in the original trajectory logs. We therefore recovered them with deterministic eval-mode float64 CPU forward passes using the frozen model, split, preprocessing, and smooth residual definition $r = \sqrt{e^2 + 10^{-12}}$. The original training logger records its epoch loss before the optimizer step while a checkpoint is saved after that step; the recovery audit therefore compares checkpoint t to stored loss $t + 1$. Across the 100 dataset–seed–checkpoint combinations, recovered matched loss and block shares agree exactly with the stored quantities under this offset convention.

The recovered population contains 26,900 Energy checkpoint–sample observations and 203,500 Parkinsons checkpoint–sample observations, for 230,400 observations in total. Multiple observations share datasets, seeds, checkpoints, and samples and are not independent replicates. Pooled summaries are state-weighted, with 11.68% of observations from Energy and 88.32% from Parkinsons; dataset-stratified conclusions should therefore be preferred whenever uncertainty is reported. Heterogeneity is measured by $R(r) = \log(r_{\max}/r_{\min})$; the pooled tertile cutpoints are 0.160245 and 0.477407. At $q = 1, m = 2$ with $\alpha = \beta = (1/2, 1/2)$, the two-block shares give the exact statewise identity

$$\|s^N - s^C\|_1 = \tanh\left(\frac{R(r)}{2}\right).$$

Accordingly, the pooled low/medium/high ordering at $q = 1$ is determined by $R(r)$ itself, while the population composition determines only the empirical occupancy and numerical effect-size summaries of the heterogeneity bins. These summaries instantiate the identity on observed residuals; they are not independent tests of it. The stored sweep artifacts contain compact pooled summaries for these bins.

The primary q-sweep is counterfactual: the residual tensor is fixed and only the native exponent changes. All 15 q values are evaluated on every eligible state using the recovered residuals directly. The exact-law and bound checks use exhaustive vectorized computations with compact stored summaries.

The exact allocation-law audit has maximum absolute error 1.776×10^{-15}. The first inequality in Eq. (13) has zero violations over 3,456,000 state–q evaluations. The remainder inequality in Eq. (13) has zero violations over 1,382,400 state–δ evaluations. For the local scaling diagnostic, six signed near-boundary deltas are used in per-state OLS fits after normalizing the remainder by $R(r)^2$. The resulting slope distribution has median 1.9983, interquartile range [1.9978, 1.9985], and 10–90% range [1.9966, 1.9985]. A separate synthetic weight perturbation uses the same residual evidence to check the weight-plus-exponent decomposition in Eq. (8); its maximum identity error is 1.78×10^{-15}. These numerical equality and inequality checks audit implementation and floating-point behavior; they are not statistical evidence for the proved statements.

C.8 Proposition 1 reversal-region occupancy

The calibration repeats the prespecified fixed-center diagnostic protocol on 100Leaves, Caltech101-7, Handwritten, MSRC-v1, ORL, and UCI-Digit with seeds 0–9. Because the original residual tensors were not persisted, candidate residuals are reconstructed by re-executing the existing standardized-view fixed-feature fit protocol (`max_iter=3`) and then evaluating the prespecified reversal diagnostic with uniform weights on the fixed seed set.

Across the resulting 60 dataset–seed states, there are 2,201,040 native-winner/challenger pairs. As a code-to-theory consistency check, the algebraically equivalent pointwise criterion $0 < \Delta_{ab} < H_{ab}$

identifies the same 1,268 pairwise reversals, with zero false positives and zero false negatives. The empirical mechanism is summarized by the margin–dispersion distribution below:

Configuration	mean top-1 ADR	$H = 0$	$0 < H \leq \Delta$	$0 < \Delta < H$
100Leaves	0.54%	65.29%	34.70%	0.009%
Caltech101-7	3.04%	84.42%	15.03%	0.544%
Handwritten	1.75%	86.53%	13.26%	0.213%
MSRC-v1	0.86%	85.37%	14.48%	0.143%
ORL	0.40%	95.15%	4.84%	0.010%
UCI-Digit	1.08%	95.36%	4.51%	0.124%

The last three columns are pair-level fractions among native-winner/challenger comparisons; sample-level ADR is reported separately. Using within-run native-margin quantile bins, the average pairwise reversal rate is 1.738% in the lowest 10% margin bin and zero in the four higher bins. The observed occupancy therefore characterizes the margin/dispersion distribution around the exact pointwise reversal region; Eq. (S.9) remains a distribution-dependent upper-bound statement.

C.9 Learned-representation performance context

For context, the following tables report downstream performance across the learned-representation controls. Matched heads are identical, while the nonmatched performance ordering varies with representation and dataset.

Frozen learned embeddings

Dataset	Matched $q = 2$ pair	Native $q = 1$	Direct-sum $q = 1, m = 2$
MSRC-v1	0.5458	**0.5611**	0.5483
Handwritten	0.7121	**0.7695**	0.7192
Caltech101-7	**0.3282**	0.3228	0.3237

Values are mean ARI over ten paired seeds. The matched pair is identical in all 30 runs.

End-to-end reconstruction

Dataset	H1/H2 matched $q = m = 2$	H3 native $q = 1$	H4 direct-sum $q = 1, m = 2$
MSRC-v1	**0.2946**	0.0231	0.0496
Handwritten	0.7815	0.8017	**0.8121**
Caltech101-7	0.3192	**0.3363**	0.3359

Five of ten native and four of ten direct-sum nonmatched MSRC-v1 runs have ARI below 0.01, while other datasets contain nonmatched settings that perform better than the matched control. The resulting performance ordering is therefore dataset- and representation-dependent.

End-to-end contrastive training

Dataset	H1/H2 matched $q = m = 2$	H3 native $q = 1$	H4 direct-sum $q = 1, m = 2$
MSRC-v1	**0.2855**	0.0039	0.0027
Handwritten	**0.7047**	0.6680	0.6760
Caltech101-7	**0.2598**	0.2289	0.2152

These values are not used to claim a performance optimum for the matching condition.

C.10 CelebA-40/ResNet-18 common-state transmission

Data, model, and training. CelebA (Liu et al., 2015) provides 40 binary face attributes. We use the aligned/cropped images and the official partition annotations, retaining a fixed 100,000-image subset of the official training split selected with data seed 2027. The diagnostic panel is a fixed set of 128 official-validation images selected with seed 2028. Images are converted to RGB, resized deterministically to 128×128, converted to tensors, and normalized with the ImageNet statistics associated with the pretrained ResNet-18 weights. No random crop, flip, augmentation, MixUp/CutMix, label smoothing, class reweighting, or task-balancing method is used. The shared encoder is an ImageNet-pretrained ResNet-18 (He et al., 2016), fully fine-tuned, followed by a $512 \to 40$ linear attribute head. The three optimization seeds are 0, 1, and 2. Training uses AdamW with learning rate 10^{-4}, weight decay 10^{-4}, no scheduler or gradient clipping, AMP enabled, actual batch size 32 for all formal seeds, and five epochs. Checkpoints are saved at epochs 0, 1, 2, 3, and 5.

Residual map and common-state counterfactual. For minibatch size B, attribute v, logit z_{iv}, and label y_{iv}, let

$$\ell_v = \frac{1}{B}\sum_{i=1}^{B} \mathrm{BCEWithLogits}(z_{iv}, y_{iv}), \qquad r_v = \sqrt{\ell_v + 10^{-12}}.$$

We set $m = 2$, $\alpha_v = \beta_v = 1/40$, and $\kappa = 1$. Hence

$$\mathcal{N}_2(r) = \mathcal{C}_{2,2}(r) = \frac{1}{40}\sum_v \ell_v + 10^{-12},$$

so the trained matched trajectory is the standard equal-weight mean BCE trajectory up to an additive constant. At each saved checkpoint, diagnostics use the same 32 minibatches of size four from the fixed validation panel and disable AMP. For each common state we evaluate $q \in \{1, 1.5, 2, 2.5, 3\}$ without retraining, producing $3\times5\times32\times5 = 2{,}400$ formal states. We record $R(r)$, $\eta = |q-m|R(r)$, $D_s = \frac{1}{2}\|s^N - s^C\|_1$, the relative shared-encoder gradient gap

$$D_g = \frac{\|g_N - g_C\|_2}{\|g_N\|_2 + 10^{-30}},$$

the encoder-gradient cosine, and the maximum pairwise allocation log-odds distortion. The shared encoder excludes the final $512 \to 40$ attribute head.

Integrity and aggregate calibration. The full $3 \times 5 \times 32 \times 5$ factorial table has no missing or duplicate states. Recomputing η from $|q - m|R(r)$ gives maximum absolute error below 5×10^{-16}. For all 480 matched $q = m = 2$ states, $D_s = D_g = 0$, the maximum $|1-\cos(g_N, g_C)|$ is 2.9×10^{-13}, and the matched objective difference is zero. The exact maximum pairwise log-odds identity agrees with $|q - m|R(r)$ to relative error below 10^{-6}. Across the three optimization seeds, Spearman $\rho(\eta, D_s)$ is 0.782, 0.778, and 0.781, while $\rho(\eta, D_g)$ is 0.797, 0.782, and 0.823. The corresponding median nonmatched effect sizes are $D_s = 0.149$ and $D_g = 0.317$.

Fixed-q/checkpoint decomposition. We compute Spearman association across the 32 diagnostic minibatches separately for each seed–checkpoint–nonmatched-q stratum, so that exponent mismatch and training stage are fixed within each correlation. This yields 60 strata. The median $\rho(R, D_g)$ is 0.324, 56/60 strata are positive, and each of $q = 1, 1.5, 2.5, 3$ has 14/15 positive strata. The per-seed medians are 0.362, 0.209, and 0.324. As a complementary rank-based adjustment, we replace R and D_g within each checkpoint by within-checkpoint rank percentiles and pool the five checkpoints separately for each seed and q; all 12 resulting associations are positive (range 0.18–0.41). These are descriptive repeated-state associations rather than IID significance tests. The same decomposition gives a positive D_s–D_g association in all 60 fixed strata (median Spearman 0.577), further supporting transmission and motivating the Jacobian-conditioned decomposition below.

Jacobian-conditioned transmission. The chain rule clarifies how a scalar share discrepancy D_s is transformed into the shared-encoder gradient discrepancy D_g. At a fixed learned state, let $c_v^N = \partial L_N/\partial r_v$ and $c_v^C = \partial L_C/\partial r_v$, and let $\phi_v = \nabla_\theta r_v$ denote the shared-encoder parameter gradient of

the same block residual scalar. Then

$$g_N - g_C = \sum_v (c_v^N - c_v^C)\phi_v.$$

With $\Delta c = c^N - c^C$ and the state-dependent Gram matrix $K_{uv} = \langle \phi_u, \phi_v \rangle$, this gives the exact identity

$$\boxed{\|g_N - g_C\|_2^2 = \Delta c^\top K \Delta c, \qquad \langle g_N, g_C \rangle = (c^N)^\top K c^C.}$$

Thus post-Jacobian amplification, cancellation, and angular rotation depend jointly on the pre-Jacobian coefficient change and the shared-parameter geometry. The normalized share gap D_s summarizes relative coefficient allocation, while common coefficient scale and cross-block geometry enter through K; accordingly, equal D_s values can map to different D_g. This common-state chain-rule decomposition complements the empirical transmission diagnostics.

Complementary angle/scale decomposition. For nonmatched states, the encoder-gradient angle has median $11.87°$ (IQR $8.54°$–$18.40°$). At fixed q and checkpoint, the median Spearman association between $R(r)$ and $1 - \cos(g_N, g_C)$ is 0.056 (36/60 positive), whereas D_g and $1 - \cos$ remain positively associated in all 60 strata (median 0.641). Because the stored diagnostics contain the relative gap and cosine but not the two branch norms separately, the norm ratio is constrained through the admissible roots below. Writing $d = D_g$, $c = \cos(g_N, g_C)$, and $a = \|g_C\|_2 / \|g_N\|_2$, the identity

$$d^2 = (1 - a)^2 + 2a(1 - c)$$

implies all admissible positive ratios satisfy $a = c \pm \sqrt{c^2 - 1 + d^2}$. We therefore report the branch-independent lower bound $\Delta_{\text{scale}}^{\min} = \min_{a>0} |a - 1|$ over the admissible roots. Its median is 0.197 (IQR 0.119–0.291); fixed-q/checkpoint $\rho(R, \Delta_{\text{scale}}^{\min})$ has median 0.333 with 57/60 positive strata, and all 12 checkpoint-adjusted seed–q associations are positive. This decomposition indicates a substantial scale-reallocation component together with genuine angular rotation. Recovering an exact scale/angle partition would require the two branch norms separately, while downstream task effects arise through the ensuing optimization trajectory.

Table 9: CelebA-40/ResNet-18 common-state summary. The η row reports medians of three seed-level Spearman correlations; the fixed-stratum row summarizes 60 nonmatched seed–checkpoint–q strata.

Diagnostic	Result
Setting	100K training subset; 40 tasks; 3 seeds; 2,400 common states
Matched control	$D_s = D_g = 0$ in all 480 matched states
Nonmatched effect size	median $D_s = 0.149$; median $D_g = 0.317$
Seed-level association with η	$\rho(\eta, D_s) = 0.781$; $\rho(\eta, D_g) = 0.797$
Fixed q and checkpoint	median $\rho(R, D_g) = 0.324$; 56/60 positive
Within-checkpoint rank adjustment	12/12 seed–q positive; range 0.18–0.41

C.11 Paired endpoint evaluation

Purpose and design. This analysis tests whether the common-state allocation and encoder-gradient differences are followed by measurable task-level changes after optimization. It is a re-analysis of 80 completed runs rather than additional training: two datasets, ten initialization seeds, and four objectives (native and composed at the nonmatched $q = 1, m = 2$ setting and the matched $q = m = 2$ control). Within each dataset and seed, objectives use the same initialization, fixed data split, and full-batch update schedule. The shared MLP maps the input through 64- and 32-unit SiLU layers to two scalar heads. Training uses Adam, learning rate 10^{-3}, 200 epochs, float64 CPU computation, and split seed 2027. For a target-standardized scalar error e_v, the positive residual is $r_v = \sqrt{e_v^2 + 10^{-12}}$; weights are $\alpha = \beta = (1/2, 1/2)$ and $\kappa = 1$. Architecture and preprocessing details agree with Appendix C.4.

Estimands and uncertainty. The prespecified fixed-budget estimand is test NRMSE at epoch 200. We report both task errors, their macro average, their maximum (worst task), and the absolute task

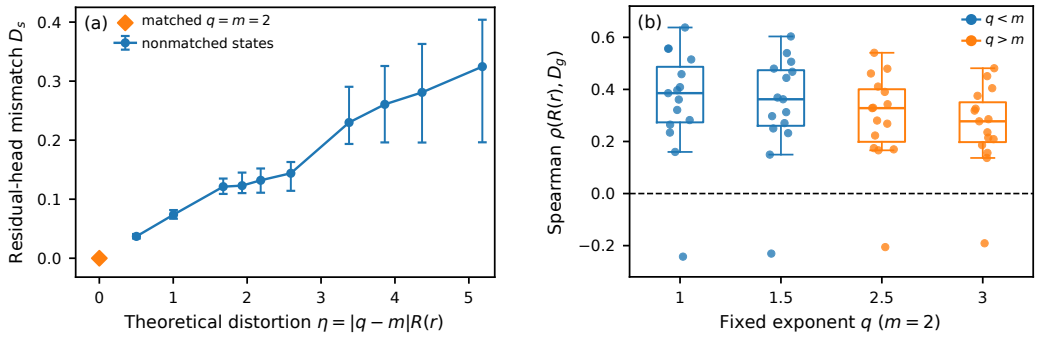


Figure 3: **CelebA-40 common-state audit.** (a) Residual-head mismatch D_s versus theoretical distortion η. (b) Fixed seed–checkpoint–q correlations between residual heterogeneity $R(r)$ and shared-encoder gradient mismatch D_g.

gap. All differences are paired as native minus composed. The validation-selected sensitivity analysis chooses the checkpoint minimizing validation macro-NRMSE within each run, with the earliest epoch breaking ties. Each 95% interval is the percentile interval from 10,000 paired bootstrap resamples of the ten initialization seeds. These intervals quantify initialization variability conditional on the fixed split; they do not quantify variation across data splits or draws from a target population.

Table 10: Full epoch-200 endpoint results. Positive $\Delta = N - C$ favors composed; negative Δ favors native.

Dataset	Metric	Native	Composed	Δ [paired 95% CI]
Energy	Task 1	0.323931	0.329232	-0.005301 [-0.006520, -0.003997]
	Task 2	0.443374	0.435629	$+0.007745$ [$+0.006668$, $+0.008715$]
	Macro	0.383652	0.382430	$+0.001222$ [$+0.000677$, $+0.001860$]
	Worst task	0.443374	0.435629	$+0.007745$ [$+0.006668$, $+0.008715$]
	Absolute task gap	0.119444	0.106397	$+0.013046$ [$+0.011113$, $+0.015014$]
Parkinsons	Task 1	1.156809	1.185059	-0.028250 [-0.032093, -0.024339]
	Task 2	1.399528	1.389512	$+0.010016$ [$+0.004768$, $+0.014998$]
	Macro	1.278168	1.287285	-0.009117 [-0.012305, -0.006007]
	Worst task	1.399528	1.389512	$+0.010016$ [$+0.004768$, $+0.014998$]
	Absolute task gap	0.242719	0.204453	$+0.038266$ [$+0.031951$, $+0.044873$]

At epoch 200, task-1 differences favor native in all ten seeds of both datasets. Task-2 differences favor composed in 10/10 Energy seeds and 9/10 Parkinsons seeds. The absolute task gap is smaller under composed in every seed on both datasets. In contrast, macro-NRMSE favors composed in 9/10 Energy seeds but native in all ten Parkinsons seeds. These sign patterns show systematic reallocation between tasks while excluding a dataset-invariant macro winner.

Table 11: Validation-selected sensitivity. Selection minimizes validation macro-NRMSE separately within each run.

Dataset	Metric	Native	Composed	Δ [paired 95% CI]
Energy	Task 1	0.3239	0.3292	-0.0053 [-0.0065, -0.0040]
	Task 2	0.4434	0.4356	$+0.0077$ [$+0.0067$, $+0.0087$]
	Macro	0.3837	0.3824	$+0.0012$ [$+0.0007$, $+0.0019$]
Parkinsons	Task 1	0.927573	0.924834	$+0.002739$ [0, 0.008217]
	Task 2	1.153105	1.154610	-0.001505 [-0.004516, 0]
	Macro	1.040339	1.039722	$+0.000617$ [0, 0.001850]
	Worst task	1.153105	1.154610	-0.001505 [-0.004516, 0]
	Absolute task gap	0.225532	0.229776	-0.004244 [-0.012733, 0]

Energy selection is effectively the fixed endpoint: nine native runs select epoch 200 and one selects epoch 197, while all composed runs select epoch 200. On Parkinsons, all native runs select epoch 0; nine composed runs select epoch 0 and one selects epoch 6. The validation-selected Parkinsons task and macro intervals therefore touch zero, with nine paired seeds tied. Training error continues to

decrease while validation macro-NRMSE rises from approximately 1.461 at initialization to 1.681 (native) and 1.676 (composed) at epoch 200. We consequently interpret the epoch-200 Parkinsons contrast as a continued-optimization trajectory effect under the fixed compute budget, not as evidence that either endpoint would be selected for deployment.

Trajectory summaries and behavioral separation. The integrated validation macro-NRMSE across 200 epochs is slightly smaller for composed: native-minus-composed AUC differences are 0.1067 [0.0540, 0.1628] on Energy and 0.6587 [0.4394, 0.8850] on Parkinsons. These differences are about 0.14% and 0.20% of the respective AUC levels; on Parkinsons, they mean slightly less accumulated degradation rather than faster improvement. No run produces a NaN or divergence. At epoch 200, the mean absolute native–composed prediction gaps, averaged over the two tasks, are 0.1535 in original target units (0.0158 normalized) on Energy and 0.4462 (0.0513 normalized) on Parkinsons, confirming that path separation reaches the learned predictor.

Controls and mechanism–endpoint bridge. For $q = m = 2$, native and composed objective values, gradients, parameter trajectories, predictions, and endpoints are bitwise identical for all 20 paired dataset–seed combinations. In the nonmatched comparison, both encoder-gradient direction and magnitude change: at epoch 200 the mean native-owner/composed-owner common-state angles are $67.36°/61.23°$ on Energy and $50.35°/36.34°$ on Parkinsons. Because there is no scale-matched retraining control, endpoint differences cannot be attributed solely to relative allocation. Finally, exploratory correlations between common-state mechanism summaries and seed-level endpoint differences are not stable: all 28 paired-bootstrap intervals include zero. We therefore use the bridge analysis only as a negative calibration result and do not claim that the measured mechanism statistic predicts which seed obtains the better endpoint.

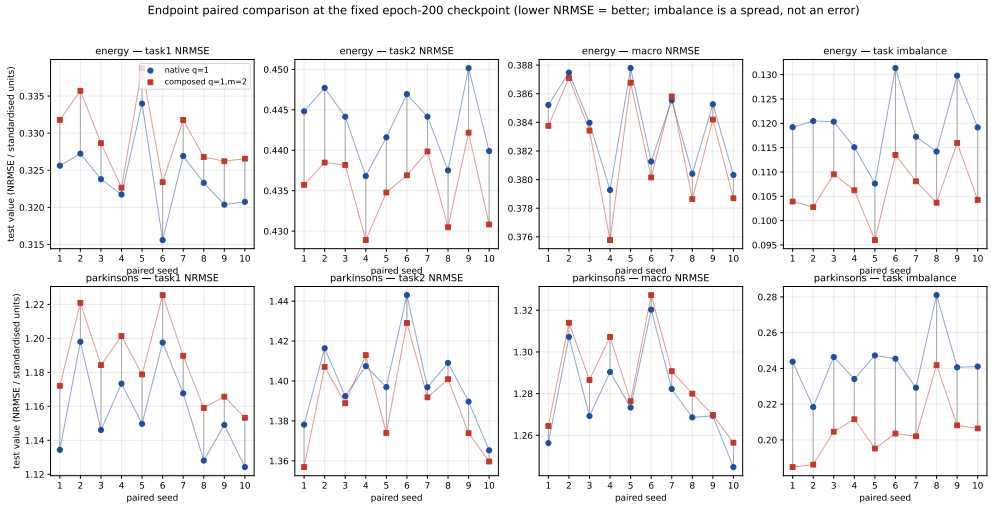


Figure 4: Paired epoch-200 endpoint NRMSE by initialization seed. Grey segments pair native and composed runs sharing initialization and data; panels separate datasets and task-level or aggregate metrics.

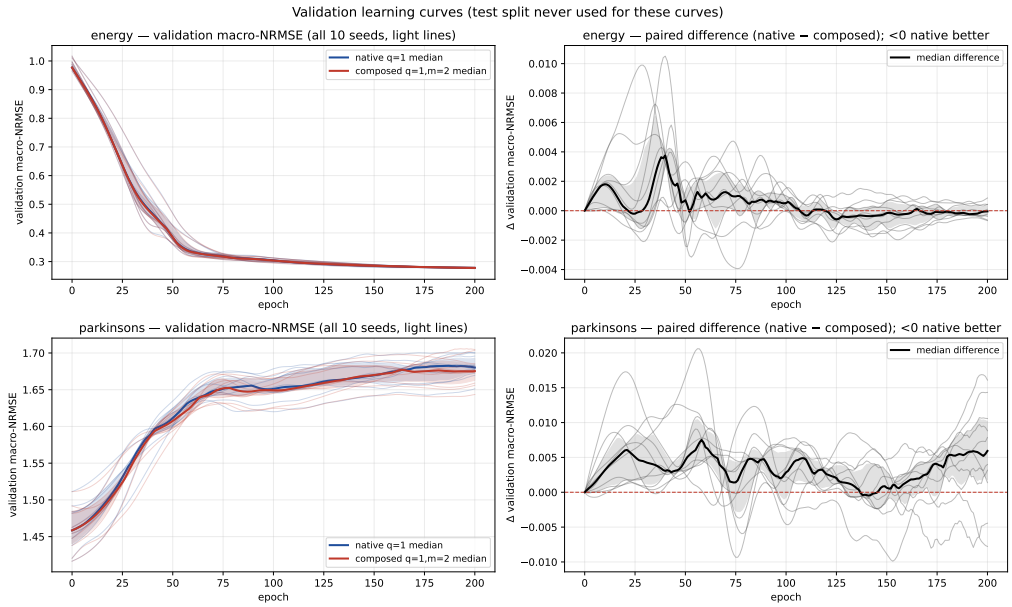


Figure 5: Training and validation macro-NRMSE trajectories. Bands summarize variation over the ten paired initialization seeds. Parkinsons validation performance worsens with continued optimization, motivating the checkpoint-selection sensitivity analysis.

APPENDIX D. DETAILED MSRC-V1 REPRESENTATION DIAGNOSTICS

MSRC-v1 has $n = 210$, $V = 5$, and $K = 7$. Frozen embeddings remain comparatively stable, while trainable nonmatched $q = 1$ heads show substantial instability under the reconstruction and contrastive protocols.

Backbone	Head	ARI mean ± std	Collapsed seeds (ARI < 0.01)
Reconstruction	H1/H2 matched $q = 2$	0.2946 ± 0.0820	0/10
Reconstruction	H3 native $q = 1$	0.0231 ± 0.0345	5/10
Reconstruction	H4 direct-sum $q = 1, m = 2$	0.0496 ± 0.0573	4/10
Contrastive	H1/H2 matched $q = 2$	0.2855 ± 0.1588	0/10
Contrastive	H3 native $q = 1$	0.0039 ± 0.0090	8/10
Contrastive	H4 direct-sum $q = 1, m = 2$	0.0027 ± 0.0095	8/10

The pattern is consistent with an interaction among sample size, alternating deep clustering, and clustering-head geometry, motivating stability-aware scheduling and regularization for nonmatched heads in future studies.