LOW-INTERACTION-RANK LEARNING: UNIFYING MULTIPLICATIVE DUAL-ENCODER HEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

A multiplicative dual-encoder network computes a real-valued output for a pair of inputs as the inner product of their separate encodings. This architecture has been developed independently in operator learning, bipartite matching, contrastive vision-language models, retrieval, and other areas, yet no unified theory guides the basic design decisions: how many interaction modes to represent, how to normalize the encoders, and when the architecture should be avoided. We provide such a foundation by introducing the class of functions of low interaction rank, a class whose intrinsic complexity is measured by its interaction spectrum. Within this framework, approximation error decomposes into a spectral truncation term and an encoder-realization term; sample complexity is governed by the sum of the two encoder complexities rather than their product; and a usability criterion based on spectral decay determines when the architecture can succeed. The same framework exposes a central identifiability problem: the encoders are defined only up to a linear gauge symmetry that leaves the learned coordinates arbitrary. We show that normalization is gauge fixing and that whitening pins the interaction modes up to permutation and sign, thereby explaining the uninterpretability of contrastive dimensions and providing a constructive remedy. Experiments on synthetic kernels, operator learning, and CLIP models validate the theoretical predictions: spectral decay rates match the predicted scaling, whitening recovers the true modes, and independently trained CLIP models are related by a single rotation which, after removal by whitening, exposes interpretable concept axes. The code of this paper is provided at anonymous repository `https://anonymous.4open.science/r/Mul-Net-FDF0`.

1 INTRODUCTION

A common architectural pattern has emerged independently across several machine learning communities: computing a real-valued output for a pair of inputs as the inner product of their separately encoded representations. Contrastive models use it for cross-modal similarity (Radford et al., 2021; Zhao et al., 2025); operator networks evaluate learned maps via branch-trunk inner products (Lu et al., 2021); retrieval systems rank queries against documents with two-tower encoders (Karpukhin et al., 2020); goal-conditioned reinforcement learning and successor features represent values as inner products of encodings (Schaul et al., 2015; Hong et al., 2021; Barreto et al., 2017); and knowledge-graph completion, linear attention, factorization machines, and multiplicative matching networks all share the same underlying structure (Yang et al., 2014; Trouillon et al., 2016; Katharopoulos et al., 2020; Rendle, 2010; Zhao & Li, 2026b;a).

In each domain, this structure provides tangible computational and representational benefits. Yet the corresponding methods have been developed largely in isolation, even though they are all instances of the same fundamental architecture, differing only in the nature of the two inputs and how the output is consumed. Consequently, each community independently confronts the same design questions: how many interaction modes to retain, how to normalize the encoders to ensure stable training and meaningful representations, and when the architecture is fundamentally ill-suited.

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.

We now formalize this shared architecture: a multiplicative dual-encoder head computes

$$\boxed{F(u,v) \approx \langle f_\theta(u), g_\varphi(v)\rangle = \sum_{k=1}^{d} f_k(u)\, g_k(v)} \tag{1}$$

where f_θ and g_φ are learned encoders mapping the two inputs into a common space $\mathbb{R}^d$. We study the class

$$\mathcal{M}_d = \{\, F :\ F(u,v) = \langle f(u), g(v)\rangle,\ f : U \to \mathbb{R}^d,\ g : V \to \mathbb{R}^d \,\} \tag{2}$$

of functions representable by a rank-d head. The complexity of a target F within this class is measured by its interaction spectrum $\{\sigma_k\}_{k\geq 1}$, the singular values of the integral operator whose kernel is F. Its interaction rank is the number of nonzero singular values.

Within this framework, we answer four questions about approximation error, encoder identifiability, sample complexity, and fundamental architectural limitations.

- **Approximation:** The error of any rank-d head decomposes into a spectral truncation term, unavoidable by any encoder, and an encoder-realization term. Target smoothness controls the decay rate. (Section 2)
- **Identifiability:** The representation (1) is invariant under a linear gauge symmetry acting on the pair of encoders, so the encoders of any trained model are defined only up to this symmetry. We show that the normalizations used across these domains are precisely gauge-fixing choices, and we prove that whitening pins the interaction modes up to permutation and sign. This explains why the individual dimensions of contrastive models lack inherent semantic meaning and provides a constructive remedy. (Section 3)
- **Estimation:** Sample complexity is governed by the sum of the two encoder complexities rather than their product. For smooth targets, the optimal rank grows only logarithmically in the sample budget. (Section 4)
- **Usability:** A flat interaction spectrum forces every rank-d head to a relative error floor of $1 - d/N$, where N is the size of the discrete input domain. This barrier is escaped by early-interaction networks and yields a practical criterion based on the measured spectrum. (Section 4)

To further validate our conclusions, we conducted experiments on synthetic kernels, operator learning, and CLIP models, confirming the predicted spectral behavior, demonstrating unique mode recovery under whitening, and verifying the flat-spectrum error floor. (Section 5)

2 The Low-Interaction-Rank Class: Unification and Approximation

Section 1 introduced the class $\mathcal{M}_d$ of rank-d heads. We now associate with every target a spectral object that measures its intrinsic complexity within this class, show that ten existing method families are instances of the same architecture, and develop an approximation theory that quantifies how well a rank-d head can represent a given target.

2.1 The class and the interaction spectrum

Let $\mathcal{U}$ and $\mathcal{V}$ be compact metric spaces with probability measures μ_U, μ_V, and let $F \in L^2(\mu_U \otimes \mu_V)$ be the target function of two arguments. A rank-d head is a pair of encoders $f : \mathcal{U} \to \mathbb{R}^d$, $g : \mathcal{V} \to \mathbb{R}^d$ with coordinate functions $f_k \in L^2(\mu_U)$, $g_k \in L^2(\mu_V)$.

Definition 1 (Interaction spectrum)**.** *For $F \in L^2(\mu_U \otimes \mu_V)$, the* interaction operator *is the integral operator $T_F : L^2(\mu_V) \to L^2(\mu_U)$ defined by*

$$(T_F h)(u) = \int_{\mathcal{V}} F(u,v)\, h(v)\, d\mu_V(v). \tag{3}$$

Since $F \in L^2(\mu_U \otimes \mu_V)$, this operator is Hilbert-Schmidt, and F admits the Schmidt decomposition

$$F = \sum_{k\geq 1} \sigma_k\, a_k \otimes b_k, \qquad (a_k \otimes b_k)(u,v) = a_k(u)\, b_k(v), \tag{4}$$

with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ and orthonormal systems $\{a_k\} \subset L^2(\mu_U)$, $\{b_k\} \subset L^2(\mu_V)$. The sequence $\{\sigma_k\}$ is the interaction spectrum *of F; the number* $\text{i-rank}(F) = \#\{k : \sigma_k > 0\}$ *is its* interaction rank*; and the functions a_k, b_k are its* interaction modes*. The class introduced in Section 1 satisfies $\mathcal{M}_d = \{F : \text{i-rank}(F) \leq d\}$, and the spectrum depends on the reference measures.*

By the isometry between $L^2(\mu_U \otimes \mu_V)$ and the Hilbert-Schmidt class, a function is representable by a rank-d head if and only if $\text{rank}(T_S) \leq d$, hence the parametric class (2) coincides with $\mathcal{M}_d$ (Lemma 1, Appendix B). The spectral truncation term $\sum_{k>d} \sigma_k^2$ depends only on the target and the reference measures, not on the encoders: no head, regardless of its capacity, can represent F with error below this quantity. The bound is attained in L^2 by the truncated Schmidt series, and is unique when $\sigma_d > \sigma_{d+1}$ (Corollary 1, Appendix B). The decay rate of this tail, which is controlled by the smoothness of F (Theorem 3), determines how many interaction modes d a good head must retain.

This is the operator-theoretic analogue of classical low-rank matrix approximation: the Schmidt decomposition is the infinite-dimensional Eckart-Young theorem (Schmidt, 1907; Eckart & Young, 1936), and the truncation term is exactly the Hilbert-Schmidt error of the best rank-d approximation of T_F. The spectrum is relative to the reference measures μ_U, μ_V, which fix the inner product in which the modes are orthonormal. The same target may have different spectra under different measures. All subsequent claims therefore hold for the measures under which a head is trained and evaluated.

2.2 Ten families as instances of the head

Appendix A collects ten method families from different communities in Table 5. Each computes a real-valued output for a pair of inputs as the inner product of two separately learned encodings. The questions of how many interaction modes d to keep, how to normalize the two encoders (Section 3), and how many samples the head requires (Section 4) are answered independently by each community. Within the present framework, they are the same questions.

2.3 Approximation and separation from single-tower models

We now study how well $\mathcal{M}_d$ approximates a fixed target. Corollary 1 isolates the part of the error that no encoder can avoid. The remaining part is attributable to the encoders themselves. We assume that the encoder classes contain functions approaching the scaled modes $\sigma_k a_k$ and b_k with errors $\eta_{f,k}, \eta_{g,k}$ as in (20) (Assumption 1, Appendix B). For neural encoders, these errors follow standard ReLU approximation rates (Yarotsky, 2017).

Theorem 1 (Approximation error decomposition)**.** *Let $F \in L^2(\mu_U \otimes \mu_V)$ have interaction spectrum $\{\sigma_k\}$. Under Assumption 1, with $M_f = \max_k \|\hat{f}_k\|_{L^2(\mu_U)}$ and $M_g = \max_k \|\hat{g}_k\|_{L^2(\mu_V)}$,*

$$\inf_{\substack{f \in \mathcal{F} \\ g \in \mathcal{G}}} \|F - \langle f, g \rangle\|^2_{L^2(\mu_U \otimes \mu_V)} \;\leq\; \underbrace{\sum_{k>d} \sigma_k^2}_{\textit{truncation term}} \;+\; \underbrace{C \sum_{k \leq d} \big(\sigma_k^2\, \eta_{g,k}^2 + M_g^2\, \eta_{f,k}^2\big)}_{\textit{realization term}}, \tag{5}$$

where C is an absolute constant.

Proof idea. The error is split around the truncated Schmidt series, and the tensor-product product rule is applied term by term. Orthonormality of the ideal modes diagonalizes the two leading sums and avoids a factor of d. The remaining term is bounded by Cauchy-Schwarz. The full calculation, including the explicit constant, is deferred to Appendix B. □

The decomposition is tight up to the constant: the truncation term is intrinsic to the target, while the realization term depends on the model. Dense encoder classes therefore approach the lower bound of Corollary 1. Whether training actually finds the coordinate alignment required by (20) is precisely the identifiability question addressed in Section 3.

Target smoothness controls the truncation term: an s-smooth target has spectrum $\sigma_k = O(k^{-s/m_V})$, while real-analytic targets decay exponentially. The precise tail bounds are given in Theorem 3 (Appendix B).

The dual-encoder head separates from single-tower models on both the approximation and the computational axes. On the approximation axis, it avoids the curse of the joint dimension: approximating a rank-d target within ε with ReLU encoders requires

$$N_{\mathrm{dual}} = O\big(d(\varepsilon^{-m_U/s} + \varepsilon^{-m_V/s})\big) \tag{6}$$

parameters, whereas any single-tower model that treats F as a general $(m_U + m_V)$-dimensional H^s function requires $\Omega(\varepsilon^{-(m_U+m_V)/s})$ parameters in the worst case, hence $N_{\mathrm{dual}}/N_{\mathrm{single}} \to 0$ as $\varepsilon \to 0$ (Proposition 2). On the compute axis, evaluating the head on all pairs costs $O((n+m)\,C_{\mathrm{net}} + nm\,d)$ against quadratic joint processing, recovering the complexity arguments of QK-attention matching (Zhao & Li, 2026a), linear attention (Katharopoulos et al., 2020), and DeepONet (Lu et al., 2021) (Proposition 1, Appendix B).

3 Identifiability as Gauge-Fixing

The theory of Section 2 characterizes what a rank-d head can represent, but it does not specify which representation a training algorithm will actually find. The representation is far from unique: for any invertible matrix A, the pair $(Af,\ A^{-\top}g)$ represents exactly the same function. Consequently, the encoders of any trained model are defined only up to a symmetry of dimension d^2. This redundancy is not purely formal. It renders the population risk constant along large orbits, so the Hessian at a minimizer is highly degenerate and optimization becomes ill-posed. At the same time, it leaves the learned coordinates essentially arbitrary. Normalizations were introduced precisely to control these instabilities (Radford et al., 2021; Zhao & Li, 2026b). A normalization makes the encoders meaningful exactly to the extent that it removes the symmetry.

We now formalize this view. Each normalization is a gauge-fixing choice, a rule for selecting one representative from each equivalence class. The residual symmetry that survives a constraint is precisely what remains arbitrary under it. The interaction spectrum constitutes the invariant content of the symmetry, and the gaps in that spectrum control how much identifiability any normalization can provide.

Definition 2 (Gauge action)**.** *For $A \in \mathrm{GL}_d$, the gauge transformation acts on the encoders by*

$$\Phi_A(f, g) = \big(Af,\ A^{-\top}g\big), \tag{7}$$

where $(Af)(u) = A\,f(u)$ and $\big(A^{-\top}g\big)(v) = A^{-\top}g(v)$. The set of all such transformations is the gauge group of the head.

Gauge transformations leave the represented function, and therefore the population risk, unchanged:

$$\langle Af(u), A^{-\top}g(v)\rangle = \langle f(u), g(v)\rangle. \tag{8}$$

The second moments $\Sigma_f = \mathbb{E}_u[ff^\top]$ and $\Sigma_g = \mathbb{E}_v[gg^\top]$ transform by congruence, so the eigenvalues of the product $\Sigma_f\Sigma_g$ are gauge invariants (Lemma 2, Appendix C). Unfixed gauge also makes optimization ill-posed. Along the orbit of a minimizer the risk is flat, so the Hessian has at least $d^2 - \dim S$ zero directions. Without normalization this is generically d^2 (Proposition 4). A normalization pins exactly the subgroup whose residual freedom it eliminates, so a smaller residual group leaves fewer degenerate directions and yields a better-conditioned problem.

Normalizations as sections of the orbit. Two levels of normalization occur in practice, and only one affects identifiability. Embedding-level normalization constrains the encoders themselves and determines the residual gauge. Output-level normalization, such as the contrastive softmax in CLIP or top-k selection in retrieval, acts on the score matrix after the inner product. It does not constrain the encoders and leaves the embedding-level gauge untouched. Table 6 in Appendix C classifies the embedding-level schemes used in practice.

What each normalization identifies. The residual symmetry of a normalization indicates which transformations can act on a trained model without changing its output. The normalizations used in practice were adopted on heuristic grounds, for training stability or empirical feature decorrelation, rather than from an analysis of which symmetry they remove. The gauge view supplies the missing analysis and lets us rank them by residual symmetry. We characterize the two-sided schemes used in

practice, in order of increasing identifiability, and find that only whitening removes the symmetry completely.

Cosine normalization is the weakest two-sided scheme (Radford et al., 2021). Two-sided unit normalization $\tilde{f} = f/\|f\|$, $\tilde{g} = g/\|g\|$ with scores $\cos\angle(f(u), g(v))$, as employed by contrastive models such as CLIP (Radford et al., 2021), leaves a residual gauge exactly equal to the orthogonal group: for every rotation $R \in \mathrm{O}$, the encodings Rf and Rg give pointwise identical cosine scores, and no other transformation does (Theorem 4). Consequently, the individual coordinates of a cosine-normalized embedding carry no intrinsic meaning. Only rotation-invariant quantities (norms, pairwise inner products, and the eigenvalues of $\Sigma_f \Sigma_g$) are well-defined. This provides a theorem-level explanation of the widely reported uninterpretability of contrastive embedding dimensions, and it identifies the whitening scheme of Theorem 2 as the remedy (Corollary 2, Appendix C).

Two-sided nonnegativity, $f(u) \succeq 0$ and $g(v) \succeq 0$ (realized by a softplus or exponential output layer (Zhao & Li, 2026b)), shrinks the residual group to the monomial group $\mathrm{Perm} \ltimes \mathrm{D}_+$ of coordinate permutations composed with positive diagonal scalings. Under separability conditions (Donoho & Stodden, 2003; Arora et al., 2012), the encoders are identified up to permutation and scaling. The cost is an expressivity limitation: negatively correlated interaction modes are excluded (Theorem 5, Appendix C).

Whitening is the unique scheme among those used in practice that removes the continuous gauge symmetry entirely. Cosine and nonnegativity each leave a nontrivial residual group and were adopted largely heuristically. By contrast, we prove that whitening admits a full identifiability guarantee: under a spectral gap it pins the interaction modes up to permutation and sign, recovering both the modes and the spectrum itself.

Theorem 2 (Whitening fixes the gauge up to permutation and sign)**.** *Let the interaction singular values be distinct, $\sigma_1 > \cdots > \sigma_d > \sigma_{d+1}$, and impose the whitening constraints*

$$\Sigma_g = I_d, \qquad \Sigma_f = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \quad \lambda_1 \geq \cdots \geq \lambda_d \geq 0, \tag{9}$$

with Λ diagonal and nonincreasing. Constraints of this form are standard in self-supervised representation learning, where they are used to decorrelate features (Ermolov et al., 2021; Zbontar et al., 2021; Bardes et al., 2022). Then every global minimizer of the population risk satisfies

$$f_k = \pm\, \sigma_k\, a_k, \qquad g_k = \pm\, b_k, \qquad \lambda_k = \sigma_k^2, \qquad k = 1, \ldots, d, \tag{10}$$

with signs paired so that the product $f_k g_k = \sigma_k a_k b_k$ is uniquely determined, and with the coordinate labeling fixed by the ordering convention in (9)*. The k-th coordinate of the whitened encoders recovers the k-th interaction mode of the target, and the diagonal of Λ recovers the interaction spectrum. Equivalently, the encoders are identified up to permutation and sign.*

Proof idea. The rank-truncation bound of Corollary 1 implies that every minimizer attains the minimal risk. When $\sigma_d > \sigma_{d+1}$, this forces the represented function to equal the unique truncation F_d. The whitening constraints then require the coefficient matrices to satisfy $CC^\top = I$ and $CS^2C^\top = \Lambda$, where $S = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ has distinct eigenvalues. This forces C to be a signed permutation. The full argument is in Appendix C. □

Implementing the whitening gauge. Hard constraints (9) can be enforced by minibatch whitening. Alternatively, a quadratic penalty inherits the identifiability of the hard constraint and improves conditioning monotonically with the penalty weight (Theorem 6, Appendix C).

In practice, the second moments are estimated from a finite sample. With probability $1 - \delta$, the identified modes carry error of order r/Δ, where $r = O\big(B^2\sqrt{\log(d/\delta)/n}\big)$, B bounds the encoder outputs, and $\Delta = \min_{1 \leq k < d}(\sigma_k^2 - \sigma_{k+1}^2)$ is the spectral gap. Achieving accuracy ε therefore requires

$$n \gtrsim \frac{B^4 \log(d/\delta)}{\Delta^2 \varepsilon^2} \tag{11}$$

samples (Appendix E). The same gap thus governs optimization conditioning, identifiability, and estimation simultaneously (Corollary 3).

Section 5 measures the gauge, conditioning, and expressivity consequences of the various schemes on controlled targets. The two real systems, branch-trunk operator learning (Lu et al., 2021) and independently trained CLIP models, are treated in Appendix F (Theorem 4). Appendix C lists the measurable quantities on which the schemes are compared.

4 SAMPLE COMPLEXITY AND WHEN NOT TO USE THE HEAD

The interaction spectrum controls both estimation error and fundamental limitations. The excess risk of a head fitted from n samples decomposes into an approximation term that decreases as the rank d grows, and an estimation term that increases with d through the complexity of the encoder classes. Balancing these two terms yields a rank-selection rule. When the spectrum is flat, however, every rank-d head suffers a relative-error floor that no amount of capacity or data can overcome. This yields a practical criterion for abandoning the architecture.

Setup and notation. We observe n independent samples (u_i, v_i, y_i) with $y_i = F(u_i, v_i) + \varepsilon_i$ and fit the head by empirical risk minimization over the class

$$\mathcal{H}_d = \big\{\, (u,v) \mapsto \langle f(u), g(v)\rangle : \; f \in \mathcal{F},\; g \in \mathcal{G} \,\big\} \tag{12}$$

of rank-d heads. The encoders are bounded pointwise, $\|f(u)\| \le B_f$ and $\|g(v)\| \le B_g$, so the scores are bounded by $B = B_f B_g$. These pointwise radii are distinct from the L^2 mode norms M_f, M_g of Theorem 1. Let $\mathfrak{R}_n$ denote the Rademacher complexity, extended to vector-valued classes coordinatewise. The class complexity collapses to the sum of the marginal complexities:

$$\mathfrak{R}_n(\mathcal{H}_d) \le 2d\big(B_g\,\mathfrak{R}_n(\mathcal{F}) + B_f\,\mathfrak{R}_n(\mathcal{G})\big), \tag{13}$$

for shared per-coordinate classes (Lemma 3, Appendix D).

The excess risk of the fitted head therefore decomposes into the approximation terms of Theorem 1 plus an estimation term controlled by this complexity:

$$\|\hat{F} - F\|^2_{L^2(\mu_U \otimes \mu_V)} \;\lesssim\; \sum_{k>d} \sigma_k^2 + \mathcal{E}_{\mathrm{real}} + B\,d\big(\mathfrak{R}_n(\mathcal{F}) + \mathfrak{R}_n(\mathcal{G})\big) + B^2\sqrt{\frac{\log(1/\delta)}{n}}, \tag{14}$$

where $\mathcal{E}_{\mathrm{real}} = C\sum_{k\le d}(\sigma_k^2\eta_{g,k}^2 + M_g^2\eta_{f,k}^2)$ and $\lesssim$ hides absolute constants, with probability $1-\delta$ (Corollary 4, Appendix D).

Linear encoders anchor the rate; rank selection. With linear encoders $f(u) = U^\top u$, $g(v) = V^\top v$ on inputs $u \in \mathbb{R}^p$, $v \in \mathbb{R}^q$, the head computes $u^\top \Theta v$ with $\Theta = UV^\top$ of rank at most d. Estimation reduces to low-rank matrix sensing of the ground-truth matrix $\Theta^\star$, whose minimax rate, with σ_ε^2 the noise variance of ε_i, is known exactly:

$$\mathbb{E}\|\hat{\Theta} - \Theta^\star\|_F^2 \asymp \sigma_\varepsilon^2\, d(p+q)/n. \tag{15}$$

This rate provides the baseline for the general bound (Theorem 7, Appendix D).

The rank d is the primary design choice for the head. It faces two opposing forces: the truncation term $\sum_{k>d}\sigma_k^2$ decreases with d, while the estimation term grows linearly with d. The optimal rank balances these two contributions. When the spectrum decays exponentially, $\sigma_k^2 = O(e^{-2ck})$, balancing gives

$$d^\star \asymp \frac{1}{2c}\log\frac{n}{p+q}, \tag{16}$$

and yields a near-parametric rate

$$\|\hat{F} - F\|^2_{L^2} \lesssim \frac{(p+q)\log(n/(p+q))}{n}. \tag{17}$$

Fast spectral decay thus eliminates the curse of dimensionality (Corollary 5, Appendix D). The three terms in (14) form the complete error ledger: fast decay keeps $d^\star$ small, whereas slow decay makes the head expensive in every respect. The rest of this section analyzes the regime in which even the optimal $d^\star$ is inadequate.

Identifiability has a sample cost. Identifiability, as characterized in Section 3, is a population-level statement. Realizing it from finite samples incurs an additional cost, detailed in Appendix E (Corollary 6). The total mode error of a fitted and subsequently whitened head (Proposition 6) decomposes into the excess risk of Corollary 4 plus the empirical whitening error. Both terms are divided by the spectral gap Δ and decay as $n^{-1/2}$.

Late versus early interaction. We now compare the multiplicative head, which we refer to as a late-interaction model, with early-interaction models in which the two inputs are combined during encoding. By Corollary 1, every rank-d head pays at least the spectral tail, so the spectrum determines the outcome.

Define the *effective interaction rank*

$$d_\varepsilon(F) = \min\Big\{ d : \frac{\sum_{k>d} \sigma_k^2}{\sum_{k\geq 1} \sigma_k^2} \leq \varepsilon \Big\} \tag{18}$$

as the smallest embedding dimension for which a rank-d head achieves relative error ε. The spectrum is called *flat* at scale N when $d_\varepsilon(F) = \Omega((1-\varepsilon)N)$.

The equality function realizes this worst case exactly. On $\mathcal{U} = \mathcal{V} = \{\pm 1\}^m$ with $N = 2^m$, the interaction operator satisfies $T_{F_{\mathrm{eq}}} = \frac{1}{N}\mathrm{Id}$. The singular values are all equal to $1/\sqrt{N}$, the interaction rank is N, and every rank-d head achieves relative error exactly $1 - d/N$. Thus, attaining relative error ε requires $d \geq (1-\varepsilon)2^m$ embedding dimensions, exponential in the input size (Theorem 8, Appendix D).

By contrast, an early-interaction predictor with $O(m)$ parameters represents F_{eq} exactly. The identity $\langle u, v\rangle = m - 2\,\mathrm{Ham}(u, v)$ allows a simple threshold on the inner product to recover the equality indicator, whether implemented by a cross-attention layer with identity projections or by a joint MLP with $O(m)$ ReLUs. Combined with Theorem 8, this yields an exponential separation: $O(m)$ parameters suffice under early interaction, while any late-interaction head requires $\Omega(2^m)$ dimensions to reach constant relative error (Theorem 9, Appendix D).

These observations lead to a concrete usability criterion (Theorem 10, Appendix D). When the effective interaction rank $d_\varepsilon(F)$ is small, a late-interaction head is appropriate: its approximation and estimation errors are controllable, and its computational and statistical advantages are preserved. When the spectrum is flat, the truncation floor forces every rank-d head to relative error at least ε, while early-interaction models can still reach ε with polynomial resources when the target has low-dimensional joint structure. The deciding quantity, the spectral decay rate, can be estimated from data by a weighted singular value decomposition of the empirical kernel matrix, making the criterion actionable before either model is trained.

The criterion is a boundary, not a verdict. The early-interaction upper bounds assume low-dimensional joint structure. A target that is both high-rank and unstructured is difficult for both architectures. Moreover, the parameter counts in Theorem 9 are statements about representational existence, not about learnability. Section 5 instantiates both sides of the dichotomy: on gapped spectra the head identifies and estimates well, while flat and narrow-band targets drive it to the exponential-rank floor.

5 Experiments

5.1 Overview

Our experiments are designed for mechanism validation rather than benchmark chasing: each study isolates a quantity that the theory predicts exactly. The controlled synthetic setting (Section 5.2) supplies exact ground truth and is reported here in full, every conclusion accompanied by the table or figure that supports it. Two real architectures instantiate the multiplicative head directly, DeepONet and CLIP; page limits confine their tables and figures to Appendix F, and we state here only the conclusions they establish. On DeepONet, post-hoc whitening lifts the learned trunk basis toward the analytic operator eigenbasis, from alignment 0.5 to 0.6 up to 0.78 to 0.94, on a model trained end to end. On two pretrained CLIP pairs, the gauge-invariant spectral summary is shared across models while per-dimension concept probes are not, and fitting the single residual rotation reconciles them.

5.2 Controlled synthetic study

Setup. This study validates the identifiability and usability claims of Sections 3 and 4 in a setting with exact ground truth. We fix a target $F(u, v) = \sum_k \sigma_k a_k(u) b_k(v)$ with shifted Legendre modes and a prescribed spectrum, fit linear dual encoders over a fixed Legendre feature map, and record four quantities per run: the fit error, the alignment of the recovered modes with the true a_k, b_k, the

Table 1: The seven normalization schemes, four seeds each. Residual gauge: the subgroup surviving the constraint. Align.: mean mode alignment of the two encoders after optimal permutation and sign matching. Cond: the gauge-dependent $\max(\operatorname{cond}\Sigma_f, \operatorname{cond}\Sigma_g)$. Cross-seed: mean pairwise gauge distance between seeds.

Normalization	Residual gauge	Fit MSE	Align. (f,g)	Cond	Cross-seed
none	GL_d	0.0004	0.605, 0.619	13.8	0.326
l2_single	$\approx \mathrm{GL}_d$	0.0092	0.561, 0.420	16.8	0.284
per_coord	O+shear	0.0004	0.607, 0.593	18.7	0.330
clip_cosine	O	0.0568	0.529, 0.550	305.9	0.212
nonneg	$\mathrm{Perm} \ltimes \mathrm{D}_+$	4.251	0.576, 0.446	121.0	0.280
whiten_soft	$\to \mathrm{Perm} \ltimes \mathrm{D}_\pm$	0.0169	0.651, 0.549	105.5	0.348
whiten_hard	$\mathrm{Perm} \ltimes \mathrm{D}_\pm$	0.0004	1.000, 1.000	9.5	0.000

Table 2: Gap and sample-size sweeps of the finite-sample whitening estimator (Appendix E), reporting the mean mode-identification error. Left: the spectral gap $\Delta = \min_{1\le k<d}(\sigma_k^2 - \sigma_{k+1}^2)$ swept by scaling a base spectrum at fixed $n = 3200$. Right: sample size swept at fixed $\Delta = 0.900$.

Fixed $n = 3200$			Fixed $\Delta = 0.900$	
Scale	Δ	Mode err	n	Mode err
1.00	0.900	0.0431	200	0.1979
0.60	0.648	0.0782	800	0.0956
0.35	0.417	0.1258	3200	0.0470
0.20	0.230	0.2458	12800	0.0236

encoder-covariance conditioning, and the cross-seed gauge distance. The encoder-realization error of Assumption 1 is zero, so normalization is the only variable across runs. The four paragraphs below test, in turn, Theorem 2, Corollaries 3 and 6, Theorem 7, and Theorem 8.

Normalization as gauge-fixing. The seven schemes of Table 1 separate cleanly by residual group. Only whitening attains perfect mode alignment (1.000) with zero cross-seed gauge distance and drives the conditioning diagnostic $\max(\operatorname{cond}\Sigma_f, \operatorname{cond}\Sigma_g)$ to its theoretical floor $\sigma_1^2/\sigma_d^2 \approx 9.5$ (Corollary 3). Every looser gauge remains stuck at alignment 0.53 to 0.65, its residual group leaving the modes rotationally entangled (Theorem 4). The gauge-invariant diagnostic $\operatorname{cond}(\Sigma_f\Sigma_g)$ cannot separate the schemes (Lemma 2). The two-sided nonnegative gauge incurs error 4.25, the predicted out-of-class cost of excluding negatively correlated modes (Theorem 5). Normalization is therefore gauge fixing, and whitening is the only scheme that removes the symmetry completely, exactly as Theorem 2 predicts.

The gap as the single unifying quantity. The spectral gap is the single quantity governing identification. With the finite-sample whitening estimator of Appendix E, we sweep the gap at fixed sample size and the sample size at fixed gap, eight configurations in all (Table 2; Figure 3 in Appendix F replots the same eight points). The mode-identification error scales as $1/\Delta$ at fixed sample size and as $n^{-1/2}$ at fixed gap, where $\Delta = \min_{1\le k<d}(\sigma_k^2 - \sigma_{k+1}^2)$, and the two sweeps collapse onto the single relation

$$\mathrm{err} \cdot \Delta \cdot \sqrt{n} \approx 2.6 \tag{19}$$

with coefficient of variation 0.12 about the mean 2.62 (median 2.48, range 2.19 to 3.20): gap and sample size enter the identification error only through their product, the concrete face of Corollaries 3 and 6. The residual spread is not noise but a mild gap dependence: the sample-size sweep is flat, 2.39 to 2.52, while the gap sweep drifts from 2.19 at the widest gap to 3.20 at the narrowest, the direction the first-order perturbation argument predicts.

The sample-complexity rate law. In the linear regime the head is exactly a rank-d matrix sensing problem, and Theorem 7 predicts $\mathbb{E}\|\hat{\Theta} - \Theta^\star\|_F^2 \asymp \sigma_\varepsilon^2 d(p+q)/n$. We sweep each of the three variables with the others held fixed and fit log-log slopes, the three sweeps shown in Figure 1. The measured slopes are -1.03 in n, $+0.89$ in d, and $+1.24$ in $p+q$, against the theoretical $-1, +1, +1$. The excess in the last slope is the expected Marchenko-Pastur finite-sample correction, and gradient

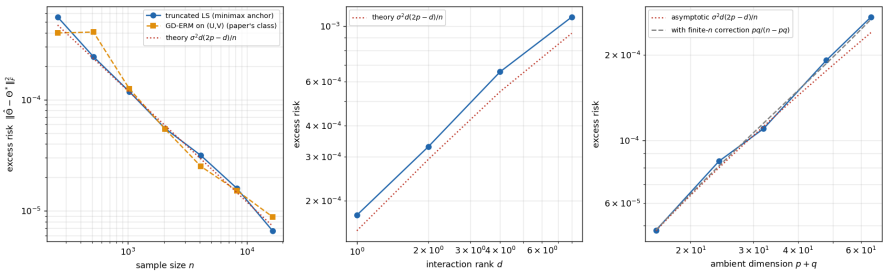


Figure 1: Log-log sweeps of the Frobenius excess risk against sample size n (left), interaction rank d (center), and ambient dimension $p+q$ (right). Points are measured risk; the dotted line is the theoretical rate $\sigma^2 d(2p-d)/n$, with the finite-n corrected form (dashed) at right.

Table 3: The boolean equality kernel $F = \mathbf{1}[u = v]$ on $\{\pm 1\}^m$, $N = 2^m$. Rank d is the embedding dimension at which the head first reaches 10% relative error; the early-interaction comparator is exact with $2m + 1$ parameters.

m	3	4	5	6	7	8
N	8	16	32	64	128	256
Rank d	8	15	29	58	116	231
d/N	1.00	0.94	0.91	0.91	0.91	0.90
Early params	7	9	11	13	15	17

Table 4: Narrowband periodic-Gaussian kernel of bandwidth h on the circle: interaction rank d_ε for $\varepsilon = 10\%$ error.

$1/h$	5	10	20	40	80
d_ε	3	4	8	15	30

descent on the factored pair matches the least-squares slope. These sweeps reject the loose general bound of Remark 6, which would predict slope $-1/2$: estimation is governed by the sum of the marginal encoder complexities, and the $1/n$ dependence is genuine.

The usability criterion. A flat spectrum floors the head regardless of capacity. We test two flat and near-flat regimes, Tables 3 and 4. On the Boolean equality kernel $F = \mathbf{1}[u = v]$ on $\{\pm 1\}^m$, which has a completely flat spectrum, a rank-d head incurs relative error exactly $1 - d/N$ and requires $d \approx 0.9 \cdot 2^m$ dimensions to reach 10% error, while an early-interaction threshold on $\langle u, v \rangle$ is exact with $2m + 1$ parameters (Theorems 8 and 9). Smooth narrow-band kernels give the polynomial counterpart $d_\varepsilon = \Theta(1/h)$ (Remark 7). A trained-model anchor on a block-boxcar kernel realizes the same floor (Table 10 and Figure 5, Appendix F).

6 Conclusion

This paper has examined a single architectural pattern, the multiplicative dual-encoder head, through a unifying lens: the interaction spectrum of the target. Through theoretical analysis and numerical experiments, we answer four central questions. (i) Approximation error decomposes into a truncation term intrinsic to the target and a realization term intrinsic to the encoders, with target smoothness controlling the decay rate. (ii) Sample complexity is governed by the sum of the two marginal encoder complexities, with the linear regime anchoring the minimax rate. (iii) Usability is determined by the decay rate itself: a flat spectrum imposes an error floor that no rank-d head can escape, while early-interaction models can, and the deciding quantity is measurable from data before either model is trained. (iv) Identifiability, the problem that the framework brings into focus, is gauge fixing: every normalization used in practice is a section of the gauge orbit; whitening pins the interaction modes up to permutation and sign; and the spectral gap simultaneously controls optimization conditioning, identifiability, and estimation.

AI USE STATEMENT

In this work, we used generative AI tools for paper polishing, including grammar and wording checks, verification and support of mathematical derivations, and coding. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the principles outlined in the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The code for this paper is available at `https://anonymous.4open.science/r/Mul-Net-FDF0`.

REFERENCES

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 145–162, 2012.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado Van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57 (4):2342–2359, 2011.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in neural information processing systems*, 16, 2003.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International conference on machine learning*, pp. 3015–3024. PMLR, 2021.

Zhang-Wei Hong, Ge Yang, and Pulkit Agrawal. Bi-linear value networks for multi-goal reinforcement learning. In *International Conference on Learning Representations*, 2021.

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pp. 3163–3172. PMLR, 2019.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Graham Little and John B Reade. Eigenvalues of analytic kernels. *SIAM journal on mathematical analysis*, 15(1):133–136, 1984.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

Sahand N. Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. In *International Conference on Machine Learning*, 2009. URL https://api.semanticscholar.org/CorpusID:1004801.

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.

John B Reade. Eigenvalues of positive definite kernels. *SIAM Journal on Mathematical Analysis*, 14 (1):152–157, 1983.

Steffen Rendle. Factorization machines. In *2010 IEEE International conference on data mining*, pp. 995–1000. IEEE, 2010.

Angelika Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39:887–930, 2009. URL https://api.semanticscholar.org/CorpusID:88512332.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320. PMLR, 2015.

Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen: I. teil: Entwicklung willkürlicher funktionen nach systemen vorgeschriebener. *Mathematische Annalen*, 63(4):433–476, 1907.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pp. 2071–2080. PMLR, 2016.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks*, 94: 103–114, 2017.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Zijian Zhao and Sen Li. The impacts of data privacy regulations on food-delivery platforms. *Transportation Research Part C: Emerging Technologies*, 181:105364, 2025.

Zijian Zhao and Sen Li. Discriminatory order assignment and payment-setting of on-demand food-delivery platforms: A multi-action and multi-agent reinforcement learning framework. *Transportation Research Part E: Logistics and Transportation Review*, 208:104653, 2026a.

Zijian Zhao and Sen Li. Triple-bert: Do we really need marl for order dispatch on ride-sharing platforms? In *International Conference on Learning Representations*, volume 2026, pp. 153780–153808, 2026b.

Zijian Zhao, Tingwei Chen, Zhijie Cai, Xiaoyang Li, Hang Li, Qimei Chen, and Guangxu Zhu. Crossfi: A cross domain wi-fi sensing framework based on siamese network. *IEEE Internet of Things Journal*, 12(12):20138–20155, 2025.

APPENDIX CONTENTS

A Derivations for the Unification Table **13**

B Proofs for Section 2 **15**

C Proofs for Section 3 **18**

D Proofs for Section 4 **22**

E Finite-Sample Technical Lemmas **26**

E.1 Empirical whitening and mode identification 26

E.2 Approximate separability in the NMF route 28

E.3 Coupling identification with estimation 28

F Experimental Details and Additional Results **29**

F.1 Controlled synthetic study 29

F.2 DeepONet 31

F.3 CLIP 33

G Index of results **37**

A Derivations for the Unification Table

This appendix spells out, for each of the ten families of Table 5, the identification that makes it an instance of the multiplicative dual-encoder head (1): the two input spaces, the encoder maps, the score, the reference measures that define the interaction spectrum, and the boundary cases where an additional output-level normalization acts after the inner product and therefore leaves the gauge untouched. Table 5 collects the ten families, each computing a real-valued output for a pair of inputs as the inner product of two separately learned encodings; reading across the rows makes visible what is shared, and the paragraphs below identify the interaction operator and its spectrum for each family, including the boundary cases.

CLIP. Inputs are an image I and a text T; the encoders are the image tower f and the text tower g, both ℓ_2-normalized; the score is $\mathrm{sim}(I, T) = \langle f(I), g(T)\rangle$, rescaled by a temperature and fed into a softmax over the batch. The reference measures are the empirical distributions over images and over texts in the training batch. The temperature and the softmax act on the similarity matrix after the inner product, so they are output-level and do not affect the embedding-level gauge. The unit-norm constraints are embedding-level and leave exactly the residual rotation O of Theorem 4.

DeepONet. Inputs are an input function a (on a sensor grid) and a query point y; the encoders are the branch net b and the trunk net t; the score is $\mathcal{G}(a)(y) \approx \langle b(a), t(y)\rangle$ (Lu et al., 2021). The reference measures are the distribution over input functions and the measure over query points. For a linear operator with kernel $\kappa = \sum_k \sigma_k \beta_k(y)\gamma_k(x)$ and a white input measure, the interaction spectrum coincides with the operator singular values and the trunk's ideal basis with the left singular functions, which is the ground truth used in Appendix F.2. The boundary case is the scale imbalance between branch and trunk outputs, familiar in practice. It is the scale redundancy of the gauge group, diagnosed in Section 3. POD-DeepONet substitutes a precomputed orthogonal POD basis for the learned trunk (Lu et al., 2021), motivated by the same nonuniqueness. Our treatment is complementary, characterizing when the learned basis becomes identifiable instead of replacing it (Appendix F.2).

Two-tower retrieval. Inputs are a query q and a document d; the encoders are two towers φ, ψ; the score is a relevance measure $\mathrm{rel}(q, d) = \langle \varphi(q), \psi(d)\rangle$ (Karpukhin et al., 2020). The reference measures are the query distribution and the document distribution, and the interaction spectrum is the relevance spectrum of the underlying ranker. The top-k nearest-neighbor search is output-level.

UVFA (universal value function approximators). Inputs are a state s and a goal g; the encoders are the state encoder $\phi(s)$, a learned map from states to features, and the goal encoder $\psi(g)$, a learned map from goals to features; the score is the goal-conditioned value $V(s, g) = \langle \phi(s), \psi(g)\rangle$ (Schaul et al., 2015). The reference measures are the state-visitation distribution and the goal distribution. Goal-conditioned value functions are naturally low-rank when goals and states share a low-dimensional feature structure, which is exactly the case in which the representation (1) is sample-efficient (Section 4). The interaction spectrum also gives these methods a rank-selection rule they currently lack: the number of features in a successor-feature representation or the width of a value-network embedding is typically a tuned hyperparameter, whereas Corollary 5 ties it to the decay of the spectrum.

BVN (bilinear value networks). Inputs are a state-action pair (s, a) and a state-goal pair (s, g); the encoders are the MLP $f(s, a)$ encoding the state-action pair and the MLP $\varphi(s, g)$ encoding the state-goal pair; the score is $Q(s, a, g) = \langle f(s, a), \varphi(s, g)\rangle$ (Hong et al., 2021). The reference measures are the joint distributions over (s, a) and (s, g). Note that the two arguments share the state s: the head's theory applies to the function Q of the pair of arguments, and the shared state is exactly the coupling that makes Q non-separable in (s, a, g) directly, which is why the two-input factorization matters.

Successor features. Inputs are a state s and a reward weight vector w; the encoder on the state side is the successor-feature map $\phi(s)$ and the encoder on the weight side is the identity w (a singleton encoder class, not learned); the score is $V_w(s) = \langle \phi(s), w\rangle$ (Barreto et al., 2017). This is the head

with $\mathcal{G} = \{\mathrm{id}\}$, whose Rademacher complexity is zero, so the sample-complexity bounds of Section 4 specialize to the standard guarantee for successor features: only the state-side class matters.

DistMult and ComplEx. Inputs are a head entity h and a tail entity t; the encoders are entity embeddings $e_h, e_t \in \mathbb{R}^d$; the score is the bilinear form $\langle e_h, r \circ e_t \rangle$ of DistMult or the real part of the complex product $e_h^\top \operatorname{diag}(r)\, \overline{e_t}$ of ComplEx, where r is the relation embedding (Yang et al., 2014; Trouillon et al., 2016). Both are rank-d bilinear forms on the entity embeddings, i.e., heads with linear encoders and a fixed middle matrix. The interaction spectrum is governed by the Gram structure of the entity embeddings under the relation. The reference measure is the empirical distribution over observed triples.

Linear attention. Inputs are a query q and a key k in a common space; the encoders are the feature map ϕ applied to each; the attention weight is $\langle \phi(q), \phi(k) \rangle$ (Katharopoulos et al., 2020). The aggregation step $\sum_j \phi(k_j) v_j^\top$ makes the compute separation of Proposition 1 exact: the pairwise inner products are never formed, giving the $O(n^2) \to O(n)$ complexity. The reference measure is the distribution over keys (or key-value pairs). Finite-dimensional feature maps realize the head exactly. Softmax attention with random-feature maps is the limiting case of an infinite-dimensional ϕ.

Factorization machines. Inputs are a feature index i and a feature index j; the encoders are the embedding maps v_i, v_j; the pairwise term is $x_i x_j \langle v_i, v_j \rangle$ (Rendle, 2010). The reference measure is the empirical distribution over feature pairs weighted by the data. The interaction spectrum of the pairwise term is the spectrum of the Gram matrix of the embeddings, and the rank selection of Section 4 is the standard FM capacity control viewed through the interaction spectrum.

QK-attention matching. Inputs are a worker w and an order o; the encoders are a BERT-style encoder f and an MLP g; the score is the match value $\langle f(w), g(o) \rangle$ forming the Q-matrix of a bipartite assignment problem, with a one-sided softplus and unit-norm constraint on the order encoder (Zhao & Li, 2026b;a; 2025). The reference measure is the empirical distribution over worker-order pairs. The matching layer that consumes the score matrix is output-level. The one-sided constraint is embedding-level but, as Remark 4 explains, it is a stability fix rather than an identifiability scheme. The two papers differ only in the downstream use of the same head, which is why they share the identifiability properties derived in this paper.

Positioning against prior work. The paragraphs above identify each family and its interaction operator. We now position the paper's claims against the closest prior work, none of which treats the families as instances of a shared architecture or formulates the identifiability question. On the estimation side, the linear case of the head is rank-d matrix sensing, and the paper relies on the classical minimax theory, upper bounds via restricted-isometry-type analyses (Candes & Plan, 2011; Negahban & Wainwright, 2009) and matching lower bounds (Rohde & Tsybakov, 2009). Bilinear bandits (Jun et al., 2019) study the online counterpart. The contribution relative to this line is a translation rather than a new bound: the estimation problem of a trained head is identified with matrix sensing, and the minimax rate is connected to the rank-selection rule of Section 4 through the decay of the interaction spectrum, which the matrix-sensing literature does not formulate because it treats the rank as given. On the identifiability side, the direct precursor of our remedy is self-supervised whitening: W-MSE applies hard whitening to the embeddings (Ermolov et al., 2021), and Barlow Twins and VICReg implement its soft relaxations, pushing the cross-correlation matrix toward the identity and regularizing variance and covariance (Zbontar et al., 2021; Bardes et al., 2022). It is also known that whitening reduces a general linear gauge to an orthogonal ambiguity, a polar-decomposition observation. Against this background, our contribution is not the whitening operation itself, which we do not claim, but three things: the interpretation of normalization as a section of the gauge orbit and the classification of the residual subgroups across the normalizations used in practice (Table 6, Appendix C); the theorem that whitening together with a spectral gap pins the modes down to permutation and sign, strictly beyond the orthogonal ambiguity that decorrelation alone leaves (Theorem 2); and the statement that the interaction spectrum is the unique gauge-invariant content of a trained model, an interpretability claim with a theorem behind it rather than an empirical observation. The nonnegative variant of the bilinear form has its own classical identifiability theory, separability-type conditions for uniqueness up to permutation and scale (Donoho & Stodden, 2003; Arora et al., 2012), which the gauge analysis of Section 3 subsumes as a special case (Theorem 5).

Table 5: Ten families of methods as instances of the multiplicative dual-encoder head (1): for each family, the two inputs, the encoder maps, the output, and the downstream use.

Instance	Inputs	Encoders	Output	Downstream use
CLIP (Radford et al., 2021)	image, text	ViT, text transformer	normalized similarity	contrastive softmax over a batch
DeepONet (Lu et al., 2021)	function a, point y	branch net, trunk net	$\langle \mathrm{branch}(a), \mathrm{trunk}(y) \rangle$	operator output $\mathcal{G}(a)(y)$
Two-tower retrieval (Karpukhin et al., 2020)	query, document	two towers	relevance	top-k ANN search
UVFA (Schaul et al., 2015)	state s, goal g	$\phi(s)$, $\psi(g)$	$V(s, g)$	value function
BVN (Hong et al., 2021)	(s, a), (s, g)	two MLPs	$Q(s, a, g)$	multi-goal RL
Successor features (Barreto et al., 2017)	state s, weights w	$\phi(s)$, w	$V_w(s)$	transfer
DistMult / ComplEx (Yang et al., 2014; Trouillon et al., 2016)	head, tail entity	entity embeddings	bilinear / complex product	link prediction
Linear attention (Katharopoulos et al., 2020)	query q, key k	feature map ϕ	$\langle \phi(q), \phi(k) \rangle$	$O(n^2) \to O(n)$ aggregation
Factorization machines (Rendle, 2010)	feature i, feature j	embedding v_i	$\langle v_i, v_j \rangle x_i x_j$	pairwise interaction
QK-attention matching (Zhao & Li, 2026b;a)	worker w, order o	BERT, MLP	$\langle f(w), g(o) \rangle$	Q-matrix for ILP matching

B Proofs for Section 2

We prove the results of Section 2 in order. Throughout, $\|\cdot\|$ denotes the $L^2(\mu_U \otimes \mu_V)$ norm, and $\|\cdot\|_{L^2(\mu_U)}$, $\|\cdot\|_{L^2(\mu_V)}$ the marginal norms.

Assumption 1 (Encoder realizability)**.** *The encoder classes $\mathcal{F}$ and $\mathcal{G}$ contain functions $\hat{f}, \hat{g}$ such that, for some nonnegative errors $\eta_{f,k}, \eta_{g,k}$,*

$$\|\hat{f}_k - \sigma_k a_k\|_{L^2(\mu_U)} \le \eta_{f,k}, \qquad \|\hat{g}_k - b_k\|_{L^2(\mu_V)} \le \eta_{g,k}, \qquad k = 1, \dots, d. \tag{20}$$

Remark 1 (What the decomposition buys)**.** *The truncation term is intrinsic to the problem, the realization term to the model. When the encoder classes are dense in L^2 the bound approaches the lower bound of Corollary 1, so the decomposition is tight up to the constant. Whether training finds the coordinate alignment of* (20) *is the question of identifiability addressed in Section 3.*

Proposition 1 (Compute separation)**.** *Evaluating a rank-d head on all $n \times m$ input pairs costs*

$$O\big((n+m)\, C_{\mathrm{net}} + nm\, d\big), \tag{21}$$

where C_{net} is the cost of one encoder forward pass, whereas a model that processes each pair jointly costs $O(nm\, C_{\mathrm{net}})$. When the encoder forward pass dominates the inner product, the head is cheaper by a factor C_{net}/d.

Remark 2 (Known complexity results as instances)**.** *Proposition 1 recovers, as special cases, the multiplicative-to-additive complexity argument of the QK-attention matching networks (Zhao & Li,*

2026b), the $O(n^2) \to O(n)$ aggregation of linear attention (Katharopoulos et al., 2020), where the attention weight is written as the inner product of feature maps, and the efficient branch-trunk evaluation of DeepONet (Lu et al., 2021).

Remark 3 (Where low interaction rank comes from)**.** *Three structural sources make the truncation term small: exact low rank ($F = \sum_{j\le r} \psi_j(u)\chi_j(v)$ implies $\text{i-rank}(F) \le r$), tensor-product structure (for F in a tensor-product RKHS the spectrum is governed by the marginal kernels), and smoothness (Theorem 3 gives explicit rates). Conversely, a flat spectrum, developed in Section 4, means the truncation term does not decay and no choice of d helps.*

Lemma 1 (Rank equivalence)**.** *A function $S \in L^2(\mu_U \otimes \mu_V)$ can be written as $S(u,v) = \langle f(u), g(v)\rangle$ for some encoders $f : \mathcal{U} \to \mathbb{R}^d$ and $g : \mathcal{V} \to \mathbb{R}^d$ if and only if $\operatorname{rank}(T_S) \le d$. In particular, the parametric class of eq.* (2) *coincides with $\mathcal{M}_d = \{F : \text{i-rank}(F) \le d\}$, and we use the two descriptions interchangeably.*

Proof of Lemma 1 (Rank equivalence). The forward direction is immediate: if $S = \sum_{k\le d} f_k \otimes g_k$, then the interaction operator is a sum of at most d rank-one operators,

$$T_S = \sum_{k\le d} \langle \cdot, g_k\rangle_{L^2(\mu_V)}\, f_k, \tag{22}$$

so $\operatorname{rank}(T_S) \le d$. Conversely, if $\operatorname{rank}(T_S) = r \le d$, the Schmidt decomposition of S reads $S = \sum_{k\le r} \sigma_k\, a_k \otimes b_k$ with $\sigma_k > 0$, and the encoders $f = (f_1, \dots, f_d)$, $g = (g_1, \dots, g_d)$ defined by $f_k = \sigma_k a_k$ and $g_k = b_k$ for $k \le r$, and $f_k = g_k = 0$ for $k > r$, satisfy $S = \langle f, g\rangle$.

Corollary 1 (Spectral truncation lower bound)**.** *For every rank-d head (f, g),*

$$\|F - \langle f, g\rangle\|^2_{L^2(\mu_U\otimes\mu_V)} \;\ge\; \sum_{k>d} \sigma_k^2, \tag{23}$$

and the bound is attained in L^2 by the truncated Schmidt series $F_d = \sum_{k\le d} \sigma_k\, a_k \otimes b_k$, which is the unique best rank-d approximation whenever $\sigma_d > \sigma_{d+1}$.

Proof of Corollary 1 (Spectral truncation lower bound). The map $S \mapsto T_S$ is an isometric isomorphism between $L^2(\mu_U \otimes \mu_V)$ and the Hilbert-Schmidt class $\mathrm{HS}(L^2(\mu_V), L^2(\mu_U))$ equipped with the inner product $\langle T_S, T_{S'}\rangle = \operatorname{tr}(T_S^* T_{S'})$. Indeed $\|T_S\|^2_{\mathrm{HS}} = \sum_{k\ge1} \sigma_k^2 = \|S\|^2_{L^2}$. Under this isometry, the rank constraint of Lemma 1 becomes the operator rank constraint, and the Eckart-Young-Mirsky theorem for compact operators (see e.g. Gohberg and Krein) gives that the best rank-d approximation of T_F is the truncation to its leading d singular triples, with squared error $\sum_{k>d} \sigma_k^2$. Uniqueness: when $\sigma_d > \sigma_{d+1}$, the leading d-dimensional singular subspace of T_F is unique, hence $F_d = \sum_{k\le d} \sigma_k a_k \otimes b_k$ is the unique minimizer.

Proof of Theorem 1 (Approximation error decomposition). Fix realizations $\hat f \in \mathcal{F}$, $\hat g \in \mathcal{G}$ satisfying Assumption 1, and write $f_k^\star = \sigma_k a_k$, $g_k^\star = b_k$ for the ideal modes, so that $F_d = \sum_{k\le d} f_k^\star \otimes g_k^\star = \langle f^\star, g^\star\rangle$. Set $\varepsilon_k = f_k^\star - \hat f_k$ and $\delta_k = g_k^\star - \hat g_k$, with $\|\varepsilon_k\|_{L^2(\mu_U)} \le \eta_{f,k}$ and $\|\delta_k\|_{L^2(\mu_V)} \le \eta_{g,k}$. Split the error around F_d:

$$\|F - \langle \hat f, \hat g\rangle\| \le \underbrace{\|F - F_d\|}_{=\sqrt{\mathcal{E}_{\mathrm{trunc}}}} + R, \qquad R := \|F_d - \langle \hat f, \hat g\rangle\|, \tag{24}$$

where $\mathcal{E}_{\mathrm{trunc}} := \sum_{k>d} \sigma_k^2$ by Corollary 1. For each coordinate, the product rule for tensor products gives

$$f_k^\star \otimes g_k^\star - \hat f_k \otimes \hat g_k = \varepsilon_k \otimes b_k + \sigma_k a_k \otimes \delta_k - \varepsilon_k \otimes \delta_k. \tag{25}$$

Sum over $k \le d$ and take norms. The first sum is diagonalized by the orthonormality of $\{b_k\}$:

$$\Big\|\sum_{k\le d} \varepsilon_k \otimes b_k\Big\|^2 = \sum_{k\le d} \|\varepsilon_k\|^2_{L^2(\mu_U)} = \sum_{k\le d} \eta_{f,k}^2, \tag{26}$$

and likewise the orthonormality of $\{a_k\}$ diagonalizes the second sum:

$$\Big\|\sum_{k\le d} \sigma_k a_k \otimes \delta_k\Big\|^2 = \sum_{k\le d} \sigma_k^2 \|\delta_k\|^2_{L^2(\mu_V)} \le \sum_{k\le d} \sigma_k^2 \eta_{g,k}^2. \tag{27}$$

This is the step that avoids a factor of d: bounding the sums coordinatewise and applying Cauchy-Schwarz over the d coordinates would pay a factor d, whereas the orthonormality of the ideal modes makes the two sums orthogonal. The third sum is bounded by Cauchy-Schwarz:

$$\Big\|\sum_{k\le d}\varepsilon_k\otimes\delta_k\Big\| \le \sum_{k\le d}\|\varepsilon_k\|\,\|\delta_k\| \le \Big(\sum_{k\le d}\eta_{f,k}^2\Big)^{1/2}\Big(\sum_{k\le d}\eta_{g,k}^2\Big)^{1/2}. \tag{28}$$

With $X=(\sum_k \eta_{f,k}^2)^{1/2}$, $Y=(\sum_k \sigma_k^2\eta_{g,k}^2)^{1/2}$, $Z=(\sum_k \eta_{g,k}^2)^{1/2}$, the three bounds give $R\le X+Y+XZ$, hence by $(x+y+z)^2\le 3(x^2+y^2+z^2)$,

$$R^2 \le 3\sum_{k\le d}\eta_{f,k}^2 + 3\sum_{k\le d}\sigma_k^2\eta_{g,k}^2 + 3\Big(\sum_{k\le d}\eta_{f,k}^2\Big)\Big(\sum_{k\le d}\eta_{g,k}^2\Big). \tag{29}$$

In the informative regime the mode errors are small: assuming $\sum_{k\le d}\eta_{g,k}^2 \le 1$ (the modes are realized with bounded total error, the only regime in which the bound is informative) and $\eta_{g,k}\le 1/2$ for all k (so that $M_g=\max_k\|\hat g_k\|_{L^2(\mu_V)}\ge \min_k(\|b_k\|-\|\delta_k\|)\ge 1/2$), the cross term absorbs into the first term and $\sum_k \eta_{f,k}^2 \le 4M_g^2\sum_k \eta_{f,k}^2$, giving

$$R^2 \le 3\sum_{k\le d}\sigma_k^2\eta_{g,k}^2 + 24M_g^2\sum_{k\le d}\eta_{f,k}^2. \tag{30}$$

Finally, since $\|F-\langle\hat f,\hat g\rangle\|\le\sqrt{\mathcal{E}_{\mathrm{trunc}}}+R$,

$$\|F-\langle\hat f,\hat g\rangle\|^2 \le 2\mathcal{E}_{\mathrm{trunc}} + 2R^2 \le 2\mathcal{E}_{\mathrm{trunc}} + 6\sum_{k\le d}\sigma_k^2\eta_{g,k}^2 + 48M_g^2\sum_{k\le d}\eta_{f,k}^2. \tag{31}$$

As $2\mathcal{E}_{\mathrm{trunc}}\le 48\mathcal{E}_{\mathrm{trunc}}$ trivially, the right-hand side is at most 48 times the sum of the truncation term and the realization term of the theorem. Taking the infimum over $f\in\mathcal{F}$, $g\in\mathcal{G}$, the bound holds with $C\le 48\le 144$. We do not optimize the constant, which is what the statement of the theorem records.

Theorem 3 (Smoothness controls the truncation term)**.** *Let $\mathcal{U}\subset\mathbb{R}^{m_U}$ and $\mathcal{V}\subset\mathbb{R}^{m_V}$ be bounded domains, and let μ_U,μ_V have densities bounded above and below with respect to Lebesgue measure. If F is s-smooth in its second argument, $F\in H^s(\mathcal{U}\times\mathcal{V})$, then*

$$\sigma_k = O\big(k^{-s/m_V}\big), \qquad \sum_{k>d}\sigma_k^2 = O\big(d^{1-2s/m_V}\big) \quad \textit{for } 2s>m_V. \tag{32}$$

If F is real analytic on $\mathcal{U}\times\mathcal{V}$, then $\sigma_k=O(e^{-ck})$ for some $c>0$ and $\sum_{k>d}\sigma_k^2=O(e^{-2cd})$.

Proof of Theorem 3 (Smoothness controls the truncation term). The squared singular values of T_F are the eigenvalues of the self-adjoint, positive semidefinite integral operator $T_F^*T_F$ on $L^2(\mu_V)$ with kernel

$$K(v,v') = \int_{\mathcal{U}} F(u,v)\,F(u,v')\,d\mu_U(u), \tag{33}$$

which is positive definite and inherits the smoothness of F in its second argument: differentiating under the integral, an s-smooth kernel on a bounded m_V-dimensional domain has eigenvalues decaying as $O(k^{-2s/m_V})$. This is the classical eigenvalue decay for positive definite kernels (Reade, 1983), and the analytic case decays exponentially (Little & Reade, 1984). Taking square roots gives $\sigma_k=O(k^{-s/m_V})$, and summing the tail,

$$\sum_{k>d}\sigma_k^2 = O\Big(\int_d^\infty x^{-2s/m_V}\,dx\Big) = O\big(d^{1-2s/m_V}\big) \quad \text{for } 2s>m_V, \tag{34}$$

which is finite exactly when $2s>m_V$. In the analytic case, $\sigma_k=O(e^{-ck})$ gives $\sum_{k>d}\sigma_k^2=O(e^{-2cd})$ by the geometric series.

Proposition 2 (Approximation separation)**.** *Let $F\in\mathcal{M}_d$ with interaction modes $a_k\in H^s(\mathcal{U})$, $b_k\in H^s(\mathcal{V})$. To approximate F within ε in $L^2(\mu_U\otimes\mu_V)$, a dual-encoder head whose encoders are ReLU networks requires*

$$N_{\mathrm{dual}} = O\Big(d\big(\varepsilon^{-m_U/s}+\varepsilon^{-m_V/s}\big)\Big) \tag{35}$$

parameters, whereas any single-tower model that approximates F as a general (m_U+m_V)-dimensional H^s function requires $\Omega(\varepsilon^{-(m_U+m_V)/s})$ parameters in the worst case. Hence $N_{\mathrm{dual}}/N_{\mathrm{single}}\to 0$ as $\varepsilon\to 0$.

Table 6: Embedding-level normalizations as sections of the gauge orbit. Each row shows a constraint on the encoders, the part of the gauge it pins, the residual group that survives, and what the scheme identifies. See Proposition 3 and Theorems 4 to 2 for the statements, and Appendix C for proofs.

Normalization	Constraint	Pins	Residual gauge	Identifies to
None	none	nothing	GL_d	nothing
One-sided L2 and softplus (Zhao & Li, 2026b)	$\|g(v)\| = 1$, $g(v) \succeq 0$	pointwise scale of g	not controlled	no guarantee
Per-coordinate standardization	$\mathrm{diag}(\Sigma_g) = I$	coordinate scales	$d^2 - d$, shears and rotations	no guarantee
Cosine, two towers (Radford et al., 2021)	$\|f\| = \|g\| = 1$	angles	O	up to a rotation
Two-sided nonnegative	$f \succeq 0$, $g \succeq 0$	signs, shears, rotations	$\mathrm{Perm} \ltimes \mathrm{D}_+$	permutation and scale
Whitening	$\Sigma_g = I$, $\Sigma_f = \Lambda$	all continuous gauge	$\mathrm{D}_\pm$, order fixed	permutation and sign

Proof of Proposition 2 (Approximation separation). Since $F \in \mathcal{M}_d$ with modes $a_k \in H^s(\mathcal{U})$, $b_k \in H^s(\mathcal{V})$, the truncation term of Theorem 1 vanishes for the rank-d head, and the entire approximation error is the realization term. Approximate each of the $2d$ modes by ReLU networks with the standard H^s rates (Yarotsky, 2017): to reach mode error ε on a marginal space of dimension m requires $O(\varepsilon^{-m/s})$ parameters. Setting the mode errors to ε makes the realization term of Theorem 1 of order $O(\varepsilon^2)$, so a total of

$$N_{\mathrm{dual}} = O\Big(d\big(\varepsilon^{-m_U/s} + \varepsilon^{-m_V/s}\big)\Big) \tag{36}$$

parameters suffice. For the lower bound, a single-tower model that treats F as a general $(m_U + m_V)$-dimensional H^s function must contend with the Kolmogorov width of the Sobolev class, whose L^2 approximation is bounded below by $\Omega(\varepsilon^{-(m_U+m_V)/s})$ parameters in the worst case (classical Sobolev width results). Since $m_U + m_V > m_U, m_V$, the exponent is additive in the dimensions, so $N_{\mathrm{dual}}/N_{\mathrm{single}} = O\big(\varepsilon^{m_{\min}/s}\big) \to 0$ as $\varepsilon \to 0$, where $m_{\min} = \min\{m_U, m_V\}$.

Proof of Proposition 1 (Compute separation). Counting operations: the head requires n forward passes of f, m forward passes of g, and nm inner products of dimension d, i.e., nmd multiply-adds, for a total of $O((n+m)C_{\mathrm{net}} + nmd)$. A joint model processes each of the nm pairs separately, costing $O(nm\, C_{\mathrm{net}})$. When $C_{\mathrm{net}} \gg d$, the head is cheaper by a factor of order C_{net}/d in the regime where $nm \gg n + m$.

C Proofs for Section 3

We prove the results of Section 3 in order. Throughout, $\mathcal{L}(f, g) = \mathbb{E}_{u,v}[(F(u, v) - \langle f(u), g(v)\rangle)^2]$ is the population risk. Table 6 collects the normalization taxonomy moved from Section 3. Proposition 3 and Remark 4 complete the picture.

The gauge group GL_d stratifies into per-coordinate scalings (D_+, d dimensions), rotations (O, $d(d-1)/2$), permutations and signs (discrete), and the remaining shears. The guarantees of Section 3 come from two-sided constraints, which act on both encoders and can pin the joint freedom of the pair. One-sided constraints cannot do this in general, and the schemes that use them are best understood as stability fixes. Normalization schemes are therefore compared on four measurable quantities: the residual gauge subgroup of Table 6, the Hessian condition number on the constrained manifold, related to the gap by Corollary 3, the expressivity cost of the constraint, for example the bounded range $[-1, 1]$ of cosine normalization, and the output semantics, which determines which decoder and loss can be attached.

Remark 4 (One-sided constraints fix scale, not identifiability)**.** *The scheme used by the QK-attention matching networks applies a softplus and a pointwise unit-norm constraint to one encoder only (Zhao & Li, 2026b). It removes the global scaling redundancy of the representation, the parameter α in $(\alpha f)\,(g/\alpha)$, whose growth caused training instabilities in early versions of those networks,*

and it leaves the other encoder unconstrained. The constraint therefore provides stability without identifiability: the residual freedom of the pair depends on the geometry of the trained encodings and is not controlled by the constraint. The per-coordinate standardization of Proposition 3 has the same character, pinning the d coordinate scales of one encoder and nothing else.

Proposition 3 (Per-coordinate standardization pins the scales)**.** *Consider the constraint $\mathrm{diag}(\Sigma_g) = I$, which standardizes each coordinate of one encoder to unit variance. It pins the positive diagonal subgroup D_+ of dimension d. Generically the residual gauge has dimension $d^2 - d$ and contains the shears and rotations that mix coordinates.*

Lemma 2 (Gauge invariance)**.** *For every $A \in \mathrm{GL}_d$ and every pair of inputs (u, v), $\langle Af(u), A^{-\top} g(v)\rangle = \langle f(u), g(v)\rangle$, so Φ_A leaves the represented function and hence the population risk unchanged. With the second moments $\Sigma_f = \mathbb{E}_u[f(u)f(u)^\top]$ and $\Sigma_g = \mathbb{E}_v[g(v)g(v)^\top]$, the action transforms them by $\Sigma_{Af} = A\Sigma_f A^\top$ and $\Sigma_{A^{-\top}g} = A^{-\top}\Sigma_g A^{-1}$, so the eigenvalues of the product $\Sigma_f\Sigma_g$ are gauge invariants.*

Proof of Lemma 2 (Gauge invariance). The identity is the cancellation

$$\langle Af(u), A^{-\top} g(v)\rangle = f(u)^\top A^\top A^{-\top} g(v) = f(u)^\top g(v) = \langle f(u), g(v)\rangle, \tag{37}$$

so the represented function and hence the risk are unchanged. For the second moments, substitution gives

$$\Sigma_{Af} = \mathbb{E}_u[Af(u)f(u)^\top A^\top] = A\Sigma_f A^\top, \qquad \Sigma_{A^{-\top}g} = A^{-\top}\Sigma_g A^{-1}, \tag{38}$$

and the product transforms by the similarity $\Sigma_{Af}\Sigma_{A^{-\top}g} = A\,\Sigma_f\Sigma_g\,A^{-1}$, whose eigenvalues are invariant. At a global minimizer, the represented function equals the rank-d truncation F_d of the target when $\sigma_d > \sigma_{d+1}$ (Corollary 1), and the coefficient computation of the proof of Theorem 2 below shows that the eigenvalues of $\Sigma_f\Sigma_g$ are then the squared singular values σ_k^2, which is the claim of Remark 5.

Remark 5 (The invariant content of a trained head)**.** *The gauge-invariant content of a trained head is exactly the function it represents, summarized by its interaction spectrum. Lemma 2 adds a finite-dimensional shadow: at any global minimizer of the population risk, the eigenvalues of $\Sigma_f\Sigma_g$ are the squared interaction singular values. This quantity is well defined under every normalization, and it reappears in Corollary 2.*

Proposition 4 (Ill-posedness from unfixed gauge)**.** *Let $(f^\star, g^\star)$ be a global minimizer of the risk and let S be its stabilizer, the set of $A \in \mathrm{GL}_d$ with $\Phi_A(f^\star, g^\star) = (f^\star, g^\star)$. Then the Hessian of the risk at $(f^\star, g^\star)$ has at least*

$$\dim\ker \nabla^2\mathcal{L}(f^\star, g^\star) \;\geq\; d^2 - \dim S \tag{39}$$

zero directions, coming from the tangent space of the gauge orbit. Without normalization this is generically d^2 degenerate directions, so the second-order problem is rank-deficient and the optimization problem is ill-posed.

Proof of Proposition 4 (Ill-posedness from unfixed gauge). Let $(f^\star, g^\star)$ be a global minimizer and S its stabilizer. By Lemma 2, the risk is constant along the orbit $\mathcal{O} = \{\Phi_A(f^\star, g^\star) : A \in \mathrm{GL}_d\}$, so for every $X \in \mathfrak{gl}_d$ the curve $A(t) = \exp(tX)$ gives a path $\gamma_X(t) = \Phi_{A(t)}(f^\star, g^\star)$ along which $\mathcal{L} \circ \gamma_X \equiv \mathcal{L}(f^\star, g^\star)$. The velocity is

$$\dot\gamma_X(0) = (Xf^\star,\, -X^\top g^\star) \in T_{(f^\star, g^\star)}, \tag{40}$$

and since the risk is constant along the path, its second derivative in this direction vanishes. Because $(f^\star, g^\star)$ is a global minimizer, $\nabla^2\mathcal{L} \succeq 0$, and a semidefinite quadratic form vanishes in a direction if and only if that direction lies in the kernel, so $\dot\gamma_X(0) \in \ker\nabla^2\mathcal{L}$ for every X. The map $X \mapsto \dot\gamma_X(0)$ has kernel equal to the Lie algebra of the stabilizer S, hence its image, which lies entirely in the kernel of the Hessian, has dimension $d^2 - \dim S$. This proves the bound. Without normalization the stabilizer at a generic minimizer is trivial, so the Hessian has at least d^2 zero directions.

Proof of Proposition 3 (Per-coordinate standardization pins the scales). The constraint $\mathrm{diag}(\Sigma_g) = I$ is d scalar conditions that fix the d coordinate scales of g: for the transformation $A^{-1} = \mathrm{diag}(s_1, \ldots, s_d) \in \mathrm{D}_+$ acting on g, the condition forces $s_k^2[\Sigma_g]_{kk} = 1$, pinning each s_k. Transformations A that mix coordinates generically change $\mathrm{diag}(A^{-\top}\Sigma_g A^{-1})$ unless they also

rotate Σ_g to a new matrix with the same diagonal. The residual group is the set of A preserving the diagonal condition, which contains the orthogonal group conjugated by any diagonal scale and the shears that preserve the diagonal of Σ_g. Counting dimensions, the constraint fixes exactly the d-dimensional subgroup D_+, and the generic residual dimension is $d^2 - d$.

Theorem 4 (Cosine normalization leaves a full rotation)**.** *Consider two-sided unit normalization, $\tilde{f} = f/\|f\|$ and $\tilde{g} = g/\|g\|$, with scores $\cos\angle(f(u), g(v))$, as in the contrastive models of Table 5 (Radford et al., 2021). A temperature that rescales the logits changes no gauge. If the encodings are nondegenerate, in that the ranges of f and g contain open subsets of $\mathbb{R}^d$, the residual gauge group is exactly the orthogonal group,*

$$R = \left\{ A \in \mathrm{GL}_d : \; A^\top A = I \right\} = \mathrm{O}. \tag{41}$$

For every rotation $R \in \mathrm{O}$, the encodings Rf and Rg give pointwise identical cosine scores, and no transformation outside O does.

Corollary 2 (Contrastive dimensions are uninterpretable)**.** *Under the residual symmetry of Theorem 4, the individual coordinates of a cosine-normalized embedding carry no intrinsic meaning: replacing (f, g) by (Rf, Rg) for any rotation R yields the same model and the same loss, and only rotation-invariant quantities are well defined, among them norms, pairwise inner products, and the eigenvalues of $\Sigma_f\Sigma_g$, which by Remark 5 are the squared interaction singular values at any optimum. This is a theorem-level explanation of the widely reported observation that contrastive embedding dimensions are uninterpretable. The remedy is the whitening scheme of Theorem 2.*

Proof of Theorem 4 (Cosine normalization leaves a full rotation). Let R be the residual group of the two-sided unit normalization, i.e., the set of $A \in \mathrm{GL}_d$ such that, after renormalizing, the cosine scores are unchanged.

($\supseteq$) For $R \in \mathrm{O}$, $\|Rf(u)\| = \|f(u)\|$ and $\|Rg(v)\| = \|g(v)\|$ by orthogonality, and $\langle Rf(u), Rg(v)\rangle = \langle f(u), g(v)\rangle$, hence

$$\cos\angle(Rf(u), Rg(v)) = \frac{\langle Rf(u), Rg(v)\rangle}{\|Rf(u)\|\,\|Rg(v)\|} = \frac{\langle f(u), g(v)\rangle}{\|f(u)\|\,\|g(v)\|} \tag{42}$$

pointwise, so $\mathrm{O} \subseteq R$.

($\subseteq$) Suppose A preserves every cosine between the encodings. Since the ranges of f and g contain open subsets of $\mathbb{R}^d$, the encodings realize a full-dimensional family of directions, so A preserves all angles between vectors. A linear map that preserves every angle is a similarity: by the classical characterization of conformal linear maps, $A = cR$ for some $c > 0$ and $R \in \mathrm{O}$. The scalar c is absorbed by the unit renormalization of both encoders, so within the cosine score A acts identically to the rotation R, giving $R \subseteq \mathrm{O}$.

Corollary 2 follows immediately. For any rotation $R \in \mathrm{O}$ the pair (Rf, Rg) is a different encoder pair with pointwise identical cosine scores and identical loss, so the individual coordinates carry no intrinsic meaning. The rotation-invariant quantities are the norms, the inner products, and the eigenvalues of $\Sigma_f\Sigma_g$. By Remark 5 these are the squared interaction singular values at any optimum. Corollary 2 is the statement of these two sentences at the theorem level.

Theorem 5 (Two-sided nonnegativity gives NMF identifiability)**.** *Suppose both encoders are constrained to nonnegative outputs, $f(u) \succeq 0$ and $g(v) \succeq 0$, as obtained by applying a softplus or exponential nonlinearity to raw encodings. On a finite set of inputs the model is the nonnegative matrix factorization $M = FG^\top$ of a nonnegative matrix M. If the encodings are nondegenerate, in that their ranges generate the positive orthant, the residual gauge group shrinks to the monomial group,*

$$R = \left\{ A \in \mathrm{GL}_d : \; A \succeq 0, \; A^{-\top} \succeq 0 \right\} = \mathrm{Perm} \ltimes \mathrm{D}_+, \tag{43}$$

the group of coordinate permutations composed with positive diagonal scalings. If the factorization is separable in the sense of Donoho & Stodden (2003), with an anchor row for each latent factor as in Arora et al. (2012), then it is essentially unique and the encoders are identified up to permutation and scaling.

Proof of Theorem 5 (Two-sided nonnegativity gives NMF identifiability). On a finite set of inputs, the nonnegative encoders form matrices $F \in \mathbb{R}^{n\times d}_{\geq 0}$, $G \in \mathbb{R}^{m\times d}_{\geq 0}$, and the model is the nonnegative factorization $M = FG^\top$. The residual group consists of $A \in \mathrm{GL}_d$ such that $Af \succeq 0$ and $A^{-\top} g \succeq 0$ for all nonnegative encodings. Since the ranges generate the positive orthant, $Af \succeq 0$ for all nonnegative f forces $A \succeq 0$ entrywise, and likewise $A^{-\top} \succeq 0$. A real matrix that is nonnegative and whose inverse transpose is also nonnegative is a generalized permutation matrix: its inverse transpose being nonnegative forces each row and column of A to have exactly one nonzero entry, and nonnegativity makes that entry positive. Hence $R \subseteq \mathrm{Perm} \ltimes \mathrm{D}_+$, and the reverse inclusion is immediate, so $R = \mathrm{Perm} \ltimes \mathrm{D}_+$. Under separability in the sense of Donoho & Stodden (2003), each latent factor has an anchor row whose encoding is nonzero in exactly one coordinate, which identifies the extreme rays of the latent cone and hence the factor directions. The constructive algorithm of Arora et al. (2012) then yields the decomposition uniquely up to permutation and scaling. The degradation of this guarantee when separability holds only approximately is analyzed in Appendix E.

Proof of Theorem 2 (Whitening fixes the gauge up to permutation and sign). The proof has four steps.

(1) Every global minimizer represents the rank-d truncation. By Corollary 1 the minimal risk is $\sum_{k>d} \sigma_k^2$, attained exactly by the truncated series $F_d = \sum_{k\leq d} \sigma_k a_k \otimes b_k$, which is unique since $\sigma_d > \sigma_{d+1}$. Hence any minimizer satisfies $\langle f(u), g(v)\rangle = F_d(u, v)$ almost everywhere, i.e.,

$$\sum_{k=1}^{d} f_k(u)\, g_k(v) = \sum_{k=1}^{d} \sigma_k\, a_k(u)\, b_k(v). \tag{44}$$

(2) The span of the encoders is pinned. As rank-d representations of the same tensor, the left factor spaces and right factor spaces coincide: $\mathrm{span}\{f_1, \ldots, f_d\} = \mathrm{span}\{a_1, \ldots, a_d\}$ and $\mathrm{span}\{g_1, \ldots, g_d\} = \mathrm{span}\{b_1, \ldots, b_d\}$. So there exist coefficient matrices $B, C \in \mathbb{R}^{d\times d}$ with $f = B\, a_{[d]}$ and $g = C\, b_{[d]}$, where $a_{[d]} = (a_1, \ldots, a_d)^\top$ and likewise for b. Substituting into the identity above and using the orthonormality of the modes gives

$$B^\top C = \mathrm{diag}(\sigma_1, \ldots, \sigma_d) =: S. \tag{45}$$

(3) The whitening constraints force C to be a signed permutation. By the orthonormality of $\{a_k\}$,

$$\Sigma_f = \mathbb{E}_u[f(u)f(u)^\top] = B\, B^\top, \qquad \Sigma_g = C\, C^\top. \tag{46}$$

The constraint $\Sigma_g = I$ gives $CC^\top = I$, so $C \in \mathrm{O}$. The constraint $\Sigma_f = \Lambda$ diagonal gives $BB^\top = \Lambda$. From $B^\top C = S$ and $C \in \mathrm{O}$ we get $B^\top = SC^{-1} = SC^\top$, so $B = CS$, and therefore

$$BB^\top = C\, S^2\, C^\top = \Lambda. \tag{47}$$

Thus the orthogonal matrix C diagonalizes the diagonal matrix $S^2 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. Since the σ_k are distinct, the eigenspaces of S^2 are the coordinate axes, and any orthogonal matrix diagonalizing S^2 maps each axis to itself up to sign, so $C \in \mathrm{Perm} \ltimes \mathrm{D}_\pm$ and $\Lambda = S^2$, i.e., $\lambda_k = \sigma_k^2$ with the ordering convention matching the decreasing σ_k. With the ordering fixed, $C = \mathrm{diag}(\pm 1, \ldots, \pm 1)$ and $B = CS = \mathrm{diag}(\pm\sigma_1, \ldots, \pm\sigma_d)$.

(4) Substitute back. With the coordinate order fixed, $g_k = \pm b_k$ and $f_k = \pm\sigma_k a_k$, the signs being paired so that the product $f_k g_k = \sigma_k a_k b_k$ is unique. This is exactly (10), and the encoders are identified up to permutation and sign.

Corollary 3 (The spectral gap controls conditioning and convergence)**.** *On the manifold defined by the constraints* (9)*, every global minimizer is isolated up to the discrete group of signs and permutations, and the Hessian of the risk is nondegenerate in the tangent directions of the manifold, with smallest nonzero curvature bounded below by a constant multiple of the spectral gap* $\Delta = \min_{1\leq k<d}(\sigma_k^2 - \sigma_{k+1}^2)$*. A projected gradient method that maintains the constraints therefore converges locally at a linear rate* $1 - \mu/L$*, where* μ *is bounded below by a constant multiple of* Δ *and* L *is the local smoothness constant. Wider gaps make the problem both better conditioned and faster to converge.*

Proof of Corollary 3 (The spectral gap controls conditioning and convergence). Isolation: by Theorem 2, on the manifold defined by (9) every global minimizer is unique up to the discrete group of signs and the permutation, and the represented function F_d is a strict minimum of the risk by the strict inequality $\sigma_d > \sigma_{d+1}$. The Hessian of the risk is therefore nondegenerate in the tangent directions of the manifold, and its smallest nonzero curvature is positive. The quantitative bound: the Hessian curvature in the tangent directions is controlled by the second derivative of the risk as the encoders move within the constraint manifold, and by the Davis-Kahan $\sin\Theta$ theorem the perturbation of a leading singular subspace under a perturbation of size ϵ is at most ϵ/Δ, where $\Delta = \min_{1\le k<d}(\sigma_k^2 - \sigma_{k+1}^2)$. Inverting this stability relation, the curvature of the risk in directions that move the encoders off the ideal modes is bounded below by a constant multiple of Δ: a displacement of size t in function space changes the risk by at least $c\,\Delta\,t^2$ for the truncation-optimal function. A projected gradient method that maintains the constraints therefore converges locally at a linear rate $1 - \mu/L$ with $\mu \ge c\,\Delta$ and L the local smoothness constant, as stated.

Theorem 6 (Soft whitening recovers the hard solution)**.** *For $\rho > 0$, consider the penalized risk*

$$\mathcal{L}_\rho(f,g) = \mathcal{L}(f,g) + \rho\Big(\big\|\Sigma_g - I_d\big\|_F^2 + \big\|\mathrm{offdiag}(\Sigma_f)\big\|_F^2\Big). \tag{48}$$

As $\rho \to \infty$, any convergent sequence of minimizers of $\mathcal{L}_\rho$ converges to a solution satisfying the constraints of Theorem 2 up to the ordering convention, and hence to the recovered modes (10)*. For finite ρ, the residual gauge directions that break the constraint $\Sigma_g = I$ carry curvature at least $2\rho\,c$ for a constant $c > 0$ that depends on the second moments, so conditioning improves monotonically with ρ.*

Proof of Theorem 6 (Soft whitening recovers the hard solution). (i) Convergence of minimizers. The zero set of the penalty

$$\rho\Big(\|\Sigma_g - I\|_F^2 + \|\,\mathrm{offdiag}(\Sigma_f)\|_F^2\Big) \tag{49}$$

is exactly the whitening constraint (9) up to the ordering convention. Since $\mathcal{L}_\rho \ge \mathcal{L}$, any limit point of a convergent sequence of minimizers satisfies the constraint (otherwise the penalty, and with it $\mathcal{L}_\rho$, would diverge) and is a global minimizer of $\mathcal{L}$ subject to the constraint, to which Theorem 2 applies.

(ii) Curvature of the penalty. Consider a gauge generator $X \in \mathfrak{gl}_d$ acting on the g-side as $g(t) = \exp(-tX^\top)g$. The second moment evolves as $\Sigma_{g(t)} = \exp(-tX^\top)\Sigma_g\exp(-tX)$. At a constraint-satisfying point $\Sigma_g = I$, the second derivative of $\|\Sigma_{g(t)} - I\|_F^2$ at $t = 0$ is $2\|X + X^\top\|_F^2$: expanding the exponential gives $\Sigma_{g(t)} = I - t(X^\top + X) + t^2X^\top X + O(t^3)$, so the squared deviation has leading term $t^2\|X + X^\top\|_F^2$ multiplied by 2. Directions that break $\Sigma_g = I$ have nonzero symmetric part, hence carry penalty curvature at least $2\rho\,c$ with $c > 0$ determined by the second moments through the contribution of $\mathrm{offdiag}(\Sigma_f)$. Directions with X antisymmetric are rotations, which preserve $\Sigma_g = I$ and are exactly the directions whose residual freedom is fixed by the ordering convention. Conditioning therefore improves monotonically with ρ, as stated.

D Proofs for Section 4

We prove the results of Section 4. Throughout, $\mathcal{R}(h) = \mathbb{E}_{u,v}[(F(u,v) - h(u,v))^2]$ is the population risk and $\hat{\mathcal{R}}$ its empirical version over the n samples. Corollary 6 and Proposition 6 are proved in Appendix E, as indicated in the main text.

Remark 6 (The general bound is conservative in d)**.** *Specializing Corollary 4 to the linear class gives $\mathfrak{R}_n(\mathcal{F}) = O(\sqrt{pd/n})$ and $\mathfrak{R}_n(\mathcal{G}) = O(\sqrt{qd/n})$, hence an estimation error of order $Bd^{3/2}\sqrt{(p+q)/n}$, a factor $\sqrt{d}$ worse than the minimax anchor* (56)*. The gap is structural: Lemma 3 sums the d coordinates as independent classes and cannot exploit the low-rank geometry that the matrix analysis of Theorem 7 uses. The general bound is therefore tight in its dependence on n and on the marginal dimensions, and conservative in the power of d. The linear case shows the achievable d-dependence is linear, and local Rademacher arguments close the gap. In the sequential setting, where (u_t, v_t) are chosen adaptively to maximize an unknown score $u^\top\Theta^\star v$, the same scaling appears as regret: bilinear bandits achieve $\tilde{O}(d\sqrt{(p+q)T})$ (Jun et al., 2019), again linear in d and in the marginal dimensions.*

Remark 7 (The separation is not a discrete artifact)**.** *The same separation holds for smooth kernels. On the circle, the narrowband kernel $F_h(u, v) = \rho_h(u-v)$ with a periodic Gaussian ρ_h of bandwidth h has a Fourier-diagonal interaction operator with singular values of order $e^{-2\pi^2 h^2 \xi^2}$, so $d_\varepsilon(F_h) = \Theta(1/h)$: the embedding dimension diverges as the kernel approaches the diagonal. Early interaction computes the difference $t = u - v$ exactly and approximates the one-dimensional smooth function ρ_h with $O(\mathrm{poly}(1/h))$ parameters (Yarotsky, 2017), a polynomial separation in the same direction. The strength of the separation is set by how the target depends jointly on its two arguments: through a comparison in the discrete case, through a difference in the smooth case, and not at all when the target is both high-rank and unstructured.*

Lemma 3 (Rademacher collapse of the inner-product class)**.** *For the class* (12) *with per-coordinate classes $\mathcal{F}^{(k)}, \mathcal{G}^{(k)}$,*

$$\mathfrak{R}_n(\mathcal{H}_d) \;\le\; \sum_{k=1}^{d} \mathfrak{R}_n\big(\mathcal{F}^{(k)} \cdot \mathcal{G}^{(k)}\big) \;\le\; 2\sum_{k=1}^{d}\Big(B_g\, \mathfrak{R}_n(\mathcal{F}^{(k)}) + B_f\, \mathfrak{R}_n(\mathcal{G}^{(k)})\Big), \tag{50}$$

so in particular $\mathfrak{R}_n(\mathcal{H}_d) \le 2d\big(B_g\, \mathfrak{R}_n(\mathcal{F}) + B_f\, \mathfrak{R}_n(\mathcal{G})\big)$ for shared per-coordinate classes: the complexity of the head is a sum of marginal complexities, not a complexity of the joint space.

Proof of Lemma 3 (Rademacher collapse of the inner-product class). Decompose the score as $\langle f(u), g(v)\rangle = \sum_{k=1}^{d} f_k(u) g_k(v)$. Rademacher complexity is subadditive over sums of function classes, so

$$\mathfrak{R}_n(\mathcal{H}_d) \le \sum_{k=1}^{d} \mathfrak{R}_n\big(\mathcal{F}^{(k)} \cdot \mathcal{G}^{(k)}\big), \tag{51}$$

where $\mathcal{F}^{(k)} \cdot \mathcal{G}^{(k)} = \{(u, v) \mapsto f_k(u) g_k(v)\}$ is the pointwise product class. For each coordinate, the product map $(f_k, g_k) \mapsto f_k g_k$ is Lipschitz in its first argument with constant B_g and in its second with constant B_f on the bounded range. The vector contraction inequality of Maurer (2016) then bounds the complexity of the product class by the sum of the two marginal complexities,

$$\mathfrak{R}_n\big(\mathcal{F}^{(k)} \cdot \mathcal{G}^{(k)}\big) \le 2\Big(B_g\, \mathfrak{R}_n(\mathcal{F}^{(k)}) + B_f\, \mathfrak{R}_n(\mathcal{G}^{(k)})\Big), \tag{52}$$

the factor 2 absorbing the contraction constants. Summing over the d coordinates and using $\mathfrak{R}_n(\mathcal{F}) = \max_k \mathfrak{R}_n(\mathcal{F}^{(k)})$ gives the displayed bound. The key structural point is that no complexity of the joint space $\mathcal{U} \times \mathcal{V}$ appears: the head is paid for by its two marginal classes.

Corollary 4 (Excess risk of empirical risk minimization)**.** *Let $\hat{F}$ be the empirical risk minimizer over $\mathcal{H}_d$. With probability at least $1-\delta$ over the sample, the excess risk decomposes into an approximation part, the truncation and realization terms of Theorem 1, and an estimation part through the class complexity:*

$$\|\hat{F} - F\|^2_{L^2(\mu_U \otimes \mu_V)} \;\lesssim\; \sum_{k>d} \sigma_k^2 + \mathcal{E}_{\mathrm{real}} \;+\; B\, d\big(\mathfrak{R}_n(\mathcal{F}) + \mathfrak{R}_n(\mathcal{G})\big) \;+\; B^2\sqrt{\frac{\log(1/\delta)}{n}}, \tag{53}$$

where $\mathcal{E}_{\mathrm{real}} = C\sum_{k\le d}(\sigma_k^2 \eta_{g,k}^2 + M_g^2 \eta_{f,k}^2)$ is the realization term and $\lesssim$ hides absolute constants.

Proof of Corollary 4 (Excess risk of empirical risk minimization). The regression oracle decomposition: for any $h \in \mathcal{H}_d$, writing $\hat{F}$ for the empirical risk minimizer,

$$\|F - \hat{F}\|^2_{L^2} \le \|F - h\|^2_{L^2} + 2 \sup_{h' \in \mathcal{H}_d} \big|\hat{\mathcal{R}}(h') - \mathcal{R}(h')\big|. \tag{54}$$

The uniform deviation is controlled as follows. The squared loss $\ell(y, h) = (y-h)^2$ is $4B$-Lipschitz in h on $|y|, |h| \le B$, so by the Talagrand contraction inequality the Rademacher complexity of the loss class is at most $4B\, \mathfrak{R}_n(\mathcal{H}_d)$ (Bartlett & Mendelson, 2002). Symmetrization bounds the expected uniform deviation by twice the Rademacher complexity of the loss class, and McDiarmid's inequality, with the bounded loss $|y-h|^2 \le 4B^2$, gives the concentration term:

$$\sup_{h' \in \mathcal{H}_d} \big|\hat{\mathcal{R}}(h') - \mathcal{R}(h')\big| \lesssim B\, \mathfrak{R}_n(\mathcal{H}_d) + B^2\sqrt{\frac{\log(1/\delta)}{n}} \tag{55}$$

with probability at least $1-\delta$. Choosing $h = h^\star$, the population optimum, makes $\|F - h^\star\|^2$ exactly the approximation term of Theorem 1, and Lemma 3 bounds $\mathfrak{R}_n(\mathcal{H}_d) \le 2d(B_g\mathfrak{R}_n(\mathcal{F}) + B_f\mathfrak{R}_n(\mathcal{G}))$. Absorbing the Lipschitz and contraction constants into the absolute constant gives (14). For parameterized encoder classes with p_f and p_g parameters under norm control, the standard Rademacher bound for parameterized classes gives $\mathfrak{R}_n(\mathcal{F}) = O(\sqrt{p_f/n})$ and $\mathfrak{R}_n(\mathcal{G}) = O(\sqrt{p_g/n})$, and the estimation term becomes $O(Bd\sqrt{(p_f+p_g)/n})$.

Theorem 7 (Linear heads are low-rank matrix sensing). *Let $\Theta^\star \in \mathbb{R}^{p\times q}$ have rank d and $\|\Theta^\star\|_F \le B$, and suppose the sensing matrices $u_i v_i^\top$ are drawn from an isotropic sub-Gaussian family and the noise is sub-Gaussian with variance σ_ε^2. The rank-constrained estimator attains*

$$\mathbb{E}\big\|\hat\Theta - \Theta^\star\big\|_F^2 \;\asymp\; \sigma_\varepsilon^2\,\frac{d(p+q)}{n}, \tag{56}$$

and this rate is minimax optimal. Equivalently, $\mathbb{E}\|\hat F - F\|_{L^2}^2 \asymp \sigma_\varepsilon^2\, d(p+q)/n$.

Proof of Theorem 7 (Linear heads are low-rank matrix sensing). The head computes $u^\top\Theta v$ with $\Theta = UV^\top$, and the observations are $y_i = \langle\Theta^\star,\, u_i v_i^\top\rangle_F + \varepsilon_i$, the standard form of low-rank matrix sensing with sensing matrices $A_i = u_i v_i^\top$ and rank constraint $\mathrm{rank}(\Theta) \le d$. Upper bound: the rank-$d$ manifold has $d(p+q-d)$ free parameters, and under restricted strong convexity of the sensing operator, which holds for isotropic sub-Gaussian sensing matrices, the constrained estimator satisfies $\mathbb{E}\|\hat\Theta - \Theta^\star\|_F^2 \lesssim \sigma_\varepsilon^2\, d(p+q)/n$. This is the robust error bound of Candes & Plan (2011) together with the high-dimensional rate of Negahban & Wainwright (2009). Lower bound: a Fano argument over a packing of the rank-d manifold gives the matching minimax lower bound $\Omega(\sigma_\varepsilon^2\, d(p+q)/n)$ (Rohde & Tsybakov, 2009). The equivalence with the function-space rate follows from the isotropy of u and v: $\mathbb{E}_{u,v}[(u^\top\Delta v)^2] \asymp \|\Delta\|_F^2$ for sub-Gaussian isotropic arguments, so $\mathbb{E}\|\hat F - F\|_{L^2}^2 \asymp \sigma_\varepsilon^2\, d(p+q)/n$.

Corollary 5 (Optimal rank selection). *If the interaction spectrum decays exponentially, $\sigma_k^2 = O(e^{-2ck})$, and the encoders are linear, the total error is minimized at*

$$d^\star \;\asymp\; \frac{1}{2c}\log\frac{n}{p+q}, \tag{57}$$

and the achieved error is $\|\hat F - F\|_{L^2}^2 \lesssim (p+q)\log(n/(p+q))/n$, a near-parametric rate in which fast spectral decay eliminates the curse of dimensionality.

Proof of Corollary 5 (Optimal rank selection). With exponential decay and linear encoders the total error of Theorem 7 and the truncation term balance as

$$\mathrm{err}(d) \;\lesssim\; e^{-2cd} + \frac{d(p+q)}{n}. \tag{58}$$

Setting the two terms equal, $e^{-2cd^\star} \asymp d^\star(p+q)/n$, and solving gives $d^\star \asymp (1/2c)\log(n/(p+q))$, where the logarithmic factor in d is absorbed by the constant. Substituting back, $e^{-2cd^\star} \asymp (p+q)/n$, so the achieved error is

$$\|\hat F - F\|_{L^2}^2 \lesssim \frac{(p+q)}{n}\Big(1 + \log\frac{n}{p+q}\Big) \asymp \frac{(p+q)\log(n/(p+q))}{n}, \tag{59}$$

the near-parametric rate stated in the corollary.

Theorem 8 (Flat spectrum forces an exponential embedding). *Let $\mathcal{U} = \mathcal{V} = \{\pm1\}^m$ with uniform measure and $N = 2^m$, and consider the equality function $F_{\mathrm{eq}}(u,v) = \mathbf{1}[u = v]$. Its interaction operator is $T_{F_{\mathrm{eq}}} = \frac{1}{N}\mathrm{Id}$, so all N singular values equal $1/N$ and the interaction rank is N. Every rank-d head therefore satisfies*

$$\frac{\inf_{f,g}\big\|F_{\mathrm{eq}} - \langle f, g\rangle\big\|_{L^2}^2}{\big\|F_{\mathrm{eq}}\big\|_{L^2}^2} \;=\; 1 - \frac{d}{N}, \tag{60}$$

so achieving relative error ε requires $d \ge (1-\varepsilon)2^m$, an embedding dimension exponential in the input size.

Proof of Theorem 8 (Flat spectrum forces an exponential embedding). Compute the interaction operator of F_{eq} on the uniform measure over $\{\pm 1\}^m$ with $N = 2^m$:

$$(T_{F_{\mathrm{eq}}} h)(u) = \sum_{v \in \{\pm 1\}^m} \mathbf{1}[u = v]\, h(v)\, \frac{1}{N} = \frac{h(u)}{N}, \tag{61}$$

so $T_{F_{\mathrm{eq}}} = \frac{1}{N}\mathrm{Id}$ and all N singular values equal $1/N$. Hence $\|F_{\mathrm{eq}}\|_{L^2}^2 = \sum_{k \geq 1} \sigma_k^2 = N/N^2 = 1/N$ and $\sum_{k>d} \sigma_k^2 = (N-d)/N^2$ for $d < N$. By Corollary 1, the relative squared error of the best rank-d head is

$$\frac{\inf_{f,g} \|F_{\mathrm{eq}} - \langle f, g\rangle\|_{L^2}^2}{\|F_{\mathrm{eq}}\|_{L^2}^2} = \frac{(N-d)/N^2}{1/N} = 1 - \frac{d}{N}, \tag{62}$$

which is at most ε only if $d \geq (1-\varepsilon)N = (1-\varepsilon)2^m$.

Theorem 9 (Early interaction escapes)**.** *For the equality function F_{eq} on $\{\pm 1\}^m \times \{\pm 1\}^m$, an early-interaction predictor with $O(m)$ parameters represents F_{eq} exactly, with zero error. Together with Theorem 8 this is an exponential separation: $O(m)$ parameters suffice under early interaction, while any late-interaction head needs $\Omega(2^m)$ embedding dimensions to reach constant relative error.*

Proof of Theorem 9 (Early interaction escapes). The construction uses the identity $u_i v_i = 1$ if and only if $u_i = v_i$, so that

$$\langle u, v\rangle = \sum_{i=1}^{m} u_i v_i = m - 2\,\mathrm{Ham}(u, v), \tag{63}$$

and $\langle u, v\rangle = m$ if and only if $u = v$, with the next possible value $m-2$. Hence $F_{\mathrm{eq}}(u, v) = \mathbf{1}[\langle u, v\rangle \geq m-1]$. A cross-attention layer with identity projections computes the query-key inner product $\langle u, v\rangle$ with $O(m)$ parameters, and a threshold unit $\phi(s) = \mathrm{ReLU}(s-(m-1)) - \mathrm{ReLU}(s-(m+1))$ evaluates to 1 exactly on $s = m$ and to 0 elsewhere on the discrete range, using $O(1)$ parameters. A joint MLP computes the same quantity bit by bit: each product $u_i v_i = ((u_i + v_i)^2 - 2)/2$ costs $O(1)$ ReLU units, the sum costs $O(m)$ additions, and the same threshold completes the network, again $O(m)$ parameters in total. A cross-encoder, being a strict superset of cross-attention, inherits the construction. Together with Theorem 8, this is the exponential separation $O(m)$ versus $\Omega(2^m)$.

Theorem 10 (Usability criterion)**.** *Given a problem (F, μ_U, μ_V), a target relative error ε, and a sample budget n: if the effective interaction rank $d_\varepsilon(F)$ is small, a late-interaction head is appropriate, since its approximation and estimation errors are then small and it enjoys the compute and statistical separations of Proposition 1 and Lemma 3. If the spectrum is flat, $d_\varepsilon(F) = \Omega(N)$, the truncation floor of Corollary 1 locks every rank-d head to relative error at least ε regardless of capacity, samples, or training, while early-interaction models reach ε polynomially when the target admits a low-dimensional joint structure. The deciding quantity, the spectral decay rate, is estimable from data by a weighted singular value decomposition of the empirical kernel matrix, so the criterion is actionable before training either model.*

Proof of Theorem 10 (Usability criterion). First case: if $d_\varepsilon(F)$ is small, take $d \gtrsim d_\varepsilon(F)$. The approximation error of Theorem 1 is at most $\varepsilon\|F\|^2$ plus the realization term, and the estimation error of Corollary 4 scales as $d(\mathfrak{R}_n(\mathcal{F}) + \mathfrak{R}_n(\mathcal{G}))$, small in d. The compute separation of Proposition 1 and the marginal-complexity collapse of Lemma 3 apply without further conditions. Second case: if the spectrum is flat with $d_\varepsilon(F) = \Omega(N)$, then for every $d < d_\varepsilon(F)$ the truncation floor of Corollary 1, in the explicit form of Theorem 8, locks the relative error of every rank-d head above ε regardless of capacity, samples, or training, while the constructive bounds of Theorem 9 and Remark 7 show that early-interaction models reach ε with a polynomial parameter budget when the target depends on its arguments through a low-dimensional joint statistic. Actionability: the empirical kernel matrix $\hat{M}_{ij} = y(u_i, v_j)$ on a grid of input pairs, weighted by the reference measures, has a weighted singular value decomposition whose values estimate $\{\sigma_k\}$. A log-linear fit of the decay of $\hat{\sigma}_k$ against k (exponential) or $\log k$ (polynomial) decides the decay type and reads off d_ε before either model is trained.

E FINITE-SAMPLE TECHNICAL LEMMAS

This appendix collects the finite-sample lemmas used in the main text. Section A makes the whitening identification of Theorem 2 quantitative: concentration of empirical second moments (Lemma 4), stability of the whitening map (Lemma 5), and a Davis-Kahan propagation step (Theorem 11) deliver the sample complexity of Corollary 6. Section B quantifies the degradation of the NMF route when separability holds only approximately (Theorem 12). Section C couples identification with estimation in one ledger (Proposition 5), which is the proof of Proposition 6. Throughout, $\|\cdot\|$ on matrices is the spectral norm and $\|\cdot\|_F$ the Frobenius norm.

E.1 EMPIRICAL WHITENING AND MODE IDENTIFICATION

Lemma 4 (Matrix Bernstein concentration of empirical moments)**.** *Let $\{u_i\}_{i=1}^n$ be drawn i.i.d. from μ_U and let f have bounded outputs, $\|f(u)\| \le B_f$. The empirical second moment $\hat{\Sigma}_f = \frac{1}{n}\sum_i f(u_i)f(u_i)^\top$ satisfies, for every $t > 0$,*

$$\Pr\big[\|\hat{\Sigma}_f - \Sigma_f\| \ge t\big] \le 2d\,\exp\Big(\frac{-nt^2/2}{\nu_f + B_f^2 t/3}\Big), \qquad \nu_f := B_f^2\|\Sigma_f\|, \tag{64}$$

and hence, with probability at least $1-\delta$,

$$\|\hat{\Sigma}_f - \Sigma_f\| \le \sqrt{\frac{2\nu_f \log(2d/\delta)}{n}} + \frac{2B_f^2\log(2d/\delta)}{3n} =: r_f(n,\delta), \tag{65}$$

which is $O\big(B_f^2\sqrt{\log(d/\delta)/n}\big)$ for large n. The symmetric statement holds for g with radius $r_g(n,\delta)$.

Proof. The matrices $X_i = f(u_i)f(u_i)^\top - \Sigma_f$ are independent, zero mean, and symmetric, with $\|X_i\| \le \|f(u_i)f(u_i)^\top\| + \|\Sigma_f\| \le 2B_f^2$. For the variance parameter, since $\mathbb{E}[X_i^2] \preceq \mathbb{E}\|f(u_i)\|^2 f(u_i)f(u_i)^\top \preceq B_f^2\Sigma_f$, we have $\big\|\sum_i \mathbb{E}[X_i^2]\big\| \le nB_f^2\|\Sigma_f\| = n\nu_f$. The matrix Bernstein inequality (Tropp, An Introduction to Matrix Concentration Inequalities) applied to $\sum_i X_i$ gives the tail bound, and solving the tail for t at the level δ gives the displayed high-probability bound. □

Lemma 5 (Stability of the whitening map)**.** *Let $\Sigma_g \succeq \gamma I$ for $\gamma > 0$, and suppose $\|\hat{\Sigma}_g - \Sigma_g\| \le r_g \le \gamma/2$. Then*

$$\big\|\hat{\Sigma}_g^{-1/2} - \Sigma_g^{-1/2}\big\| \le \frac{C_0\, r_g}{\gamma^{3/2}}, \tag{66}$$

for an absolute constant C_0.

Proof. The map $A \mapsto A^{-1/2}$ is Fréchet differentiable on the cone $A \succeq \gamma I$ with derivative bounded by $\frac{1}{2}\gamma^{-3/2}$: using the integral representation $A^{-1/2} = \frac{1}{\pi}\int_0^\infty (A+sI)^{-1}s^{-1/2}\,ds$, the derivative at A applied to H is $-\frac{1}{\pi}\int_0^\infty (A+sI)^{-1}H(A+sI)^{-1}s^{-1/2}\,ds$, whose norm is at most $\|H\|\cdot\frac{1}{\pi}\int_0^\infty(\gamma+s)^{-2}s^{-1/2}\,ds = \|H\|\frac{1}{2}\gamma^{-3/2}$. The condition $r_g \le \gamma/2$ keeps $\hat{\Sigma}_g \succeq \frac{\gamma}{2}I$ by Weyl's inequality, so the mean value theorem applies along the segment and gives the bound with C_0 absorbing the integration constant. □

Theorem 11 (Finite-sample mode identification)**.** *Under the assumptions of Theorem 2, suppose the encoders represent the population optimum exactly, $\langle f, g\rangle = F_d$, and whitening is performed on the empirical moments of a sample of size n, with encoder outputs bounded by B. Let $r = \max(r_f, r_g)$ be the moment radii of Lemma 4, let $\Delta = \min_{1\le k<d}(\sigma_k^2 - \sigma_{k+1}^2)$ be the spectral gap, and assume $r \le c_1\min(\gamma,\Delta)$ for a small absolute constant c_1. Then with probability at least $1-\delta$, the empirically whitened encoders $(\hat{f},\hat{g})$ recover the modes up to*

$$\big|\sin\angle(\hat{g}_k, b_k)\big| \le \frac{C\,r}{\Delta}, \qquad \|\hat{g}_k - (\pm b_k)\| \le \frac{C'\,r}{\Delta}, \qquad |\hat{\Lambda}_{kk} - \sigma_k^2| \le C''\,r, \tag{67}$$

where $\hat{\Lambda}_{kk}$ is the estimated k-th interaction strength. In particular the identification error is $O\big(B^2\sqrt{\log(d/\delta)/n}/\Delta\big)$, vanishing as $n^{-1/2}$ and amplified by the inverse spectral gap.

Proof. (1) Empirical whitening followed by an eigendecomposition of the whitened second moment is equivalent to a singular value decomposition of the empirical interaction operator. Since the encoders represent $F_d = \sum_{k \le d} \sigma_k a_k \otimes b_k$ exactly, the population singular subspaces are spanned by $\{b_k\}$ on the g-side and $\{a_k\}$ on the f-side. (2) The perturbation of the empirical operator relative to its population version is bounded by $C_3\, r/\gamma^{3/2} \le C_4\, r$: the moment error $\|\hat{\Sigma}_f - \Sigma_f\| \le r_f$ from Lemma 4, the whitening error from Lemma 5, and the remaining factors are absorbed into the constant, which for brevity absorbs the γ dependence as well. (3) The k-th empirical mode is the k-th singular vector of the perturbed operator. By the Davis-Kahan $\sin\Theta$ theorem, the perturbation of a singular vector is bounded by the operator perturbation divided by the gap between the corresponding singular value squared and its neighbors, which is at least Δ:

$$\sin\angle(\hat{g}_k, b_k) \le \frac{\|E\|}{\min(\sigma_{k-1}^2 - \sigma_k^2,\ \sigma_k^2 - \sigma_{k+1}^2)} \le \frac{C\, r}{\Delta}. \tag{68}$$

Choosing the sign so that the inner product is nonnegative, $\|\hat{g}_k - (\pm b_k)\| \le \sqrt{2}\, \sin\angle \le C' r/\Delta$. (4) For the strengths, Weyl's inequality gives $|\hat{\Lambda}_{kk} - \sigma_k^2| \le \|E\| \le C'' r$. □

Corollary 6 (Sample complexity of whitening)**.** *Under the assumptions of Theorem 11, whitening the empirical second moments of a sample of size n identifies the interaction modes to accuracy ε with probability at least $1 - \delta$ provided*

$$n \ \gtrsim\ \frac{B^4\, \log(d/\delta)}{\Delta^2\, \varepsilon^2}, \tag{69}$$

where $\Delta = \min_{1 \le k < d}(\sigma_k^2 - \sigma_{k+1}^2)$ is the spectral gap.

Proof of Corollary 6 (Sample complexity of whitening). Set $Cr/\Delta \le \varepsilon$ with $r = O\big(B^2\sqrt{\log(d/\delta)/n}\big)$ from Lemma 4 and solve for n:

$$n \ \gtrsim\ \frac{B^4\, \log(d/\delta)}{\Delta^2\, \varepsilon^2}. \tag{70}$$

The sample budget scales as $1/\Delta^2$ in the gap and $1/\varepsilon^2$ in the target accuracy, matching (69).

Corollary 7 (Block stability under small gaps)**.** *Adopt the setting and notation of Theorem 11, and let $S = \{k_0, \dots, k_1\} \subseteq \{1, \dots, d\}$ be a contiguous block of indices. Write*

$$\delta_S = \min\big(\sigma_{k_0-1}^2 - \sigma_{k_0}^2,\ \sigma_{k_1}^2 - \sigma_{k_1+1}^2\big) \tag{71}$$

for the separation of the block from the rest of the spectrum, with the convention that a term is $+\infty$ when its index falls outside $\{1, \dots, d\}$. Let P_S be the population spectral projector onto $\mathrm{span}\{b_k : k \in S\}$ and $\hat{P}_S$ its empirical counterpart. If the perturbation E of the whitened empirical operator, bounded in step (2) of the proof of Theorem 11, satisfies $\|E\| \le \delta_S/2$, then

$$\|\hat{P}_S - P_S\| \ \le\ \frac{C\, r}{\delta_S}. \tag{72}$$

Only the separation of the block from its complement enters; the gaps inside S do not appear. Two consequences follow. When a pair of singular values nearly coincides, $\sigma_k^2 - \sigma_{k+1}^2 \to 0$, the single-mode bound of Theorem 11 diverges and the individual coordinates become unidentifiable, while the block $S = \{k, k+1\}$ remains stable, so identification degrades to uniqueness up to rotation within the block, the quantitative version of the degenerate-case remark of Section 3. Taking $S = \{1, \dots, \hat{r}\}$, the leading $\hat{r}$-dimensional subspace is identified at rate $r/(\sigma_{\hat{r}}^2 - \sigma_{\hat{r}+1}^2)$ whatever the internal spectrum does, so a truncated head can be well posed at the subspace level even when no individual mode is.

Proof of Corollary 7 (Block stability under small gaps). Step (1) of the proof of Theorem 11 identifies the population object as a symmetric operator whose eigenvalues are $\{\sigma_k^2\}$ in decreasing order and whose eigenvectors are the interaction modes $\{b_k\}$; step (2) bounds the perturbation of its empirical counterpart by $\|E\| \le C_4\, r$. The block S occupies the eigenvalue interval $[\sigma_{k_1}^2, \sigma_{k_0}^2]$, and by construction every eigenvalue outside S lies at distance at least δ_S from that interval. Weyl's inequality moves each eigenvalue by at most $\|E\| \le \delta_S/2$, so under the stated condition the empirical eigenvalues indexed by S remain separated from the rest and $\hat{P}_S$ is well defined as the spectral

projector of the same index set. The Davis-Kahan $\sin\Theta$ theorem in its spectral-projector form then gives

$$\|\hat{P}_S - P_S\| \;\le\; \frac{2\|E\|}{\delta_S} \;\le\; \frac{2C_4\, r}{\delta_S}, \tag{73}$$

which is (72) after absorbing the constant. The within-block gaps are absent because P_S is invariant under rotations inside $\mathrm{span}\{b_k : k \in S\}$: only the separation that defines the block as a spectral set is used. Specializing to $S = \{k, k+1\}$ gives the two-dimensional statement, and to $S = \{1, \dots, \hat{r}\}$ the leading-subspace statement, in which $\delta_S = \sigma_{\hat{r}}^2 - \sigma_{\hat{r}+1}^2$ since the upper end of the interval is unconstrained.

E.2 Approximate separability in the NMF route

Definition 3 (α-approximate separability)**.** *A nonnegative factor matrix $F \in \mathbb{R}^{n\times d}_{\ge 0}$ with rows $f(u_i)$ is α-separable if for every latent factor k there is a row $i(k)$ with*

$$F_{i(k),k} \ge 1, \qquad \sum_{\ell \ne k} F_{i(k),\ell} \le \alpha, \tag{74}$$

i.e., a near-anchor row dominated by coordinate k with residual mass at most α. The case $\alpha = 0$ is exact separability (Donoho & Stodden, 2003; Arora et al., 2012).

Theorem 12 (Degradation of approximately separable NMF)**.** *Let $M = FG^\top$ be a population nonnegative rank-d factorization in which F is α-separable and the true factor cone is well conditioned, with the smallest angle between its extreme rays at least $\theta_0 > 0$. Then every nonnegative optimal factorization $(\hat{F}, \hat{G})$ satisfies, in a suitable permutation and scaling gauge,*

$$\min_{P\in\mathrm{Perm},\ D\in\mathrm{D}_+} \|\hat{F}\, P\, D - F\| \le \frac{C\,\alpha}{\sin\theta_0}\,\|F\|, \tag{75}$$

so the identification error is linear in the separability deficit α and amplified by the cone condition number $1/\sin\theta_0$. As $\alpha \to 0$ the bound recovers the exact uniqueness of Theorem 5.

Proof sketch. Identification in separable NMF is recovery of the extreme rays of the latent cone from the data points: under exact separability the anchor rows are exactly the extreme rays, and the constructive algorithm of Arora et al. (2012) identifies them uniquely up to permutation and scaling. Under α-approximate separability the near-anchor rows deviate from the true extreme rays by at most α, and by the stability of vertex recovery in convex cones, a perturbation of the vertex set by α moves the recovered rays by $O(\alpha/\sin\theta_0)$ when the adjacent rays are separated by angle at least θ_0. This is the robust version of separable NMF (Arora et al., 2012) following the noise-robustness analysis of Gillis and Vavasis. The two measurable quantities α and θ_0 therefore provide an identifiability budget for the NMF route, which is what Section 5 reports for its nonnegativity condition. □

E.3 Coupling identification with estimation

Proposition 5 (Identification and estimation in one ledger)**.** *Let $\hat{F}$ be the empirical risk minimizer of Corollary 4 with encoder capacity fixed and whitening applied on the empirical moments. Then the total identification error of the k-th mode decomposes as*

$$\|\hat{g}_k - (\pm b_k)\| \le O\Big(\frac{\sqrt{\|\hat{F} - F\|_{L^2}^2}}{\Delta}\Big) + O\Big(\frac{r}{\Delta}\Big), \tag{76}$$

where the first term converts the excess risk of the estimate into subspace error via Davis-Kahan and the second is the empirical whitening error of Theorem 11. Both terms scale as $1/\Delta$ and decay as $n^{-1/2}$.

Proof. Split the error around the population whitened solution by the triangle inequality. The second piece vanishes: at the population optimum the whitening theorem (Theorem 2) identifies the modes exactly. The first piece is the deviation of the estimate: $\|\hat{F} - F_d\|_{L^2}^2 \le \|\hat{F} - F\|_{L^2}^2$ by optimality, and the Davis-Kahan step of Theorem 11 converts an L^2 deviation of the represented function into a singular-vector deviation divided by the gap, giving the first term. □

Proposition 6 (Mode error ledger)**.** *Let $\hat{F}$ be the empirical risk minimizer of Corollary 4 and let $\hat{f}, \hat{g}$ be its whitened encoders. The error in recovering the k-th interaction mode satisfies*

$$\max\{\|\hat{f}_k - f_k\|,\ \|\hat{g}_k - g_k\|\} \ = \ O\left(\frac{\sqrt{\|\hat{F} - F\|_{L^2}^2} + r}{\Delta}\right), \tag{77}$$

with $r = O\big(B^2\sqrt{\log(d/\delta)/n}\big)$, both terms being inversely proportional to the gap and decaying as $n^{-1/2}$.

Proof of Proposition 6 (Mode error ledger). Substitute the excess-risk bound of Corollary 4 for $\|\hat{F} - F\|_{L^2}^2$ into the first term of Proposition 5, and the moment radius $r = O\big(B^2\sqrt{\log(d/\delta)/n}\big)$ of Lemma 4 into the second. The result is exactly (77):

$$\max\{\|\hat{f}_k - f_k\|,\ \|\hat{g}_k - g_k\|\} = O\Big(\frac{\sqrt{\|\hat{F} - F\|_{L^2}^2} + r}{\Delta}\Big). \tag{78}$$

Both terms are inversely proportional to the gap Δ and decay as $n^{-1/2}$, which is the statement that identification and estimation are governed by the same spectral gap.

F Experimental Details and Additional Results

This appendix records the complete setups of the experiments of Section 5 and the additional results that the main text summarizes. All numbers are produced by the scripts in the companion repository, whose data files are grouped per study. We state the metrics alongside the tables, because a recurring theme is that gauge-invariant diagnostics cannot separate normalizations while gauge-dependent ones can.

F.1 Controlled synthetic study

Targets and training. We fix $F(u, v) = \sum_{k \le d} \sigma_k\, a_k(u) b_k(v)$ with a_k, b_k shifted Legendre polynomials, orthonormal on $[0, 1]$, and a prescribed spectrum $\{\sigma_k\}$ chosen distinct and separated. Encoders are linear maps over a fixed Legendre feature map, $f(u) = U^\top \varphi(u)$ with $U \in \mathbb{R}^{p \times d}$, so the encoder realization error of Assumption 1 is zero and the normalization is the only variable across runs. Points u, v are drawn on a common grid of n sites per marginal. Training uses Adam with global-norm gradient clipping. The bare bilinear objective is nonconvex and stiff, and plain gradient descent diverges. Analytic gradients are checked against finite differences to about 10^{-9}. Every scheme runs four seeds.

The seven normalization schemes. The main text reports the headline of this study. Table 1 in Section 5.2 gives the full panel, shown graphically in Figure 2. Fit MSE is the squared regression error. Alignment is the mean over modes of $|\langle \hat{f}_k, a_k \rangle|$ after optimally matching permutation and sign, reported for the two encoders. Cond is the gauge-dependent diagnostic $\max(\operatorname{cond} \Sigma_f, \operatorname{cond} \Sigma_g)$. Cross-seed is the gauge distance between solutions from different seeds, the mean over pairs of the Procrustes-optimal embedding distance. Only whitening reaches alignment 1.000 with cross-seed distance 0.000, matching Theorem 2. Every looser gauge stalls at alignment 0.53 to 0.65 with cross-seed drift 0.28 to 0.35, the residual O of cosine and the full GL_d of none leaving the modes rotationally entangled (Theorem 4). Whitening drives Cond to its floor $\sigma_1^2/\sigma_d^2 \approx 9.5$ (Corollary 3). The two-sided nonnegative gauge stays at Fit MSE 4.25: positivity forces $\langle f, g \rangle \ge 0$, the sign-changing target is out of class (Theorem 5), and identifiability is bought at the cost of excluding negative interactions. One measurement subtlety, exploited throughout: $\operatorname{cond}(\Sigma_f \Sigma_g)$ is gauge invariant, its eigenvalues equal the σ_k^2 by Lemma 2, so it cannot separate normalizations and is not reported as a diagnostic.

The gap collapse. The gap sweeps use the finite-sample whitening estimator of Appendix E directly, empirical second-moment whitening followed by eigendecomposition, rather than a trained optimizer: with Legendre features the target is represented exactly, and the training identification error has a

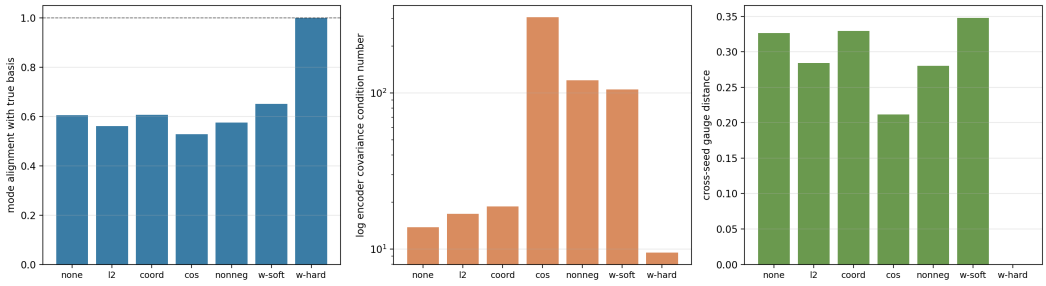


Figure 2: The seven normalization schemes of Table 1, means over four seeds. Left: mode alignment with the true basis (dashed line: 1). Middle: the conditioning diagnostic $\max(\operatorname{cond}\Sigma_f, \operatorname{cond}\Sigma_g)$ on a log scale, whose floor is $\sigma_1^2/\sigma_d^2 \approx 9.5$. Right: cross-seed gauge distance (the short tick marks the exact zero of hard whitening). Abbreviations: l2 = L2 normalization, coord = per-coordinate standardization, cos = cosine normalization, nonneg = two-sided nonnegativity, w-soft = soft whitening, w-hard = hard whitening.

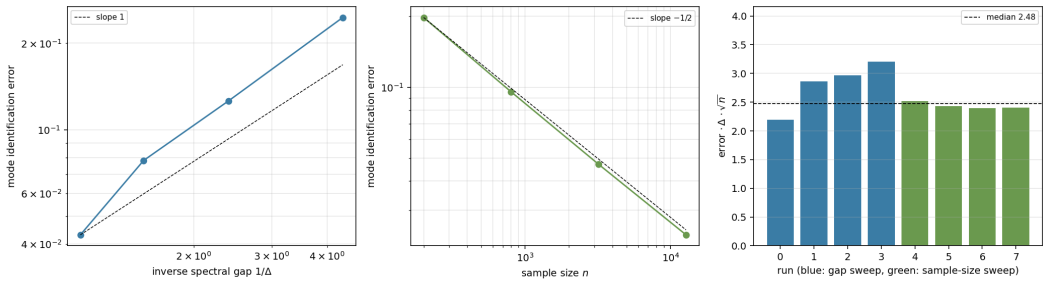


Figure 3: The gap collapse. Mode identification error of the finite-sample whitening estimator: against the inverse spectral gap $1/\Delta$ at fixed n (left) and against sample size n at fixed Δ (center). Right: the product error $\cdot\,\Delta\cdot\sqrt{n}$ per run across both sweeps.

sharp threshold around $n \approx 300$ that would hide the $1/\Delta$ law, whereas the perturbation mechanism of the estimator exposes it cleanly. Table 2 in Section 5.2 reports the two sweeps. The eight points collapse onto a single constant, $\mathrm{err}\cdot\Delta\cdot\sqrt{n} \approx 2.6$, mean 2.62 with coefficient of variation 0.12 (median 2.48, range 2.19 to 3.20): gap and sample size enter the identification error only through the product, the concrete face of Corollaries 6 and 3. Within the spread the sample-size sweep is flat, 2.39 to 2.52, and the drift sits in the gap sweep, 2.19 at the widest gap to 3.20 at the narrowest, so the constant is mildly gap dependent in the direction the first-order perturbation argument predicts. Figure 3 shows the collapse.

The sample-complexity rate law. The linear setting is rank-d matrix sensing, $y_i = u_i^\top \Theta^\star v_i + \varepsilon_i$ with $\operatorname{rank}\Theta^\star = d$, and Theorem 7 predicts $\mathbb{E}\|\hat{\Theta} - \Theta^\star\|_F^2 \asymp \sigma_\varepsilon^2 d(p+q)/n$. We sweep each variable with the others fixed and fit log-log slopes: n at $p = q = 4$, $d = 2$ gives -1.03 (theory -1); d at $p = q = 16$, $n = 2048$ gives $+0.89$ (theory $+1$); $p + q$ at $n = 10^4$, $d = 4$ gives $+1.24$ (theory $+1$). The excess of the last slope is accounted for by the Marchenko-Pastur finite-sample correction of the Gaussian design: measured ratios to $\sigma_\varepsilon^2 d(2p-d)/n$ are $1.00, 1.05, 0.98, 1.09, 1.13$ across $p = q = 8, \ldots, 32$, tracking the correction factor $pq/(n-pq)$, which grows from 1.01 to 1.11 over the same range. Gradient descent on the factored pair (U, V), the estimator class of this paper, is warm-started from the truncated least-squares factors and reproduces the least-squares slope, -1.01, on the same grid. The sweep rejects the loose general bound of Remark 6, which scales as $d^{3/2}\sqrt{(p+q)/n}$ and predicts slope $-1/2$ in n: the $1/n$ dependence on samples is real (Figure 1).

The negative results. Tables 3 and 4 in Section 5.2 give the two regimes. Figure 4 shows both graphically. On the boolean equality kernel $F = \mathbf{1}[u = v]$ on $\{\pm 1\}^m$, the embedding rank needed for 10% error tracks $d \approx 0.9 \cdot 2^m$, exponential in m, while the early-interaction threshold on $\langle u, v\rangle$

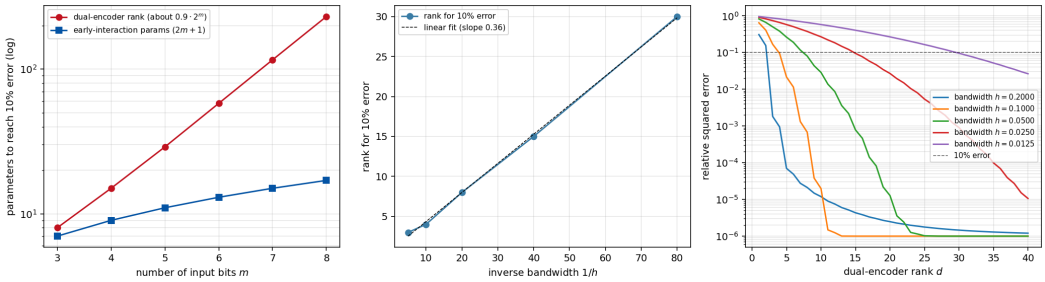


Figure 4: The negative results. Left: on the boolean equality kernel, parameters needed to reach 10% relative error against the number of input bits m, for the dual-encoder head (rank d) and the early-interaction threshold. Center: for the narrowband periodic-Gaussian kernel, the dual-encoder rank needed for 10% error against the inverse bandwidth $1/h$, with a linear fit. Right: relative squared error of the dual-encoder head against its rank d, one curve per bandwidth h, with the 10% line marked.

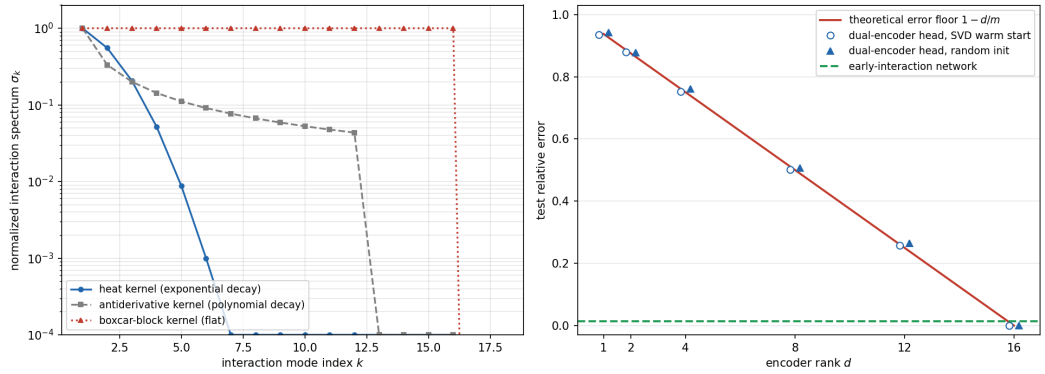


Figure 5: The flat-spectrum error floor on a trained DeepONet head. Left: normalized interaction spectra σ_k for three operators, from exponentially decaying (heat) to polynomially decaying (antiderivative) to exactly flat (boxcar block operator). Right: test relative error of a rank-d dual-encoder head versus d, from an SVD warm start and from random initialization, against the predicted floor $1 - d/m$. The dashed line is an early-interaction MLP baseline.

represents F exactly with $2m + 1$ parameters (Theorems 8 and 9). On the narrowband periodic-Gaussian kernel $F_h(u, v) = \rho_h(u - v)$ of bandwidth h on the circle, the required rank d_ε grows as $\Theta(1/h)$ as the kernel approaches the diagonal (Remark 7).

F.2 DeepONet

This study carries the identifiability results of Section 3 onto a real dual-encoder architecture. We train branch-trunk DeepONet heads end to end on four operators and ask three questions, one per paragraph below: whether the rank-selection error curve tracks the decay of the operator spectrum (Corollary 5); whether post-hoc whitening, applied after training without touching the loss, lifts the learned trunk basis to the analytic operator eigenmodes and collapses the cross-seed gauge distance (Theorem 2); and whether the whitened gauge improves out-of-distribution error (Corollary 3). Two operators, the heat semigroup and the Volterra antiderivative, have closed-form spectra and provide exact ground truth for the recovered basis; two, viscous Burgers and Darcy flow, have only empirical spectral proxies and are used for the qualitative claims. The main-text Section 5 reports the conclusions; the tables and figures below give the full numbers.

Architecture and training. The branch and trunk networks are MLPs of width 128 and depth 3, outputting embeddings of dimension $d = 32$. The rank-selection sweep of Table 7 varies d from 2

Table 7: Rank selection, test relative L^2 error versus trunk width d, two seeds per width.

Operator	$d=2$	$d=4$	$d=8$	$d=16$	$d=32$	$d=64$	Drop $4 \to 8$
heat	.070	.056	.053	.052	.055	.056	-5%
antiderivative	.106	.081	.056	.055	.070	.060	-31%
burgers	.237	.135	.134	.133	.133	.136	flat
darcy	.167	.150	.149	.153	.158	.169	flat

Table 8: Basis recovery per gauge, five seeds. Alignment: mean mode alignment against the analytic left singular functions, raw and after post-hoc whitening. Cross-seed: mean pairwise gauge distance, raw and after whitening.

Operator / gauge	Align	Whitened align	Seed dist	Whitened dist
heat / none	0.632	0.915	0.442	0.039
heat / cosine	0.529	0.799	0.313	0.007
heat / whiten_soft	0.517	0.780	0.479	0.140
antiderivative / none	0.582	0.940	0.282	0.064
antiderivative / cosine	0.576	0.823	0.245	0.011
antiderivative / whiten_soft	0.587	0.837	0.414	0.018

to 64. Training runs 300 epochs with Adam at learning rate 10^{-3} and cosine annealing, batch size 256, and five seeds per configuration for the gauge comparison; the rank-selection sweep, which is 6 widths per operator, runs two seeds per width. Inputs are Gaussian random fields. The four operators are the heat semigroup $e^{t\Delta}$ (analytic singular values $\sigma_k = e^{-(k\pi)^2 t}$, exponential), the antiderivative or Volterra operator ($\sigma_k = 1/((k-\frac{1}{2})\pi)$, polynomial), and viscous Burgers and Darcy flow, whose spectra are empirical proxies, labeled as such. Whitening is realized in two stages: the stable soft-whitening penalty during training, the relaxation of Theorem 6 and closely related to the Barlow Twins objective, and hard whitening post hoc on the active rank-r subspace. The two-stage design is forced by the data: the interaction spectra decay fast, leaving an effective rank of 3 to 8 against an embedding width of 32, and in-forward hard whitening inverts about 25 near-zero covariance eigenvalues and backpropagates through a degenerate eigendecomposition, exactly the small-gap failure predicted by the block-stability analysis of Corollary 7 in Appendix E. Because whitening leaves $\langle b, t\rangle$ unchanged, applying it after training is exact, and projecting to the active subspace first avoids inverting dead directions. The empirical interaction spectrum is estimated as the square roots of the eigenvalues of $\Sigma_b^{1/2}\Sigma_t\Sigma_b^{1/2}$, computed with $\mathrm{eigvalsh}$. This avoids the numerically unstable SVD of the ill-conditioned product $\Sigma_b^{1/2}\Sigma_t^{1/2}$ and agrees with the direct computation to 10^{-9} on well-conditioned instances.

Spectra and rank selection. The normalized empirical spectra reproduce the dichotomy qualitatively, heat $1, 0.22, 0.028, \ldots$ and antiderivative $1, 0.14, 0.042, \ldots$. Because the network allocates no capacity to near-null modes, the learned spectra decay faster than the analytic ones, so we claim recovery of the dominant modes and of the decay type, not a mode-by-mode match. Table 7 gives the rank sweep. The curves are flat past $d \approx 8$ and the argmin is dominated by noise, so we do not report a single optimal rank. The honest quantity is the drop from $d = 4$ to $d = 8$, 5% for heat (early saturation, small optimal rank) versus 31% for antiderivative (more modes needed), consistent with Corollary 5. This is qualitative support, and the figure is labeled accordingly.

Basis recovery. Table 8 gives the full per-gauge numbers behind Figure 6. Alignment is the mean over modes of $|\langle \hat{t}_k, \beta_k\rangle|$ after optimal permutation and sign matching against the analytic left singular functions, before and after the post-hoc whitening gauge fix; cross-seed is the mean pairwise gauge distance between seeds. Whitening lifts the loose gauges from 0.5 to 0.6 to 0.78 to 0.94 and collapses the cross-seed distance by roughly an order of magnitude. For heat the recovered modes are the analytic Fourier basis $\sqrt{2}\sin(k\pi y)$, a hard ground truth. On the nonlinear benchmarks, where no analytic basis exists, whitening still tightens the cross-seed distance for Burgers, $0.30 \to 0.08$; Darcy is the weakest case, $0.56 \to 0.33$, and its proxy spectrum is the noisiest, which the figure notes.

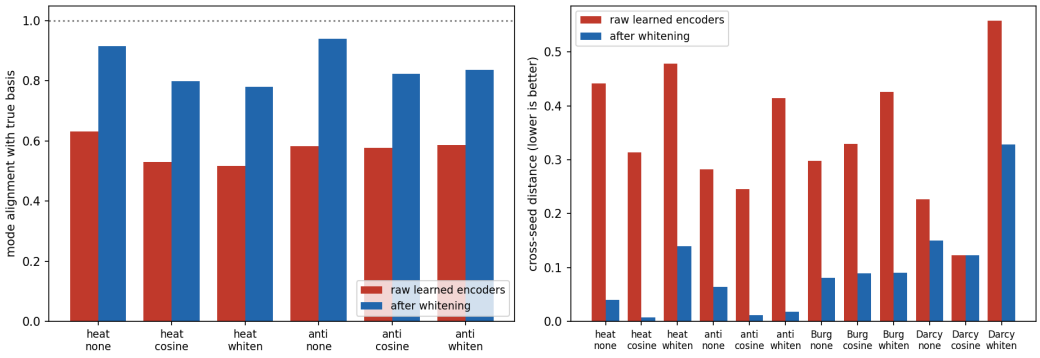


Figure 6: Trunk-encoder basis quality before and after post-hoc whitening. Each group on the horizontal axis corresponds to one operator (heat, anti = antiderivative, Burg = Burgers, Darcy) trained under one gauge constraint (none, cosine, or whiten = soft whitening). Left: alignment of the learned trunk basis with the analytic operator eigenmodes (1 is exact). Right: distance between bases learned from different random seeds (lower indicates greater stability).

Table 9: The payoff of gauge-fixing, five seeds: relative L^2 error in distribution and out of distribution, none versus soft whitening.

Operator	none, in-dist.	whiten, in-dist.	none, OOD	whiten, OOD	OOD gain
heat	0.051	0.019	0.038	0.016	2.4×
antiderivative	0.057	0.039	0.046	0.031	1.5×
burgers	0.122	0.112	0.113	0.103	1.1×
darcy	0.151	0.137	0.142	0.120	1.2×

The payoff of gauge-fixing. Table 9 gives the in-distribution and out-of-distribution relative L^2 errors of the none and soft-whitening gauges. The out-of-distribution test set draws inputs from a different smoothness regime of the same operator family. Soft whitening attains the lowest error in all eight cells, with the out-of-distribution gain largest, $2.4\times$ for heat. Two caveats are recorded with the table. First, soft whitening does not reduce seed-to-seed error variance. If anything it increases it, heat standard deviation $0.0005 \to 0.0036$ and Darcy $0.0017 \to 0.0181$, so the stability it confers is in the identified basis of Table 8, not in the error. Second, the soft-whitening training loss carries the decorrelation penalty and its curve is not directly comparable to the unpenalized gauges, so we make no claim about convergence speed.

The flat-spectrum anchor. The block-boxcar operator $G(a)(y) = c_{j(y)}$, the output is the block mean of the input field at the block containing y, with $m = 16$ blocks, has an exactly flat interaction spectrum: sixteen equal singular values, the seventeenth at 10^{-16}. Table 10 gives the trained relative error of the dual-encoder head with linear encoders, from both the SVD warm start and random initialization, against the floor $1 - d/m$ of Theorem 8. The worst deviation is 0.014, the finite-sample excess, and gradient descent reaches the floor from scratch. The early-interaction comparator, a ReLU MLP on $\mathrm{concat}(a, \mathrm{onehot}(y))$ with 12.4k parameters, four times the budget of the $d = 16$ head, reaches 0.013 ± 0.001 on the same test set (Theorem 9). One caveat belongs in the setup: with only 8k training samples the comparator memorizes, training MSE 4×10^{-5} and test error 1.7. At 60k samples, where parameters are far fewer than samples, memorization is impossible and the escape shown in Figure 5 is an honest generalization result. The comparison deliberately pits the floor against a well-trained early model.

F.3 CLIP

This study tests the gauge view on heterogeneous modalities and on pretrained models that never saw each other. It proceeds in two stages. The white-box synthetic kernel supplies exact ground truth for the mechanism: under linear encoders and white inputs the interaction spectrum equals the

Table 10: The flat-spectrum anchor: trained relative error of a linear-encoder head on the block-boxcar operator, versus the floor $1 - d/m$.

d	Trained error	Floor $1 - d/m$
1	0.936	0.938
2	0.879	0.875
4	0.752	0.750
8	0.502	0.500
12	0.258	0.250
16	0.000	0.000

Table 11: Synthetic multimodal kernel, ground truth known.

Quantity	Value
Spectrum invariant across gauges	3×10^{-16}
Spectrum equals true ρ	4×10^{-16}
Per-dimension probe transfer	0.43 (best axis 0.65)
Text mode recovery, raw	0.554
Text mode recovery, cosine (isotropization)	0.581
Text mode recovery, whitening plus gap	1.000
Cross-encoder frame stability, raw $\rightarrow$ whitened	$0.684 \rightarrow 1.000$

singular values of $W_f^\top W_g$, which is gauge invariant, and two independently trained encoders are two random gauges of the same target. On this kernel we verify that the spectrum is invariant and equals the truth, that per-dimension concept probes do not transfer, and that whitening plus gap recovers the text modes exactly. The real CLIP pairs then answer whether the same statements hold for contrastive models trained in the wild: is the gauge-invariant spectral summary shared across independently pretrained models, are per-dimension probes not, and is the residual relationship a single rotation that whitening removes, exposing interpretable concept axes (Theorem 4)? The main-text Section 5 reports the conclusions; the tables and figures below give the full numbers.

The synthetic multimodal kernel. The white-box counterpart is $F^\star(u, v) = \sum_k \rho_k \, (p_k^\top u)(q_k^\top v)$ with $u \sim \mathcal{N}(0, I_p)$, $v \sim \mathcal{N}(0, I_q)$, orthogonal canonical directions P, Q, and distinct, gapped canonical correlations $\rho = (0.9, 0.75, 0.6, 0.45, 0.3, 0.15)$. The key identity: under linear encoders and white inputs the interaction spectrum equals the singular values of $B = W_f^\top W_g$, which is explicitly gauge invariant, $f \rightarrow Af$, $g \rightarrow A^{-\top} g$ leaves $W_f^\top W_g$ unchanged, and two independently trained encoders are two random gauges A_1, A_2. Table 11 collects the numbers: the spectrum is invariant across gauges to 3×10^{-16} and equals the true ρ to 4×10^{-16}; per-dimension concept probes transfer at 0.43, the best-matched axis reaching only 0.65; whitening plus gap recovers the true text modes at $|\cos| = 1.000$, against 0.554 raw and 0.581 under cosine, which is stuck at the residual O; and cross-encoder frame stability rises from 0.684 to 1.000. We deliberately do not report an alignment of the canonical subspaces here. A gauge acts by an invertible map on the encoder outputs, so it leaves the row space of the encoder untouched, and any subspace-level alignment between two gauges of the same target is an algebraic identity rather than a measurement. The informative quantities are the coordinatewise ones above, where the gauge does act. The distance between the 0.58 of cosine and the 1.00 of whitening is, numerically, the contribution of whitening beyond the decorrelation that self-supervised methods already perform, precisely the quantity Theorem 2 isolates.

Real CLIP. Two backbone pairs, each two independent gauges of one similarity function: ViT-B/32 (OpenAI) versus ViT-B/32 (LAION-2B), the same architecture with different pretraining data, and ViT-B/32 (OpenAI) versus RN50 (OpenAI), different architectures. Image features come from CIFAR-100 test images, 49 fine classes spanning animals, vehicles, plants, and household objects, 64 images per class, text features are prompt-ensembled class embeddings plus attribute contrast directions, and both towers are projected to a common 32-dimensional working space, the image PCA top-d directions. Whitening is realized via canonical correlation analysis with shrinkage covariance, and the relation of this estimator to Theorem 2 deserves a precise statement. Canonical correlation analysis whitens the two towers separately, so it is invariant under the larger group $(f, g) \mapsto (Af, Bg)$

Table 12: Canonical correlation spectra across independently pretrained CLIP models, the gauge-invariant summary described in the text. Pearson correlation and normalized L1 difference of the two spectra, with per-dimension probe transfer raw and after fitting the single residual O rotation.

Backbone pair	Spectrum Pearson	Norm. L1
ViT-B/32 (OpenAI) vs ViT-B/32 (LAION)	0.99	0.065
ViT-B/32 (OpenAI) vs RN50 (OpenAI)	0.98	0.096

Table 13: Group-signal concentration on the reference backbone: R^2 of regressing true group labels on the top-6 mode activations, whitened canonical modes versus cosine principal coordinates. All three group splits we tested are listed, including the one on which whitening does not help. Cross-backbone frame stability is reported in the text.

Group split	Whitened	Cosine
animate / rest	0.92	0.88
natural / manmade	0.90	0.48
large / small	0.68	0.72

with A and B invertible independently, whereas the gauge group of Definition 2 is the coupled subgroup $B = A^{-\top}$. Invariance under the larger group implies invariance under the subgroup, so the canonical frame is a valid section of the gauge orbit and the comparisons reported below are gauge-invariant comparisons in the sense of Section 3. The price is that the canonical correlations returned by this estimator are not the interaction spectrum: the larger group also quotients out the relative scaling between the two towers, which is precisely the information the interaction spectrum records. The two spectra coincide in the synthetic kernel above, where both towers carry $\sqrt{\rho_k}$ over orthonormal canonical directions and the inputs are white, which is why that study can be checked against exact ground truth. No such normalization holds for pretrained CLIP towers, so the quantity reported in this paragraph is the canonical correlation spectrum, and what it establishes is the shared gauge-invariant structure of the pair, not a numerical estimate of $\{\sigma_k\}$. The direct estimator of $\{\sigma_k\}$, the square roots of the eigenvalues of $\Sigma_f^{1/2}\Sigma_g\Sigma_f^{1/2}$, is the one used in Appendix E and in the DeepONet study.

Table 12 and Figure 7 give the canonical correlation spectra across backbone pairs: near-identical spectra, agreement best at the top. Per-dimension probe transfer is 0.28 without alignment and 0.92 after fitting a single global orthogonal map: the two embeddings are related by one O rotation, the residual group Theorem 4 predicts for the cosine gauge, and no coordinatewise repair exists short of discovering that rotation. Table 13 gives the group-signal regression of true labels on the top-6 mode activations: the whitened span contains the animate and natural axes, the cosine span contains only the animate axis, and per-mode statistics show why, whitening concentrates the animate signal in a single mode ($|t| = 3.1$) while cosine spreads it across several. The third split is a counterexample and we report it as one: on large versus small, whitening is slightly worse, 0.68 against 0.72. Object size in CIFAR-100 is confounded by cropping and is the weakest of the three anchors, so we do not read the reversal as evidence against the mechanism, but neither do we claim a uniform gain. Figure 8 shows the canonical text modes as concept axes: mode 1 separates animals (leopard, bear, wolf, camel) from flowers and fruit, mode 2 is the vehicle axis (bus, streetcar, pickup truck, train), mode 3 separates plants from household objects, and mode 5 the food axis (orange, apple, pear, sweet pepper). The cosine principal coordinates are semantic mixtures by comparison. Cross-backbone frame stability, matching modes by their backbone-independent class-signature profiles, favors whitening in both pairs, 0.77 versus 0.72 for the same-architecture pair and 0.84 versus 0.62 across architectures.

What whitening buys for CLIP in practice. None of the steps above requires retraining: whitening is a post-hoc second-moment transform, it leaves the inner-product scores unchanged, and at inference time it costs one fixed linear map. It can therefore be applied as an analysis layer on any deployed CLIP model, and the results above suggest three concrete payoffs. First, model alignment: because independently pretrained pairs are related by a single residual rotation, whitening both models into the

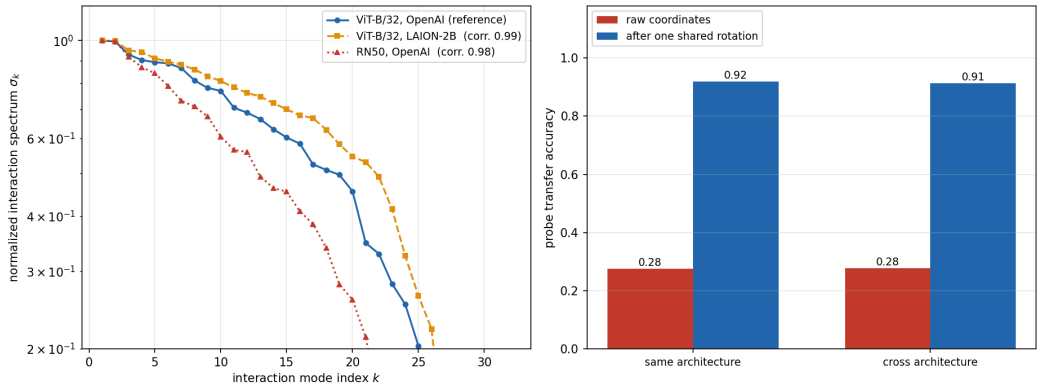


Figure 7: Shared structure across independently trained CLIP models. Left: normalized canonical correlation spectra for three CLIP encoders (a ViT-B/32 reference against a LAION-trained ViT-B/32 and an RN50), with Pearson correlations to the reference in the legend. Right: accuracy of a linear concept probe trained on one model and applied to the other, using raw coordinates and after fitting a single shared rotation, for same-architecture and cross-architecture pairs.

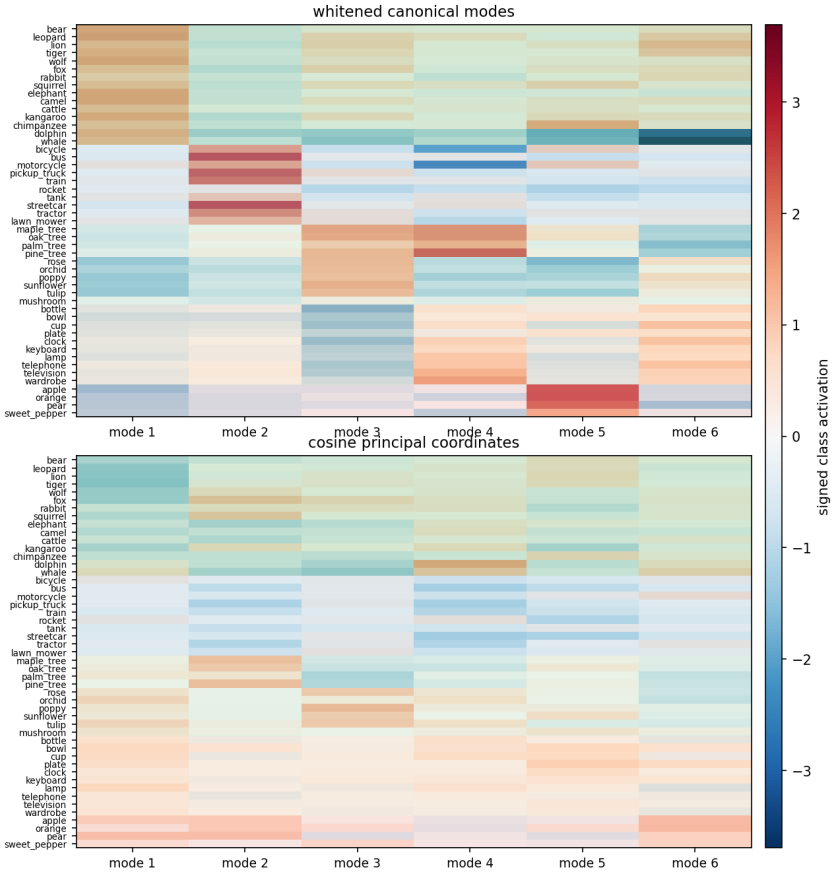


Figure 8: Whitened canonical text modes (rows) over the 49 CIFAR-100 classes (columns) of the reference backbone: modes read as concept axes (animals, vehicles, plants, food), while the cosine principal coordinates are semantic mixtures.

shared canonical frame makes their coordinates comparable dimension by dimension, a prerequisite

for model merging, distillation, and cross-model probing that raw cosine coordinates do not offer. Second, interpretability with known limits: the whitened axes read as concept axes (animals, vehicles, plants, and food in the CIFAR-100 probe), and the spectral gap quantifies in advance how much interpretability any basis can carry, so an analyst knows where the axes are trustworthy and where the small-gap failure of Appendix E leaves them partially mixed. Third, a spectral summary of the model itself: the interaction spectrum is a gauge-invariant description of what a CLIP pair can discriminate, and its decay predicts the effective interaction rank (Corollary 5) and hence how many concept axes are worth reading out at all. A natural next step is to use the whitened canonical frame as the initialization or constraint for fine-tuning, so that new axes are learned in coordinates that are already interpretable and transferable across models.

Summary. The synthetic study establishes the three claims where ground truth is exact: whitening uniquely fixes the gauge, the gap controls identification through the product $\Delta\sqrt{n}$ (collapse constant ≈ 2.6), and flat spectra force exponential late-interaction cost. DeepONet carries identifiability onto a real architecture, lifting basis alignment from 0.5 to 0.6 up to 0.78 to 0.94 against the analytic eigenbasis, with an out-of-distribution payoff up to $2.4\times$. CLIP extends the claim to heterogeneous modalities, the gauge-invariant spectral summary shared across independently pretrained models while per-dimension probes are not. On a flat spectrum the head is floored at $1 - d/m$ while an early-interaction model reaches 0.013: the interaction spectrum decides whether a multiplicative dual-encoder head is the right tool. On a flat spectrum the head is floored at $1 - d/m$ while an early-interaction model reaches 0.013: the interaction spectrum decides whether a multiplicative dual-encoder head is the right tool.

G INDEX OF RESULTS

Table 14 collects the paper's numbered results with a one-line statement of what each establishes and where it is stated or proved. The results are grouped by the four questions the paper answers, with the finite-sample lemmas of Appendix E last. Within each group the numbering follows the order of appearance, and the table doubles as a reading map for the appendices. Remarks and the technical lemmas internal to proofs are not indexed separately. They are cited where they are used.

Table 14: Index of the paper's numbered results.

Result	What it establishes	Where
Unification and approximation (Section 2)		
Definition 1	Interaction spectrum, rank, and modes as the singular data of the target	Section 2
Theorem 1	Approximation error decomposes into a truncation term and an encoder-realization term	Section 2
Assumption 1	Encoder classes contain approximations of the scaled interaction modes	Appendix B
Proposition 1	Evaluating a rank-d head on all pairs costs $O((n+m)\,C_{\mathrm{net}} + nm\,d)$	Appendix B
Lemma 1	Representability of F by a rank-d head iff $\mathrm{rank}(T_F) \le d$	Appendix B
Corollary 1	The truncation term is the best possible error of any rank-d head	Appendix B
Theorem 3	Target smoothness controls the decay of the truncation tail	Appendix B
Proposition 2	The dual head escapes the curse of the joint input dimension	Appendix B
Identifiability as gauge-fixing (Section 3)		
Definition 2	Gauge action of GL_d on encoder pairs leaves the represented function unchanged	Section 3
Theorem 2	Under a spectral gap, whitening identifies the modes up to permutation and sign	Section 3
Proposition 3	Per-coordinate standardization pins the coordinate scales but no more	Appendix C
Lemma 2	The eigenvalues of $\Sigma_f\Sigma_g$ are gauge invariants, equal to the σ_k^2 at any optimum	Appendix C

Index of results (continued)

Result	What it establishes	Where
Proposition 4	Unfixed gauge renders the risk flat along the orbit and optimization ill-posed	Appendix C
Theorem 4	Cosine normalization leaves the full orthogonal group as residual gauge	Appendix C
Corollary 2	Individual coordinates of a cosine-normalized embedding carry no intrinsic meaning	Appendix C
Theorem 5	Two-sided nonnegativity gives NMF-style identifiability, at the cost of excluding negative modes	Appendix C
Corollary 3	The spectral gap controls conditioning, identifiability, and convergence simultaneously	Appendix C
Theorem 6	A quadratic whitening penalty inherits the identifiability of the hard constraint	Appendix C
Estimation and rank selection (Section 4)		
Lemma 3	The Rademacher complexity of the inner-product class collapses to the sum of the encoder complexities	Appendix D
Corollary 4	Excess risk of empirical risk minimization on the head class	Appendix D
Theorem 7	Linear heads are rank-d matrix sensing with minimax rate $\sigma_\varepsilon^2 d(p+q)/n$	Appendix D
Corollary 5	Optimal rank grows logarithmically in the sample budget for smooth targets	Appendix D
The usability criterion (Section 4)		
Theorem 8	A flat spectrum forces every rank-d head to relative error $1 - d/N$	Appendix D
Theorem 9	Early-interaction models escape the flat-spectrum floor	Appendix D
Theorem 10	Usability criterion: decide by the measured spectrum whether the head can succeed	Appendix D
Finite-sample identification (Appendix E)		
Lemma 4	Matrix Bernstein concentration for empirical second moments	Appendix E
Lemma 5	Stability of the empirical whitening map under moment perturbation	Appendix E
Theorem 11	Finite-sample mode identification with error of order r/Δ	Appendix E
Corollary 6	Sample complexity of the whitening estimator: $n \gtrsim B^4 \log(d/\delta)/(\Delta^2 \varepsilon^2)$	Appendix E
Corollary 7	Block stability of mode recovery under small gaps	Appendix E
Definition 3	α-approximate separability as a quantitative relaxation of NMF conditions	Appendix E
Theorem 12	Approximately separable targets degrade identifiability gracefully	Appendix E
Proposition 5	Identification and estimation errors combine into one ledger	Appendix E
Proposition 6	Mode error as the bridge between risk and identified basis	Appendix E