RANK FRONTIERS FOR SHARED LINEAR ENCODERS: PAIRED TASK FILTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

For paired task filters, we derive the exact information-loss frontier of a rank-limited shared linear encoder. At equal overlap, the global optimizer changes exactly from pairing rank units within output coordinates to spreading them across coordinates at $s^\star = 1 + \sigma^2 - \sqrt{\sigma^2(\sigma^2+2)}$. This allocation law is fragile to unequal overlap: an explicit three-output projector beats every block-aligned allocation, while the complete two-output unequal-overlap frontier remains solvable at all five ranks. A held-out learned-subspace study recovers the predicted transition within .0233 nats. These are results for a paired-filter linear–Gaussian model, not for nonlinear encoders or practical architecture selection. Common whitening covers one shared nonspherical covariance; distinct covariances receive only the stated residual results. Capacity-comparison signs are accounting-dependent, and all public-data selectors abstain, so neither supplies an architecture recommendation. Finite temporal histories are one covariance-structured application of the static Gaussian-vector result; we make no new claim about temporal dynamics.

1 INTRODUCTION

1.1 SHARED LINEAR ENCODING WITH PAIRED TASK FILTERS

Applying one linear map to both modalities imposes a common rank-limited subspace. We study how that subspace allocates rank across paired task filters for a common target. Contrastive vision–language training aligns representations (Radford et al., 2021) but does not thereby establish shared encoder weights, so it is motivation rather than an instance of our intervention. We measure neither the prevalence of weight sharing nor how often it is audited; the object analyzed here is the stated linear–Gaussian model.

Only **different task-filter subspaces** are instantiated and formally defined in Section 2. The filters have no assumed "local" or "sequential" semantics. Finite temporal histories fit the model when treated as covariance-structured input vectors (Corollary 1); the frontier does not describe state evolution, forecasting dynamics, or learned time-series architectures.

The central question is: *How do rank, paired task-filter overlap, and covariance determine shared-linear-encoder information loss, and what remains after capacity matching?* Our answer is formal only within the explicit restricted model. Noncoincident task subspaces plus a shared linear rank bottleneck imply a quantified loss; temporal labels alone do not. Separately, a fixed finite-pool probe may abstain from routing, but it is not a certificate of the theorem premises.

1.2 MAIN CONTRIBUTIONS

Our contributions are deliberately separated by logical strength:

1. **Exact pairing-to-spreading transition.** Proposition 2 gives the global unrestricted frontier at every rank for equal principal cosine. A Fantope and principal-angle reduction proves maximal pairing below $s^\star$ and maximal spreading above it; Proposition 4 is the disjoint specialization.
2. **An exact unequal-overlap boundary.** Proposition 3 gives a closed-form three-output projector that beats every block-aligned allocation by 0.0218684394 nats. Proposition 8 solves

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

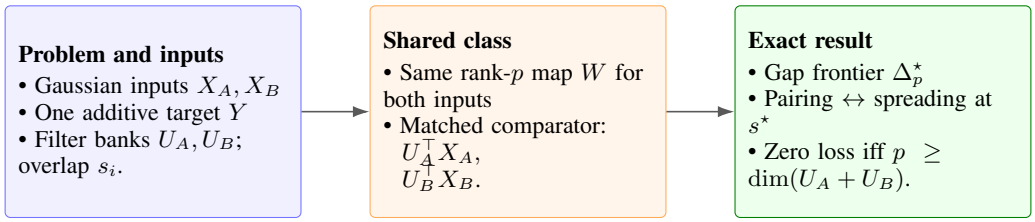

Figure 1: **Whole-paper map.** The input model and tying intervention lead to exact rank frontiers under explicit assumptions. The bottom box is a scope boundary, not an empirical architecture claim.

Table 1: Claim–evidence boundary.

Claim	Evidence	Explicit exclusion
Exact paired-filter variational frontier	Proposition 1; closed forms in Propositions 2, 4	General higher-output unequal-overlap closed form; nonlinear theorem
Distinct-covariance raw tying	Proposition 5; 1,600-projector audit	General closed-form nonzero-rank frontier beyond the diagonal rank-one case
Capacity-matched tied/untied frontier	Appendix B; exhaustive rank audit	Statistical/optimization capacity; convention-independent reuse advantage (App. P)
Conditional omission gap	Lemma 1	Derivation of its premises for learned encoders
Fixed-pool selection	Proposition 9; synthetic calibration; public abstention	i.i.d. coverage for dependent windows or architecture search
Applicability boundary	finite witness PASS; ETTh1 FAIL/NOT_TESTED	Real-data theorem invocation

all five ranks for two outputs, while one bounded three-output cell remains open. Thus the equal-overlap allocation rule is exact but does not silently generalize.

3. **A theorem-targeted learned check.** Randomly initialized shared linear subspaces recover the complete rank-loss curve within .0233 nats and switch from paired to spread allocation across the predicted transition, using held-out data and no analytic initialization.

The variational, covariance, multi-modality, capacity-accounting, and transfer-boundary results are retained as scope analyses in the appendix. In particular, capacity signs reverse under alternative declared conventions, every public selector abstains, and none of those audits supports a practical architecture recommendation.

Scope. The positive theorem and learned experiment cover one common additive target, independent Gaussian modalities, paired task filters, and one shared linear encoder subspace. "Paired" means that $U_A^\top U_B$ is diagonal in the target's fixed output coordinates; it is not a harmless coordinate choice. Common covariance admits the exact rank frontiers after one shared whitening map. Distinct covariances admit the stated raw-weight objective, bounds, zero-loss rank, and diagonal rank-one case, but not a general closed-form nonzero-rank frontier. Nonlinear encoders, independently rotated outputs, and architecture choice on real heterogeneous time series remain outside the result. Public-data runs test only invocation boundaries and stay in the appendix.

Related work. Section 4 positions the restricted linear theorem and fixed-pool selector against multimodal fusion, architecture search, multi-view information theory, and modern time-series encoders. Section 4.1 states the closest prior theorems (Ky Fan, reduced-rank regression, the Gaussian information bottleneck, common principal components, CCA, and shared-subspace multi-task learn-

ing) in this notation and records, result by result, which statements here are corollaries and which are not; Proposition 10 quantifies the excess of the tied frontier over the unconstrained reduced-rank optimum. We make no architecture novelty or field-priority claim.

2 THEORETICAL FOUNDATION

2.1 WHAT COUNTING ALONE PROVES

Fix the standard basis $(e_1, \dots, e_d)$. For an index set $J \subseteq [d]$, write $U_J = \operatorname{span}\{e_j : j \in J\}$ and $Z_J = (Z_j)_{j \in J}$. Thus J is a coordinate set, U_J is a subspace, $|J|$ is a cardinality, and $\dim(U_J \cap V)$ is a rank; these objects are never interchanged.

Lemma 1 (Coordinate-omission lemma)**.** *Let Z_A^*, Z_B^* be matched representations and let $R_m \subseteq J_m \subseteq [d]$ be retained and informative coordinate sets. Suppose shared outputs $\widetilde{Z}_m$ are measurable functions of Z^*_{m,R_m} for both m, and for one modality m,*

$$I(Z^*_{m,J_m \setminus R_m}; Y \mid Z^*_{m,R_m}, Z^*_{\bar{m}}) \geq |J_m \setminus R_m| \kappa. \tag{1}$$

Then

$$I(Z_A^*, Z_B^*; Y) - I(\widetilde{Z}_A, \widetilde{Z}_B; Y) \geq |J_m \setminus R_m| \kappa. \tag{2}$$

Proof. Bilateral measurability and data processing give $I(\widetilde{Z}_A, \widetilde{Z}_B; Y) \leq I(Z^*_{m,R_m}, Z^*_{\bar{m}}; Y)$. The chain rule makes the difference from $I(Z^*_m, Z^*_{\bar{m}}; Y)$ at least the conditional mutual information in the premise; coordinates outside J_m contribute an additional nonnegative conditional-information term. This lemma is coordinate counting after explicit omission; it makes no architectural or temporal claim. □

2.2 A COMMON-TARGET BOTTLENECK FOR A RESTRICTED ARCHITECTURE CLASS

Let $U_A, U_B \in \mathbb{R}^{d \times r}$ have orthonormal columns ($2r \leq d$), with principal cosines $s_1, \dots, s_r$ and mean $\bar{\rho} = r^{-1} \sum_i s_i$. Draw independent modality input vectors $X_A, X_B \sim \mathcal{N}(0, I_d)$ and define the *single common fusion target*

$$Y = U_A^\top X_A + U_B^\top X_B + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_r), \quad \sigma^2 > 0. \tag{3}$$

A shared rank-p encoder applies one $W \in \mathbb{R}^{d \times p}$, $W^\top W = I_p$, to both input vectors. Matched encoders use $U_A^\top X_A$ and $U_B^\top X_B$. Thus both encoded modalities jointly predict the same target, as in multimodal fusion.

Theorem 1 (Common-target shared linear bottleneck)**.** *For every shared rank-r encoder W in the model above,*

$$I(U_A^\top X_A, U_B^\top X_B; Y) - I(W^\top X_A, W^\top X_B; Y) \geq \frac{r}{4}(1 - \bar{\rho}) \log(1 + 2\sigma^{-2}). \tag{4}$$

The bound is positive for noncoincident task subspaces and vanishes honestly at full overlap.

Proof. Put $P_W = WW^\top$ and $P_m = U_m U_m^\top$. Conditional on the shared representations, the target covariance is

$$\Sigma_{Y|W} = \sigma^2 I_r + 2I_r - U_A^\top P_W U_A - U_B^\top P_W U_B. \tag{5}$$

Matched encoders leave covariance $\sigma^2 I_r$, so the information gap is $\frac{1}{2} \log\det(\Sigma_{Y|W}/\sigma^2)$. For $C = 2I_r - U_A^\top P_W U_A - U_B^\top P_W U_B$, the eigenvalues lie in $[0, 2]$. Concavity of $g(x) = \log(1 + x/\sigma^2)$ gives the chord lower bound $g(x) \geq \frac{x}{2} g(2)$. Hence the gap is at least

$$\frac{1}{4} \operatorname{tr}(C) \log(1 + 2\sigma^{-2}). \tag{6}$$

The nonzero eigenvalues of $P_A + P_B$ are $1 + s_i$ and $1 - s_i$. Ky Fan gives $\operatorname{tr}[P_W(P_A + P_B)] \leq r(1 + \bar{\rho})$, hence $\operatorname{tr}(C) \geq r(1 - \bar{\rho})$. □

Output-coordinate alignment for exact overlap formulas. The exact principal-cosine frontiers and calculations invoking them below additionally assume $U_A^\top U_B = \mathrm{diag}(s_1, \ldots, s_r)$ in the fixed output coordinates of the common target. This is a paired-principal- filter model, not a without-loss-of-generality change of basis: independent right rotations of the two filters generally change the target. The preceding trace lower bound and the residual-covariance identity do not need this alignment; disjoint filters satisfy it automatically. Appendix E gives a same-subspace counterexample outside this assumption.

Proposition 1 (Exact paired-filter arbitrary-angle variational frontier)**.** *For $1 \le p \le 2r$, put $D_\pm = \mathrm{diag}(1 \pm s_i)$. The global optimum over all shared rank-p projectors is*

$$\Delta_p^\star = \frac{1}{2} \min_{\substack{0 \preceq A, B \preceq I \\ \mathrm{tr}(A+B)=p}} \log\det\Big[I + \sigma^{-2}\{2I - D_+^{1/2} A D_+^{1/2} - D_-^{1/2} B D_-^{1/2}\}\Big]. \tag{7}$$

A minimizer can be chosen with A, B projectors, so this is exact rather than a block-aligned relaxation. It recovers Propositions 2, 4, and 3. Zero gap occurs exactly when $p \ge \dim(U_A + U_B)$.

Proof. Appendix E.

Proposition 2 (Global equal-overlap rank frontier)**.** *Let all principal cosines equal $s \in [0, 1)$, so U_A and U_B decompose into r mutually orthogonal two-dimensional blocks V_i. Define*

$$c_0 = \tfrac{1}{2}\log(1 + 2\sigma^{-2}), \qquad c_1(s) = \tfrac{1}{2}\log(1 + (1-s)\sigma^{-2}),$$

and let t denote the number of blocks receiving two rank units. Over all *rank-p shared projectors, not only block-aligned ones, the exact minimum gap is*

$$\Delta_p^\star = \min_{\max\{0, p-r\} \le t \le \lfloor p/2 \rfloor} \{(r - p + t)c_0 + (p - 2t)c_1(s)\}.$$

Consequently a block that receives no rank unit contributes $\frac{1}{2}\log(1 + 2\sigma^{-2})$ to the gap, one unit along $U_{A,i} + U_{B,i}$ contributes $\frac{1}{2}\log(1 + (1-s)\sigma^{-2})$, and two units contribute 0; cross-block mixing cannot improve these allocations. The minimizing endpoint is $t = \lfloor p/2 \rfloor$ for $s < s^\star$, $t = \max\{0, p - r\}$ for $s > s^\star$, and every feasible t ties at

$$s^\star(\sigma^2) = 1 + \sigma^2 - \sqrt{\sigma^2(\sigma^2 + 2)}, \tag{8}$$

so the allocation pairs as much as possible below $s^\star$ and spreads as much as possible above it. The optimal value is nevertheless nonincreasing in s. At $s = 0$ this reduces to Proposition 4.

Proof. The Fantope reduction gives projector pairs A, B with captured covariance $(1+s)A + (1-s)B$. A principal-angle block contributes determinant factor $M^2 - 2M + (1 - s^2)\sin^2\theta$, where $M = \sigma^2 + 2$. It is minimized by aligned subspaces, giving the displayed allocations. Their affine slope in t, $c_0 - 2c_1(s)$, gives the transition; each allocation cost is nonincreasing in s. The complete extreme-point and principal-angle proof is in Appendix F. □

Proposition 3 (Cross-block mixing can help under unequal overlap)**.** *Let $r = p = 3$, $\sigma^2 = 1/4$, and principal cosines $(s_1, s_2, s_3) = (.05, .45, .85)$. The explicit rank-three projector in Appendix G has gap 1.3117456638 nats, strictly below the exhaustive block-aligned minimum 1.3336141033 nats. Thus cross-block mixing improves on the block-aligned optimum by at least 0.0218684394 nats, and the equal-cosine hypothesis in Proposition 2 is necessary.*

Proof. Direct substitution gives the two-block determinant minimized by $x^\star$; Appendix G gives the exact polynomial and the complete block-allocation enumeration. □

Proposition 4 (Exact disjoint-filter rank frontier)**.** *If $U_A^\top U_B = 0$ and $0 \le p \le 2r$, let $q = \lfloor p/2 \rfloor$ and $b = p - 2q \in \{0, 1\}$. Among all shared rank-p maps, the exact minimum common-target information gap is*

$$\Delta_p^* = \frac{1}{2}\left[(r - q - b)\log(1 + 2\sigma^{-2}) + b\log(1 + \sigma^{-2})\right]. \tag{9}$$

Thus each two rank units completely recover one target coordinate across both modalities; an odd unit can only halve the next coordinate's residual variance.

Proof. The captured eigenvalues lie in $[0, 2]$ and sum to at most p. Concavity and monotonicity reduce the relaxed minimum to q twos, one b and zeros. Taking q complete modality pairs and one additional direction when $b = 1$ realizes that spectrum, so the relaxation is tight. Appendix R.2 preserves the full proof. □

Corollary 1 (Dependent finite-horizon temporal processes)**.** *Let the independent modality histories satisfy $X_m \sim \mathcal{N}(0, \Sigma)$ for one positive-definite, possibly nonspherical Σ. Suppose $U_m^\top \Sigma U_m = I$ and constrain the shared map by $W^\top \Sigma W = I$. With $Q_m = \Sigma^{1/2} U_m$ and $V = \Sigma^{1/2} W$, every theorem and exact frontier above holds unchanged, with principal cosines given by the singular values of $Q_A^\top Q_B$; the exact paired-filter frontiers additionally require $Q_A^\top Q_B = \mathrm{diag}(s_1, \ldots, s_r)$ in the fixed target coordinates.*

Proof. Whitening sends each history to $\widetilde{X}_m = \Sigma^{-1/2} X_m \sim \mathcal{N}(0, I)$, each task score to $Q_m^\top \widetilde{X}_m$, and both encoded modalities to the same map $V^\top \widetilde{X}_m$. Thus it is a bijective reparameterization of the stated model. If the modality covariances differ, their whitening maps generally differ, so equal raw weights become different whitened weights; the tying intervention is then not preserved. □

Accounting boundary. The capacity comparator and every convention-sensitive count are moved to Appendices B and P. Their signs belong to declared accounting rules, not to statistical or optimization capacity, and are not architecture evidence.

Proposition 5 (Distinct covariances with one raw shared map)**.** *Let $X_m \sim \mathcal{N}(0, \Sigma_m)$ independently, with $U_m^\top \Sigma_m U_m = I$, and apply the same full-rank raw matrix W to both modalities. Define*

$$C(W) = 2I - \sum_{m \in \{A,B\}} U_m^\top \Sigma_m W (W^\top \Sigma_m W)^{-1} W^\top \Sigma_m U_m.$$

Then

$$\Delta_p^\star = \min_{\mathrm{rank}(W)=p} \tfrac{1}{2} \log\det(I + \sigma^{-2} C(W)).$$

If $C_p^\star = \min_{\mathrm{rank}(W)=p} \mathrm{tr}\, C(W)$, then

$$\tfrac{1}{4} \log(1 + 2\sigma^{-2}) C_p^\star \le \Delta_p^\star \le C_p^\star / (2\sigma^2).$$

Moreover $\Delta_p^\star = 0$ iff a rank-p subspace contains $\mathrm{span}(U_A, U_B)$, so the zero-loss rank remains exactly $\dim(U_A + U_B)$ for arbitrary positive-definite Σ_A, Σ_B.

Proof. Appendix C.

Corollary 2 (Exact diagonal rank-one frontier)**.** *Let $r = 1, d = 2$, $\Sigma_A = \mathrm{diag}(a, c)$, $\Sigma_B = \mathrm{diag}(d, b)$, $U_A = e_1/\sqrt{a}$, $U_B = e_2/\sqrt{b}$, and $W \propto (\sqrt{t}, 1)^\top$. The residual is*

$$h(t) = 2 - \frac{at}{at + c} - \frac{b}{dt + b}.$$

Its global minimum is the least value at $t = 0$, $t = \infty$, and, when its denominator is nonzero and its value is positive,

$$t^\star = \frac{\sqrt{bd}\, c - \sqrt{ac}\, b}{\sqrt{ac}\, d - \sqrt{bd}\, a}.$$

Thus $\Delta_1^\star = \frac{1}{2} \log(1 + \min_t h(t)/\sigma^2)$ is closed form. If $\Sigma_A = \Sigma_B$ ($a = d, c = b$), then $h(t) \equiv 1$, exactly recovering the common-covariance disjoint frontier.

Proof. Appendix D.

2.3 Geometry and a one-direction example

Figure 1 summarizes the model and its scope; at the orthogonal endpoint $\bar{\rho} = 0$, the target subspaces occupy $2r$ directions, but a shared encoder supplies one rank-r projector to each modality. The trace identity $\mathrm{tr}[P_W(P_A + P_B)] \le r$ is therefore an allocation budget: alignment assigned to U_A cannot also align with the orthogonal U_B. This argument is about shared rank, not optimization failure.

For $r = 1$, let $U_A = e_1$, $U_B = e_2$, $Y = X_{A,1} + X_{B,2} + \varepsilon$, and $w = \cos\theta\, e_1 + \sin\theta\, e_2$. Conditioning on $(w^\top X_A, w^\top X_B)$ leaves target variance $\sigma^2 + 2 - \cos^2\theta - \sin^2\theta = \sigma^2 + 1$. Thus every shared rank-one allocation has gap $\frac{1}{2}\log(1+\sigma^{-2})$, while rank two spans both filters and has zero gap. This is the common-target fusion conflict without matrix notation.

Scope of the bridge. Theorem 1 is a static result for one additive common target, independent Gaussian input vectors, paired linear task filters, and one shared linear subspace. Corollary 1 applies it to finite-history vectors with one common covariance by whitening; neither statement models state evolution or establishes a new temporal-dynamics law. The result does not follow merely from autocorrelation labels, does not cover modality-specific covariance whitening or arbitrary nonlinear encoders, and does not equate a raw parameter budget with retained information. Table 16 in the same appendix retains every theorem-invocation gate; the ETTh1 representation cannot invoke the law.

2.4 A SAMPLE-SPLIT FIXED-FEATURE PROBE

The fixed-pool selector is separate from the rank-frontier theorem. Appendix R.3 retains its sample split, Hoeffding guarantee, 5,000-seed calibration and engineering fallback. Failure to certify a winner does not establish near-optimality of a symmetric encoder.

3 THEOREM-TARGETED LEARNED SHARED-SUBSPACE FRONTIER

This experiment targets Proposition 2 directly: it learns the same shared linear subspace, varies only rank and equal principal cosine across the proved transition, and compares held-out information loss to the exact frontier. It does not compare practical architecture families or learn temporal dynamics. Each observation consists of two independent six-dimensional Gaussian vectors with the same Toeplitz covariance $\Sigma_{ij} = .6^{|i-j|}$, which can represent a six-step AR(1) history but is sampled independently here. Known rank-three task banks have equal principal cosine .25 or .8, on opposite sides of the pairing/spreading transition. For every rank 1–6 and five seeds, 4,000/1,500/4,000 train/validation/test observations are disjoint.

For reproducibility, a raw $6 \times p$ parameter matrix A is mapped to $Q = \mathrm{qr}(A)$ and applied as $M = \mathrm{blockdiag}(Q, Q)$. For each of four random Gaussian starts, an unrestricted common-target readout B is fit from training second moments with ridge $10^{-9}I$, and Powell minimizes $\frac{1}{2}\log\det\widehat{\Sigma}_{Y-MB,\mathrm{train}}$ (at most 400 iterations, `xtol=1e-7`, `ftol=1e-9`). Validation log determinant selects the restart; test is evaluated once. The reported held-out statistic is $\widehat{g} = \frac{1}{2}[\log\det\widehat{\Sigma}_{\mathrm{shared,test}} - \log\det\widehat{\Sigma}_{\mathrm{matched,test}}]$, where both readouts are fit on training moments. Rank six uses full-rank QR without optimization. No analytic projector initializes training.

The stable anonymous entry point `artifact/v22_learned_frontier_recovery.py` regenerates `artifact/v22_learned_frontier_recovery_results.json`, including all 60 seed gaps; `artifact/README.md` gives the command and environment. Seeds are $11, 23, 37, 53, 71$. The reported intervals use the five seed gaps, $\bar{g} \pm 2.776\, s_g/\sqrt{5}$ (rounded Student-t critical value, four degrees of freedom). Normality across optimizer seeds is not established, so these are descriptive t-reference intervals, not a validated population-coverage guarantee.

The maximum mean learned–exact discrepancy is .0233 nats. At rank three, the learned minus trace is 1.0025 below the transition and $.0124$ above it, recovering the predicted paired versus spread allocation. The analytic projector is used only as a post-training comparator; its largest held-out mean discrepancy is .0152 nats. This controlled study validates optimization within the theorem class, not a population claim for public nonlinear models.

Transfer boundary. The legacy public controls and fail-closed premise audits remain in the appendix. They establish no theorem applicability or architecture recommendation and are not used as support for Table 2.

Table 2: Learned held-out information gaps (mean and 95% seed interval) versus the exact frontier. m_- is learned mass in the minus bank.

	$s = .25$ (paired regime)			$s = .8$ (spread regime)		
p	Exact	Learned [95% CI]	m_-	Exact	Learned [95% CI]	m_-
1	2.890	2.897 [2.863,2.932]	.002	2.491	2.492 [2.454,2.530]	.000
2	2.197	2.209 [2.170,2.248]	1.000	1.686	1.683 [1.658,1.707]	.001
3	1.792	1.805 [1.773,1.837]	1.002	.882	.905 [.848,.962]	.012
4	1.099	1.106 [1.069,1.142]	2.000	.588	.592 [.583,.601]	1.001
5	.693	.699 [.682,.716]	2.002	.294	.300 [.294,.307]	2.000
6	0	.002 [.000,.003]	3.000	0	.002 [.001,.003]	3.000

4 Related Work and Precise Positioning

Multimodal representation and fusion. CLIP (Radford et al., 2021), MulT (Tsai et al., 2019), multimodal VAEs (Wu & Goodman, 2018; Shi et al., 2019), MISA (Hazarika et al., 2020), MultiBench (Liang et al., 2021), and HighMMT (Liang et al., 2023) demonstrate the value of modality-specific processing and cross-modal fusion. These systems motivate the design question but do not imply our lower bound. Conversely, Theorem 1 does not establish superiority over those systems: it bounds the cost of one rank-limited linear subspace continuously in the Gaussian target-subspace overlap.

Architecture search and routing. Multimodal architecture search such as MFAS (Pérez-Rúa et al., 2019) selects components using observed validation performance. Our fixed probe is closer to constrained model selection than to unsupervised structure discovery: the candidate set, bounded score, split, simultaneous radius, and abstention rule are fixed. Proposition 9 certifies a population winner only inside that pool. It neither searches globally nor provides an advantage over directly validating complete shared and matched models. The Italy outcome makes this limitation concrete.

Multi-view information theory. Multi-view representation theory often exploits shared content or redundancy across views (Xu et al., 2013; Federici et al., 2020; Tosh et al., 2021). Lemma 1 instead requires conditional nonredundancy of what is omitted. This is not a priority claim; it identifies why familiar redundancy assumptions cannot be silently reused in the proof. The general-overlap bound uses standard Gaussian, chord, and Ky Fan ingredients. Proposition 4 adds a sharper structural result at maximal heterogeneity: every two rank units recover one complete cross-modality target coordinate, while an odd unit leaves the intermediate $\log(1+\sigma^{-2})$ term. We claim this exact disjoint-filter allocation law, not a general closed-form unequal-angle solution or a nonlinear theorem; the exact variational formula is for the stated paired-filter model.

Joint/private and shared-subspace geometry. JIVE separates joint and individual low-rank variation (Lock et al., 2013), and AJIVE uses principal-angle geometry to estimate joint score subspaces (Feng et al., 2018); shared linear representation theory likewise studies recovery and transfer when tasks admit one common subspace under diversity or coverage conditions (Tripuraneni et al., 2021). Thus neither shared/private decomposition nor principal-angle analysis is new here. Our narrower object is the exact population log-determinant loss allocation when the *same* rank-limited map acts on both modalities and the two filter banks are paired in fixed output coordinates. JIVE/AJIVE are unsupervised decompositions rather than this supervised tied-map optimization, while shared-representation recovery bounds do not determine its pairing-to-spreading transition. Conversely, our frontier supplies no estimator or transfer guarantee for those settings.

Time-series encoders and foundation models. N-BEATS (Oreshkin et al., 2020), PatchTST (Nie et al., 2023), TimesNet (Wu et al., 2023), Crossformer (Zhang & Yan, 2023), iTransformer (Liu et al., 2024), S4 (Gu et al., 2022), and Mamba (Gu & Dao, 2023) span convolutional, patching, attention, and state-space biases. TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), and Mamba-2 (Dao & Gu, 2024) further expand capacity and pretraining. None is ruled out by our linear theorem, and raw parameter count alone does not imply its rank bottleneck.

Table 3: Separation from adjacent contributions. Checkmarks describe what is established here, not field priority.

Work type	Encoder architecture	Search/selection	Encoder lower bound	Premise abstention
MulT/MISA/MultiBench	✓	—	—	—
MFAS / constrained NAS	✓	✓	—	—
Multi-view information theory	—	—	content-dependent	—
Time-series foundation models	✓	pretraining	—	—
This work	excluded/design only	fixed-pool	linear rank-r class	✓

Application-facing forecasting. Clinical time-series benchmarks (Harutyunyan et al., 2019), TFT (Lim et al., 2021), DeepAR (Salinas et al., 2020), and battery/RUL studies (Severson et al., 2019; Saxena et al., 2008; Li et al., 2018) motivate heterogeneous temporal inputs. Our Beijing, Italy, Appliances, and ETTh1 results are not leaderboard entries. They serve different audit roles: three preserve mixed model-selection outcomes, while ETTh1 demonstrates that a synthetic premise certificate does not automatically transfer to a learned real-data representation.

4.1 Closest prior theorems: what is and is not a corollary

The trace bound is Ky Fan's principle (Fan, 1949; Overton & Womersley, 1992) plus a chord bound, and the shared hypothesis class is classical (Ando & Zhang, 2005). Table 4 (Appendix A) separates these corollaries, together with existing joint/private and shared-representation geometry, from the exact log-determinant frontiers. Unlike a stacked reduced-rank predictor, tying must retain both task directions. Proposition 10 gives the excess explicitly in Appendix R; no new generic regression or information-bottleneck result is claimed.

5 Discussion, Limitations, and Conclusion

The headline result is the exact equal-overlap allocation law and its pairing/spreading transition, with the disjoint case as a specialization. Proposition 8 additionally closes all encoder ranks for two outputs with unequal overlap. Higher-output unequal overlap remains variational: Appendix Q certifies 18 of 24 tested $p = r$ cells as block-aligned suboptimal; the refinement proves five others optimal and leaves one three-output cell open. An additional spectral/trace bound narrows that cell's certified interval to $[.458145, .594113]$ nats (rounded outward), but its width remains above .01 nats; no global optimum is established (Appendix S). All of this concerns the linear-Gaussian model of Section 2, and every experiment above draws its data from that model; no result here shows the frontier guiding a real-world architecture choice.

The limitation specific to this law. In the synthetic construction, exhaustive finite witnesses satisfy the required premises and the conditional-mutual-information cross-check has maximum error 0.0 bits. Real-data origins do not inherit those premises from temporal labels or encoder tying: ETTh1 fails or cannot test its invocation gates, making the law unavailable before rank allocation. All public selectors abstain without temporal coverage; ETTh1 cannot invoke the frontier.

The appendix preserves the complete adverse public record: no eight-seed family-wise comparison separates; the capacity-aware aggregate has noncommensurate raw MSE units; the duplicate-collapsed ratios are descriptive; and the requested overlap axis is NONINFORMATIVE because it never changes the public inputs or targets. These are transfer-boundary failures, not experiments supporting the frontier.

Conclusion. The exact result is an allocation law, not an architecture prescription. In the paired-filter linear–Gaussian model, equal overlap produces a sharp pairing-to-spreading transition at $s^\star$; unequal overlap invalidates the blockwise rule through an explicit strict counterexample, even though the complete two-output frontier is still solvable. The theorem-targeted learned study recovers both sides of that transition. Higher-output unequal overlap retains one explicitly bounded open cell, capacity signs depend on accounting, and every public-data route abstains or fails an invocation gate. No evidence here supports choosing a practical encoder architecture from the frontier.

Reproducibility Statement

Section 3 states the theorem-targeted study's data generation, parameterization, objective, held-out statistic, splits, ranks, overlaps, optimizer, restarts, seeds, and interval convention. The anonymous entry point is `artifact/v22_learned_frontier_recovery.py`; its frozen result record and focused test are packaged beside it. The appendix gives proofs, controlled checks, and the full negative transfer record. Raw UCI files are excluded from redistribution; reproducing public-data boundary audits requires local copies of the named data.

References

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

Abdul Fatir Ansari, Lorenzo Stella, Caner Türkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Luis M. Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140:81–97, 2017. Source paper for the UCI Appliances Energy Prediction dataset, `https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction`.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

Song Chen. Beijing PM2.5 data set. UCI Machine Learning Repository, 2015. Hourly observations 2010–2014 at the US Embassy, Beijing. Available at `https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data`.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 10041–10071. PMLR, 2024.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 10148–10167. PMLR, 2024.

Saverio De Vito. Air quality data set. UCI Machine Learning Repository, 2008. Hourly observations from a single Italian city station, March 2004 – February 2005. Available at `https://archive.ics.uci.edu/ml/datasets/Air+Quality`.

Ky Fan. On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

Qing Feng, Meilei Jiang, Jan Hannig, and J. S. Marron. Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 166:241–265, 2018. doi: 10.1016/j.jmva.2018.03.008.

Bernhard N. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984. doi: 10.1080/01621459.1984.10477108.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:96, 2019.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

Xiang Li, Qian Ding, and Jian-Qiao Sun. Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, 172:1–11, 2018.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. High-modality multimodal transformer: Quantifying modality and interaction heterogeneity for high-modality representation learning. *Transactions on Machine Learning Research*, 2023.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, 2013. doi: 10.1214/12-AOAS597.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

Michael L. Overton and Robert S. Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992.

Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6966–6975, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Gregory C. Reinsel and Raja P. Velu. *Multivariate Reduced-Rank Regression: Theory and Applications*, volume 136 of *Lecture Notes in Statistics*. Springer, New York, 1998.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Abhinav Saxena, Kai Goebel, Don Simon, and Neil Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. *International Conference on Prognostics and Health Management*, pp. 1–9, 2008.

Kristen A Severson, Peter M Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H Chen, Muratahan Aykol, Patrick K Herring, Dimitrios Fraggedakis, et al. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5):383–391, 2019.

Yuge Shi, N Siddharth, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, 2019.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206, 2021.

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10434–10443. PMLR, 2021. URL https://proceedings.mlr.press/v139/tripuraneni21a.html.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Association for Computational Linguistics*, pp. 6558–6569, 2019.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5575–5585, 2018.

Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023. OpenReview: https://openreview.net/forum?id=vSVLM2j9eie.

A SEPARATION DERIVATIONS FOR THE CLOSEST PRIOR THEOREMS

This appendix supports Table 4 and Proposition 10. Throughout, $X = (X_A; X_B) \in \mathbb{R}^{2d}$ denotes the stacked history, so $\Sigma_X = I_{2d}$ and $Y = B^\star X + \varepsilon$ with $B^\star = [U_A^\top \; U_B^\top]$.

Table 4: Closest prior theorems in this notation. "Corollary" means the listed statement follows from the prior result plus elementary steps; "not implied" is supported by the separation in the last column.

Prior result	Statement in this notation	Corollary here	Not implied, with separation
Ky Fan maximum principle (Fan, 1949; Overton & Womersley, 1992)	$\max_{P_W} \operatorname{tr}[P_W(P_A + P_B)] = r(1+\bar{\rho})$ over rank-r projectors; the Fantope is the convex hull of these projectors	Theorem 1 (with one chord bound)	Propositions 1–4: the objective is $\log\det$, not linear, so Ky Fan gives only the chord bound, tight at $s = 0$ and even p, and at full recovery; Proposition 3 shows no Ky Fan-type block-aligned rule exists for unequal cosines
Reduced-rank regression (Izenman, 1975; Reinsel & Velu, 1998); Gaussian information bottleneck (Chechik et al., 2005)	rank-at-most-p deterministic predictor from the stacked history: $\Delta_{\mathrm{RRR}}(p) = (r - \min\{p, r\})c_0$; the Gaussian bottleneck shares relevance eigendirections but has an additional rate/noise tradeoff	none	The stacked optimum is blind to s and spends one rank unit per target coordinate; the tied optimum depends on s and needs two. Proposition 10 gives the excess in closed form; zero-loss rank r versus $\dim(U_A + U_B)$ (Proposition 5)
Common principal components (Flury, 1984)	one orthogonal basis simultaneously diagonalizing Σ_A, Σ_B	none	Unsupervised: it does not involve Y, U_A, or U_B. Under $\Sigma_A = \Sigma_B = I$ every basis is a CPC basis while the frontier is nontrivial
Canonical correlation (Hotelling, 1936)	$\max \operatorname{corr}(a^\top X_A, b^\top X_B)$	none	All canonical correlations vanish under the independence assumed here; the only shared structure is the common target
Shared-subspace multi-task learning (Ando & Zhang, 2005; Argyriou et al., 2008)	one shared $\Theta = W^\top$ applied to each stream with stream-specific readouts	the hypothesis class (two streams, joint readout, no direct term)	Their results are algorithms and generalization bounds; no population frontier, pairing/spreading transition (8), or capacity-matched sign theorem (Proposition 6) is stated there

Reduced-rank regression is blind to the principal cosines. The rank-p reduced-rank regression problem $\min_{\mathrm{rank}(B)\le p} \mathbb{E}\|Y - BX\|^2$ is solved, because $\Sigma_X = I$, by the rank-p truncated singular value decomposition of $B^\star$ (Izenman, 1975; Reinsel & Velu, 1998). Since $B^\star B^{\star\top} = U_A^\top U_A + U_B^\top U_B = 2I_r$, every singular value equals $\sqrt{2}$ regardless of $U_A^\top U_B$. Writing Y in the left singular basis, each retained coordinate is $\sqrt{2}\, v_i^\top X + \varepsilon_i$ with $v_i = (U_{A,i}; U_{B,i})/\sqrt{2}$ a unit vector, hence predicted exactly, while each dropped coordinate keeps variance $2 + \sigma^2$. The gap is therefore $(r - \min\{p, r\}) \cdot \frac{1}{2}\log(1 + 2\sigma^{-2}) = (r - \min\{p, r\})c_0$. In particular the zero-loss rank of the stacked predictor is r, whereas Proposition 5 shows the tied zero-loss rank is $\dim(U_A + U_B) \in [r, 2r]$.

Gaussian information bottleneck. For jointly Gaussian (X, Y) the optimal bottleneck projections are eigenvectors of $\Sigma_{X|Y}\Sigma_X^{-1}$ ordered by increasing eigenvalue (Chechik et al., 2005). Here $\Sigma_{X|Y} = I - B^{\star\top}\Sigma_Y^{-1}B^\star$ with $\Sigma_Y = (2 + \sigma^2)I_r$, so $\Sigma_{X|Y}$ has eigenvalue $\sigma^2/(2 + \sigma^2)$ on the r-dimensional row space of $B^\star$ and 1 on its orthogonal complement. A deterministic projection onto $p \le r$ of those directions attains the reduced-rank value. This does not equate that value with the finite-rate Gaussian information bottleneck, which also optimizes encoder noise and scale under an information tradeoff. A nonconstant noiseless linear projection of a continuous Gaussian input has infinite $I(X; Z)$, so it is not a finite-rate bottleneck. Neither unconstrained projection is required to have the block form $\mathrm{diag}(W^\top, W^\top)$, which is the constraint the present paper analyzes.

Why tying costs a factor of two and creates the s dependence. A tied projector P_W acts in $\mathbb{R}^d$ and enters the gap only through $H = U_A^\top P_W U_A + U_B^\top P_W U_B$. Recovering target coordinate i exactly requires $H_{ii} = 2$, i.e. both $U_{A,i}$ and $U_{B,i}$ in the range of P_W. Those two unit vectors span a two-dimensional subspace unless $s_i = 1$; s_i describes the corresponding pair of task directions, not an angle to other coordinate blocks. Proposition 4 is the case $s = 0$, where the $2r$ vectors are orthonormal and two rank units per coordinate is exactly attainable; Proposition 2 shows how the intermediate option of one unit along $U_{A,i} + U_{B,i}$, worth $c_1(s)$ instead of c_0, becomes preferable above $s^\star$. Subtracting $(r - p)c_0$ gives Proposition 10.

Ky Fan, the Fantope, and the log-determinant objective. Ky Fan's maximum principle (Fan, 1949) gives $\max_{P_W} \operatorname{tr}[P_W(P_A + P_B)] = \sum_{i \le r}(1 + s_i) = r(1 + \bar{\rho})$ over rank-r projectors, and the convex hull of rank-p projectors is the Fantope (Overton & Womersley, 1992). Theorem 1 uses exactly these two facts plus one chord bound and is stated only as a bridge. The exact frontiers minimize $\frac{1}{2}\log\det(I + \sigma^{-2}(2I - H))$, which is concave in P_W; Ky Fan controls $\operatorname{tr} H$ but not the spectrum of H, and the chord bound is tight when the residual spectrum is $\{0, 2\}$-valued, including the disjoint even-rank allocation and full recovery. The extreme-point and principal-angle computation in Appendix F is what replaces the missing spectral control, and Proposition 3 shows that no block-aligned (Ky Fan-type) rule survives unequal cosines.

Common principal components and canonical correlation. Common principal components (Flury, 1984) seek one orthogonal matrix simultaneously diagonalizing Σ_A and Σ_B; the target Y and the task subspaces U_A, U_B do not enter, and under $\Sigma_A = \Sigma_B = I$ every basis qualifies while the frontier is nontrivial. Canonical correlation (Hotelling, 1936) maximizes $\operatorname{corr}(a^\top X_A, b^\top X_B)$, which is identically zero under the independence assumed here. Both are unsupervised statements about the inputs; the shared structure in this model is supervised, through the common target only.

Shared-subspace multi-task learning. The tied class coincides with the shared-structure hypothesis class of Ando & Zhang (2005) restricted to the shared term (no stream-specific direct term), with the two modality streams playing the role of tasks and one joint readout replacing per-task outputs; trace-norm multi-task feature learning (Argyriou et al., 2008) relaxes the same shared map. Those works supply estimators and generalization bounds for the class. They do not state a population optimum over the shared map, and none of Propositions 2, 4, or 6 appears there or follows from their bounds.

B CAPACITY-ACCOUNTING COMPARATOR AND PROOF

Proposition 6 (Capacity-matched tied/untied frontier)**.** *For input dimension d, target rank r, and unrestricted linear readouts, a tied rank-p model has effective dimension $\nu_T(p) = p(d - p) + 2rp$. Separate ranks (q_A, q_B) have*

$$\nu_U = q_A(d - q_A) + q_B(d - q_B) + r(q_A + q_B).$$

Restricting to integers $0 \le q_A, q_B \le r$, at those fixed untied ranks the exact minimum gap is

$$\Delta_U(q_A, q_B) = (r - \max\{q_A, q_B\})c_0 + |q_A - q_B|c_{1/2}, \quad c_{1/2} = \tfrac{1}{2}\log(1 + \sigma^{-2}).$$

The exact capacity-matched comparator is the finite minimum of Δ_U over $\nu_U \le \nu_T(p)$. For arbitrary principal-cosine spectrum $\mathbf{s}$, let $\Delta_T^\star(\mathbf{s}, p)$ be Proposition 1's exact tied optimum and define

$$\Gamma = \min_{\nu_U \le \nu_T(p)} \Delta_U - \Delta_T^\star(\mathbf{s}, p).$$

Then $\Gamma < 0$, $= 0$, or > 0 means untied, equal, or tied advantage only under this accounting rule. For equal cosine, the finite closed form in Proposition 2 removes continuous optimization. The declared 2,160-cell grid contains 221/740/1,199 untied/equal/tied cells.

Proof. Only subspaces matter after quotienting basis rotations into the readouts, so a rank-p shared subspace contributes $p(d - p)$ dimensions and its two readouts contribute $2rp$; the untied count follows separately. At ranks (q_A, q_B), concavity of $\frac{1}{2}\log(1 + x/\sigma^2)$ makes the two recovered coordinate sets nest at the optimum. The residual eigenvalues are therefore zero $\min(q_A, q_B)$ times, one $|q_A - q_B|$ times, and two $r - \max(q_A, q_B)$ times, proving the formula; task-coordinate subspaces attain it. Exhausting $0 \le q_A, q_B \le r$ under the displayed budget proves the finite untied minimum. Subtracting the exact tied optimum proves the sign characterization. For equal cosine, the tied optimum is a finite minimum over allocation index k, so the comparison is finite on both sides and nondecreasing in s. The 2,160-cell audit `v20_general_capacity_sign.py` exhausts all declared ranks and verifies the untied finite minimum and overlap monotonicity with zero violations. □

The sign is not statistical or optimization capacity. Appendix P preserves the five-convention sensitivity analysis and all reversals.

C PROOF OF THE DISTINCT-COVARIANCE RESULT

Gaussian conditioning gives $C(W)$ exactly. Each modality's conditional covariance term is positive semidefinite and at most I, so the eigenvalues of $C(W)$ lie in $[0, 2]$. Apply $\log(1 + x/\sigma^2) \geq (x/2)\log(1 + 2/\sigma^2)$ and $\log(1 + x/\sigma^2) \leq x/\sigma^2$, then minimize, to obtain the sandwich. Because Σ_m is invertible, the mth conditional covariance vanishes iff $U_m^\top X_m$ is measurable from $W^\top X_m$, equivalently $\mathrm{span}(U_m) \subseteq \mathrm{span}(W)$. Both vanish exactly under the stated joint-span condition.

D PROOF OF THE DIAGONAL RANK-ONE FRONTIER

Direct Gaussian conditioning gives $h(t)$. Its derivative is

$$h'(t) = -\frac{ac}{(at + c)^2} + \frac{bd}{(dt + b)^2}.$$

Taking positive square roots and solving $h'(t) = 0$ gives $t^\star$; if $ab \neq cd$, any interior minimum must be this stationary point; if $ab = cd$, then $h(t) \equiv 1$ and the endpoints already attain the minimum. Under common covariance the two recovered fractions sum to one for every t.

E PROOF OF THE PAIRED-FILTER VARIATIONAL FRONTIER

Why independent output rotations are not harmless. Let $d = 4, r = 2, p = 1, \sigma^2 = 1$, $U_A = (e_1, e_2)$, and $U_B = (e_2, -e_1)$. Both task subspaces coincide and both principal cosines are one. For a rank-one encoder whose projection into that plane has squared norm $t \leq 1$, the two recovered covariance terms sum to tI_2; hence the minimum residual is I_2 and the minimum gap is $\log 2$. With $U_B = U_A$ instead, the minimum rank-one residual has eigenvalues $(0, 2)$ and the gap is $\frac{1}{2}\log 3$. Identical subspace principal cosines therefore do not determine the frontier for arbitrarily oriented output filters. The covariance identity verifies these values directly; the paired-filter assumption is substantive.

Under the paired-filter assumption $U_A^\top U_B = \mathrm{diag}(s_i)$, choose orthonormal banks Q, T so that

$$U_A = 2^{-1/2}(QD_+^{1/2} + TD_-^{1/2}), \qquad U_B = 2^{-1/2}(QD_+^{1/2} - TD_-^{1/2}).$$

For $P = P_W$, cancellation of the cross terms gives

$$U_A^\top P U_A + U_B^\top P U_B = D_+^{1/2} A D_+^{1/2} + D_-^{1/2} B D_-^{1/2}, \quad A = Q^\top P Q,\ B = T^\top P T.$$

Relax P to the Fantope $\{0 \preceq P \preceq I : \mathrm{tr}\, P = p\}$. The log-determinant gap is concave in P, so a minimum occurs at an extreme point, exactly a rank-p projector. Any trace outside $\mathrm{span}(Q, T)$ can be moved into that span, increasing the captured matrix in positive-semidefinite order. Inside the span the objective depends only on A, B, and every pair $0 \preceq A, B \preceq I$ with $\mathrm{tr}(A + B) = p$ is realized by the block-diagonal Fantope element $\mathrm{diag}(A, B)$. This proves Eq. (7).

Finally, a minimizing pair can be chosen at an extreme point of this pair set. If either spectrum had a value strictly between zero and one, integrality of the total trace would force a second fractional eigenvalue; opposite small perturbations along their eigenvectors preserve feasibility and trace, contradicting extremality. Thus A, B can be chosen as projectors, whose direct sum is itself a rank-p projector and maps back to a feasible shared encoder. When all $D_\pm$ are scalar, the principal-angle reduction in Appendix F gives the closed form. When $D_+ = D_- = I$, it gives Proposition 4. Substituting $(.05, .45, .85)$ and the two mixed rank-one projectors of Appendix G gives the strict counterexample. For completeness, sorting $\{1 + s_i, 1 - s_i\}$ as $\lambda_1 \geq \cdots \geq \lambda_{2r}$ and applying Ky Fan plus the chord bound also yields $\Delta_p^\star \geq \frac{1}{4}(2r - \sum_{j \leq p} \lambda_j)\log(1 + 2\sigma^{-2})$. Finally, zero gap forces both positive-semidefinite summands $I - U_m^\top P_W U_m$ to vanish, which is possible exactly when W contains $U_A + U_B$, i.e. $p \geq \dim(U_A + U_B)$.

F FULL PROOF OF THE GLOBAL EQUAL-OVERLAP FRONTIER

We fill in the reduction used in Proposition 2. Relax the shared projector to

$$\mathcal{F}_p = \{P : 0 \preceq P \preceq I,\ \mathrm{tr}\, P = p\}.$$

The gap is a concave function of P, because it is a log determinant of a positive affine map. A minimum over this compact convex set therefore occurs at an extreme point, and the extreme points of $\mathcal{F}_p$ are exactly the rank-p projectors. Thus the relaxation has no value gap. The objective depends only on the compression to $\operatorname{span}(Q, T)$. If that compression has trace below p, its missing trace can be added inside this $2r$-dimensional span while preserving $0 \preceq P \preceq I$; this increases H in positive-semidefinite order and cannot increase the gap. Hence an optimum uses the full trace in the useful span.

Writing the compression in the $Q \oplus T$ coordinates, the objective depends only on its diagonal blocks A and B. Conversely every pair $0 \preceq A, B \preceq I$ with $\operatorname{tr} A + \operatorname{tr} B = p$ is realized by the Fantope element $\operatorname{diag}(A, B)$. Concavity again puts a minimum at an extreme pair. At such a pair both matrices are projectors: if a spectrum contained a value strictly between zero and one, integrality of the total trace would force another fractional eigenvalue across the two spectra, and opposite small perturbations along their eigenvectors would preserve the box constraints and total trace, contradicting extremality.

Let $a = 1 + s$, $b = 1 - s$, and $M = \sigma^2 + 2$. The standard principal-angle decomposition of two projectors splits $aA + bB$ into common, unmatched, and two-dimensional angle blocks. In an angle block its eigenvalues

$$\lambda_\pm = \frac{a + b \pm \sqrt{(a-b)^2 + 4ab\cos^2\theta}}{2}$$

satisfy

$$(M - \lambda_+)(M - \lambda_-) = M^2 - 2M + ab\sin^2\theta.$$

The factor is increasing in $\sin^2\theta$, so every nonzero angle can only increase the objective; the smaller projector range is contained in the larger. If $\operatorname{rank} A < \operatorname{rank} B$, swapping the ranks puts the unmatched directions under weight $a \geq b$ and cannot increase the gap. Thus take $B \preceq A$, set $t = \operatorname{rank} B$ and $\operatorname{rank} A = p - t$. Feasibility is exactly $\max\{0, p - r\} \leq t \leq \lfloor p/2 \rfloor$, and the spectrum claimed in the proposition follows. Its gap is $(r - p + t)c_0 + (p - 2t)c_1(s)$, affine in t; the sign of $c_0 - 2c_1(s)$ gives the two endpoint allocations and Eq. 8.

G Exact unequal-overlap counterexample

Define the normalized directions q_i, t_i as in Proposition 2, put

$$x^\star = \frac{-263 + \sqrt{127281}}{256}, \quad u = \sqrt{x^\star}q_1 + \sqrt{1 - x^\star}q_2, \quad v = \sqrt{1 - x^\star}t_1 + \sqrt{x^\star}t_2,$$

and take the projector onto $\operatorname{span}\{u, q_3, v\}$. For $(s_1, s_2) = (1/20, 9/20)$, direct evaluation of the two-coordinate factor of $\det((\sigma^2 + 2)I - H)$ gives

$$d(x) = \frac{(\sqrt{127281} + 391)x^2 - (\sqrt{127281} + 191)x + 208}{200}.$$

It is strictly convex, and $d'(x) = 0$ gives $x^\star = (-263 + \sqrt{127281})/256 \in (0, 1)$. Adding the selected plus direction q_3 yields the exact gap

$$\frac{1}{2}\log\left(\frac{9\sqrt{127281} + 3681}{500}\right) = 1.311745663832762\ldots.$$

For a block-aligned rank-three projector each of the three blocks receives zero, one plus, one minus, or two units, with costs respectively c_0, $c_1(s_i)$, $\frac{1}{2}\log(1 + (1 + s_i)/\sigma^2)$, and zero. Enumerating the coefficient of z^3 in the product of these three four-choice polynomials gives the minimum $c_0 + c_1(.85) = 1.333614103290977\ldots$, attained by both $(0, 2, 1)$ and $(2, 0, 1)$ units on blocks $(1, 2, 3)$. The witnessed improvement is $0.021868439458215\ldots$ nats.

H Exact disjoint frontier for more than two modalities

The disjoint-filter law of Proposition 4 extends verbatim from two modalities to any number of independent modalities sharing one additive target and one encoder. Remark 1 records what does not extend with it.

Proposition 7 (Disjoint-filter frontier for M modalities)**.** *Let $M \geq 2$ and $Mr \leq d$, let $X_1, \ldots, X_M \sim \mathcal{N}(0, I_d)$ be independent, and let $U_1, \ldots, U_M \in \mathbb{R}^{d\times r}$ have orthonormal columns with mutually orthogonal ranges. Put*

$$Y = \sum_{m=1}^{M} U_m^\top X_m + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_r), \quad \sigma^2 > 0,$$

apply one shared $W \in \mathbb{R}^{d\times p}$ with $W^\top W = I_p$ to every modality, and compare against the matched encoders $U_m^\top X_m$. Write $q = \lfloor p/M \rfloor$, $b = p - qM$, and $b_+ = \min\{b, 1\}$. For $0 \leq p \leq Mr$ the exact minimum common-target information gap over all shared rank-p maps is

$$\Delta_p^\star = \tfrac{1}{2}\big[(r - q - b_+)\log(1 + M\sigma^{-2}) + b_+ \log(1 + (M-b)\sigma^{-2})\big],$$

and $\Delta_p^\star = 0$ for $p \geq Mr$. Thus every M rank units recover one complete target coordinate across all modalities, while a remainder of b units only reduces the next coordinate's residual variance from $\sigma^2 + M$ to $\sigma^2 + M - b$. At $M = 2$ this is Proposition 4.

Proof. Put $P = WW^\top$ and $H = \sum_m U_m^\top P U_m$. Matched encoders determine every $U_m^\top X_m$ and leave target covariance $\sigma^2 I_r$. Given the shared encodings $(W^\top X_m)_{m\leq M}$ the modalities remain independent, and Gaussian conditioning gives $\mathrm{Cov}(U_m^\top X_m \mid W^\top X_m) = U_m^\top (I - P) U_m$, so

$$\Sigma_{Y|W} = \sigma^2 I_r + \sum_{m=1}^{M} U_m^\top (I - P) U_m = \sigma^2 I_r + M I_r - H.$$

Since $I(Z;Y) = h(Y) - h(Y \mid Z)$ and $h(Y)$ is common to both encoders, the gap is $\frac{1}{2}\log\det(\Sigma_{Y|W}/\sigma^2)$.

Feasible spectra. Each summand obeys $0 \preceq U_m^\top P U_m \preceq I$, so $0 \preceq H \preceq M I_r$. Because the ranges of the U_m are mutually orthogonal, $\Pi = \sum_m U_m U_m^\top$ is an orthogonal projector, hence $\mathrm{tr}\, H = \mathrm{tr}(P\Pi) \leq \mathrm{tr}\, P = p$. The eigenvalue vector h of H therefore lies in $\mathcal{H} = \{h \in [0, M]^r : \sum_i h_i \leq p\}$.

Relaxed minimum. Minimize $f(h) = \frac{1}{2}\sum_i \log(1 + (M - h_i)\sigma^{-2})$ over $\mathcal{H}$. Each summand has second derivative $-(\sigma^2 + M - h_i)^{-2} < 0$, so f is concave and attains its minimum over this compact polytope at an extreme point. An extreme point has r active constraints, so either every coordinate lies in $\{0, M\}$, or $\sum_i h_i = p$ and at most one coordinate lies strictly inside $(0, M)$. As f is strictly decreasing in each coordinate, a minimizing extreme point has $\sum_i h_i = p$ whenever $p \leq Mr$. Such a point has k coordinates equal to M, one equal to $p - kM \in [0, M]$ and the rest zero; $p - kM \geq 0$ forces $k \leq p/M$ and $p - kM \leq M$ forces $k \geq p/M - 1$, so $k = q$ (when M divides p the two admissible values of k describe the same vector). Up to permutation the minimizer is therefore unique, $h^\star = (\underbrace{M, \ldots, M}_{q}, b, 0, \ldots, 0)$, with value the displayed $\Delta_p^\star$.

Attainment. Let W have range $\mathrm{span}\{U_{m,i} : m \leq M,\ i \leq q\} \cup \{U_{m,q+1} : m \leq b\}$, a set of $qM + b = p$ mutually orthonormal vectors, since disjointness makes the whole family $\{U_{m,i}\}$ orthonormal. Then P is the orthogonal projector onto a subset of that family, so $(U_m^\top P U_m)_{ij} = \delta_{ij}\mathbf{1}\{U_{m,i} \in \mathrm{range}\, W\}$ and $H = \mathrm{diag}(h^\star)$. The relaxed lower bound is thus attained by a feasible shared encoder and is the exact frontier. For $p \geq Mr$, any W whose range contains every $\mathrm{span}(U_m)$ gives $H = M I_r$ and gap zero. □

Remark 1 (Where the overlap frontier stops)**.** *Disjointness and equal modality weights are both load bearing. For two overlapping banks, the reduction behind Proposition 1 writes $U_A = 2^{-1/2}(QD_+^{1/2} + TD_-^{1/2})$ and $U_B = 2^{-1/2}(QD_+^{1/2} - TD_-^{1/2})$ and relies on the cancellation*

$$U_A^\top P U_A + U_B^\top P U_B = D_+^{1/2} A D_+^{1/2} + D_-^{1/2} B D_-^{1/2}, \qquad A = Q^\top P Q, \quad B = T^\top P T,$$

in which the off-diagonal block $X = D_+^{1/2} Q^\top P T D_-^{1/2}$ enters the two modalities with opposite signs. If the modalities enter the target with unequal weights, $Y = \alpha U_A^\top X_A + \beta U_B^\top X_B + \varepsilon$, the same algebra gives

$$\alpha^2 U_A^\top P U_A + \beta^2 U_B^\top P U_B = \tfrac{\alpha^2+\beta^2}{2}\big(D_+^{1/2} A D_+^{1/2} + D_-^{1/2} B D_-^{1/2}\big) + \tfrac{\alpha^2-\beta^2}{2}\big(X + X^\top\big),$$

and X does not vanish on the feasible set: the rank-one projector onto $(q_1 + t_1)/\sqrt{2}$ has $X_{11} = \frac{1}{2}\sqrt{1 - s_1^2} > 0$ for $s_1 < 1$. Unless $\alpha^2 = \beta^2$ the residual is therefore not a function of (A, B) alone, and the two-variable problem (7) *is no longer equivalent to the encoder problem. For three or more overlapping banks there is no pairwise $Q \oplus T$ splitting to start from. Dependent modalities break the reduction one step earlier: when $\mathrm{Cov}(X_A, X_B) \neq 0$ the conditional target covariance acquires the cross term $U_A^\top \mathrm{Cov}(X_A, X_B \mid W^\top X_A, W^\top X_B) U_B$ and its transpose, which is not a function of H. We state no frontier in any of these cases.*

I SEED-LEVEL LEARNED-CONTROL AUDIT

Family-wise sign inference. The primary exact sign test targets the null $\Pr(D > 0 \mid D \neq 0) = \Pr(D < 0 \mid D \neq 0) = 1/2$ for independent paired training-seed differences D. It requires neither Gaussian differences nor symmetry of their magnitudes, but independence and fair signs remain assumptions about the seed-generating procedure. Holm adjustment controls the three-dataset family under arbitrary cross-dataset dependence. Beijing's 7/8 tied-lower directions give raw $p = .070$ and adjusted $p = .211$; Italy has 6/8 in the opposite direction ($.289/.578$) and Appliances splits 4/8 ($1.000/1.000$). All three therefore remain unresolved, and there is no family-wise architecture recommendation. With eight pairs the minimum attainable two-sided value is $.0078$, so 7/8 can never clear even unadjusted $.05$; this is an explicit inferential-resolution ceiling, not evidence of equality. These descriptive public runs measure neither the theorem's Gaussian task filters, rank, nor principal overlap.

Capacity-aware negative accounting record. An eight-seed family-wise control does not separate on any dataset. The capacity-aware follow-up has a tied-minus-untied MSE of 76.67 with seed-cluster 95% CI $[23.87, 148.62]$, but raw MSE units are not commensurate across traces, so an unweighted aggregate over them has no common scale. We therefore treat it only as protocol output, not as a cross-dataset effect.

A descriptive re-analysis with one unit per source–rank–seed gives a geometric mean tied/untied MSE ratio of 2.16. The per-trace ratios are 6.46 (Beijing), 1.56 (Italy), and 1.01 (Appliances), with untied lower in 27/36 units. A classical ridge or mean baseline wins at least one row in all but 6 units. These ratios are descriptive: four seeds recur across ranks and traces, so we attach no new confidence interval or p-value.

The public-residual experiment is uninformative by construction: its inputs and targets do not use the requested overlap, so all three settings return identical values (NONINFORMATIVE). It supplies neither theorem-premise validation nor a real-data extension of the synthetic law.

Architecture and training. Each modality-specific learned adapter maps every input channel to width 8 at each of 12 evenly spaced past-only tokens. Each temporal branch flattens the resulting 12-by-8 array, applies tanh layers $96 \to 16 \to 8$, and a linear head maps the concatenated 16 features to the standardized target. Tied and untied runs have identical nominal slots (including the inactive duplicate slot), paired initialization, adapters, head, full-batch Adam ($.01$), chronological 60/15/25 splits, train-only standardization, at most 4,096 deterministic-grid training windows and 2,048 validation windows, 200 epochs, validation patience 20, and minimum improvement 10^{-6}. The test set is evaluated once at the validation-selected epoch. The only intervention is whether branch B reuses branch A's temporal-layer weights; adapters remain modality-specific because channel counts differ.

Table 5: Seed-level public learned-control audit. Each cell is tied/untied standardized-target test MSE for seeds 11, 23, 37, 53, 71, 89, 107, and 131. These descriptive values are separate from the theorem-targeted experiment; selected epochs are listed below.

Dataset	11	23	37	53	71	89	107	131
Beijing	.06836/.07286	.06747/.06932	.06790/.06792	.06684/.06742	.06618/.06759	.06663/.06962	.06919/.06896	.06785/.06863
Italy	.20034/.19163	.20202/.19634	.19761/.20162	.19784/.19904	.19493/.19317	.20238/.19808	.18920/.18877	.19342/.18909
Appliances	.66923/.66383	.73640/.71104	.65320/.65865	.67514/.66535	.68543/.71127	.66967/.74452	.73681/.73143	.65060/.66555

Selected tied/untied epochs, in seed order, are Beijing: 200/198, 200/184, 177/200, 200/184, 200/200, 200/193, 200/198, 185/200; Italy: 200/189, 172/151, 162/148, 172/95, 187/174, 167/156, 156/116, 195/160; and Appliances: 2/2, 59/5, 7/6, 5/19, 48/36, 6/34, 46/40, 20/18. Appliances'

selected epochs are early and spread widely across seeds; we report this dispersion without attributing it to a specific cause, and it supports neither tying choice. The primary exact sign-family audit is executable in `v11_sign_family_audit.py`; its input is the frozen full-precision `v9_learned_tying_results.json`. No real-data theorem premise is inferred.

J Fixed-Feature Probe Algorithm and Guarantees

Algorithm 1 Sample-split fixed-feature probe

1: Split training data into independent construction and score subsets.
2: On construction data only, fit or specify each candidate predictor h_e; then freeze it.
3: **for** $e = 1, \dots, K$ **do**
4: On score examples compute fixed bounded values $S_{i,e} = s(h_e(X_i), Y_i) \in [0, 1]$.
5: Set $\widehat{\mu}_e = n^{-1} \sum_i S_{i,e}$.
6: **end for**
7: Let $\widehat{e} = \arg\max_e \widehat{\mu}_e$ and compute its runner-up margin $\widehat{\Delta}$.
8: Certify $\widehat{e}$ only if $\widehat{\Delta} > 2\sqrt{\log(2kK/\delta)/(2n)}$; otherwise abstain.

The score split is never used to train candidates, choose coordinates, estimate a random variance denominator, or tune the score. Proposition 9 therefore analyzes the algorithm as implemented. The older BH-selected variance-ratio procedure is retired; BH and a fixed correlation cutoff are neither equivalent nor used by this probe.

J.1 Calibration protocol and artifact

The executable local audit is `code/aegis/probe_calibration_fixed.py`; its machine-readable output is `results/fixed_probe_calibration_20260823.json`. The seed is 20260823 and trials are independent. Clopper–Pearson 95% intervals are $[0, 0.000738]$ for 0/5000 and $[0.999262, 1]$ for 5000/5000; these quantify Monte Carlo uncertainty.

Executed fixed-probe record. The protocol fixes K, k, δ, candidates, and bounded score before the independent score split. Public runs use frozen recent/mean/trend ridge predictors and non-overlapping score windows; every route abstains. Temporal independence is unestablished, so i.i.d. coverage is not invoked. The threshold is Algorithm 1's displayed value; no BH step, correlation cutoff, selected-coordinate mask, random denominator, or score-split fitting is used. Synthetic calibration uses 5,000 trials, $n = 512$, $K = k = 2$, $\delta = .05$, and Gaussian noise SD .5. Older public-probe files remain legacy evidence, not instances of this protocol. Exact settings and outputs are in the calibration paths above; the public executable and result are `public_fixed_probe_audit.py` and `public_fixed_probe_results.json`.

J.2 Retired reference architecture

The previous gated-attention/heteroscedastic reference architecture consisted of standard components and had no end-to-end evidence. It is removed to keep the contribution centered on the proved linear bottleneck and the analyzable selection test.

Remark 2 (Probe/theorem separation)**.** *The fixed probe performs finite-pool predictive selection. It neither estimates the overlap/rank conditions of Theorem 1 nor P1–P4 of Lemma 1. Its guarantee is exactly Proposition 9; no adaptive BH or variance-ratio claim remains.*

Remark 3 (Architecture boundary)**.** *Theorem 1 is not architecture-class agnostic. It covers shared rank-r linear maps. Large hybrid and nonlinear encoders are outside its scope, whether or not their parameter count resembles dr.*

K Architecture and Ethical Scope

The theorem's architecture is the shared rank-p linear encoder class stated in Section 2. Foundation-scale nonlinear encoders are outside that class; this is an exclusion, not a violated premise of

Lemma 1 or evidence that our design dominates them. The learned control is a small tanh temporal-network comparator with no family-wise architecture recommendation, not a foundation-model result.

All empirical inputs are existing public environmental and energy time series; no human-subject, clinical, private, or demographic data is used. We make no fairness, labor, deployment, privacy, or indirect-effect measurement claim. Any high-stakes deployment would require separate governance, subgroup evaluation, security, and human-oversight evidence.

L Proof Audit

All theorem proofs appear with their statements. The general bound uses Gaussian conditional covariance, a chord inequality, and Ky Fan. The exact disjoint result instead minimizes the concave log-determinant over the feasible retained-eigenvalue polytope and supplies an attaining paired-direction construction. It does not imply an exact general-overlap or nonlinear result.

L.1 Falsification and applicability

We enumerate the ways this work can be refuted so that later references are unambiguous. *Signature 1, scaling collapse.* Theorem 1 is refuted inside its class by any W whose measured Gaussian mutual-information gap is smaller than the displayed bound, in particular if the predicted order-$\sqrt{d}$ scaling collapses. *Signature 2, premise failure.* Outside that class, P1–P4 audit only whether Lemma 1 may be invoked: one FAIL or NOT_TESTED forbids invocation. *Signature 3, probe abstention.* Abstention merely says that a finite-pool score difference is not separable at that pool size; it refutes nothing. *Signature 4, outcome reversal.* A predictive outcome reversal challenges routing utility but is not a counterexample to either information statement. ETTh1 is premise failure, not theorem confirmation or refutation.

L.2 Fallback regimes

In the equal-cosine frontier, increasing temporal-filter overlap cannot increase the linear shared-encoder gap; rank at least $\dim(U_A + U_B)$ erases the gap exactly for arbitrary overlap. A weak probe margin triggers abstention only; it supplies no near-optimality claim.

M Theorem Checks and Transfer-Boundary Audits

The positive checks in this section target stated linear–Gaussian identities; the nonlinear and public-data studies are retained only as negative scope and transfer audits. The finite discrete check audits Lemma 1; the Gaussian regression check shares Theorem 1's orthogonal geometry but measures prediction error rather than mutual information. None is a real-world validation of the frontier.

M.1 Common-target temporal lower-bound audit

The local script `v7_common_target_audit.py` constructs length-eight Gaussian histories and rank-two DCT filter banks with noise variance .25. At each $p = 1, \ldots, 4$, it selects the last p eigenvectors returned by NumPy's `eigh` for $U_A U_A^\top + U_B U_B^\top$ and evaluates that projector's common-target gap against Proposition 1's bound. The three cases isolate disjoint, partly shared, and coincident bands. Degenerate eigenspaces need not select a gap-minimizing projector: these are pointwise evaluations, not global optima or learned-representation evidence. All 12 general-overlap points satisfy the lower bound to 2.3×10^{-15} nats. The separate rank-four disjoint audit obtains exact minima $(4.101, 3.296, 3.002, 2.197, 1.903, 1.099, .805, 0)$ for ranks 1–8; constructive error is at most 1.5×10^{-15} nats, and all 2,048 seeded random maps lie above the formula. Random search is only a regression check; Proposition 4 is the proof.

Dependent-history audit. `v17_temporal_gaussian_frontier.py` generates two independent stationary Gaussian AR(1) trajectories of length 40,007, forms 40,000 overlapping length-eight histories, and uses chronological 60/15/25 splits. It crosses $\phi \in \{0, .5, .9\}$, equal principal

Table 6: Common-target audit at the specified eigenvector projectors (nats), not an optimum frontier. Gap/bound pairs span $p = 1, \ldots, 4$; tiny negative values at zero are roundoff.

Filters A/B	Principal cosines	Evaluated gaps / bounds	Zero rank
$\{1,2\}/\{5,6\}$	$(0,0)$	1.920/1.648; 1.602/1.099; .881/.549; 0/0	4
$\{1,2\}/\{2,3\}$	$(1,0)$	1.609/1.099; 1.034/.549; 0/0; 0/0	3
$\{1,2\}/\{1,2\}$	$(1,1)$	1.099/1.099; 0/0; 0/0; 0/0	2

cosines $s \in \{0, .4, .8\}$, and every shared rank $p = 1, \ldots, 6$ for 54 cells. Raw task filters and the shared map are $\Sigma^{-1/2}$ times their whitened bases, so this is a direct audit of Corollary 1, not a learned-encoder claim.

Table 7: Dependent Gaussian temporal audit. "Emp." is the largest absolute chronological held-out gap discrepancy over 18 overlap–rank cells.

ϕ	empirical lag 1	formula	raw coordinates	Emp.
0	.00243	2.7×10^{-15}	1.4×10^{-15}	.01811
.5	.50125	2.7×10^{-15}	6.3×10^{-15}	.01672
.9	.90012	2.7×10^{-15}	3.3×10^{-14}	.01723

Distinct-covariance audit. `v19_heterogeneous_covariance_frontier.py` crosses four covariance mismatch levels and raw ranks 1–4, evaluating 100 deterministic shared projectors per cell. All 1,600 pointwise sandwiches hold; common-covariance raw/whitened gaps agree to 3.8×10^{-15} nats. The explicit rank-four span has gap at most 3.4×10^{-15} nats at every mismatch, while holding raw W fixed across covariance levels changes a gap by as much as .9404 nats. These are controlled algebraic checks, not global optimization claims for the random candidate bank.

`v21_diagonal_covariance_frontier.py` independently exhausts $a, c, d, b \in \{.25, .5, 1, 2, 4\}$ and compares the closed form with 20,001 encoder angles in each of 625 cells. The maximum gap discrepancy is 2.0×10^{-9} nats; all 25 common-covariance cells reduce exactly, and 270/355 cells have interior/boundary optima.

A rank-four equal-overlap audit at $s \in \{0, .25, .5, .6, .8, .95\}$ and every shared rank checks Proposition 2 the same way: each closed-form optimum is realised by an explicit orthonormal encoder to 6.9×10^{-15} nats, and none of the 512 seeded random rank-p maps per cell beats it, the smallest observed slack being .0268 nats. The regime change is visible in the allocation itself: at shared rank four the optimum uses two paired blocks at $s = .25$ and four single blocks at $s = .8$, while the optimal value decreases with s. The chord bound reproduces the frontier at $s = 0$ only at even shared rank: at $p = 2, 4, 6, 8$ the two agree to machine precision, while at $p = 1, 3, 5, 7$ the bound is low by exactly .255413 nats, the value of $\frac{1}{2}\log(1+\sigma^{-2}) - \frac{1}{4}\log(1+2\sigma^{-2})$ at $\sigma^2 = .25$. An earlier version of this paragraph asserted tightness at $s = 0$ without qualification; that is false at odd rank and is withdrawn. Away from the disjoint endpoint the bound falls below the frontier by .549, 1.099, and .736 nats at $s = .25, .5, .8$ for $p = 4$, so its looseness peaks at the regime change rather than growing with overlap.

Symmetry averaging alone does not prove block alignment: concavity means that diagonalising H can raise the objective. Proposition 2 instead uses the Fantope extreme-point and principal-angle argument above. As an independent regression check, a six-restart Powell search over twelve cells crossing $s \in \{.25, .5, .8\}$ and shared ranks 2–5 found no unrestricted projector below the formula: maximum improvement was exactly zero and the recovered optimum's relative off-diagonal mass was 1.6×10^{-15}.

Exact-PASS versus learned-representation non-invocation. For independent fair-bit modality blocks of size q, $d = q^2$, and a shared coordinate-selector class with at most q retained coordinates, P1–P4 and the coordinate-omission bound were checked for every admissible selector at $q = 2, \ldots, 8$: exactly 52,475 selectors (11, 42, 163, 638, 2,510, 9,908, and 39,203 by q), all PASS. At $q = 2$ only, an independent probability-table implementation enumerated all 16 states and 11 selectors; its maximum absolute difference from the analytic conditional MI was exactly 0.0 bits. Thus

"0.0 bits" denotes deterministic equality in this small discrete cross-check—not sampling error, estimator calibration, numerical precision on ETTh1, or probability-table verification of all 52,475 selectors. In direct contrast, the frozen ETTh1 stress test returns P4=FAIL and P2=P3=NOT_TESTED for the tested learned representations. This rejects automatic transfer and supplies no real-data theorem claim.

Scaling sanity check. For each $q \in \{8, 12, 16, 24, 32\}$ we set $d = q^2$ and $r = \sqrt{d} = q$. We sample $X_A, X_B \in \mathbb{R}^d$ with iid standard-normal coordinates and define private informative blocks $S_A = \{1, \ldots, r\}$ and $S_B = \{r+1, \ldots, 2r\}$. The target is vector-valued recovery, $y = (X_A[S_A], X_B[S_B]) + \epsilon$ with Gaussian noise $\epsilon \sim \mathcal{N}(0, 0.1^2 I)$. The bounded shared encoder uses the same rank-r paired projection $(e_i + e_{r+i})/\sqrt{2}$ for both modalities, while the matched encoder uses modality-specific rank-r projections onto S_A and S_B. A ridge readout is fit from encoded features to y using 2,000 training and 2,000 test samples for each seed in $\{42, 123, 456, 789, 1024\}$.

Table 8: Controlled regression stress test sharing the theorem's orthogonal-subspace geometry. Values are total test squared error across recovered private factors, mean ± standard deviation over five seeds. This is prediction-error evidence, not a mutual-information estimate.

d	$\sqrt{d}$	Shared MSE	Matched MSE	Gap / $\sqrt{d}$
64	8	8.231 ± 0.085	0.160 ± 0.001	1.009 ± 0.011
144	12	12.381 ± 0.060	0.241 ± 0.002	1.012 ± 0.005
256	16	16.618 ± 0.059	0.324 ± 0.002	1.018 ± 0.004
576	24	25.037 ± 0.071	0.491 ± 0.001	1.023 ± 0.003
1024	32	33.674 ± 0.110	0.663 ± 0.004	1.032 ± 0.003

Two controls check that the effect is tied to the theorem's premises. When there is no mismatch ($S_B = S_A$), the maximum absolute mean gap across dimensions is 0.000. When the shared projection is given enough capacity (rank $2r$), the maximum absolute mean gap is 0.022. Thus the observed gap is specific to the controlled mismatch-plus-bottleneck construction rather than to the ridge readout or synthetic sampling procedure.

Nonlinear stress test (not theorem evidence). We also apply a componentwise $\tanh$ after the same fixed linear projections. The resulting numerical scaling is a robustness observation only: Theorem 1 does not cover nonlinear encoders, and no averaged-Jacobian span is used to infer retained information.

Table 9: Non-linear-encoder robustness check for Theorem 1: same construction as Table 8, with $\tanh$ applied componentwise to the encoder output before the linear ridge readout. Mean ± standard deviation over five seeds. Both the shared MSE and the matched MSE rise (matched is no longer at the noise floor because linear ridge cannot fully invert $\tanh$), but the shared-minus-matched gap divided by $\sqrt{d}$ remains order-1 in this construction; this does not establish a learned-nonlinear result.

d	$\sqrt{d}$	Shared MSE	Matched MSE	Gap / $\sqrt{d}$
64	8	8.790 ± 0.092	1.270 ± 0.018	0.940 ± 0.011
144	12	13.232 ± 0.065	1.950 ± 0.031	0.940 ± 0.005
256	16	17.731 ± 0.064	2.587 ± 0.022	0.946 ± 0.004
576	24	26.759 ± 0.073	3.902 ± 0.044	0.952 ± 0.002
1024	32	35.965 ± 0.086	5.233 ± 0.029	0.960 ± 0.003

The non-linear gap/$\sqrt{d}$ ratio is slightly smaller than the linear ratio but remains in the same order-1 band; this run does not isolate whether the difference comes from tanh saturation, the ridge readout, or their interaction. In this fixed tanh-plus-ridge pipeline, the observed prediction-error gap remains approximately proportional to $\sqrt{d}$ over the tested dimensions; this does not test Falsification Signature 1. We do not interpret this as evidence of real-world performance. The check applies one fixed componentwise tanh to the same prescribed projections before a linear ridge readout; it is a single additional pipeline, not an optimized or provably larger encoder class containing the linear one. The script and raw outputs are retained in the local audit materials at `code/synthetic_validation.py` and `results/synthetic_validation/`; the

non-linear table is generated alongside the linear table by the same script and stored at `results/synthetic_validation/synthetic_validation_table_nonlinear.tex`.

N PUBLIC-DATA ILLUSTRATION: REMARK 4 ON UCI BEIJING PM2.5

To complement the controlled synthetic checks of Appendix M with a single real-data illustration, we evaluate the shared/matched comparison on a public multimodal time-series, the UCI Beijing PM2.5 hourly dataset (Chen, 2015). **This is one dataset, not a comprehensive benchmark, and the result is reported as-is rather than after architecture or hyperparameter search.** The headline finding is that, for this particular bimodal pair, the matched configuration does not measurably beat the forced-symmetric configuration; we read this as a real-data illustration of fallback behavior, not as verification of Remark 4's population conditions or validation of the order-$\sqrt{d}$ bound of Theorem 1.

Setup. The dataset has 43,824 hourly observations (2010–2014) of meteorological covariates and PM2.5 concentration at one Beijing station. We use four meteorological covariates as Modality A (DEWP, TEMP, PRES, I_{ws}) on a 48-hour lookback; we use the PM2.5 history itself as Modality B on the same 48-hour lookback. The target is next-hour PM2.5. We split by time: train on 2010–2012, validate on 2013-H1, test on 2013-H2 plus 2014. All inputs and the target are standardized using training statistics; after de-standardization, test MSE is reported in original target units squared.

Configurations. Three small models with comparable parameter counts and identical training budget (Adam lr 10^{-3}, batch 128, 30 epochs, gradient clip 1.0):

- **Shared-CNN (forced symmetry):** a single rank-32 1D-CNN encoder is applied to both modalities (encoder weights tied, with a per-modality input projection so the channel counts agree). 14,909 parameters.
- **Matched (CNN+LSTM, heuristic pair):** 1D-CNN for the meteorological modality, LSTM for the PM2.5 history modality. 20,417 parameters.
- **Transformer (modern reference baseline):** 2-layer Transformer encoder over the concatenated bimodal token sequence with a 1-bit modality embedding. 18,465 parameters.

The forced-symmetric configuration is a comparison that Theorem 1 bounds only when all latent mismatch premises hold; those premises are not verified here. The matched CNN+LSTM configuration is a heuristic comparison, not the output of the bilateral gate on this dataset; the Transformer is one parameter-matched modern reference.

Table 10: Public-data illustration on UCI Beijing PM2.5 next-hour forecasting. Test MSE is reported in original target units squared (PM2.5 concentration units squared) and as standardized MSE; mean $\pm$ standard deviation over five seeds $\{42, 123, 456, 789, 1024\}$. The paired shared-minus-matched difference is -6.62 with 95% t interval $[-19.10, 5.87]$ and magnitude-randomization sensitivity $p = .25$; the Transformer result is retained only as an under-tuned reference at this fixed budget. We do not interpret this single dataset as validating the order-$\sqrt{d}$ bound; the relevant theory anchor is Remark 4.

Model	Params	Test MSE (orig. units squared)	Test MSE (standardized)
Shared-CNN (forced symmetry)	14,909	409.94 $\pm$ 2.82	0.0527 $\pm$ 0.0004
Matched (CNN + LSTM, heuristic pair)	20,417	416.55 $\pm$ 9.24	0.0535 $\pm$ 0.0012
Transformer (modern baseline)	18,465	4288.61 $\pm$ 193.39	0.5508 $\pm$ 0.0248

Honest reading. Three observations and three caveats.

- *Observation 1.* Matched is not separated from Shared-CNN: the paired difference (shared minus matched) is -6.62, with 95% t interval $[-19.10, 5.87]$ and magnitude-randomization sensitivity $p = .25$. This does *not* corroborate Theorem 1 on this dataset.

- *Observation 2.* The result is descriptively compatible with the protocol's fallback branch: both meteorological covariates and PM2.5 history may share daily-cycle structure at the 48-hour lookback, but the run does not measure geometric separability, retained informative dimensions, or conditional nonredundancy. We therefore do not infer whether Remark 4's population conditions hold. The corrected bilateral probe rule selects fallback on all five seeds; it is retained as a diagnostic, not promoted to a contribution.
- *Observation 3.* At the matched parameter and 30-epoch budget, both CNN/LSTM-based configurations have roughly $10\times$ lower test MSE than the Transformer baseline on this dataset. Because the Transformer was not trained to independently optimized convergence, this is a measured budget-specific difference with no inductive-bias or architecture-superiority interpretation.
- *Caveat 1 (single dataset).* One real dataset is not a comprehensive benchmark; we deliberately avoid claiming generality from this one comparison.
- *Caveat 2 (single-dataset probe).* The retired adaptive diagnostic was run on this data (see Section N.1 below); the corrected bilateral rule recommends the symmetric fallback on all five seeds; on one seed only one modality clears its margin, which is insufficient for the matched route. We treat this as one self-consistency point, not as validation across datasets.
- *Caveat 3 (Transformer training budget).* The Transformer baseline is reported at the same 30-epoch budget, not at independently optimized convergence. Equal epochs do not imply equal optimization opportunity, so we assign this baseline no comparative evidentiary weight.

What this changes for the paper's claims. The synthetic-sanity-check results (Tables 8 and 9) remain the only evidence for the order-$\sqrt{d}$ scaling itself, and they are still confined to constructions in which the mismatch and capacity premises hold by design. The Beijing PM2.5 result adds one real-data fallback observation without establishing the corollary's population premises; the selector remains a diagnostic rather than a contribution or an unconditional asymmetric-architecture recommendation. The script, raw seed-level outputs, and aggregated table are retained in the local audit materials at `code/public_benchmark.py` and `results/public_benchmark/`.

N.1 Retired adaptive probe summary

The adaptive variance-ratio diagnostic predates the repaired fixed-score protocol and is not covered by Proposition 9. We preserve its executable files and complete seed rows, but compress it here because it neither tests theorem premises nor identifies a sharing-only effect. Its conservative bilateral gate returned fallback on Beijing 5/5 seeds, Italy 5/5, and Appliances 4/5; the sole Appliances non-fallback was one asymmetric route. Italy is the decisive utility failure: the diagnostic abstained despite the confounded matched model being lower in all five seeds. Candidate identities also changed across modalities and datasets, and uniformly high Beijing scores made margins weakly discriminative. These failures remain evidence against routing utility, not evidence for symmetry. Exact scores, parameters, margins, and per-seed decisions remain in `results/probe_validation/`, `results/probe_validation_2/`, and the Appliances result artifact.

N.2 Second-Dataset Replication: UCI Air Quality (Italy)

To inspect cross-dataset behavior of the shared/matched comparison and the retired adaptive diagnostic, we replicate the same procedure on a second public dataset, the UCI Air Quality (Italy) hourly record (March 2004 – February 2005, single Italian station) (De Vito, 2008). The dataset is independent of Beijing PM2.5 in country, time period, pollutant, and target variable; we use it as a cross-domain check, not as evidence of generality. Modality A is meteorological (T, RH, AH; 3 features); Modality B is the target's own history (CO(GT) lookback); the target is next-hour CO(GT). The split is chronological 60/15/25 train/val/test; lookback is 48 hours; all other hyperparameters match Appendix N.

Direct shared-vs-matched comparison. Unlike Beijing PM2.5, on Italy the matched configuration is descriptively lower in all five seeds: matched test MSE is 0.414 ± 0.021 versus shared-CNN test MSE 0.505 ± 0.020 in squared CO(GT) target units. The mean gap of 0.091 is roughly $4.5\times$

the average per-config seed standard deviation 0.021, and the matched configuration wins in every one of the five seeds. The Transformer baseline at the same training budget reaches 1.712 ± 0.027, again roughly $4\times$ worse than either CNN-based configuration.

Table 11: Second public-data illustration on UCI Air Quality (Italy) next-hour CO(GT) forecasting. Test MSE is reported in original target units squared (CO(GT) units squared) and as standardized MSE; mean $\pm$ standard deviation over five seeds. Same training budget and architecture configurations as Table 10. **Absolute MSE values are not comparable to Table 10 because the targets differ (PM2.5 vs CO(GT) concentration); only the within-table matched-vs-shared pattern is meaningful.** *Unlike Beijing PM2.5, on Italy the matched configuration is lower in all five paired seeds, but the two-sided exact sign test is* $p = 0.0625$; this within-table gap is consistent with, but does not establish, the theorem's latent mismatch conditions, because the bilateral probe gate falls back on every seed.

Model	Params	Test MSE (orig. units squared)	Test MSE (standardized)
Shared-CNN (forced symmetry)	14,803	0.505 ± 0.020	0.2783 ± 0.0110
Matched (CNN + LSTM, heuristic pair)	20,321	$\mathbf{0.414 \pm 0.021}$	$\mathbf{0.2282 \pm 0.0117}$
Transformer (modern baseline)	18,433	1.712 ± 0.027	0.9435 ± 0.0149

Italy retired-probe outcome. The adaptive diagnostic selected different candidates most often (LSTM for meteorology and CNN for CO history), but the CO-history margin missed its threshold in every seed, so the bilateral gate fell back 5/5. This choice also disagreed with the heuristic matched mapping that was descriptively lower in all five direct runs. Because architecture family, tying, and parameter count change together, neither observation identifies a sharing effect or theorem regime; complete scores remain in the artifact.

Joint reading across Beijing and Italy.

- *Beijing PM2.5*: matched is not statistically separated from shared in the direct comparison; the corrected bilateral probe says symmetric in all five seeds. Both observations are compatible with fallback on this dataset, but they do not verify the corollary's population assumptions.
- *UCI Air Quality (Italy)*: matched beats shared by $\sim 18\%$ in direct comparison, while the retired bilateral diagnostic falls back in all five seeds because one modality lacks a reliable margin. Thus the direct comparison and probe rule disagree on this illustration; we report that disagreement as a limitation rather than claiming a validated mismatch-branch selection.
- *Combined*: Beijing gives an internally consistent fallback illustration. Italy shows that a direct matched-vs-shared comparison can favor matching even when the retired adaptive diagnostic declines to route to it. We do not claim generality from two datasets, probe-to-empirical-winner agreement, or end-to-end validation of the protocol. A third public-data illustration with two modern baselines (iTransformer, DLinear) follows in Section N.3 and is integrated in the joint reading there. Italy scripts and raw outputs: `code/public_benchmark_italy.py`; `code/probe_validation_italy.py`; `results/public_benchmark_2/`; `results/probe_validation_2/`.

N.3 Third-Dataset Replication With Modern Baselines: UCI Appliances Energy

To check whether the two-branch reading above generalizes beyond the air-quality domain and to compare against two additional modern baselines under a matched parameter budget, we run the same shared/matched comparison plus the retired adaptive diagnostic on a third independent public dataset, the UCI Appliances Energy Prediction record (Candanedo et al., 2017). The dataset contains 19,735 rows at 10-minute resolution from 2016-01-11 to 2016-05-27 and is independent of Beijing PM2.5 and UCI Air Quality (Italy) in country, sensor stack, and target physics. Modality A is the per-room indoor sensor block (temperatures T_1–T_9 and relative humidities RH_1–RH_9, 18 channels); Modality B is the outdoor weather block ($T_{\mathrm{out}}, \mathrm{RH}_{\mathrm{out}}, P, \mathrm{Wind}, \mathrm{Vis}, T_{\mathrm{dew}}$, 6 channels).

The target is next-step appliance energy use (10 minutes ahead). The split is chronological 60/15/25 train/val/test; the lookback is 144 steps (24 hours).

Configurations. We add two modern reference baselines on top of the three configurations used in Sections N–N.2: a variable-token Transformer baseline in the style of iTransformer (Liu et al., 2024) and a decomposition-linear baseline in the style of DLinear (Zeng et al., 2023). Both are compact reimplementations with the same forward signature as our reference implementations so that they share the harness. Encoder dimension, batch size, optimizer, and epoch budget match Appendix N. The DLinear branch carries a much larger parameter count because it linearizes both the trend and the seasonal residual over the full lookback per channel; we report it as-is rather than after architecture search.

Direct shared-vs-matched-vs-modern comparison. Table 12 reports test MSE in original target units squared and as standardized MSE, mean $\pm$ standard deviation across the same five seeds.

Table 12: Third public-data illustration on UCI Appliances Energy Prediction: shared-vs-matched comparison plus two modern baselines at the same training budget. Test MSE in original target units squared (squared appliance-energy units) and as standardized MSE; mean $\pm$ standard deviation over five seeds. **Absolute MSE values are not comparable to Tables 10 or 11 because the targets differ; only within-table patterns are meaningful.** The shared-minus-matched paired difference is 801.6 with 95% t interval $[-571.4, 2174.6]$ and magnitude-randomization sensitivity $p = .125$, so it is not separable.

Model	Params	Test MSE (orig. units squared)	Test MSE (standardized)
Shared-CNN (forced symmetry)	16,693	8970.73 $\pm$ 1062.99	0.8018 $\pm$ 0.0950
Matched (CNN + LSTM, heuristic pair)	22,401	**8169.14 $\pm$ 200.43**	**0.7302 $\pm$ 0.0179**
Transformer (modern baseline)	19,073	8590.44 $\pm$ 306.43	0.7678 $\pm$ 0.0274
iTransformer (modern baseline)	22,881	8326.66 $\pm$ 471.59	0.7442 $\pm$ 0.0422
DLinear (modern baseline)	223,425	13883.41 $\pm$ 4017.59	1.2409 $\pm$ 0.3591

Honest reading on Appliances Energy.

- *Matched is numerically best, but the paired evidence does not separate.* Shared-minus-matched is 801.6 with 95% t interval $[-571.4, 2174.6]$ and magnitude-randomization sensitivity $p = .125$. Matched wins in 4 of 5 seeds; in seed 1024 Shared-CNN narrowly wins. We read this as a Beijing-style fallback observation rather than evidence that the theorem's mismatch premises hold.
- *iTransformer is competitive with matched at this budget; DLinear is not.* The iTransformer baseline lands within Matched's seed band (8327 vs 8169); the DLinear baseline is roughly 70% worse at the same training budget despite carrying ten times the parameters. We do not interpret this as a general statement about iTransformer or DLinear; it is one dataset at one training budget.
- *The probe again disagrees with the manuscript's heuristic mapping.* The 5-seed retired adaptive probe run on the same splits selects LSTM as the per-modality top-1 candidate for both modalities in 4 of 5 seeds, including the indoor block where the former heuristic mapping would predict CNN. The conservative margin gate ($C = 0.5$) recommends symmetric fallback in 4 of 5 seeds, with a single asymmetric recommendation on seed 123. This pattern is closest to Beijing's, with the same heuristic–probe discrepancy that Italy reported.

Joint reading across three datasets. The third dataset lands on the same conservative branch as Beijing (symmetric fallback): the probe says fallback on 4 of 5 seeds and the paired empirical interval crosses zero. Italy is the only illustration whose paired t interval excludes zero, but its exact sign sensitivity is $p = .0625$ and the run does not establish the assumptions needed to privilege the t interval; we therefore do not call the gap statistically established. Its bilateral gate also falls back in every seed, so it is not an end-to-end validation of the mismatch branch. The matched configuration's numerical lead on Appliances is not statistically separated, and we do not claim it

as positive evidence for the order-$\sqrt{d}$ bound. The bound supplement retains the three scripts, raw outputs, and run guide; its inclusion manifest gives exact paths.

A frozen run that produced no usable evidence. A CPU-only bridge emitted no endpoint or output. The Italy trace had 9,471 raw rows (9,357 usable), below the 14,368-row split; the process exited nonzero with 0/960 cells and no aggregation. It supports no model comparison and no rank or overlap claim.

Capacity-aware completion. A separate protocol used a 4,096/1,024/4,096 split with lookback 32, ranks $\{1, 4, 8\}$, overlap labels $\{0, .5, 1\}$, and seeds $\{42, 123, 456, 789\}$. Italy supplied 9,357 usable rows against 9,248 required. The run completed 108 recorded cells and 216 CNN fits. The raw output is 76.67 MSE (seed-cluster 95% CI $[23.87, 148.62]$), while the duplicate-collapsed audit has 36 source–rank–seed units, 27 untied wins, and geometric tied/untied ratio 2.16. Because overlap labels do not alter public inputs or targets, they are duplicate records rather than independent cells and support no overlap effect.

O CONDITIONAL STABILITY UNDER OUTPUT ROTATION

Assume independent $X_A, X_B \sim \mathcal{N}(0, I_d)$ and independent noise $\varepsilon \sim \mathcal{N}(0, vI_r)$, where $v > 0$, $2r \le d$, and $0 \le p \le d$ is integer. Let $U_A, U_B \in \mathbb{R}^{d\times r}$ have orthonormal columns with $U_A^\top U_B = \operatorname{diag}(s_1, \dots, s_r)$. For orthogonal $R \in \mathbb{R}^{r\times r}$, replace the target by $Y_R = U_A^\top X_A + R^\top U_B^\top X_B + \varepsilon$. This preserves task subspaces and principal cosines, not the target.

For a rank-p orthogonal projector P, define the residual covariance and information gap, in nats, by

$$S_R(P) = (v+2)I_r - U_A^\top P U_A - R^\top U_B^\top P U_B R, \qquad \Delta_R(P) = \tfrac{1}{2}\log\det(S_R(P)/v).$$

Write $\Delta^\star_{R,p} = \inf_P \Delta_R(P)$; I denotes the identity rotation. The nuclear norm $\|\cdot\|_*$ is the sum of singular values, and $\|\cdot\|_{\mathrm{op}}$ is the largest singular value.

Proposition and proof: optimum stability. The conditional certificate is

$$|\Delta^\star_{R,p} - \Delta^\star_{I,p}| \le \epsilon_R, \qquad \epsilon_R = \min\left\{\frac{\|R-I\|_*}{v}, \frac{r}{2}\log(1+2/v)\right\}.$$

Indeed, $M = U_B^\top P U_B$ satisfies $0 \preceq M \preceq I_r$, and

$$R^\top M R - M = (R^\top - I)MR + M(R-I), \qquad \|R^\top M R - M\|_* \le 2\|R-I\|_*.$$

Every matrix on the segment joining $S_I(P)$ and $S_R(P)$ is bounded below by vI_r. Integrating the log-determinant derivative and using $|\operatorname{tr}(AB)| \le \|A\|_{\mathrm{op}}\|B\|_*$ gives $|\Delta_R(P) - \Delta_I(P)| \le \|R-I\|_*/v$. Taking infima over the identical projector class preserves this bound. Both gaps also lie in $[0, (r/2)\log(1+2/v)]$, proving the capped certificate.

Let $L_u^\star$ denote the optimal untied gap under the manuscript's dimension budget and set $\Gamma_I = L_u^\star - \Delta^\star_{I,p}$. Right rotations preserve the untied optimum: separately chosen captured output subspaces can still be oriented arbitrarily at each allocated rank. They also preserve the dimension count. Thus $\Gamma_I > \epsilon_R$ certifies tied advantage, while $\Gamma_I < -\epsilon_R$ certifies untied advantage. Otherwise the comparison is `UNCERTIFIED`, not equality. Small noise can make the band vacuous; paired-model ties generally receive no certificate.

Proposition and proof: exact rotating plane. Take $d = 4$, $r = 2$, $U_A = U_B = (e_1, e_2)$, where e_i are standard basis vectors, and let $R = R_\theta$ be planar rotation through θ. For $p = 1$, a unit encoder whose squared projection into this plane is $t \in [0, 1]$ captures eigenvalues $t(1 \pm |\cos\theta|)$. Increasing t cannot increase the residual determinant, so $t = 1$ attains the global minimum, including against encoders outside the plane:

$$\Delta^\star_{R_\theta,1} = \tfrac{1}{2}\log\frac{(v+1)^2 - \cos^2\theta}{v^2}.$$

The endpoints recover aligned and quarter-turn losses. At $p = 0$ the gap is $\log(1+2/v)$; at $p \ge 2$ it is zero.

Local numerical evidence and limitations. The accompanying artifact records leave sources unchanged. Exactly 240 fits returned `EVALUATED`; 800 optimizer calls retained three nonconverged starts. There were no Schur failures or below-band gaps, but 50 attained gaps exceeded the optimum band through fitting error. Maximum Schur discrepancy was $2.6645352591003757 \times 10^{-15}$; maximum full-rank gap was $1.7763568394002505 \times 10^{-15}$. Selected rank-three means below use five seeds per cell, $d = 6$, $r = 3$, $v = 0.25$, equal principal cosine s, and rotation of the first two output coordinates.

s	θ	Population	Held-out
0.25	0	1.792827	1.790228
0.25	$\pi/2$	1.762831	1.768582
0.80	0	0.901560	0.902820
0.80	$\pi/2$	0.887350	0.892327

Population fits supply attainable upper bounds, not certified optima; held-out estimates additionally reflect sampling/readout error. The band concerns the optimized infimum, not every fitted value. These local preparation artifacts are not publicly released data and establish neither a general nonaligned frontier nor public-data premises, nonlinear superiority, or independent human validation.

Reproduction record. The entry point is `v23_learned_rotation.py`. The result records for this appendix are held together in the accompanying artifact record. Source and configuration bindings accompany the local records; this does not establish independent external replication.

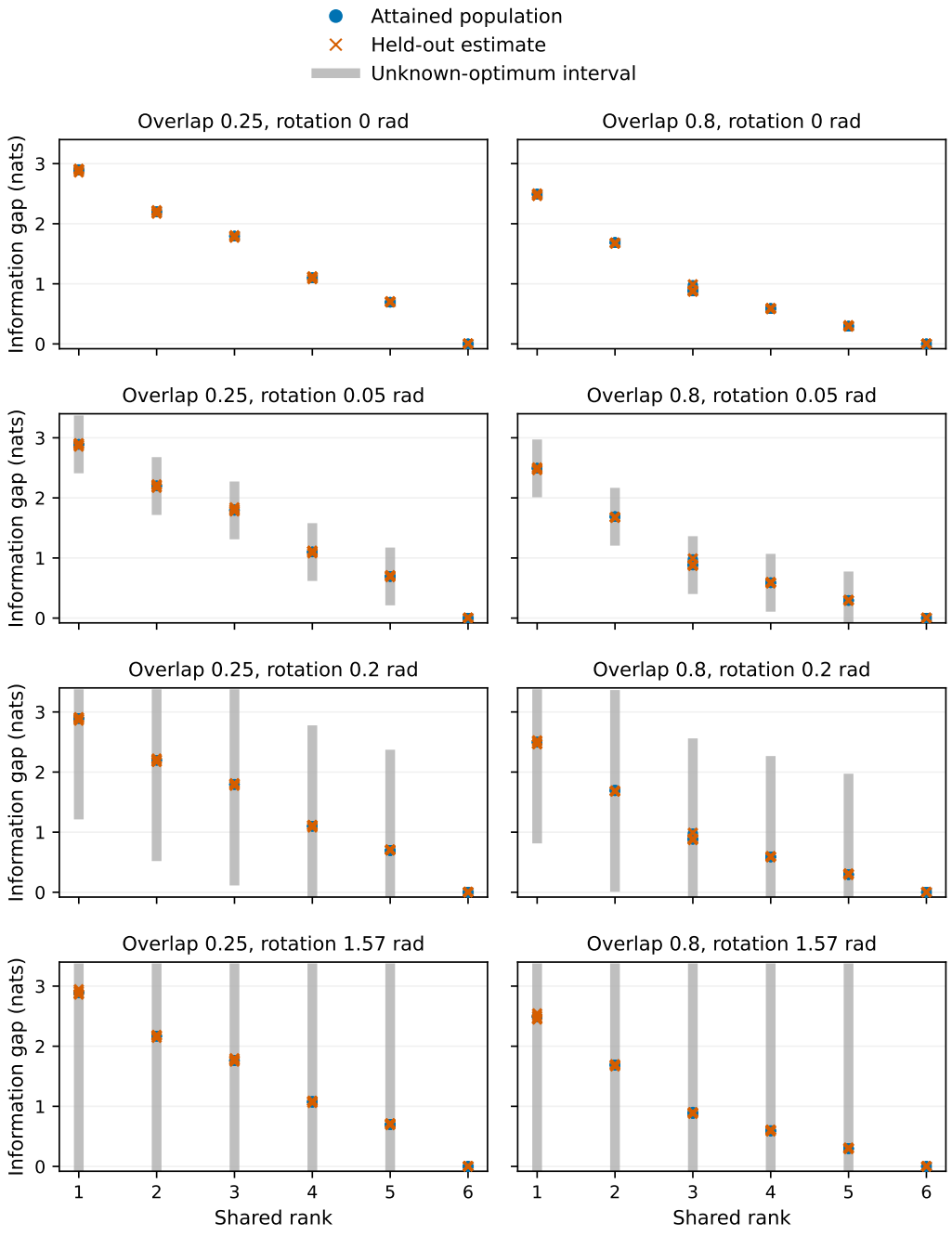

Figure 2: All 240 fits: attained population gaps (blue), held-out estimates (orange), and certified intervals for the unknown optimum (gray). Fitted values need not lie inside the optimum interval.

Capacity signs beyond exact output alignment. We apply the same certificate at $d = 6, r = 3$, $s \in \{.25, .8\}$, $v \in \{.05, .25, 1, 4\}$, all $p = 0, \ldots, 6$, and the eight angles below. The shared budget remains $p(d-p) + 2rp$; we enumerate every feasible untied rank pair under that same budget. The aligned tied optimum is the exact finite allocation minimum of Proposition 2. For rotation of the first two output coordinates, $\|R_\theta - I\|_* = 4\sin(\theta/2)$ on this angle range.

The code evaluates the finite minima and perturbation radius with 60-digit `mpmath` interval arithmetic, then rounds binary64 endpoints outward. For candidate intervals $[L_j, U_j]$, the finite minimum lies in $[\min_j L_j, \min_j U_j]$. Combining these enclosures with the uniform perturbation bound gives an interval for $L_u^\star - \Delta_{R,p}^\star$; a strict sign is reported only if that interval excludes zero. No sign is assigned from a rounded near-zero scalar.

θ	Tied-favoring	Untied-favoring	Uncertified	Width $<$ .01 nats
0	27	11	2	40
.005	25	9	6	10
.01	24	8	8	10
.02	19	8	13	0
.05	16	5	19	0
.1	9	2	29	0
.2	2	0	38	0
$\pi/2$	0	0	40	0

Each row has 40 cells with $1 \le p \le 5$; the 128 structural endpoint-tie cells at $p = 0, 6$ are excluded from this table but retained in the 448-cell artifact. The last column concerns the nonaligned-optimum bracket width, not estimation error or a statistical confidence interval. Its 20 nonzero angle cells all have $v = 4$ and $\theta \in \{.005, .01\}$; a uniform .01-nat global-optimality result is not established.

Strict signs survive in 127 of 280 interior cells at nonzero angles. For example, $(p, s, v, \theta) = (3, .8, .25, .05)$ is tied-favoring, whereas $(3, .25, .25, .01)$ is untied-favoring. The radius divided by the certified absolute aligned margin is retained for every cell whose aligned sign is nonzero; it is undefined at unresolved aligned ties. These are deterministic grid fractions, not probabilities over architectures or population data.

The entry point `v24_rotated_capacity_certificate.py` produces the accompanying artifact record. All 448 aligned calculations were cross-checked against the existing capacity implementation. Five regressions check high-precision endpoint enclosure, both strict signs, large-angle abstention, structural ties and monotone widening. The artifact records `mpmath` version, precision and source hashes. The method bounds an unknown optimum using the existing stability theorem; it does not solve the general unequal-overlap or nonaligned global optimization problem.

P BUDGET-CONVENTION SENSITIVITY OF THE CAPACITY SIGN

Which capacity signs survive a change of budget convention. Proposition 6 matches a Grassmann subspace dimension plus unrestricted target readouts. That rule is a modelling choice, so we recompute the identical exact frontiers on the same 2,160-cell equal-overlap grid under five declared conventions, changing only the feasible set of untied rank allocations. The published convention is reproduced exactly (0 disagreements with the audited implementation). The five feasible sets are nested at every one of the 90 (d, r, p) shapes, giving 7 inclusions that hold shape by shape. Because the untied frontier is a minimum over that set, a larger set can never give a worse untied optimum; across all 2,160 cells this monotonicity is violated 0 times. The least and most permissive rules, $q_A + q_B \le p$ and $q_A + q_B \le 2p$, therefore bracket every attainable sign, and the published rule lies strictly inside the bracket.

Budget convention	Tied	Equal	Untied	Reversed
$q_A + q_B \le p$	1290	870	0	91
$\nu_U \le \nu_T$ (published)	1199	740	221	0
$\max(q_A, q_B) \le p$	0	360	1800	1199
raw coefficients	889	617	654	326
$q_A + q_B \le 2p$	0	360	1800	1199

Columns count cells of the fixed grid; *Reversed* counts strict sign flips against the published convention. The direction of the comparison is not convention free: 1,290 of 2,160 cells flip strictly between the two extremes, and 1,199 cells that the published rule reports as tied-favoring become untied-favoring once the shared subspace is charged per modality. Consequently 0 cells exhibit a reuse advantage under all five conventions. Two conclusions are convention independent. First, no convention manufactures a reuse advantage at disjoint targets: exactly 0 cells with $s = 0$ are tied-favoring under any rule, and no $s = 0$ cell reverses. Second, the 360 cells on which all five rules agree are precisely the 360 saturated cells $p = 2r$, where the comparison is an exact tie. Reversals otherwise grow with overlap: $s = 0$ gives 0, $s = 0.1$ gives 180, $s = 0.4$ gives 240, $s = 0.6$ gives 270, $s = 0.8$ gives 300, $s = 0.95$ gives 300 reversing cells. This delimits Proposition 6 rather than contradicting it: under its own stated rule the theorem and its signs are unchanged. The audit is a deterministic recomputation of exact equal-overlap frontiers on one fixed grid. It establishes no nonaligned global optimum, no learned-encoder behaviour, no statistical coverage statement, and no architecture recommendation, and it does not identify which accounting rule a practitioner should prefer.

Reproduction record. The entry point `v25_budget_convention_sensitivity.py` produces the accompanying artifact record. Thirteen regressions check exact reproduction of the published signs, feasible-set nesting, monotonicity, the zero-overlap invariance and the absence of a convention-independent reuse advantage. The receipt records source hashes for the audit and both existing capacity modules.

Q CERTIFIED UNEQUAL-OVERLAP AND ROTATED FRONTIER STUDY

The unequal-overlap counterexample of Appendix G shows one point where cross-block mixing beats every block-aligned choice; the budget-convention study changes the accounting rule but keeps the equal-overlap grid. The 2026-09-12 post-run audit reuses the low-dimensional grid (`v26_unequal_rotated_frontier.py`, with its corrected evidence entry in the accompanying artifact record): eight unequal cosine spectra at target ranks $r = 2, 3$, with encoder rank $p = r$ only, crossed with $\sigma^2 \in \{.05, .25, 1\}$. A floating 41^2 search chooses a feasible encoder; its analytic objective, the finite block-aligned minimum and the Ky Fan lower bound are then enclosed using 60-digit interval arithmetic with outward export. The displayed decimals denote exact inputs. Floating orthogonality checks alone are not the certificates.

The old rank-two "exact global" value zero was imported from a different, common-plane family. It is inapplicable to these unequal spectra: their strictly positive Ky Fan lower bounds already exclude zero. This v26 bracket alone does not solve the global unequal-overlap optimum; the two-output refinement below supplies a separate applicable formula. A SUBOPTIMAL certificate requires the witness interval's upper endpoint to lie strictly below the block optimum's lower endpoint. OPTIMAL would require a global lower bound reaching the block upper bound, without a tolerance-based equality claim; all other cases remain OPEN.

Analytic witness and interval certificate. For the two orthogonal plus/minus mixtures used by the search, put $x, y \in \{0, 1/40, \dots, 1\}$ and $v = \sigma^2$. Their captured two-target covariance has

$$C_{11} = x(1 + s_1) + (1 - y)(1 - s_1),$$
$$C_{22} = (1 - x)(1 + s_2) + y(1 - s_2),$$
$$C_{12} = \sqrt{x(1 - x)(1 + s_1)(1 + s_2)} + \sqrt{y(1 - y)(1 - s_1)(1 - s_2)}.$$

The witness is orthonormal by construction, and its rank-two gap is $\frac{1}{2}\log(((v + 2 - C_{11})(v + 2 - C_{22}) - C_{12}^2)/v^2)$. At $r = p = 3$, retaining the third plus direction adds $\frac{1}{2}\log((v + 1 - s_3)/v)$. The

Ky Fan lower bound is $(r - \sum_i s_i)\log(1 + 2/v)/4$. These formulas, and finite enumeration of all block allocations, produce the intervals below. A feasible witness need not be the best grid point to certify an upper bound.

	$\sigma^2 = .05$	$\sigma^2 = .25$	$\sigma^2 = 1$
$(.05, .45)$	suboptimal $(+.022)$	suboptimal $(+.022)$	suboptimal $(+.022)$
$(.25, .75)$	suboptimal $(+.045)$	open	open
$(.05, .85)$	suboptimal $(+.121)$	suboptimal $(+.042)$	open
$(.45, .85)$	suboptimal $(+.038)$	open	open
$(.05, .45, .85)$	suboptimal $(+.022)$	suboptimal $(+.022)$	suboptimal $(+.022)$
$(.25, .50, .75)$	suboptimal $(+.009)$	suboptimal $(+.009)$	open
$(.10, .40, .90)$	suboptimal $(+.012)$	suboptimal $(+.012)$	suboptimal $(+.012)$
$(.05, .20, .80)$	suboptimal $(+.003)$	suboptimal $(+.003)$	suboptimal $(+.003)$

Table 13: Initial v26 unequal-overlap bound audit. Each entry is the certified classification with the block-optimum-minus-witness margin in nats; "open" means neither strict suboptimality nor block-aligned optimality is certified. Margins are rounded descriptive values; the interval endpoints determine the classifications. This bracket alone certifies no block-aligned optimum.

Cross-block mixing is not a single-point curiosity: at the lowest noise every spectrum certifies suboptimality of the block-aligned class, and the phenomenon persists across the grid except where high noise opens the bracket. Eighteen cells certify strict suboptimality; five rank-two cells and one rank-three cell remain open at this stage. The following two-output refinement closes the five rank-two cases.

Cosine spectrum	$\sigma^2 = .05$	$\sigma^2 = .25$	$\sigma^2 = 1$
$(.05, .45)$	$[1.3925, 1.8350]$	$[.8239, 1.0769]$	$[.4119, .5275]$
$(.25, .75)$	$[.9283, 1.8124]$	$[.5493, 1.0398]$	$[.2746, .3914]$
$(.05, .85)$	$[1.0212, 1.7359]$	$[.6042, .9773]$	$[.3021, .4038]$
$(.45, .85)$	$[.6498, 1.8190]$	$[.3845, .8166]$	$[.1922, .2891]$
$(.05, .45, .85)$	$[1.5318, 2.5281]$	$[.9063, 1.3119]$	$[.4531, .5974]$
$(.25, .50, .75)$	$[1.3925, 2.7436]$	$[.8239, 1.4361]$	$[.4119, .5942]$
$(.10, .40, .90)$	$[1.4854, 2.3942]$	$[.8788, 1.2547]$	$[.4394, .5849]$
$(.05, .20, .80)$	$[1.8103, 2.6588]$	$[1.0711, 1.3897]$	$[.5355, .6377]$

Table 14: Global-optimum enclosures in nats, rounded outward to four decimals. The upper endpoint is the smaller of the feasible witness and block-aligned upper bounds. These are deterministic mathematical enclosures, not statistical confidence intervals or exact optima.

Q.1 Exact unequal-overlap refinement for two outputs

The coarse bounds above can be closed when $r = 2$. This refinement uses the paired-principal-filter assumption, not the unrelated common-plane rotation formula that the earlier audit incorrectly imported.

Proposition 8 (Unequal two-output global frontier)**.** *Let $r = 2$, $d = 4$, $U_A^\top U_B = \mathrm{diag}(s_1, s_2)$ with $0 \le s_1 \le s_2 < 1$, and $v = \sigma^2 > 0$. Define*

$$\delta = s_2 - s_1, \quad M = v + 2, \quad a = (1 + s_1)(1 - s_2), \quad b = (1 + s_2)(1 - s_1), \quad K = \sqrt{a} + \sqrt{b},$$

$$x^\star = \mathrm{clip}_{[0,1/2]} \frac{K\sqrt{b} - M\delta}{K^2}, \qquad D_{\mathrm{mix}}(x) = M^2 - M\{2 + \delta(1 - 2x)\} + (Kx - \sqrt{b})^2.$$

The unrestricted shared-encoder optimum at every integer rank is

$$\Delta_0^\star = \log(M/v), \qquad \Delta_1^\star = \tfrac{1}{2}\log\frac{M(v + 1 - s_2)}{v^2},$$

$$\Delta_2^\star = \tfrac{1}{2}\log\frac{\min\{(v + 1 - s_1)(v + 1 - s_2), D_{\mathrm{mix}}(x^\star)\}}{v^2}, \quad \Delta_3^\star = \tfrac{1}{2}\log\frac{v + 1 - s_2}{v}, \qquad \Delta_4^\star = 0.$$

Proof. Proposition 1 reduces the global minimum to projector pairs A, B with ranks summing to p. At $p = 2$, the pairs $(I, 0)$ and $(0, I)$ are block aligned, and the latter is dominated because

$D_+ \succeq D_-$. For rank-one A, B, fixing squared coordinates x, y and aligning signs maximizes the absolute captured off-diagonal entry, thereby minimizing the two-dimensional residual determinant. It becomes

$$D(x,y) = M^2 - M\{2 + \delta(1-x-y)\} + \{\sqrt{axy} - \sqrt{b(1-x)(1-y)}\}^2.$$

When $\delta > 0$ and $u = x+y \le 1$, write $t = xy \in [0, u^2/4]$. Differentiation shows that the last square is minimized at $t_0 = a(1-u)/(b-a)$ if $t_0 \le u^2/4$, and at $u^2/4$ otherwise. In the first region the determinant is $M^2 - 2M - \delta v(1-u)$, increasing in u; its minimum is therefore at the boundary with the second region. That boundary also has $x = y$. For $u \ge 1$, $D(x,y) \ge M^2 - 2M$, already attainable by keeping one complete principal plane. Consequently a global rank-one-pair minimum occurs at $x = y \in [0, 1/2]$. The case $\delta = 0$ directly attains the same reduction at $x = y = 1/2$. Substitution gives the convex quadratic D_{mix} and the clipped minimizer above; compare it with $(I, 0)$.

At $p = 1$, the determinant lemma reduces the choice to maximizing a Rayleigh quotient of D_+ or D_-, attained by the largest plus direction $1 + s_2$. At $p = 3$, for $(A, B) = (I, zz^\top)$ the same lemma maximizes $(1-s_i)/(v+1-s_i)$ at $i = 1$, giving determinant $v(v+1-s_2)$. The alternative $(zz^\top, I)$ gives $v(v + 1 + s_1)$ and is no better. The zero and full-rank cases are immediate. All selected projectors are feasible, so the lower bounds are attained. □

Cosines	v	$\Delta_2^\star$	Global allocation
$(.05, .45)$	.05	1.83491759	mixed
$(.05, .45)$	.25	1.07674385	mixed
$(.05, .45)$	1	.52743770	mixed
$(.25, .75)$	.05	1.81180574	mixed
$(.25, .75)$	.25	1.03972077	block aligned
$(.25, .75)$	1	.39137967	block aligned
$(.05, .85)$	.05	1.73531178	mixed
$(.05, .85)$	.25	.97713803	mixed
$(.05, .85)$	1	.40379566	block aligned
$(.45, .85)$	.05	1.81859706	mixed
$(.45, .85)$	.25	.81657722	block aligned
$(.45, .85)$	1	.28900844	block aligned

Table 15: Two-output rank-two global optima (nats, rounded for display). Seven cells are strictly better than every block-aligned allocation; the five cells left open by the v26 bracket are now proved block-aligned optimal. The three-output open cell and rotated three-output failures remain open.

Executable verification and scope. `v27_rank2_global_frontier.py` evaluates all $p = 0, \ldots, 4$ on the four spectra and three variances, producing 60 outward-rounded intervals at 60-digit working precision. Their maximum width is below $3 \cdot 10^{-15}$. The accompanying artifact record binds the formulas and source hashes. Fourteen related tests include equal-overlap recovery, independent joint-Gaussian Schur conditioning at the attained encoders, positive unequal-case gaps, rank monotonicity and arbitrary projector checks. Numerical checks supplement the proof; they do not establish globality by sampling. This result covers paired output coordinates and two outputs only, not arbitrary right rotations, three-output unequal overlap or learned nonlinear representations.

Rotated rank-three brackets: explicit failure and its meaning. For the rotated rank-three cells of the output-rotation audit at angles .05, .2, and $\pi/2$ (overlaps .25, .8; ranks 1–5; the same three noise levels), the uniform perturbation bracket never tightens below .01 nats: all 90 cells are reported as explicit tightening failures, with bracket widths recorded per cell. This is failure of the available certificate to meet the requested precision, not evidence of fragile signs. A *separate* rank-two common-plane family, at encoder rank one and $\sigma^2 = 1$, has rotation effects approximately .00042 nats at angle .05 and .14384 at $\pi/2$, while the uniform bounds are .10 and 1.10. These effects subtract the zero-angle optimum; they are not the rotated gap itself. This illustrates looseness only in that common-plane control and does not measure the true rank-three effect. Where the rank-three bracket exceeds the aligned sign margin, the cell stays UNCERTIFIED, and we report it as such rather than claiming either sign.

Table 16: Theorem-invocation audit. P1–P4 refer only to Lemma 1; Theorem 1 instead derives its rank bottleneck in the stated Gaussian linear class.

Gate	Meaning	Finite selector witness	ETTh1 learned representations
P1	omitted-block conditional information	PASS	FAIL
P2	class-wide retained-coordinate bound	PASS	NOT_TESTED
P3	positive cross-modal nonredundancy	PASS	NOT_TESTED
P4	bilateral retained-block measurability	PASS	FAIL
Joint	all four	52,475/52,475 PASS	FAIL: no invocation

R SUPPLEMENTARY STATEMENTS AND PROOFS

R.1 GEOMETRY AND THEOREM-INVOCATION DETAILS

Figure 1 is placed in the main text so that the model, frontier result, and invocation boundary are visible before the formal results.

R.2 COMPLETE PROOF OF THE DISJOINT-FILTER FRONTIER

Let $H = U_A^\top P_W U_A + U_B^\top P_W U_B$. Orthogonality gives $0 \preceq H \preceq 2I$ and $\operatorname{tr} H \le p$. Therefore every feasible projector induces an eigenvalue vector in the relaxed polytope

$$\mathcal{H}_p = \{h \in [0,2]^r : \textstyle\sum_i h_i \le p\}.$$

Minimizing $f(h) = \frac{1}{2}\sum_i \log[1 + (2-h_i)/\sigma^2]$ over $\mathcal{H}_p$ lower-bounds the projector optimum. Monotonicity gives $\sum_i h_i = p$. For $0 < h_i \le h_j < 2$, transferring $\varepsilon = \min\{h_i, 2-h_j\}$ from h_i to h_j cannot increase this concave objective. Repeating leaves at most one coordinate in $(0,2)$, giving

$$h^* = (\underbrace{2,\dots,2}_{q},\ b,\ 0,\dots,0).$$

Thus $f(h(W)) \ge f(h^*)$ for every rank-p projector P_W.

For attainment, choose columns $U_{A,i}, U_{B,i}$ for $i \le q$, adding $U_{A,q+1}$ when $b = 1$. These orthonormal vectors give a rank-p map with

$$U_A^\top P_W U_A + U_B^\top P_W U_B = \operatorname{diag}(\underbrace{2,\dots,2}_{q}, b, 0,\dots,0) = \operatorname{diag}(h^*).$$

Hence the relaxed lower bound is attained by a feasible shared projector, proving global optimality; substituting h^* yields the stated formula.

R.3 A SAMPLE-SPLIT FIXED-FEATURE PROBE

For each of k modalities and candidate $e \in [K]$, a construction split fixes a feature/predictor h_e. A disjoint score split of n independent examples is used only to compute pre-specified bounded scores $S_{i,e} = s(h_e(X_i), Y_i) \in [0,1]$. Let $\widehat{\mu}_e = n^{-1}\sum_i S_{i,e}$, $\widehat{e} = \arg\max_e \widehat{\mu}_e$, and $\widehat{\Delta} = \widehat{\mu}_{\widehat{e}} - \max_{e \ne \widehat{e}} \widehat{\mu}_e$. No encoder fitting, coordinate selection, BH testing, or denominator estimation occurs on the score split.

Proposition 9 (Complete fixed-probe guarantee)**.** *With $\epsilon_n = \sqrt{\log(2kK/\delta)/(2n)}$, all kK score means are within ϵ_n of their population means with probability at least $1-\delta$. On that event, a population winner with gap greater than $2\epsilon_n$ is selected. Certifying a route only when $\widehat{\Delta} > 2\epsilon_n$ cannot certify a population tie.*

Proof. Hoeffding gives $\Pr(|\widehat{\mu}_e - \mu_e| > \epsilon) \le 2e^{-2n\epsilon^2}$. A union bound over kK fixed scores gives the event. Two triangle inequalities prove both winner preservation and tie non-certification. □

Executed calibration. Using two fixed coordinate predictors and $s(y, \widehat{y}) = \exp[-(y-\widehat{y})^2/2]$, we ran 5,000 independent Monte Carlo seeds with $n = 512$ and $\alpha = 0.05$. Under an exchangeable null (Y independent of both features), the gate certified $0/5000$ winners (measured type-I 0.000). Under the witness alternative $Y_A = X_{A,1}+\epsilon, Y_B = X_{B,2}+\epsilon, \epsilon \sim N(0, 0.5^2)$, single-modality certification and the bilateral different-winner route both occurred in $5000/5000$ trials (power 1.000). The single-modality radius was 0.06542 and the bilateral radius 0.07040. These are measured finite-run rates, not exact probabilities; the analytic guarantee is Proposition 9.

Remark 4 (Engineering fallback, not a corollary)**.** *A weak fixed-probe margin triggers a conservative shared-baseline route. It does* not *imply that a symmetric encoder is near-optimal, and failure of Lemma 1's sufficient premises has no converse consequence.*

R.4 TYING EXCESS OVER REDUCED-RANK REGRESSION

Proposition 10 (Tying excess over reduced-rank regression)**.** *Let $\Delta_{\mathrm{RRR}}(p)$ be the minimum common-target gap over linear predictors of rank at most p from the stacked history (X_A, X_B). This is a deterministic rank constraint, not a finite-rate Gaussian information-bottleneck objective. For every principal-cosine spectrum, $\Delta_{\mathrm{RRR}}(p) = (r - \min\{p, r\})\, c_0$. For equal cosine s and $1 \le p \le r$, the excess of the exact tied optimum is*

$$E(p, s) = \Delta_p^\star - \Delta_{\mathrm{RRR}}(p) = \begin{cases} \lfloor p/2 \rfloor\, c_0 + (p - 2\lfloor p/2 \rfloor)\, c_1(s), & s \le s^\star, \\ p\, c_1(s), & s \ge s^\star, \end{cases}$$

and for $r \le p \le 2r$ it equals $\Delta_p^\star$ itself. Hence $E(p, s) > 0$ whenever $1 \le p < 2r$ and $s < 1$. At $p = 2r$ it is zero for every s; for each fixed $p \ge r$ it tends to zero as $s \to 1$.

Proof. $B^\star = [U_A^\top\ U_B^\top]$ has $B^\star B^{\star\top} = 2I_r$ under identity input covariance, so the rank-p optimum drops $r - \min\{p, r\}$ coordinates of variance $2 + \sigma^2$ each, independently of $U_A^\top U_B$. Subtract this from Proposition 2; the slope in t is $c_0 - 2c_1(s)$. Appendix A gives the full computation. □

The stacked predictor spends one direction $(U_{A,i}; U_{B,i})/\sqrt{2} \in \mathbb{R}^{2d}$ per target coordinate; a tied $W \in \mathbb{R}^{d \times p}$ must contain $U_{A,i}$ and $U_{B,i}$ separately, a two-dimensional span unless $s_i = 1$. Every s-dependent statement in Section 2 lives in that difference.

S A TIGHTER CERTIFICATE FOR THE REMAINING PAIRED-PRINCIPAL RANK-THREE CELL

We select the unique unresolved three-output cell from the preserved 24-cell paired-principal frontier: shared rank $p = r = 3$, variance $v = 1$, and exact cosines $s = (1/4, 1/2, 3/4)$. This is a targeted deterministic certificate calculation on a known instance, not a new population sample or an arbitrary-output-rotation experiment.

Write $D_\pm = \mathrm{diag}(1 \pm s_i)$ and let A, B be the diagonal blocks induced by a feasible shared projector. They satisfy $0 \preceq A, B \preceq I$ and $\mathrm{tr}(A + B) = 3$. For the paired-principal residual

$$S = 3I - D_+^{1/2} A D_+^{1/2} - D_-^{1/2} B D_-^{1/2}, \qquad J = \tfrac{1}{2} \log\det S,$$

we have $I \preceq S \preceq 3I$. With $C = \mathrm{diag}(s_i)$,

$$\mathrm{tr}\, S = 6 - \mathrm{tr}\{C(A - B)\} \ge 6 - \mathrm{tr}\, C = 4.5.$$

Consequently, each eigenvalue satisfies $\lambda_i \ge 1$, and the elementary product inequality gives

$$\det S = \prod_{i=1}^{3} \{1 + (\lambda_i - 1)\} \ge 1 + \sum_{i=1}^{3} (\lambda_i - 1) \ge 5/2.$$

This spectral/trace outer relaxation contains every feasible projector; it does not assert that every relaxed eigenvalue triple is realizable. Thus $\ell = \frac{1}{2} \log(5/2)$ is a valid global lower bound.

For an upper witness, enumerate the 20 rank-three coordinate choices among the 4^3 possibilities of taking neither, plus, minus, or both directions in each principal plane. The all-plus choice is feasible and has $S = \mathrm{diag}(7/4, 3/2, 5/4)$, hence $u = \frac{1}{2} \log(105/32)$. It is the minimum of this enumerated coordinate class, not a proved optimum over all shared projectors.

Quantity	Certified value (nats)
Global lower bound ℓ	$\frac{1}{2}\log(5/2) = .4581453659\ldots$
Feasible coordinate upper bound u	$\frac{1}{2}\log(105/32) = .5941122236\ldots$
Bracket width $u - \ell$	$\frac{1}{2}\log(21/16) = .1359668577\ldots$
Prespecified closure tolerance	.01

Outward 80-digit enclosures and an independent higher-precision recomputation agree. The certified width upper endpoint, rounded upward, is $.1359668578$ nats, so the instance remains unresolved. The lower bound improves the prior chord-based floor $.4119796082\ldots$ by approximately $.0461657577$ nats. The feasible upper witness is unchanged; neither a better encoder nor global optimality follows. No architecture recommendation is inferred. The one-node pilot retained its negative closure decision without automatic expansion. Source selection, exact decimal inputs and arithmetic records are retained in the accompanying artifact record.