Objective-Specific Privileged Bases via Full-Prefix Matryoshka Learning

Arghamitra Talukder [1] **Philippe Chlenski** [1] **Itsik Pe'er** [1]

Abstract

Learned representations are often invariant to rotational transformations, leaving individual dimensions non-identifiable and interchangeable. We study how Matryoshka Representation Learning (MRL) induces a task-aligned privileged basis distinct from variance-based or regularizer-induced orderings. In the linear setting, we prove that full-prefix MRL recovers the ordered principal directions, and can be computed efficiently using shared statistics. Empirically, we demonstrate that MRL yields consistent per-dimension structure aligned with task signal, where coordinate magnitude reflects informativeness.

1. Introduction

Learned representations of high dimensional data are the backbone of modern machine learning: once trained, lower dimensional embeddings can be reused across retrieval, compression, and other tasks with wildly different computational budgets. A persistent obstacle is non-identifiability where many embedding functions appear equally effective and the representational dimensions are not interpretable. In particular, we focus on this occurring due to insensitivity of loss functions to rotational transformations of the embedded data coordinates (Baldi & Hornik, 1989).

Formally, this is orthogonal symmetry of the d-dimensional embedding space w.r.t. $\mathrm{SO}(d)$ transformations. This symmetry needs to be broken in order to distinguish individual dimensions, and to canonically order coordinate axes. Moreover, we want the distinct dimensions to be meaningful in terms of the training objective, so we extend the definition of a privileged basis from Elhage et al. (2023) to an *objective-specific privileged basis*: a coordinate system whose axes are not just unique and identifiable, but aligned with the structure of the training objective. Axis ordering can arise from regularization or optimizer artifacts (Elhage et al., 2023), and does not by itself constitute a privileged basis. Prescribing a mechanism for breaking the symmetry in a meaningful way is thus a central question and the focus of this manuscript.

Prior work. Prior methods answer this question in two structurally distinct ways. *Nested prefix methods* — nested dropout (Rippel et al., 2014; Bao et al., 2020) and the Prefix-weighted reconstruction loss (Oftadeh et al., 2020) optimize a weighted sum of mean square error (MSE) losses across nested prefixes, recovering ordered eigenvectors of the data covariance Σ. *Regularizer-based methods* — non-uniform ℓ_2 (Kunin et al., 2019; Bao et al., 2020) and variance-ordering (Zhan et al., 2026) impose ordering through monotone penalty schedules on weight norms or latent variances, again anchored to Σ. Both families have been studied primarily to optimize reconstruction loss, so the privileged directions they recover are ultimately the PCA directions of the data.

Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) reframes the prefix reconstruction idea as a practical recipe for jointly training multiscale models with any differentiable loss $\mathcal{L}$, and has since been widely adopted for efficiency-adaptive deployment across retrieval, re-ranking, and vision-language tasks (Xiao et al., 2026; Cai et al., 2025; Hu et al., 2024; Verma et al., 2025). Yet, despite this practical success, the structure of the privileged basis induced by MRL and its connection to PCA-aligned bases remains largely unexplored.

To address this gap, we formalize, prove, and measure the objective-specificity of MRL's privileged basis. Our contributions are:

1. **MRL's privileged basis is objective-specific.** We show that MRL's gradient weighting privileges dimensions in proportion to task utility, rather than reconstruction variance, and is structurally distinct from the sparsity patterns induced by regularizer-based methods.
2. **Full-Prefix MRL recovers PCA for linear autoencoder at no additional asymptotic cost.** For linear autoencoder (LAE), the Full-Prefix MRL loss recovers the ordered eigenvalue spectrum of the data covariance and exact eigenvectors under an orthonormality constraint at the same asymptotic cost as standard re-

[1]Computer Science, Columbia University. Correspondence to: Arghamitra Talukder <at3836@columbia.edu>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

construction training, via shared matrix products. We extend the analysis to Fisher discriminant objective.

2. Theoretical Results

Setup. A summary of all symbols used in this paper is provided in Table 2. Let $X \in \mathbb{R}^{p\times n}$ be zero-centered data with empirical covariance $\Sigma = XX^\top$, where n is the number of samples and p the input dimension. An LAE has encoder $B \in \mathbb{R}^{d\times p}$ and decoder $A \in \mathbb{R}^{p\times d}$ with columns $a_1, \dots, a_d$, where $d \le p$ is the latent dimension. For a nesting set $\mathcal{M} = \{m_1 < \cdots < m_K = d\} \subseteq [d]$, the latent code is $z = Bx \in \mathbb{R}^d$, with $z_{1:m}$ denoting its first m coordinates. We define the rank-1 contribution of dimension k as $y_k = z_k a_k \in \mathbb{R}^p$, so the m-th prefix reconstruction is $\hat{x}^{(m)} = \sum_{k=1}^{m} y_k$.

The standard *LAE loss* is invariant under $(A, B) \mapsto (AT, T^{-1}B)$ for $T \in \mathrm{SO}(d)$ (orthogonal transformations), a subgroup of $\mathrm{GL}(d)$ (invertible linear transformations) (Baldi & Hornik, 1989). It is unordered and assigns equal weight to every dimension k; full expansion is given in A.1.

$$\mathcal{L}_{\mathrm{AE}}(\theta) = \Big\| x - \sum_{k=1}^{d} y_k \Big\|^2, \tag{1}$$

2.1. Objective-specific Privileged Basis

MRL loss. Kusupati et al. (2022) introduce Matryoshka Representation Learning (MRL) loss that jointly trains representations across nested prefix dimensions:

$$\mathcal{L}_{\mathrm{MRL}}(\theta) = \frac{1}{n} \sum_{s=1}^{n} \sum_{m\in\mathcal{M}} \omega_m \, \mathcal{L}\Big(W^{(m)} z_{1:m}^{(s)};\, \tilde{y}_s\Big), \tag{2}$$

where $z_{1:m}^{(s)} \in \mathbb{R}^m$ denotes the first m coordinates of the encoder output for sample s, $\omega_m \ge 0$ are per-scale loss weights (we use $\omega_m = 1$ for all $m \in \mathcal{M}$), and $W^{(m)} \in \mathbb{R}^{L\times m}$ is a linear head for scale m that maps the prefix embedding to the L-dimensional task output (e.g., $L = C$ for classification, $L = p$ for reconstruction), jointly trained with the encoder. We use the weight-tied variant where, $W^{(m)} = W_{1:m}$ are the first m columns of a shared $W \in \mathbb{R}^{L\times d}$. In practice, m is chosen from a geometrically spaced set, e.g. $\mathcal{M} = \{2, 4, 8, \dots, 1024, 2048\}$, yielding $O(\log d)$ nested prefixes; we refer to this spaced variant as *Sparse MRL (S-MRL)*.

Full-prefix MRL loss for LAE.

$$\mathcal{L}_{\text{FP-MRL}}(\theta) = \sum_{m=1}^{d} \omega_m \Big\| x - \sum_{k=1}^{m} y_k \Big\|^2 \tag{3}$$

$$= \sum_{m=1}^{d} \omega_m \|x\|^2 - 2\sum_{k=1}^{d} w_k \, x^\top y_k + \mathbf{y}^\top (S_d^\omega \otimes I_p)\, \mathbf{y}, \tag{4}$$

where $\omega_m > 0$ are per-prefix weights, $w_k = \sum_{m=k}^{d} \omega_m$ is the cumulative weight on dimension k, $\mathbf{y} = [y_1^\top, \dots, y_d^\top]^\top \in \mathbb{R}^{pd}$, and $(S_d^\omega)_{ij} = w_{\max(i,j)}$. Inducing a privileged basis requires only that the cumulative weights be strictly monotonic, $w_1 > w_2 > \cdots > w_d > 0$, which holds for any positive choice of ω_m. A milder schedule such as $\omega_m \propto 1/m$ or $\omega_m \propto 1/(d - m + 1)$ flattens this ratio while preserving the monotonicity. We adopt $\omega_m = 1$ throughout this work which recovers the standard form with $w_k = d - k + 1$, where each dimension is weighted by the number of prefixes it appears in. The full expansion is given in Section A.2.

Non-uniform regularization loss. A natural approach to imposing dimension ordering is to penalize latent dimensions with monotonically varying coefficients, forcing the model to concentrate information in less-penalized coordinates. Two instantiations of this idea bracket the design space.

Inspired by Kunin et al. (2019), Bao et al. (2020) proposes a *Non-uniform ℓ_2 (NU-ℓ_2)* regularization that penalizes the weights connected to each dimension with strictly increasing ℓ_2 coefficients $0 < \lambda_1 \cdots < \lambda_d$. Heavier penalties on late dimensions force the LAE to explain higher-variance directions of X with early dimensions:

$$\mathcal{L}_{\text{NU-}\ell_2}(\theta) = \frac{1}{n}\|X - ABX\|_F^2 + \|\Lambda^{1/2}B\|_F^2 + \|A\Lambda^{1/2}\|_F^2. \tag{5}$$

We introduce as an exploratory baseline the *Monotone-decay ℓ_1 (MD-ℓ_1)* loss, which instead penalizes the embedding mass of each dimension with strictly increasing ℓ_1 coefficients lighter on early dimensions to encourage them to hold more information. To our knowledge, this specific monotone weighting scheme has not been previously studied in this context:

$$\mathcal{L}_{\text{MD-}\ell_1}(\theta) = \frac{1}{n}\|X - ABX\|_F^2 + \frac{\alpha}{n}\sum_{k=1}^{d} k \sum_{s=1}^{n} |b_k^\top x_s|, \tag{6}$$

where $\alpha > 0$ controls the penalty strength and b_k is the k-th row of the encoder B. See appendix A.3 and A.4 for the full scalar derivation with a concrete $p = 5,\ d = 3$ example.

Both losses share the same inductive bias: a monotone coefficient schedule that makes early dimensions cheap and

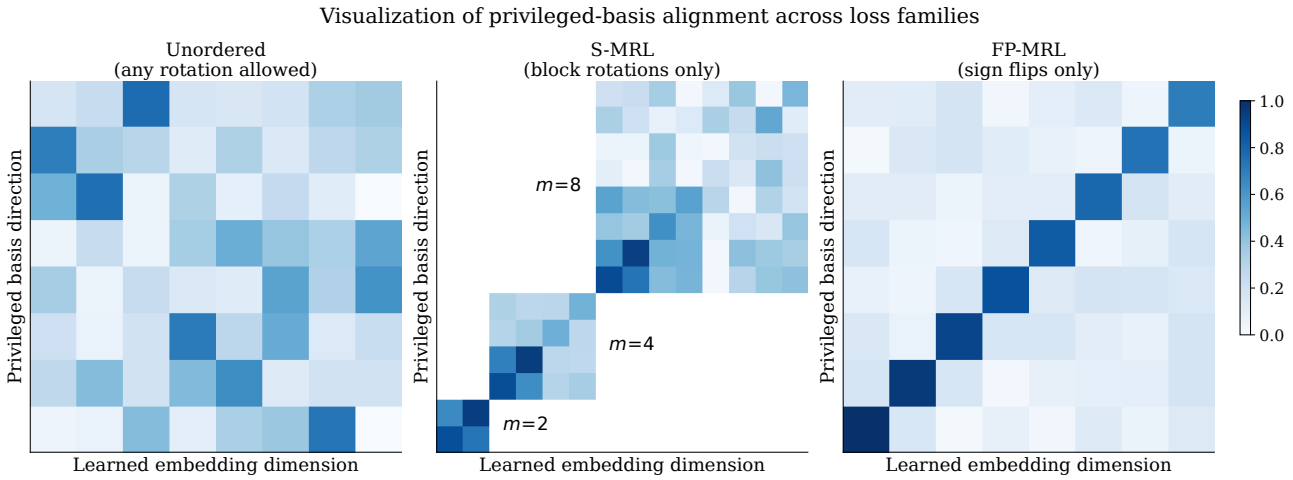


Figure 1. **Privileged-basis alignment visualization across loss families.** Illustrative heatmaps showing the relationship between learned embedding dimensions and a privileged reference basis under different symmetry-breaking mechanisms. Unordered losses (left) are invariant to arbitrary rotations of the latent basis, so any direction can map to any dimension. S-MRL (middle) supervises only a discrete set of prefix sizes (m=2, 4, 8), breaking the rotational symmetry into block-diagonal structure with arbitrary rotations remaining within each block. FP-MRL (right) supervises every prefix length, collapsing the residual symmetry to sign flips and producing a coordinate-wise identifiable basis.

late dimensions costly, pushing information toward low indices. However, the two penalties break rotational symmetry through different mechanisms. Non-uniform ℓ_2 breaks it through the monotone weighting Λ: rotating the latent space mixes rows of B that carry different penalties, so any non-trivial rotation strictly increases the regularizer. MD-ℓ_1 breaks rotational symmetry at the level of the norm itself ℓ_1 privileges the coordinate basis and the monotone coefficient schedule additionally breaks permutation symmetry. Neither consults the task labels y explicitly.

Gradient comparison. The gradient of each loss with respect to y_k reveals how it induces an ordering over the components of y: NU-ℓ_2 scales by $\lambda_k(b_k \otimes a_k)$, ordering through the weights alone; MD-ℓ_1 by $k \cdot \mathrm{sign}(b_k^\top x) \cdot x$, ordering through the input. Neither uses the task signal. Full-prefix MRL by $-2\sum_{m=k}^{d} r^{(m)}$, where $r^{(m)} = x - \hat{x}^{(m)}$ is the prefix-m residual, and the ordering simultaneously acts on the gradient toward x *(ordered linear weights)*, the self-penalty on $\|y_k\|^2$ *(ordered quadratic weights)*, and the coupling between dimensions (*ordered cross-dim. coupling)* — all three shaped by the task loss through the learned encoder-decoder, see equation (13). Unlike regularizer-based losses, which impose ordering on a single object (weights or activations), FP-MRL propagates task signal through all three channels at once. NU-ℓ_2 and MD-ℓ_1 break symmetry by penalizing embedding geometry, tying privileged directions to data statistics or embedding mass. In contrast, FP-MRL accumulates task-gradient signal, inducing a privileged basis that is specific to the training objective— the same architecture trained with a different loss $\mathcal{L}$ learns a different basis.

Figure 1 provides an intuitive visualization of how different symmetry-breaking mechanisms constrain the learned latent basis. For unordered objectives, rotational symmetry remains fully preserved, so privileged directions need not align with any coordinate axis and the learned representation is effectively unidentifiable up to arbitrary rotations. Sparse-prefix MRL (S-MRL) partially breaks this symmetry by supervising only a discrete set of prefix lengths, producing block-wise structure while still permitting arbitrary rotations within each block. In contrast, FP-MRL supervises every prefix length, progressively constraining the latent space until only sign ambiguities remain, yielding a coordinate-wise identifiable privileged basis.

2.2. PCA Recovery

The ordered weighting of Equation (4) forces a nested chain of principal subspaces. By the Eckart–Young–Mirsky Theorem, each prefix loss $\|X - A_{1:m}B_{1:m}X\|_F^2$ is minimized iff $\mathrm{colspan}(A_{1:m}) = \mathrm{span}(u_1, \ldots, u_m)$. Standard LAE training enforces this only at the full width $m = d$, which pins the column span of A to the top-d principal subspace but leaves the choice of basis within that subspace free; FP-MRL imposes it at every $m \in [d]$ simultaneously, and the unique family of subspaces satisfying all d constraints is the nested chain $\mathrm{span}(u_1) \subset \cdots \subset \mathrm{span}(u_1, \ldots, u_d)$. Equivalently, eigenvector u_k contributes σ_k^2 to the bound of every prefix $m < k$, so the summed loss weights σ_k^2 by $(k-1)$, and only the ordered chain achieves the lower bound $\sum_{k=2}^{d}(k-1)\sigma_k^2 + d\sum_{k=d+1}^{p}\sigma_k^2$. The resulting global optimum (A^*, B^*) satisfies $\mathrm{colspan}(A^*_{1:m}) = \mathrm{span}(u_1, \ldots, u_m)$ for all $m \in [d]$, determined up to right-

multiplication by $T \in \mathrm{SO}(m)$ reflecting the column-space invariance of MSE (Baldi & Hornik, 1989). Adding $A^\top A = I_d$ restricts this gauge to orthogonal transformations, and distinct eigenvalues then eliminate the remaining rotational freedom at every scale, yielding the full PCA recovery $A^* = [u_1, \ldots, u_d]$ (Rippel et al., 2014).

LDA recovery. Appendix A.5 shows that the FP-MRL construction extends naturally beyond MSE reconstruction objectives to supervised discriminative objectives based on Fisher's criterion. In this setting, the encoder is trained to maximize between-class variance relative to within-class variance in the learned embedding, yielding the classical Linear Discriminant Analysis (LDA) directions as the optimal discriminative subspace. The role of the Eckart–Young–Mirsky theorem in the PCA setting is played here by the Rayleigh–Ritz characterization of the Fisher trace-ratio objective. Specifically, the Fisher objective is optimized by the generalized eigenvectors, where the corresponding eigenvalues measure class-discriminative strength through the Fisher ratio. Just as the MSE objective recovers PCA directions ordered by explained variance, the Fisher objective recovers LDA directions ordered by discriminative power. Thus, the privileged basis induced by FP-MRL adapts to the underlying training objective itself. We confirm the emergence of these ordered LDA directions empirically in Section 3. Unlike the MSE reconstruction objective, however, $\mathcal{L}_{\mathrm{Fisher}}$ in equation (29) is non-convex. Consequently, optimization is substantially less stable in practice, exhibiting slower convergence.

Efficient loss calculation. Naively, full-prefix MRL requires d reconstruction losses, one per prefix suggesting a d-fold cost increase over standard training. In the linear setting, however, all prefix losses share the same sufficient statistics and can be assembled from $d \times d$ matrices computed once from X, A, and B. Full-prefix evaluation therefore contributes only an $O(d^2)$ additive overhead on top of the $O(npd)$ matrix multiplications already required for standard reconstruction, with no multiplicative blow-up. We note that for PCA itself a direct SVD remains the most efficient route. This result is universally applicable across linear models: even when no closed-form analytical solution exists (unlike PCA or LDA), the same sufficient-statistic structure ensures that full-prefix supervision incurs only additive overhead. A complete proof and derivation of the computational complexity is given in appendix B.

Beyond MSE. The structural basis for the gradient ordering argument above and the efficient loss evaluation result in appendix B requires only that the model's prefix-m output factorizes as a sum of per-coordinate contributions: besides a rank-1 contribution to reconstruction, this could be a vector contribution to logits or a scalar contribution to a pairwise distance. The same gradient accumulation and efficient computation arguments extend naturally to classification, regression, and distance-based losses. We leave the unified treatment for future work.

3. Empirical Validation

We evaluate PCA and LDA recovery on synthetic data, and on Fashion-MNIST in Figure 2. Figure 3 evaluates per-dimension statistics and probe accuracy on MNIST. The model architecture, synthetic data generation, and evaluation metric are detailed in appendix Section C, with hyperparameters summarized in Table 1.

Alignment with PCA and LDA. Figure 2(i–ii) shows the maximum cosine similarity between learned prefix directions and PCA eigenvectors. FP-MRL maintains near-unit alignment across most prefix sizes on both synthetic and Fashion-MNIST data, with slight degradation at larger k. This drop occurs as later eigenvalues become small and closely spaced: when $\sigma_k^2 \approx \sigma_{k+1}^2$, multiple directions explain nearly identical variance making alignment ill-conditioned. In contrast, S-MRL, NU-ℓ_2, and MD-ℓ_1 either degrade with prefix size or fail to align consistently.

When trained with a Fisher discriminant objective (appendix A.5), we observe the same qualitative behavior for LDA directions Figure 2(iii–iv). On synthetic data, FP-MRL achieves near-perfect alignment across prefix sizes, mirroring the PCA case. On real data, alignment degrades for all methods. We attribute this to optimization rather than objective mismatch: the simple linear architecture described in appendix Section C.1 used here does not reliably reach the global optimum of the LDA objective, leading to imperfect recovery of discriminative directions (Figure 4). This gap suggests an important direction for future work—whether MRL requires more expressive architectures or optimization schemes to fully realize its theoretical guarantees under supervised objectives. Regularizer-based methods show weaker and less stable alignment throughout, reinforcing that their induced bases are not driven by task signal.

Objective Specificity. To study whether MRL's induced ordering reflects the training objective itself rather than merely imposing sparsity, we move beyond the linear autoencoder setting to a nonlinear classification task on MNIST. We train a shared nonlinear encoder-classifier architecture on MNIST using either FP-MRL or MD-ℓ_1 as the ordering objective. The model consists of a two-hidden-layer ReLU MLP encoder with batch normalization and dropout, followed by a simple linear classifier. We then evaluate whether the resulting coordinate ordering aligns with class-discriminative usefulness. Concretely, for each latent coordinate z_k, we measure both its magnitude statistics and its ability to linearly separate classes in isolation via a 1D logistic probe. We additionally evaluate prefix-wise linear classification ac-

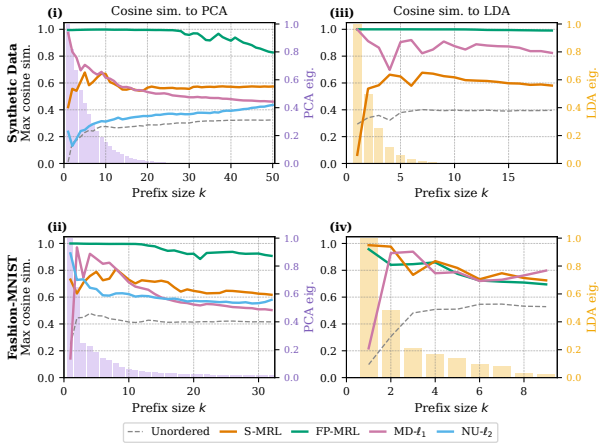


Figure 2. **Alignment with PCA and LDA bases under different symmetry-breaking mechanisms.** Maximum cosine similarity between learned prefix directions and the ordered PCA (i, ii) and LDA (iii, iv) bases, on synthetic data (top) and Fashion-MNIST (bottom). Bars show the normalized PCA/LDA eigenvalue spectra.

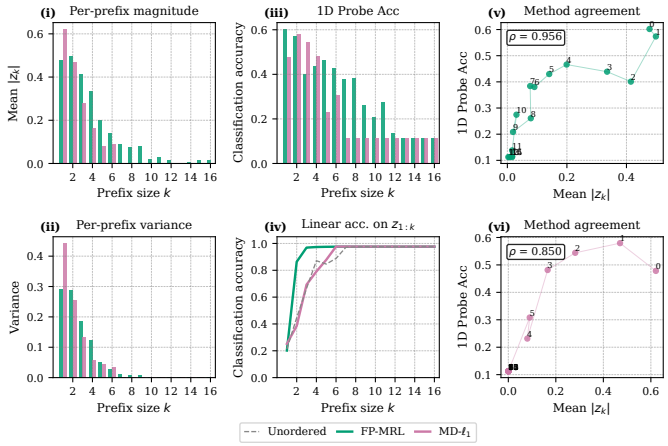


Figure 3. **Per-prefix geometry and utility on MNIST classification.** Comparing FP-MRL and MD-ℓ_1. (i, ii) Mean $|z_k|$ and variance per coordinate. (iii) Top-1 probe accuracy of each coordinate in isolation. (iv) Cumulative top-1 accuracy on the prefix $z_{1:k}$. (v, vi) Rank agreement between mean $|z_k|$ and 1D probe accuracy.

curacy using only the first k coordinates. Full architectural, training, and evaluation details are provided in appendix Section C.1.

Figure 3 isolates how each training objective shapes the geometry of the learned code along the prefix axis k. Panels (i)–(ii) show that both FP-MRL and MD-ℓ_1 produce the expected monotonic ordering of the per-dimension magnitude $\mathbb{E}[|z_k|]$ with k, concentrating energy in the leading coordinates. The two objectives, however, induce different decay profiles—MD-ℓ_1 collapses mass more aggressively into the first 2–3 dimensions, whereas FP-MRL spreads usable signal out to roughly $k \approx 8$ before the tail goes inert. Subplot (iii) shows that this geometric ordering is functionally meaningful. The 1D probe accuracy measures how well a single coordinate z_k, used in isolation, can linearly separate the classes, a direct test of whether magnitude actually reflects per-dimension informativeness. MD-ℓ_1 allocates the largest magnitude and variance to $k = 1$, yet FP-MRL achieves the higher 1D probe accuracy there, ties at $k = 2$, and is only overtaken at $k = 3$. Past $k = 3$, MD-ℓ_1's probe accuracy decays in step with its magnitude, whereas FP-MRL retains measurable signal across the remaining coordinates. Subplot (iv) reinforces this pattern at the cumulative level: the linear accuracy on the prefix $z_{1:k}$ rises sharply for FP-MRL and saturates near the full-dimensional ceiling by $k \approx 3$, while MD-ℓ_1 requires roughly twice as many coordinates ($k \approx 6$) to reach the same plateau. In deployment terms, this means FP-MRL can be truncated to half the embedding width of MD-ℓ_1 at no accuracy cost, a practical consequence of aligning magnitude with informativeness. The MD-ℓ_1 penalty inflates the leading coordinate's scale as a byproduct of its sparsity geometry, whereas FP-MRL's nested-prefix loss directly rewards each coordinate for being useful in isolation, so its magnitude ordering coincides with its information ordering. Panels (v)–(vi) confirm this at the population level: the rank correlation between $\mathbb{E}[|z_k|]$ and 1D probe accuracy is $\rho = 0.956$ for FP-MRL but only $\rho = 0.850$ for MD-ℓ_1, quantifying how much more faithfully magnitude serves as a proxy for informativeness under the nested-prefix objective. Together, these panels show that FP-MRL does not merely produce an ordered representation but an ordered representation whose magnitude axis is itself the information axis—an alignment that MD-ℓ_1's sparsity prior approximates only loosely, and that constitutes the defining specificity of the MRL objective.

4. Conclusion

We show that MRL induces an *objective-specific* privileged basis, ordering dimensions via cumulative task gradients rather than data statistics or embedding mass. For LAE, FP-MRL recovers ordered PCA directions at no extra asymptotic cost; with nonlinear encoders, this structure incurs an $O(d)$ per-step overhead. Important future directions include characterizing whether such privileged bases persist in more complex nonlinear architectures, and alternative objectives such as contrastive or generative learning. Another important direction is to explore whether specific architectures are better suited to fully extract and utilize this task-aligned ordering. While these embeddings are optimal for the training objective by construction, natural next steps include understanding how they generalize to unseen data and downstream tasks, and whether regularizer-based approaches can recover comparable objective-specific ordering.

References

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Bao, X., Lucas, J., Sachdeva, S., and Grosse, R. B. Regularized linear autoencoders recover the principal components, eventually. *Advances in Neural Information Processing Systems*, 33:6971–6981, 2020.

Cai, M., Yang, J., Gao, J., and Lee, Y. J. Matryoshka multimodal models. *Proceedings of the International Conference on Learning Representation*, 2025.

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

Elhage, N., Lasenby, R., and Olah, C. Privileged bases in the transformer residual stream. *Transformer Circuits Thread*, 24, 2023.

Hu, W., Dou, Z.-Y., Li, L. H., Kamath, A., Peng, N., and Chang, K.-W. Matryoshka query transformer for large vision-language models. *Advances in Neural Information Processing Systems*, 37:50168–50188, 2024.

Kunin, D., Bloom, J., Goeva, A., and Seed, C. Loss landscapes of regularized linear autoencoders. In *International conference on machine learning*, pp. 3560–3569. PMLR, 2019.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Mirsky, L. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1): 50–59, 1960.

Oftadeh, R., Shen, J., Wang, Z., and Shell, D. Eliminating the invariance on the loss landscape of linear autoencoders. In *International Conference on Machine Learning*, pp. 7405–7413. PMLR, 2020.

Rippel, O., Gelbart, M., and Adams, R. Learning ordered representations with nested dropout. In *International Conference on Machine Learning*, pp. 1746–1754. PMLR, 2014.

Verma, C., Timmaraju, A. S., Hsieh, C.-J., Damle, S., Bui, N., Zhang, Y., Chen, W., Liu, X., Jain, P., and Dhillon, I. Matryoshka model learning for improved elastic student models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4935–4944, 2025.

Xiao, Z., Ma, Q., Gu, M., cheng Jason Chen, C., Chen, X., Ordonez, V., and Mohan, V. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yKDqg9HwZX.

Zhan, Q., Zhou, Z., Wang, Z., and Shen, L. Learning ordered representations in latent space for intrinsic dimension estimation via principal component autoencoder. *arXiv preprint arXiv:2601.19179*, 2026.

A. Loss Expansions and Gradients

Setup. We expand each loss for a common worked example with input dimension $p = 5$ and latent dimension $d = 3$. Let $x \in \mathbb{R}^p$ be a single input. The encoder $B \in \mathbb{R}^{d\times p}$ has rows b_1, b_2, b_3, and the decoder $A \in \mathbb{R}^{p\times d}$ has columns a_1, a_2, a_3. The rank-one contribution of dimension k to the reconstruction is

$$y_k \;=\; a_k\, b_k^\top x \;\in\; \mathbb{R}^p,$$

that is, the latent activation $b_k^\top x \in \mathbb{R}$ scaled by the decoder column a_k. Stacking these contributions gives the prefix-m reconstruction and its residual:

$$\hat{x}^{(m)} \;=\; \sum_{k=1}^{m} y_k, \qquad r^{(m)} \;=\; x - \hat{x}^{(m)}.$$

Each $\hat{x}^{(m)}$ is the model's reconstruction using only the first m latent dimensions, and $r^{(m)}$ measures what those m dimensions fail to capture. The full reconstruction is $\hat{x}^{(d)} = \sum_{k=1}^{d} y_k$ and the trailing residual is $r^{(d)}$.

A.1. LAE Loss

Definition. The standard linear autoencoder loss evaluates MSE at the full reconstruction only:

$$\mathcal{L}_{\text{AE}}(\theta) = \Big\| x - \sum_{k=1}^{d} y_k \Big\|^2 = \|x - (y_1 + y_2 + y_3)\|^2. \quad (7)$$

Expansion ($p = 5,\ d = 3$).

$$\begin{aligned}\mathcal{L}_{\text{AE}}(\theta) = \|x\|^2 &- 2x^\top(y_1{+}y_2{+}y_3) \\ &+ \|y_1\|^2 + \|y_2\|^2 + \|y_3\|^2 \\ &+ 2y_1^\top y_2 + 2y_1^\top y_3 + 2y_2^\top y_3. \end{aligned} \quad (8)$$

Every y_k receives linear weight 1 toward x and every cross-term $y_i^\top y_j$ has coefficient 2, corresponding to weight matrix $\mathbf{1}\mathbf{1}^\top \otimes I_p$.

A.2. Full-Prefix MRL Loss

Definition. The FP-MRL loss sums MSE at every prefix reconstruction under $\omega_m = 1$:

$$\begin{aligned}\mathcal{L}_{\text{FP-MRL}}(\theta) = &\underbrace{\|x - y_1\|^2}_{m=1} + \underbrace{\|x - (y_1{+}y_2)\|^2}_{m=2} \\ &+ \underbrace{\|x - (y_1{+}y_2{+}y_3)\|^2}_{m=3}. \end{aligned} \quad (9)$$

Expanding each term.

$$\|x - y_1\|^2 = \|x\|^2 - 2x^\top y_1 + \|y_1\|^2, \quad (10)$$

$$\begin{aligned}\|x - (y_1{+}y_2)\|^2 = &\|x\|^2 - 2x^\top(y_1{+}y_2) \\ &+ \|y_1\|^2 + 2y_1^\top y_2 + \|y_2\|^2, \end{aligned} \quad (11)$$

$$\begin{aligned}\|x{-}(y_1{+}y_2{+}y_3)\|^2 = &\|x\|^2 - 2x^\top(y_1{+}y_2{+}y_3) \\ &+ \|y_1\|^2 + \|y_2\|^2 + \|y_3\|^2 \\ &+ 2y_1^\top y_2 + 2y_1^\top y_3 + 2y_2^\top y_3. \end{aligned} \quad (12)$$

Summing and collecting.

$$\begin{aligned}\mathcal{L}_{\text{FP-MRL}}(\theta) = 3\|x\|^2 &\\ &- \underbrace{2(3\,x^\top y_1 + 2\,x^\top y_2 + x^\top y_3)}_{\text{(1) ordered linear weights}} \\ &+ \underbrace{3\|y_1\|^2 + 2\|y_2\|^2 + \|y_3\|^2}_{\text{(2) ordered quadratic weights}} \\ &+ \underbrace{4\,y_1^\top y_2 + 2\,y_1^\top y_3 + 2\,y_2^\top y_3}_{\text{(3) ordered cross-dim. coupling}}. \end{aligned} \quad (13)$$

The quadratic part is governed by $S_3 \otimes I_5$, where

$$S_3 = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad (S_3)_{ij} = d - \max(i,j) + 1. \quad (14)$$

$(S_d)_{ij}$ counts the number of prefixes containing both dimensions i and j: a prefix of size m contains y_k iff $k \le m$, so the count is $\sum_{m=\max(i,j)}^{d} 1 = d - \max(i,j) + 1$.

A.3. Non-Uniform ℓ_2 Regularization

Definition

$$\mathcal{L}_{\text{NU-}\ell_2}(\theta) = \frac{1}{n}\|X - ABX\|_F^2 + \|\Lambda^{1/2}B\|_F^2 + \|A\Lambda^{1/2}\|_F^2, \quad (15)$$

with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ (Kunin et al., 2019; Bao et al., 2020). Let $\sigma_1^2 > \cdots > \sigma_d^2 > 0$ denote the top-d eigenvalues of the data covariance $\frac{1}{n}XX^\top$, with corresponding PCA eigenvectors $u_1, \ldots, u_d$.

For the recovery guarantee to apply, the diagonal entries of Λ must satisfy

$$0 \;<\; \lambda_1 \;<\; \lambda_2 \;<\; \cdots \;<\; \lambda_d \;<\; \sigma_d^2. \quad (16)$$

Strict ordering ($\lambda_1 < \cdots < \lambda_d$): breaks the orthogonal symmetry of the uniform ℓ_2 loss down to sign flips, making the set of global minima a discrete set of axis-aligned solutions rather than a continuous rotation orbit. Upper bound ($\lambda_d < \sigma_d^2$): ensures the shrinkage factors $(I - \Lambda S^{-2})^{1/2}$ in Theorem 2 in Bao et al. (2020) are real and positive, so every latent dimension k receives a non-trivial contribution from its assigned principal component.

Expansion ($p = 5,\ d = 3$). Writing $b_k \in \mathbb{R}^p$ for the k-th row of B and $a_k \in \mathbb{R}^p$ for the k-th column of A:

$$\begin{aligned}\mathcal{L}_{\text{NU-}\ell_2}(\theta) = \frac{1}{n}\|X - ABX\|_F^2 \\ + \lambda_1\|b_1\|^2 + \lambda_2\|b_2\|^2 + \lambda_3\|b_3\|^2 \\ + \lambda_1\|a_1\|^2 + \lambda_2\|a_2\|^2 + \lambda_3\|a_3\|^2. \quad (17)\end{aligned}$$

The reconstruction term is identical across all dimensions; ordering comes entirely from the scalar multipliers λ_k on weight norms.

A.4. Monotone-Decay ℓ_1 Loss

Definition. We penalize the activation of each latent dimension $z_k = b_k^\top x$ with strictly increasing ℓ_1 coefficients — heavier on late dimensions to discourage them from being used with $\alpha > 0$:

$$\mathcal{L}_{\text{MD-}\ell_1}(\theta) = \frac{1}{n}\|X - ABX\|_F^2 + \frac{\alpha}{n}\sum_{k=1}^{d} k \sum_{s=1}^{n} |b_k^\top x_s|, \quad (18)$$

Expansion ($p = 5,\ d = 3$).

$$\begin{aligned}\mathcal{L}_{\text{MD-}\ell_1}(\theta) = \frac{1}{n}\|X - ABX\|_F^2 \\ + \frac{\alpha}{n}\Big(1 \cdot \textstyle\sum_s |b_1^\top x_s| + 2 \cdot \sum_s |b_2^\top x_s| \\ + 3 \cdot \textstyle\sum_s |b_3^\top x_s|\Big). \quad (19)\end{aligned}$$

The penalty on z_k grows with k, pushing the model to concentrate information in early dimensions.

A.5. Fisher discriminant objective

Definition. We are given data $\{(x_s, \tilde{y}_s)\}_{s=1}^n$ with $x_s \in \mathbb{R}^p$ and class labels $\tilde{y}_s \in \{1, \dots, c\}$. Let n_c denote the number of samples in class c. We define the class means μ_c, global mean μ, and between- and within-class scatter matrices S_{BC} and S_{WC}:

$$\mu_c = \frac{1}{n_c}\sum_{s:\,\tilde{y}_s = c} x_s, \quad (20)$$

$$\mu = \frac{1}{n}\sum_{s=1}^{n} x_s, \quad (21)$$

$$S_{\text{BC}} = \sum_c \frac{n_c}{n}(\mu_c - \mu)(\mu_c - \mu)^\top, \quad (22)$$

$$S_{\text{WC}} = \sum_c \frac{1}{n}\sum_{s:\,\tilde{y}_s = c}(x_s - \mu_c)(x_s - \mu_c)^\top. \quad (23)$$

We seek directions $v \in \mathbb{R}^p$ along which classes are maximally separated relative to their internal spread, i.e. that maximize the Fisher ratio

$$J(v) \;=\; \frac{v^\top S_{\text{BC}}\, v}{v^\top S_{\text{WC}}\, v}. \quad (24)$$

We assume $S_{\text{WC}} \succ 0$. Setting $\nabla_v J = 0$, we find that the stationary points of J are exactly the solutions of the generalized eigenproblem

$$S_{\text{BC}}\, v_k \;=\; \gamma_k\, S_{\text{WC}}\, v_k, \qquad \gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p \geq 0, \quad (25)$$

with $J(v_k) = \gamma_k$. The leading direction v_1 maximizes the Fisher ratio; the top-d directions $v_1, \dots, v_d$ jointly maximize the trace-ratio objective $\operatorname{Tr}\big[(V^\top S_{\text{WC}} V)^{-1} V^\top S_{\text{BC}} V\big]$ over $V \in \mathbb{R}^{p \times d}$, with optimal value $\sum_{k=1}^d \gamma_k$. We note that, since S_{BC} is a sum of c rank-one terms whose deviations $\mu_c - \mu$ sum to zero (weighted by n_c/n), we have $\operatorname{rank}(S_{\text{BC}}) \leq c - 1$, so at most $c - 1$ of the generalized eigenvalues γ_k are nonzero. LDA thus yields at most $c - 1$ informative directions.

Training. We train the encoder $B \in \mathbb{R}^{d \times p}$ using a supervised Fisher objective. Let $z_s = Bx_s \in \mathbb{R}^d$ be the embedding of input x_s. Define between-class and within-class scatter matrices in input space:

$$S_{\text{BC}}^{(x)} = \sum_c \frac{n_c}{n}(\mu_c - \mu)(\mu_c - \mu)^\top, \quad (26)$$

$$S_{\text{WC}}^{(x)} = \sum_c \frac{1}{n}\sum_{s:\,\tilde{y}_s = c}(x_s - \mu_c)(x_s - \mu_c)^\top, \quad (27)$$

where μ_c and μ are class and global means. The corresponding scatter matrices in embedding space are

$$S_{\text{BC}}^{(z)} = B S_{\text{BC}}^{(x)} B^\top, \qquad S_{\text{WC}}^{(z)} = B S_{\text{WC}}^{(x)} B^\top. \quad (28)$$

We optimize the Fisher objective:

$$\mathcal{L}_{\text{Fisher}}(B) = -\operatorname{Tr}\left[\left(S_{\text{WC}}^{(z)} + \varepsilon I\right)^{-1} S_{\text{BC}}^{(z)}\right], \quad (29)$$

with $\varepsilon > 0$ for numerical stability. This objective maximizes class separation in the learned embedding by enlarging between-class variance relative to within-class variance. At optimum, the row space of B spans the top-d LDA directions.

Analogy with PCA and FP-MRL. The role of the Eckart–Young–Mirsky theorem in the PCA case is played here by the Rayleigh–Ritz characterization of the trace-ratio problem: each prefix loss is minimized iff $\operatorname{rowspan}(B_{1:m,:})$ equals the span of the top-m LDA directions $v_1, \dots, v_m$, defined as the leading solutions of $S_{\text{BC}}^{(x)} v = \gamma\, S_{\text{WC}}^{(x)} v$ and ordered by descending Fisher ratio γ. Summing over m forces the nested chain $\operatorname{span}(v_1) \subset \cdots \subset \operatorname{span}(v_1, \dots, v_d)$, in

parallel to the PCA recovery of Theorem B.1, but with the principal directions u_k of $XX^\top$ replaced by the LDA directions v_k of the pair $(S_{\mathrm{BC}}^{(x)}, S_{\mathrm{WC}}^{(x)})$. Unlike the MSE loss, $\mathcal{L}_{\mathrm{Fisher}}$ is non-convex in B. In practice this manifests as slower convergence and higher sensitivity to initialization and ε than the MSE case.

B. Efficient PCA recovery by FP-MRL

This section gives the formal statement and proof of the PCA recovery result sketched in Section 2.2, together with the computational complexity analysis.

Theorem B.1 (Full-prefix linear MRL and PCA recovery)**.** *Let $\Sigma = XX^\top$ have distinct top-d eigenvalues $\sigma_1^2 > \cdots > \sigma_d^2 > 0$ with eigenvectors $u_1, \ldots, u_d$.*

1. (Subspace recovery)*: every global optimum (A^*, B^*) of $\mathcal{L}_{\text{FP-MRL}}$ satisfies $\operatorname{colspan}(A^*_{1:m}) = \operatorname{span}(u_1, \ldots, u_m)$ for all $m \in [d]$, with each column a_k^* aligned with $\pm u_k$. The full-prefix structure pins the nested flag of subspaces exactly, leaving only the discrete sign-flip symmetry $\{\pm 1\}^d$.*
2. (Full PCA recovery)*: under $A^\top A = I_d$, the unique global optimum is $A^* = [u_1, \ldots, u_d]$.*

Proof. **Part (1).** By the Eckart–Young–Mirsky theorem (Eckart & Young, 1936; Mirsky, 1960), for each m:

$$\|X - A_{1:m}B_{1:m}X\|_F^2 \;\geq\; \sum_{k=m+1}^{p} \sigma_k^2,$$

with equality iff $\operatorname{colspan}(A_{1:m}) = \operatorname{span}(u_1, \ldots, u_m)$. Summing over $m = 1, \ldots, d$, each eigenvalue σ_k^2 for $k \leq d$ appears in the bound of every prefix $m < k$, contributing coefficient $(k-1)$:

$$\mathcal{L}_{\text{FP-MRL}}(A, B) \;\geq\; \sum_{k=2}^{d}(k-1)\sigma_k^2 + d\sum_{k=d+1}^{p}\sigma_k^2.$$

Equality holds simultaneously at every m iff $\operatorname{colspan}(A_{1:m}) = \operatorname{span}(u_1, \ldots, u_m)$ for all $m \in [d]$, which is achieved by the unique nested chain $\operatorname{span}(u_1) \subset \cdots \subset \operatorname{span}(u_1, \ldots, u_d)$. Within each prefix, $A^*_{1:m}$ is free up to $T \in \mathrm{SO}(m)$ since the MSE loss depends only on the column space (Baldi & Hornik, 1989).

Part (2). Under $A^\top A = I_d$, distinct eigenvalues yield a unique orthonormal minimizer at each scale, collapsing the $\mathrm{SO}(m)$ gauge to the identity and hence $A^*_{1:m} = [u_1, \ldots, u_m]$ for all m (Rippel et al., 2014). □

Computation complexity. Naively, full-prefix MRL requires evaluating d reconstruction losses, one per prefix width, suggesting a d-fold cost over standard training. In the linear setting this overhead is avoidable: every prefix loss is determined by the same three $d \times d$ summaries of the data, encoder, and decoder, computed once per forward pass. The result is that full-prefix MRL has the same $O(npd)$ asymptotic cost as a single standard reconstruction pass, with only $O(d^2)$ additive overhead for prefix accumulation.

Proof. Let $Z = BX \in \mathbb{R}^{d\times n}$. Write $a_i \in \mathbb{R}^p$ for the i-th column of A and $z_i^\top \in \mathbb{R}^{1\times n}$ for the i-th row of Z (so $z_i \in \mathbb{R}^n$). The prefix-m reconstruction decomposes as

$$\hat{X}^{(m)} = A_{:,1:m}\, Z_{1:m,:} = \sum_{i=1}^{m} a_i z_i^\top.$$

. Expanding the Frobenius norm and substituting,

$$\ell_m := \|X - \hat{X}^{(m)}\|_F^2 = \|X\|_F^2 - 2\sum_{i=1}^{m} H_{ii} + \sum_{i,j=1}^{m} G^A_{ij}\, G^Z_{ij},$$

where $Z = BX$, $H = A^\top XZ^\top$, $G^A = A^\top A$, and $G^Z = ZZ^\top$ are $d \times d$ matrices, with total cost $O(npd)$ to form. They depend on m only through prefix truncation. Forming $Z = BX$ and $XZ^\top$ costs $O(npd)$ in total. The Gram matrices G^A and G^Z contribute $O(pd^2)$ and $O(nd^2)$, which are absorbed when $n, p \gg d$. Once these are in hand, evaluating ℓ_m for all $m \in [d]$ then reduces to incremental sums over H, G^A, and G^Z, taking $O(d^2)$ more work — total cost $O(npd)$. Total cost: $O(npd + d^2) = O(npd)$. □

C. Experimental Setup

C.1. Model architecture and training dimensions

PCA/LDA recovery. All models share an identical linear autoencoder backbone for PCA/LDA recovery experiments.

$$z = Bx, \qquad \hat{x} = Az, \qquad B \in \mathbb{R}^{d\times p},\ A \in \mathbb{R}^{p\times d}, \quad (30)$$

where $x \in \mathbb{R}^p$, $z \in \mathbb{R}^d$, B is the encoder and A is the decoder. Both maps are bias-free and contain no nonlinearity, so the end-to-end map $\hat{x} = ABx$ is strictly linear. The values of the input dimension p of data x, the number of classes C, the latent dimension d, the S-MRL prefix set $\mathcal{M}$, and the optimiser/regulariser settings differ by dataset and loss family; they are given row-by-row in Table 1. The decoder is initialised with orthonormal columns ($A^\top A = I_d$) via reduced QR decomposition. When the decoder-orthogonality constraint is active (FP-MRL), a QR re-projection of A is applied after every optimiser step. Fisher-family losses use the same backbone but ignore A during training; labels enter only through the between- and within-class scatter matrices S_{BC} and S_{WC} computed on z, whose rank is bounded by $n_{\mathrm{lda}} = C - 1$ where C is the number of classes.

Table 1. **PCA/LDA recovery experiments: architecture and training parameters per dataset and loss family.** p: input dimensionality; C: number of classes; $n_{\mathrm{lda}} = C - 1$: rank of the LDA subspace; d: latent dimensionality; $\mathcal{M}$: S-MRL prefix set; σ_d^2: the d-th PCA eigenvalue of the training data. Optimiser is Adam with learning rate 10^{-3} and weight decay 10^{-4} in every row.

Dataset	Family	p	C	n_{lda}	d	$\mathcal{M}$ (S-MRL)	Batch	Regulariser
Synthetic data	MSE	69	20	19	50	$\{5, 10, 25, 50\}$	256	α=0.01, $\lambda_k \in [\frac{\sigma_d^2}{18}, \frac{\sigma_d^2}{2}; \Delta\lambda = \frac{\lambda_{\max}-\lambda_{\min}}{(d-1)}]$
Synthetic data	Fisher	69	20	19	19	$\{5, 10, 15, 19\}$	full	$\varepsilon = 10^{-4}$
Fashion-MNIST	MSE	784	10	9	32	$\{4, 8, 16, 32\}$	256	α=0.01, $\lambda_k \in [\frac{\sigma_d^2}{18}, \frac{\sigma_d^2}{2}; \Delta\lambda = \frac{\lambda_{\max}-\lambda_{\min}}{(d-1)}]$
Fashion-MNIST	Fisher	784	10	9	9	$\{2, 4, 7, 9\}$	full	$\varepsilon = 10^{-4}$

We use two datasets in the configurations of Table 1. Synthetic dataset: n_{tot}=10,000 samples (500 per class), split stratified $70/10/20$ into 7,000 train, 1,000 val, and 2,000 test. Details of synthetic data are given in appendix Section C.2. Fashion-MNIST dataset: the full n_{tot}=70,000 images, split stratified $70/10/20$ into 49,000 train, 7,000 val, and 14,000 test. In both cases the inputs are mean-centred on the training split (no variance scaling, so the noise/signal eigenvalue contrast in the synthetic data is preserved).

Classification. The classification experiment replaces the linear autoencoder with a nonlinear encoder followed by a linear classifier, trained end-to-end on MNIST ($p = 784$, $C = 10$) with a bottleneck width of $d = 16$. The encoder

$$z \;=\; \frac{f_\theta(x)}{\|f_\theta(x)\|_2}, \qquad f_\theta : \mathbb{R}^p \to \mathbb{R}^d, \tag{31}$$

is a two-hidden-layer MLP ($p \to 256 \to 256 \to d$) with batch normalisation, ReLU, and dropout (0.1) between layers, and an ℓ_2-normalised output. A single linear classifier $W \in \mathbb{R}^{C\times d}$ maps z to class logits. We train two loss families on this shared backbone: FP-MRL, and (MD-ℓ_1). All models are optimised with Adam (learning rate 10^{-3}, weight decay 10^{-4}) at batch size 128 for up to 20 epochs with early stopping (patience 5) on validation loss. The best checkpoint by validation loss is retained for evaluation. Unlike the linear autoencoder setting, the nonlinear encoder breaks the shared sufficient-statistics structure that made FP-MRL asymptotically free: each of the d prefix losses now requires a separate forward pass through the classifier head, so FP-MRL incurs an $O(d)$ overhead per step relative to MD-ℓ_1.

Prefix-wise evaluation. We report two test-set metrics for every trained model.

- *1-D probing accuracy.* For each prefix $k = 1, \ldots, d$ we fit an independent 1-D logistic regression on $z_{\mathrm{train}}[:, k]$ (restricted to a 2,000-sample subset for tractability) and score it on $z_{\mathrm{test}}[:, k]$. The resulting per-dimension accuracy profile quantifies how much class-discriminative information each coordinate carries in isolation, independently of any interaction with other dimensions.
- *Linear accuracy.* For each prefix length $k = 1, \ldots, d$ we re-use the learned classifier:

$$\mathrm{logits}_k \;=\; z_{\mathrm{test}}[:, 1{:}k]\, W_{:,1:k}^{\top} \;+\; b, \tag{32}$$

  and report top-1 accuracy against the test labels. This metric measures how well the trained classifier itself respects the prefix structure — i.e., whether the leading coordinates alone are sufficient for W to discriminate the classes.

C.2. Synthetic data generation

The synthetic dataset is designed so that the top PCA and LDA subspaces are orthogonal by construction, providing a controlled setting in which to test whether each loss recovers its objective-specific subspace. Each input $x \in \mathbb{R}^p$ is partitioned into two disjoint blocks,

$$x \;=\; [\underbrace{x_{\mathrm{noise}}}_{p_{\mathrm{noise}}} \mid \underbrace{x_{\mathrm{sig}}}_{p_{\mathrm{sig}}}], \qquad p \;=\; p_{\mathrm{noise}} + p_{\mathrm{sig}}, \tag{33}$$

drawn independently per sample from

$$x_{\mathrm{noise}} \sim \mathcal{N}(0, \Sigma_{\mathrm{noise}}), \qquad x_{\mathrm{sig}} \mid c \sim \mathcal{N}(\mu_c, \tau_{\mathrm{sig}}^2 I), \tag{34}$$

where $c \in \{1, \ldots, C\}$ is the class label. The noise block carries no class information and has large variance, so PCA eigenvectors concentrate on it. The signal block has small within-class variance but class-dependent means, so LDA directions concentrate on it. Setting $p_{\mathrm{sig}} = C - 1$ ensures that between-class scatter is full rank and that the signal block is exactly spanned by the $C - 1$ LDA discriminants.

We use $C = 20$ classes, $p_{\mathrm{noise}} = 50$, $p_{\mathrm{sig}} = C - 1 = 19$ (so $p = 69$), $\tau_{\mathrm{noise}} = 5.0$, $\tau_{\mathrm{sig}} = 0.1$, and base class-mean scale $\beta_0 = 1.0$. We draw 500 samples per class (total $n = 10{,}000$) and split $70/10/20$ into train/val/test (stratified). Features are mean-centred on the train split only; we deliberately do not standardise variance, which would destroy the variance contrast between the two blocks.

Noise block. $\Sigma_{\mathrm{noise}} \in \mathbb{R}^{p_{\mathrm{noise}}\times p_{\mathrm{noise}}}$ is diagonal with a geometrically decaying spectrum,

$$(\Sigma_{\mathrm{noise}})_{ii} \;=\; \tau_{\mathrm{noise}}^2\, \rho_{\mathrm{noise}}^{2i}, \qquad i = 0, \ldots, p_{\mathrm{noise}} - 1, \tag{35}$$

with $\rho_{\text{noise}} = 0.9$, so consecutive noise eigenvalues are well-separated and individual PCA eigenvectors (not just the subspace) are identifiable.

Signal block. Class means in the signal block are constructed to give a ground-truth LDA ordering. Let $H \in \mathbb{R}^{C \times p_{\text{sig}}}$ have orthonormal columns, each orthogonal to $\mathbf{1}_C$ (so the grand mean is zero on every signal dimension), obtained by reduced QR of a Gaussian matrix projected onto $\mathbf{1}_C^{\perp}$. Let $\bar{\mu} \in \mathbb{R}^{C \times p_{\text{sig}}}$ be the matrix whose c-th row is the class-c signal-block mean μ_c. We set

$$\bar{\mu} \;=\; H\,\mathrm{diag}(\beta), \qquad \beta_i \;=\; \beta_0\,\rho_{\text{sig}}^{\,i}, \tag{36}$$

with $\beta_0 = 1.0$ and $\rho_{\text{sig}} = 0.7$. The between-class scatter in signal space is then

$$S_{\text{BC}}^{(\text{sig})} \;=\; \tfrac{1}{C}\,\mathrm{diag}\big(\beta_0^2, \ldots, \beta_{p_{\text{sig}}-1}^2\big),$$

so the i-th LDA direction aligns with signal axis i and class separability decreases monotonically with i.

C.3. Evaluation metrics

Let $\{q_1, \ldots, q_d\}$ denote the learned prefix directions (unit-norm columns of the decoder A, or rows of the encoder B in the Fisher case), and let $\{e_1, \ldots, e_d\}$ denote the reference basis (PCA eigenvectors u_k of Σ, or LDA discriminants v_k), also unit-norm and ordered by descending eigenvalue. We use three metrics to compare the learned and reference bases.

Paired cosine similarity. Measures whether the k-th learned direction matches the k-th reference direction, capturing coordinate-wise ordering:

$$\mathrm{PairedCos}(k) \;=\; \big|\langle q_k,\, e_k\rangle\big|, \qquad k \in [d]. \tag{37}$$

Max cosine similarity. Measures whether the k-th learned direction lies along *any* reference direction, capturing subspace membership without requiring ordering:

$$\mathrm{MaxCos}(k) \;=\; \max_{j \in [d]} \big|\langle q_k,\, e_j\rangle\big|. \tag{38}$$

Mean principal angle. Measures alignment between the leading-k learned subspace and the leading-k reference subspace, independent of any per-axis correspondence. Let $Q_k = [q_1, \ldots, q_k]$ and $E_k = [e_1, \ldots, e_k]$ (with orthonormalised columns), and let $\eta_1 \geq \cdots \geq \eta_k$ be the singular values of $Q_k^{\top} E_k$. The principal angles are $\varphi_i = \arccos(\eta_i)$, and we report

$$\overline{\varphi}(k) \;=\; \frac{1}{k}\sum_{i=1}^{k} \varphi_i. \tag{39}$$

A value of $0°$ indicates the two leading-k subspaces coincide; $90°$ indicates they are orthogonal.

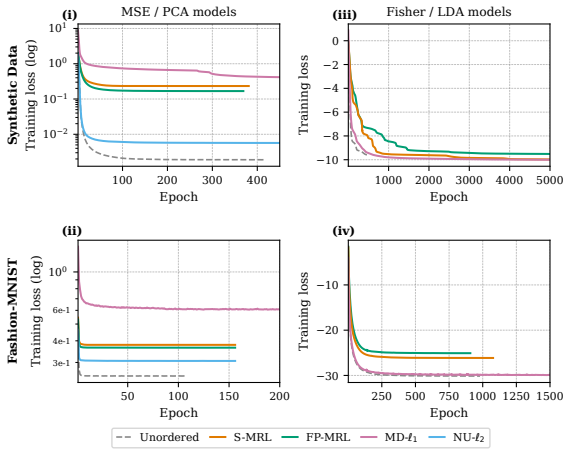


Figure 4. **Fisher/LDA optimization is harder to reach the global optimum under prefix-based losses.** Training loss curves for MSE/PCA models (i, ii) and Fisher/LDA models (iii, iv), on synthetic data (top) and Fashion-MNIST (bottom). All methods reach the global optimum on the LAE objective for both datasets, and on the LDA objective for synthetic data. On Fashion-MNIST LDA, however, S-MRL and FP-MRL plateau above the optimum reached by MD-ℓ_1, NU-ℓ_2, and the unordered baseline, indicating that with the simple linear model architecture described in Section C.1, prefix-based losses struggle to optimize the Fisher objective.

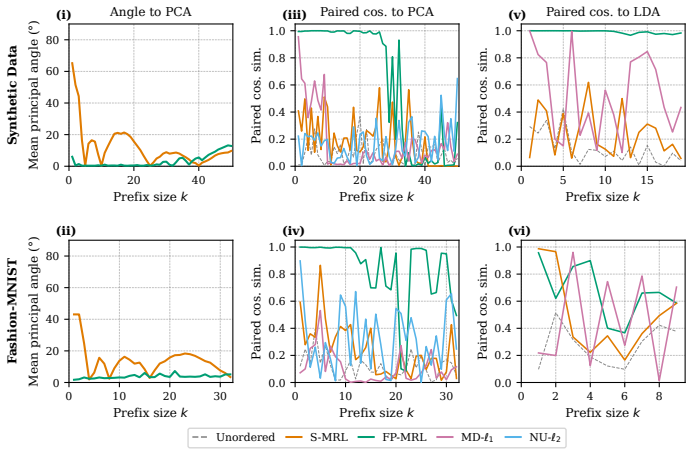
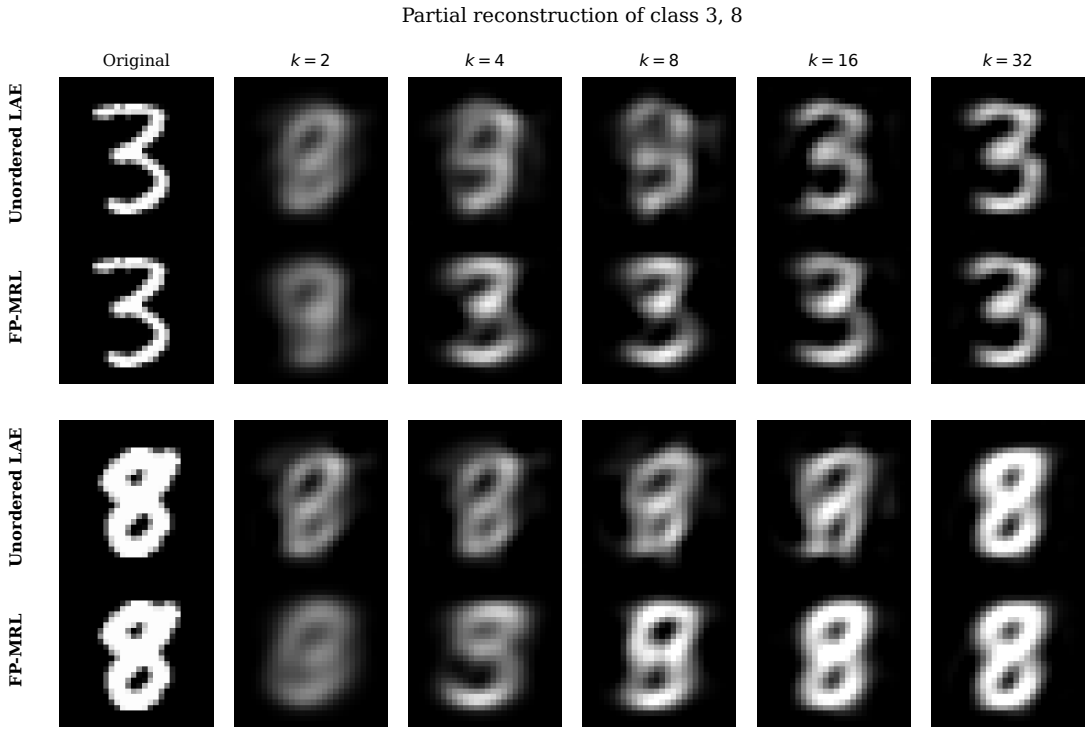
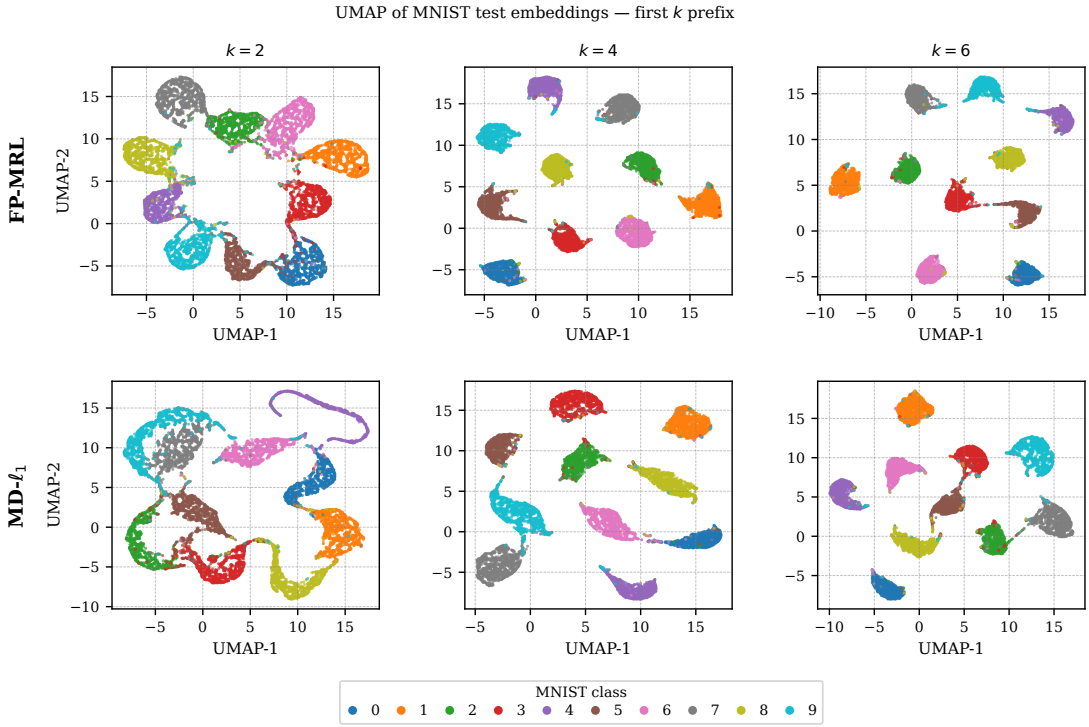


Figure 5. **Basis alignment across symmetry-breaking mechanisms.** Mean principal angle between the leading k learned prefix directions and the leading k PCA eigenvectors (i, ii), and paired cosine similarity between the k-th learned prefix direction and the k-th PCA (iii, iv) or LDA (v, vi) eigenvector, on synthetic data (top row) and Fashion-MNIST (bottom row). In (i, ii), FP-MRL tracks the ordered PCA basis closely across all prefix sizes, while S-MRL aligns only at the discrete prefix sizes evaluated during training (see Table 1) and departs from the ordered basis at intermediate k. Across rest of the panels, MD-ℓ_1, NU-ℓ_2, and the unordered baseline recover the subspace but not the coordinate-wise ordering.



(a) Reconstruction grid.



(b) UMAP clustering.

Figure 6. **Prefix-truncated representation quality on MNIST. (a)** Partial reconstructions of digits 3 and 8 using only the first $k \in \{2, 4, 8, 16, 32\}$ latent coordinates. The unordered LAE baseline produces incoherent outputs at small k, while FP-MRL yields recognizable digits even from $k = 2$, with quality improving monotonically as more coordinates are added. **(b)** UMAP projections of the first-k prefix of the test-set embeddings for $k \in \{2, 4, 6\}$. FP-MRL (top) produces visibly more compact and class-separated clusters at every prefix length than the MD-ℓ_1 baseline (bottom).

Table 2. Summary of notation used throughout the paper.

Symbol	Meaning	Symbol	Meaning
Data and dimensions			
p	Input dimensionality	C	Number of classes
d	Embedding dimensionality	L	Output dimensionality of task head
n	Number of training samples	$\Sigma = XX^\top$	Empirical data covariance
$X \in \mathbb{R}^{p\times n}$	Zero-centered data matrix	$\sigma_k^2,\ u_k$	k-th PCA eigenvalue / eigenvector of Σ
$x \in \mathbb{R}^p$	Single input sample	$\tilde{y}_s$	Target / label for sample s
$s \in [n]$	Sample index	$k \in [d]$	Latent dimension index
$i, j \in [d]$	Dimension indices for matrix entries		
Model parameters			
$B \in \mathbb{R}^{d\times p}$	Linear encoder	$a_k \in \mathbb{R}^p$	k-th column of decoder A
$A \in \mathbb{R}^{p\times d}$	Linear decoder; cols. $a_1, \ldots, a_d$	$W^{(m)} \in \mathbb{R}^{L\times m}$	Linear task head at prefix scale m
$b_k \in \mathbb{R}^p$	k-th row of encoder B		
Latent and reconstruction			
$z = Bx \in \mathbb{R}^d$	Embedding	$y_k = z_k a_k \in \mathbb{R}^p$	Rank-1 contribution of dim. k
$z_k = b_k^\top x$	k-th prefix of the embedding	$\hat{x}^{(m)} = \sum_{k=1}^m y_k$	Prefix-m reconstruction
$z_{1:m}$	First m prefix of z	$r^{(m)} = x - \hat{x}^{(m)}$	Prefix-m residual
Prefix structure			
$\mathcal{M} \subseteq [d]$	Nesting set of prefix sizes	$\omega_m > 0$	Per-prefix weight (we use $\omega_m{=}1$)
$m \in \mathcal{M}$	A prefix size	$w_k = \sum_{m=k}^d \omega_m$	Cumulative weight on dimension k
$Z = BX \in \mathbb{R}^{d\times n}$	Embedding matrix (all samples)	$\hat{X}^{(m)} = A_{:,1:m} Z_{1:m,:}$	Prefix-m reconstruction (matrix form)
$\ell_m = \lVert X - \hat{X}^{(m)} \rVert_F^2$	Prefix-m squared loss	S_d^ω	Prefix coupling matrix; $(S_d^\omega)_{ij}{=}w_{\max(i,j)}$
		$H,\ G^A,\ G^Z$	Sufficient $d{\times}d$ statistics (App. B)
Symmetry and identifiability			
$\mathrm{GL}(d)$	Invertible reparameterizations of latent space (broader than $\mathrm{SO}(d)$)		
$\mathrm{SO}(d)$	Rotations: $T^\top T{=}I,\ \det T{=}+1$	$\mathcal{G}_\mathcal{L}$	Subgroup of $\mathrm{SO}(d)$ leaving $\mathcal{L}$ invariant at θ^*
$T \cdot (A, B)$	Latent reparameterization $(AT,\ T^{-1}B)$	$\mathcal{G}_\mathcal{L} = \{I\}$	Full identifiability
Losses and regularizers			
$\mathcal{L}$	Generic differentiable training loss	$\mathcal{L}_{\text{Fisher}}$	Fisher discriminant objective
θ	All trainable parameters	$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$	NU-ℓ_2 coefficient matrix
θ^*	Optimum (minimizer) of $\mathcal{L}$	α	MD-ℓ_1 regularization strength
$\mathcal{L}_{\text{AE}}$	Standard LAE (full-width MSE) loss	ε	Fisher numerical stabilizer
$\mathcal{L}_{\text{MRL}}$	Matryoshka loss (general $\mathcal{M}$)	$\mathcal{L}_{\text{NU-}\ell_2}$	Non-uniform ℓ_2 loss (NU-ℓ_2)
$\mathcal{L}_{\text{FP-MRL}}$	Full-prefix MRL loss ($\mathcal{M} = [d]$)	$\mathcal{L}_{\text{MD-}\ell_1}$	Monotone-decay ℓ_1 loss (MD-ℓ_1)
Fisher / LDA			
$\mu_c,\ \mu$	Class-c mean / global mean	$S_{\text{BC}}^{(z)},\ S_{\text{WC}}^{(z)}$	Between- / within-class scatter (embedding)
$S_{\text{BC}}^{(x)},\ S_{\text{WC}}^{(x)}$	Between- / within-class scatter (input)	$v_k,\ \gamma_k$	k-th LDA direction / Fisher ratio
		$n_{\text{lda}} = C{-}1$	Rank of LDA subspace
Methods (abbreviations)			
LAE	Linear autoencoder	FP-MRL	Full-prefix MRL ($\mathcal{M} = [d]$)
MRL	Matryoshka Representation Learning	NU-ℓ_2	Non-uniform ℓ_2 regularization
S-MRL	Sparse MRL ($O(\log d)$ prefixes)	MD-ℓ_1	Monotone-decay ℓ_1 regularization
		PCA / LDA	Principal / Linear Discriminant Analysis
Evaluation metrics			
$q_k \in \mathbb{R}^p$	k-th learned prefix direction (unit norm)	$Q_k = [q_1, \ldots, q_k]$	Leading-k learned basis matrix
$e_k \in \mathbb{R}^p$	k-th reference basis direction (PCA u_k / LDA v_k)	$E_k = [e_1, \ldots, e_k]$	Leading-k reference basis matrix
$\mathrm{PairedCos}(k)$	Paired cosine sim. $\lvert\langle q_k, e_k\rangle\rvert$	η_i	i-th singular value of $Q_k^\top E_k$
$\mathrm{MaxCos}(k)$	Max cosine sim. $\max_j \lvert\langle q_k, e_j\rangle\rvert$	$\varphi_i = \arccos(\eta_i)$	i-th principal angle
		$\overline{\varphi}(k)$	Mean principal angle over leading-k subspaces