IMPERFECT ALIGNMENT, RELIABLE SIMILARITY: SCORE–ENCODER SEPARATION IN MULTIMODAL CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal contrastive learning is often interpreted as aligning matched representations, yet each modality may contain noise irrelevant to cross-modal comparison. We ask whether reliable similarity requires the individual encoders to be denoised. We study population gradient flow for linear encoders with a learned temperature in a noisy two-modality model, where the score is factorized through the encoders. In a regular non-collapsing regime, every nuisance interaction in the score vanishes asymptotically while shared score signal remains. Generic initialization keeps encoder nuisance nonzero at every finite time, but permits asymptotic encoder denoising. Separately, when the output dimension exceeds the shared-signal dimension by at least two, an open basin of full-signal optima retains nuisance in both encoders asymptotically. Thus score- and encoder-level denoising are distinct. Controlled linear and nonlinear experiments support this separation; a real-data compression study further separates feature reconstruction, preservation of learned comparisons, and label retrieval.

1 INTRODUCTION

Contrastive learning trains representations through comparisons between matched and unmatched observations (Radford et al., 2021; van den Oord et al., 2018; Chen et al., 2020). For cross-modal retrieval, however, the deployed object is the similarity score formed jointly by two encoders. A direction can remain in one encoder yet be invisible to the score when it is orthogonal to the output of the other encoder. We call this phenomenon *score–encoder separation*: reliable comparisons and nuisance-free representations are different properties. Alignment and uniformity describe representation geometry (Wang & Isola, 2020; Tschannen et al., 2020); the score additionally depends on how the two encoder output spaces interact. Figure 1 shows nuisance remaining in both encoders while disappearing from their score interactions.

We analyze population gradient flow for linear encoders with expected-norm normalization and learned temperature in a noisy two-modality model. General-function analyses characterize global InfoNCE minimizers assuming aligned, maximally informative representations (Gui et al., 2025); positive measurement noise rules out such alignment here. Even in this linear–Gaussian setting, the factorized encoder–temperature flow remains nonconvex: the closed-form optimal score does not determine its long-time trajectory. We must exclude score collapse and temperature escape, handle rank-deficient limits, and track the score and its factors separately. Under the stated dimension conditions, generic trajectories preserve shared score signal while eliminating every off-signal score block asymptotically. Encoder nuisance persists at every finite time under generic initialization; positive limiting nuisance is established separately on an open basin.

The separation also has a downstream consequence for compression: preserving the joint score can require fewer coordinates than reconstructing either encoder. We test this design principle on pretrained image–text representations, without treating the compression experiment as evidence that those representations follow the latent nuisance model or as a claim that one compressor is universally best.

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

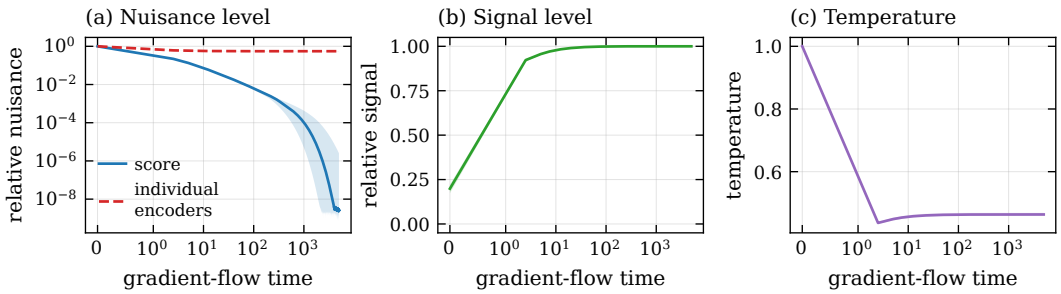

Figure 1: **Population score–encoder separation.** (a) The off-signal blocks of $M = W_B^\top W_A$ decrease while nuisance remains in W_A, W_B. (b) Encoder signal energy remains nonzero over the integration window; each trajectory is normalized by its maximum over that window. (c) The learned temperature stays bounded over the integration window, as required by the trajectory analysis. Shared dimension $r = 3$, input dimensions $d_1 = d_2 = 20$, output dimension $d = 25$, measurement-noise variance $\sigma^2 = 0.5$, modality-specific variance $\nu^2 = 0.3$, and initial temperature $\tau(0) = 1$. Curves show medians and $[10\%, 90\%]$ envelopes over 20 seeds.

Our contributions are as follows:

- Under generic initialization, nuisance persists in both encoders at every finite time; under the stated dimension conditions, trajectories almost surely preserve shared score signal while every off-signal score block vanishes asymptotically.
- We characterize the limiting scores and show that $d = r + 2$ is the tight threshold for a full-signal score and nuisance in both encoders; such optima have an open basin with uniformly positive nuisance norms.
- Controlled linear and nonlinear experiments test score–encoder separation. Separately, a real-data study shows that feature reconstruction, preservation of the teacher's comparisons, and label retrieval are distinct compression objectives.

2 Related Work

Representation-level theories characterize objective geometry or minimizers (Wang & Isola, 2020; Gui et al., 2025). Most directly, Almudévar et al. (2025) show that contrastive objectives need not remove modality-specific nuisance and regularize the encoders toward minimal, better-aligned representations. Their goal is encoder minimality; we keep the objective fixed and ask the distinct question of whether a *joint score* can become nuisance-insensitive while nuisance remains in its factors, analyzing factorized dynamics, score blocks, and output capacity rather than explicit nuisance removal. Dufumier et al. (2025) instead learn fused representations retaining redundant, unique, and synergistic information. Linear feature-recovery analyses (Ji et al., 2023; Nakada et al., 2023) do not determine the long-time two-factor, learned-temperature trajectory studied here. Appendix I gives detailed contrasts, and Appendix G.11 gives a controlled comparison with the alignment regularizer.

Dynamical analyses address spectral objectives, coordinate-wise learning, and infinite-width linearization (HaoChen et al., 2021; Tian, 2022; Simon et al., 2023), as well as encoder feature purity (Sun et al., 2025). Our score-level question differs from feature purity because nuisance can survive in both encoders while their cross-modal interactions vanish. Temperature and modality-gap dynamics have also been analyzed under a parallel-embedding parameterization (Yaras et al., 2025); that gap measures the distance between modality centers. Our centered model has zero population center gap throughout training, even as off-signal score blocks evolve. Appendix I gives further comparisons, including noisy-correspondence rectification (Yang et al., 2023).

3 PROBLEM SETUP

3.1 A NOISY TWO-MODALITY MODEL

The two modalities share a low-dimensional signal but also contain variation that is not shared across views. We use the Gaussian model

$$X = A\begin{bmatrix} z \\ \xi_x \end{bmatrix} + \varepsilon_x \in \mathbb{R}^{d_1}, \qquad Y = B\begin{bmatrix} z \\ \xi_y \end{bmatrix} + \varepsilon_y \in \mathbb{R}^{d_2}. \tag{1}$$

Here $z \sim \mathcal{N}(0, I_r)$ is the shared signal. The orthogonal input bases $A = [A_s\ A_\perp] \in \mathbb{R}^{d_1 \times d_1}$ and $B = [B_s\ B_\perp] \in \mathbb{R}^{d_2 \times d_2}$ separate the first r shared coordinates from the remaining modality-specific coordinates. The modality-specific variables and measurement noises have laws

$$\xi_x \sim \mathcal{N}(0, \nu^2 I_{d_1 - r}), \quad \xi_y \sim \mathcal{N}(0, \nu^2 I_{d_2 - r}), \quad \varepsilon_x \sim \mathcal{N}(0, \sigma^2 I_{d_1}), \quad \varepsilon_y \sim \mathcal{N}(0, \sigma^2 I_{d_2}).$$

All five sources are mutually independent. For output dimension d, the encoders satisfy $W_A \in \mathbb{R}^{d \times d_1}$, $W_B \in \mathbb{R}^{d \times d_2}$. We assume

$$\sigma^2, \nu^2 \geq 0, \quad \sigma^2 + \nu^2 > 0, \qquad d_1 > r \geq 1, \quad d_2 > r, \quad d \geq r.$$

The convention $d \geq r$ allows all shared directions to be represented; Theorem 1 and Corollary 3(i)–(ii) hold for every integer $d \geq 1$, independently of the input dimensions.

Write $\Sigma_x = \mathbb{E}[XX^\top] \in \mathbb{R}^{d_1 \times d_1}$ and $\Sigma_y = \mathbb{E}[YY^\top] \in \mathbb{R}^{d_2 \times d_2}$. Independence gives

$$\Sigma_{yx} = \mathbb{E}[YX^\top] = B_s A_s^\top \in \mathbb{R}^{d_2 \times d_1}. \tag{2}$$

Thus cross-view covariance is supported only on the shared coordinates. The Gaussian spiked model (Johnstone, 2001; Nakada et al., 2023) permits an explicit loss and flow; Appendix E.1 identifies where independence enters, and Appendix G.7 tests correlated variables.

3.2 SIMILARITY SCORE, INDIVIDUAL ENCODERS, AND SIGNAL–NUISANCE BLOCKS

For nonzero encoders $f(x) = W_A x$, $g(y) = W_B y$, define $N_A = \mathbb{E}\|W_A X\|_2$, $N_B = \mathbb{E}\|W_B Y\|_2$, and temperature $\tau > 0$. The similarity score is

$$S(X, Y) = \frac{Y^\top M X}{\tau N_A N_B}, \qquad M = W_B^\top W_A \in \mathbb{R}^{d_2 \times d_1}. \tag{3}$$

For a fixed trained model the positive scalar $1/(\tau N_A N_B)$ does not affect candidate rankings, but it matters for the training dynamics.

The population symmetrized InfoNCE objective is

$$\mathcal{L} = -2\mathbb{E}_{(X,Y)}[S(X, Y)] + \mathbb{E}_X \log \mathbb{E}_{Y'} e^{S(X,Y')} + \mathbb{E}_Y \log \mathbb{E}_{X'} e^{S(X',Y)}, \tag{4}$$

where (X, Y) is a positive pair and X', Y' are independent marginal draws. Expected-norm normalization (Gui et al., 2025) gives an analytically tractable population objective. Its normalizers are shared across observations but depend on the encoder parameters. Under the Gaussian model, the loss is quadratic in αM, where $\alpha = 1/(\tau N_A N_B)$. Appendix A.4 relates norm concentration to effective rank and compares the gradients with sample-level cosine normalization.

The encoder signal and nuisance blocks are

$$W_{A,s} = W_A A_s \in \mathbb{R}^{d \times r}, \qquad W_{A,\perp} = W_A A_\perp \in \mathbb{R}^{d \times (d_1 - r)}, \tag{5}$$

with $W_{B,s} = W_B B_s \in \mathbb{R}^{d \times r}$ and $W_{B,\perp} = W_B B_\perp \in \mathbb{R}^{d \times (d_2 - r)}$. The score blocks are

$$\widetilde{M} = B^\top M A = \begin{pmatrix} \widetilde{M}_{ss} & \widetilde{M}_{sn} \\ \widetilde{M}_{ns} & \widetilde{M}_{nn} \end{pmatrix}, \tag{6}$$

where

$$\widetilde{M}_{ss} = B_s^\top M A_s \in \mathbb{R}^{r \times r}, \qquad \widetilde{M}_{sn} = B_s^\top M A_\perp \in \mathbb{R}^{r \times (d_1 - r)},$$
$$\widetilde{M}_{ns} = B_\perp^\top M A_s \in \mathbb{R}^{(d_2 - r) \times r}, \qquad \widetilde{M}_{nn} = B_\perp^\top M A_\perp \in \mathbb{R}^{(d_2 - r) \times (d_1 - r)}.$$

The three off-signal blocks are the score-level nuisance interactions.

For example, $\widetilde{M}_{sn} = W_{B,s}^\top W_{A,\perp}$ can vanish with both factors nonzero when their output subspaces are orthogonal. We call removal of the three off-signal blocks *score-level denoising*; *encoder-level denoising* removes $W_{A,\perp}$ or $W_{B,\perp}$. Score-level denoising leaves measurement noise in the signal coordinates intact. Section 4.4 gives the capacity needed for a full-rank clean score with two nuisance-bearing encoders.

3.3 Population gradient flow

For the joint state $\theta = (W_A, W_B, \tau)$, we study Euclidean population gradient flow,

$$\dot{W}_A = -\nabla_{W_A}\mathcal{L}, \qquad \dot{W}_B = -\nabla_{W_B}\mathcal{L}, \qquad \dot{\tau} = -\partial_\tau \mathcal{L}. \tag{7}$$

Temperature is optimized directly, rather than through $\log\tau$. The state space is

$$\mathcal{U} = \{(W_A, W_B, \tau) : W_A \neq 0,\ W_B \neq 0,\ \tau > 0\}. \tag{8}$$

Let T_+ denote the maximal forward existence time. Appendix A.3 proves existence, uniqueness, and continuation on $\mathcal{U}$; for a global trajectory, set $\mathcal{L}^* = \lim_{t\to\infty}\mathcal{L}(t)$.

Set

$$\alpha(t) = \frac{1}{\tau(t)N_A(t)N_B(t)}, \qquad \widehat{M}(t) = \alpha(t)M(t) \in \mathbb{R}^{d_2\times d_1}. \tag{9}$$

The rescaled blocks are $\widehat{\widetilde{M}} = B^\top \widehat{M} A = \alpha\widetilde{M}$. Define the encoder signal and nuisance energies

$$\Phi(t) = \|W_{A,s}(t)\|_F^2 + \|W_{B,s}(t)\|_F^2, \qquad \Psi(t) = \|W_{A,\perp}(t)\|_F^2 + \|W_{B,\perp}(t)\|_F^2. \tag{10}$$

Positive encoder-rescaling invariance of $\mathcal{L}$ conserves $C_A = \|W_A(0)\|_F > 0$ and $C_B = \|W_B(0)\|_F > 0$ (Appendix B.1).

4 Main Results

Our main result separates the score from its encoder factors along the same learning trajectory. First, generic random initialization almost surely reaches a non-collapsing regime in which temperature stays bounded, shared score signal remains, and every off-signal score block vanishes asymptotically. Second, nuisance cannot disappear from either encoder at finite time. This generic statement does not require nuisance to remain positive in the limit. Third, when the output dimension reaches the tight threshold $d = r + 2$, a full-signal score can coexist with persistent nuisance in both encoders on an open basin of trajectories. Together, these results establish score–encoder separation dynamically and geometrically. The ranking and compression statements below are consequences of this central distinction.

4.1 Initialization and temperature regimes

Bounded temperature is first used as a sufficient condition for off-signal score removal. Subsequent results show that generic trajectories satisfy it while preserving shared signal.

Definition 1 (Generic initialization)**.** The initialization $(W_A(0), W_B(0))$ is *generic* if its distribution is absolutely continuous with respect to the induced surface measure on the product of Frobenius spheres

$$\{W_A \in \mathbb{R}^{d\times d_1} : \|W_A\|_F = C_A\} \times \{W_B \in \mathbb{R}^{d\times d_2} : \|W_B\|_F = C_B\}.$$

It therefore avoids each fixed proper algebraic subvariety of this product almost surely.

Independent Gaussian matrices rescaled to the prescribed Frobenius norms satisfy this condition.

We use the following temperature condition.

Definition 2 (Bounded-temperature regime)**.** A trajectory belongs to the bounded-temperature regime $\mathcal{B}_{\mathrm{bnd}}$ if

$$\mathcal{B}_{\mathrm{bnd}} = \left\{ T_+ = \infty, \quad \sup_{t\geq 0}\tau(t) < \infty, \quad \liminf_{t\to\infty}\tau(t) > 0 \right\}. \tag{11}$$

The flow is global and temperature has a positive lower bound (Theorem C.3). Every trajectory in $\mathcal{B}_{\mathrm{bnd}}$ converges to one state (Theorem B.9).

4.2 OFF-SIGNAL SCORE REMOVAL ALONG THE COUPLED FLOW

Optimizing the rescaled score matrix directly gives the minimizer $\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$, supported on the shared-signal block (Appendix E.3). This calculation identifies a desirable score, but does not show that the coupled encoder–temperature flow reaches one. The following theorem establishes off-signal block decay along that flow; convergence of the encoder factors is proved as part of the analysis.

Theorem 1 (Nuisance interactions vanish under bounded temperature)**.** *Under the noisy two-modality model* (1) *and the population gradient flow* (7)*, on every trajectory in* $\mathcal{B}_{\mathrm{bnd}}$*,*

$$\|\widetilde{M}_{sn}(t)\|_F \to 0, \qquad \|\widetilde{M}_{ns}(t)\|_F \to 0, \qquad \|\widetilde{M}_{nn}(t)\|_F \to 0.$$

Thus the score asymptotically removes all three off-signal nuisance interactions.

The mechanism: decay through the other encoder. Define

$$K_B(t) = W_B(t)\Sigma_y W_B(t)^\top \in \mathbb{R}^{d\times d}, \qquad R_A(t) = \nabla_{W_A} N_A(W_A(t)) \in \mathbb{R}^{d\times d_1},$$

and let $\Delta(t) = \tau(t)\dot{\tau}(t)$. The identities $\Sigma_{yx}A_\perp = 0$ and $\Sigma_x A_\perp = (\sigma^2+\nu^2)A_\perp$ give

$$\dot{W}_{A,\perp} = -c(t)K_B(t)W_{A,\perp} + \frac{\Delta(t)}{N_A(t)}R_A(t)A_\perp, \qquad c(t) = \frac{2(\sigma^2+\nu^2)}{(\tau(t)N_A(t)N_B(t))^2}.$$

The matrix $K_B(t)$ can be singular and changes during training; the normalizer term also remains. The vanishing cross-covariance source therefore leaves a coupled equation.

On $\mathcal{B}_{\mathrm{bnd}}$, Theorem B.9 gives convergence to one stationary state, including at rank-deficient encoders. At that state $\Delta = 0$ and $c > 0$, so the nuisance equation gives $K_B(t)W_{A,\perp}(t) \to 0$. Taking its Frobenius inner product with the bounded matrix $W_{A,\perp}(t)$ yields

$$\|\Sigma_y^{1/2}W_B^\top W_{A,\perp}\|_F^2 = \langle W_{A,\perp}, K_B W_{A,\perp}\rangle_F \to 0.$$

Positive definiteness of Σ_y gives $W_B^\top W_{A,\perp} = MA_\perp \to 0$, hence $\widetilde{M}_{sn}, \widetilde{M}_{nn} \to 0$; the symmetric argument gives $\widetilde{M}_{ns} \to 0$. Any surviving nuisance columns lie in the kernel of the other limiting encoder's transpose, making them invisible to the score. Appendices A and B.7 provide the regularity estimates and complete proof.

4.3 ENCODER PERSISTENCE AND PRESERVATION OF SHARED SIGNAL

Off-signal block decay alone also permits the entire score to collapse to zero. Negative limiting loss rules out this uninformative outcome. The next results establish almost-sure entry into this regime and distinguish the surviving score signal from the nuisance retained by each encoder.

Definition 3 (Regular non-collapsing regime)**.** The regular non-collapsing regime is

$$\mathcal{B}_{\mathrm{reg}} = \left\{ T_+ = \infty, \quad \mathcal{L}^* < 0, \quad \liminf_{t\to\infty} \tau(t) > 0 \right\}. \tag{12}$$

On a global trajectory, $\mathcal{L}^* < 0$ implies $\sup_t \tau(t) < \infty$ (Lemma B.15), so $\mathcal{B}_{\mathrm{reg}} \subset \mathcal{B}_{\mathrm{bnd}}$. This loss condition selects a nontrivial cross-view score without imposing any block-decay condition.

Corollary 2 (Almost-sure negative-loss entry)**.** *Suppose* $r \geq 1$*,* $d \geq 2$*, and* $d_1 - r, d_2 - r \geq 2$*. Fix* $C_A, C_B > 0$ *and any* $\tau_0 > 0$*. Under generic encoder initialization, the loss becomes negative at a finite time almost surely. Consequently the trajectory belongs to* $\mathcal{B}_{\mathrm{reg}}$*.*

Appendix C.9 proves entry by excluding both zero-loss convergence and infinite-temperature escape, including rank-deficient limits.

Corollary 3 (Encoder persistence and score–encoder separation)**.** *Under the noisy two-modality model* (1) *and the population gradient flow* (7) *with generic initialization (Definition 1), almost surely:*

(i) $\Phi(t) > 0$ *for every* $0 \leq t < T_+$*;*
(ii) $\|W_{A,\perp}(t)\|_F > 0$ *and* $\|W_{B,\perp}(t)\|_F > 0$ *for every* $0 \leq t < T_+$*.*

Moreover, on $\mathcal{B}_{\mathrm{reg}}$,

(iii) $\sup_{t\geq 0}\tau(t) < \infty$*;*
(iv) $\liminf_{t\to\infty}\Phi(t) > 0$*;*
(v) $\|\widetilde{M}_{sn}(t)\|_F \to 0$*,* $\|\widetilde{M}_{ns}(t)\|_F \to 0$*, and* $\|\widetilde{M}_{nn}(t)\|_F \to 0$*.*

Encoder nuisance persists because each nuisance-free set is invariant under the flow. Forward–backward Osgood uniqueness prevents a trajectory initialized outside such a set from reaching it at finite time, and generic initialization avoids both sets almost surely (Appendix B.4). This permits asymptotic encoder denoising. The open basin below establishes a stronger long-time separation: both nuisance norms remain bounded away from zero while the score converges to a full-signal optimum.

Corollary 4 (Signal preservation on the regular non-collapsing regime)**.** *Under the noisy two-modality model* (1) *and the population gradient flow* (7)*, on* $\mathcal{B}_{\mathrm{reg}}$*, every subsequential limit* $\widehat{M}^*$ *of* $\widehat{M}(t) = \alpha(t)M(t)$ *satisfies* $\|\widehat{M}^*\|_F \geq -\mathcal{L}^*/(2\|\Sigma_{yx}\|_F) > 0$*; in particular,* $\widehat{M}(t) \not\to 0$*.*

Non-collapse guarantees that some shared score signal survives; the stationary equations further identify which shared directions survive. Write the limiting state as $(\bar{W}_A, \bar{W}_B, \bar{\alpha})$. Every regular score limit has $\widehat{M}^* = (1+\sigma^2)^{-2} B_s \Pi A_s^\top$, where

$$\Pi = (1+\sigma^2)^2 \bar{\alpha}(\bar{W}_B B_s)^\top (\bar{W}_A A_s) = QQ^\top,$$
$$Q \in \mathbb{R}^{r\times k}, \qquad Q^\top Q = I_k, \qquad k = -(1+\sigma^2)^2\mathcal{L}^* \geq 1.$$

Thus Π projects onto the k retained shared directions. If $-\mathcal{L}^* > (r-1)/(1+\sigma^2)^2$, then $k = r$ and the score is the full-signal optimum (Proposition B.21).

4.4 Capacity for nuisance-bearing encoders

Finite-time nuisance persistence leaves open what survives in the limit. A full-rank clean score uses every shared direction; retaining nuisance in both encoders then requires additional output space. The following threshold determines when these properties can coexist.

Proposition 5 (Capacity for two nuisance-bearing encoders)**.** *Suppose* $\widetilde{M}_{sn} = \widetilde{M}_{ns} = \widetilde{M}_{nn} = 0$*,* $\mathrm{rank}(\widetilde{M}_{ss}) = r$*, and both* $W_{A,\perp}$ *and* $W_{B,\perp}$ *are nonzero. Then* $d \geq r+2$*. Conversely, such a configuration exists for every* $d \geq r+2$*.*

At $d = r+2$, the two encoders can share an r-dimensional signal subspace and place nuisance on two mutually orthogonal directions. Propositions E.1–E.2 construct optimal scores with this geometry and an open basin in which both nuisance norms stay uniformly positive. At $d = r$, a full-rank clean score forces both nuisance factors to vanish; at $d = r+1$, at most one can survive. On $\mathcal{B}_{\mathrm{bnd}}$, a uniformly full-rank signal block turns these geometric constraints into asymptotic encoder denoising. Appendix E.2 proves the trajectory result, and Appendix G.5 reports the corresponding pattern at $d = 3, 4, 5$ for $r = 3$.

4.5 A ranking consequence of removing nuisance interactions

Let $D = S(X_i, Y_i) - S(X_i, Y_j)$ compare a matched pair against an independent marginal distractor. Gaussian moments give $\mathbb{E}D = \alpha\,\mathrm{tr}(\widetilde{M}_{ss})$ and $\mathrm{Var}(D) = \alpha^2(V_{\mathrm{sig}} + V_{\mathrm{nui}})$, where

$$V_{\mathrm{sig}} = 2(1+\sigma^2)^2\|\widetilde{M}_{ss}\|_F^2 + \mathrm{tr}(\widetilde{M}_{ss}^2),$$
$$V_{\mathrm{nui}} = 2(1+\sigma^2)(\sigma^2+\nu^2)\big(\|\widetilde{M}_{sn}\|_F^2 + \|\widetilde{M}_{ns}\|_F^2\big) + 2(\sigma^2+\nu^2)^2\|\widetilde{M}_{nn}\|_F^2.$$

Negative loss implies $\mathrm{tr}(\widetilde{M}_{ss}) > 0$, and Cantelli gives

$$\Pr(D > 0) \geq \frac{\mathrm{tr}(\widetilde{M}_{ss})^2}{\mathrm{tr}(\widetilde{M}_{ss})^2 + V_{\mathrm{sig}} + V_{\mathrm{nui}}}.$$

At a fixed signal block, eliminating the three off-signal blocks sets $V_{\mathrm{nui}} = 0$ and improves the certificate; Appendix E.5 gives the exact rank-k limiting probability. Thus the off-signal blocks leave the mean matched–distractor gap unchanged but inflate its variance. Removing them can improve ranking reliability even when nuisance remains in both encoders.

4.6 A DOWNSTREAM CONSEQUENCE FOR REPRESENTATION COMPRESSION

Score–encoder separation implies that reconstructing the individual encoders can be strictly harder than preserving their joint comparison. The following fixed-state proposition quantifies this dimension gap. It is a consequence of the separation geometry, not an additional assumption used by the trajectory results.

Proposition 6 (Score-preserving and reconstructive compression)**.** *At a fixed state with three vanishing off-signal blocks and a rank-r signal block, put $q_A = \mathrm{rank}(W_{A,\perp})$ and $q_B = \mathrm{rank}(W_{B,\perp})$. With the original normalizers and temperature fixed, the minimum output width of two linear postprocessing maps that preserve every bilinear score is r. Exact linear reconstruction of the individual encoder outputs instead requires widths $r + q_A$ and $r + q_B$.*

With nuisance present, exact encoder reconstruction must allocate dimensions to information ignored by every cross-modal score. Appendix F proves the proposition and constructs a global optimum in which PCA retains arbitrarily close to all feature variance while deleting every cross-modal score. Thus feature-variance retention does not certify comparison preservation.

Section 5.3 tests this design principle on pretrained cosine embeddings by jointly fitting two maps to preserve bidirectional comparison distributions. The population proposition does not prescribe this finite-gallery optimizer or guarantee its approximation error. The experiment therefore evaluates a downstream implication rather than extending the nuisance theorem to pretrained CLIP features. Appendix F also gives an observable KL certificate for preserving a query's top-ranked candidate.

5 EXPERIMENTS

The experiments test three claims at progressively greater distance from the population theory. The controlled linear experiments test the predicted block-level separation directly. Nonlinear nuisance resampling asks whether more reliable comparisons can coexist with encoder sensitivity to the same variation beyond linear encoders. The real-data compression study tests only the downstream distinction between preserving features and preserving comparisons; it does not identify latent nuisance in pretrained features. Correlations and weight decay are examined in Appendices G.7 and D.

5.1 CONTROLLED LINEAR DYNAMICS

This experiment directly tests whether the score blocks and encoder nuisance follow the different trajectories predicted by the theory.

Block indicators. For the linear model, define

$$C_{\mathrm{lin}} = \frac{\|\widetilde{M}_{sn}\|_F^2 + \|\widetilde{M}_{ns}\|_F^2 + \|\widetilde{M}_{nn}\|_F^2}{\|\widetilde{M}_{ss}\|_F^2 + \epsilon},$$
$$I_{\mathrm{lin}} = \frac{1}{2}\left(\frac{\|W_{A,\perp}\|_F^2}{\|W_{A,s}\|_F^2 + \epsilon} + \frac{\|W_{B,\perp}\|_F^2}{\|W_{B,s}\|_F^2 + \epsilon}\right), \qquad \epsilon = 10^{-30}. \tag{13}$$

Score-block annihilation and signal preservation give $C_{\mathrm{lin}} \to 0$ on $\mathcal{B}_{\mathrm{reg}}$. In the open basin of Proposition E.2, conserved encoder norms and uniformly positive nuisance norms give $\inf_t I_{\mathrm{lin}}(t) > 0$, hence $C_{\mathrm{lin}}/I_{\mathrm{lin}} \to 0$. We also record component norms to distinguish nuisance attenuation from signal growth.

Population flow and finite-data training. Using the model parameters in Figure 1, we initialize Frobenius-normalized Gaussian encoders and approximate population expected norms with $50{,}000$ fixed Monte Carlo samples. Figure 1 summarizes 20 seeds; Figure 2 summarizes 10 runs per protocol for population flow and finite-data one-way cosine InfoNCE training with Adam and fixed $\tau = 0.5$.

Figure 2 shows lower score than encoder nuisance in both population flow and finite-sample training at $n = 20{,}000$, with greater separation at larger sample sizes. A separate three-seed population run confirms that the off-signal numerator contracts: its block norm falls from 0.196 to 1.8×10^{-6} by

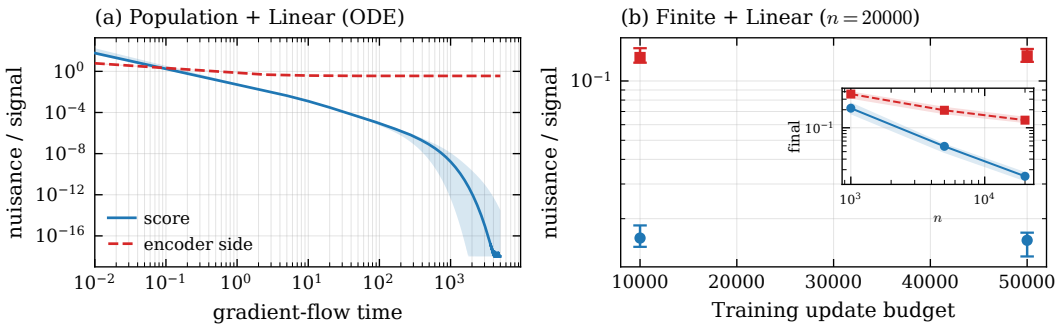

Figure 2: **Score and encoder blocks in the controlled linear model.** Both panels plot $C_{\text{lin}}, I_{\text{lin}}$ from Eq. (13). (a) Population gradient flow. (b) Finite-data measurements at two training budgets for $n = 20{,}000$, with a final-budget inset for $n \in \{1000, 5000, 20000\}$. Medians and $[10\%, 90\%]$ intervals summarize 10 runs per setting.

$t = 2000$, while the signal-block norm remains 0.42. Appendix G.5 examines how this behavior depends on output capacity.

Temperature and encoder allocation. All 980 runs in an output-dimension and measurement-noise sweep end with negative loss; recorded temperatures stay above 10^{-2}, and median nuisance–nuisance block reductions range from 3.1 to 10.5 orders of magnitude across settings (Appendix G.4). Capacity permits nuisance-bearing optima without requiring asymptotic nuisance retention on every trajectory. At $d = 25$, median nuisance energy at $T = 5000$ is 0.528, 1.27×10^{-3}, and 2.16×10^{-18} for initial temperatures $1, 3, 10$, respectively (Appendix G.3).

5.2 Nonlinear comparison and representation variability

This experiment tests the qualitative separation beyond linear encoders by intervening on nuisance and measuring both comparison reliability and encoder variability.

Pairwise reliability under nuisance resampling. Eighteen continuous-latent runs use six depths and three seeds per depth with fixed $\tau = 0.5$. Holding the shared signal and measurement noise fixed, we resample modality nuisance in query/positive/distractor triples. Mean cosine pairwise accuracy rises from 50.2% before training to 80.1%, while the probability of an ordering reversal across resamples falls from 38.3% to 4.6%. Every run improves both quantities. Resampling measurement noise as well gives the same direction of change in all runs. Encoder variability under the same nuisance-only draws decreases during training and remains nonzero at every final checkpoint. All eighteen final models have a lower normalized score indicator than encoder indicator, with ratios 0.390–0.926 (Appendix G.10).

A separate evaluation with 200 fresh signal groups reproduces the accuracy gains and reversal reductions in every run under both interventions. Group-bootstrap intervals support these changes at the fixed checkpoints (Appendix G.10). Thus the same nuisance intervention can continue to move both representations while becoming much less able to reverse their comparison.

Retrieval across datasets and depths. A complementary study uses bias-free MLPs with 2–10 hidden layers on CITE-seq RNA/ADT and frozen COCO CLIP ViT-B/32 features. Across two recipes and five seeds per setting, all 100 runs improve class and exact-pair Recall@5 from initialization; every final model also exceeds chance for within-class pair retrieval. Relative to squared conditional means, score variability is lower than encoder variability. These class-conditioned measurements include fine-grained biology or semantics and do not measure retrieval accuracy. Appendices E.7 and G.6 define and operationalize the indicators; Appendix G.8 gives the protocol and depth-wise results (Table 6).

Table 1: **Three compression objectives on the held-out cohorts.** R^2 is image/text affine reconstruction; mean R@1 and exact teacher top-1 agreement are percentages. CCA is validation-selected, and Comparisons reports the mean over three fixed maps.

Dataset	Method	R^2 I/T	Mean R@1	Teacher top-1
COCO	Pooled PCA	0.643/0.768	31.95	45.94
COCO	Regularized CCA	0.542/0.769	**43.28**	30.45
COCO	Comparisons	0.602/0.746	42.25	**60.06**
Flickr30k	Pooled PCA	0.583/0.711	60.64	59.93
Flickr30k	Regularized CCA	0.469/0.690	**76.60**	52.18
Flickr30k	Comparisons	0.541/0.680	74.00	**72.06**

5.3 A DOWNSTREAM TEST: PRESERVING COMPARISONS UNDER A STORAGE BUDGET

On frozen CLIP ViT-L/14 COCO and Flickr30k features (Radford et al., 2021), we fit two modality-specific maps on 10,000 calibration images and compare them with validation-selected pooled PCA at widths 120 and 96. Neither receives latent ranks or nuisance labels. This tests a design consequence, not the latent model; Appendix H gives the protocol.

On additional disjoint cohorts of 5,000 COCO and 1,000 Flickr30k images, the comparison maps improve mean R@1 over PCA by 10.30 and 13.36 points at 6.4- and 8-fold vector-storage reductions; every seed retains at least 90% of original R@1 in each direction. Paired 95% intervals for the three-map mean are $[9.60, 11.00]$ and $[11.87, 14.85]$ points (Table 18).

Affine decoders fitted on calibration data show that PCA reconstructs both modalities better on both datasets, despite substantially worse retrieval (Table 1); PCA + Procrustes and cross-covariance SVD show the same higher-reconstruction, lower-retrieval pattern. Conversely, validation-selected regularized CCA attains higher label R@1 than the comparison maps but preserves the teacher's top-ranked candidate substantially less often. The study therefore does not claim a universally best compressor: encoder reconstruction, teacher-comparison preservation, and label retrieval are distinct objectives (Table 17).

6 DISCUSSION

Multimodal contrastive learning can remove nuisance interactions from a score without denoising either encoder. The trajectory analysis explains this separation, and the capacity result identifies when it persists at a full-signal optimum. Controlled experiments support the distinction; comparison-preserving compression shows that feature reconstruction need not preserve retrieval comparisons, while the CCA result separates label retrieval from fidelity to the teacher's rule.

The relevant notion of denoising therefore depends on the deployment object. Retrieval exposes the joint score, whereas reconstruction or single-modality transfer exposes the encoder outputs themselves. Score audits and representation audits are complementary: the former assess comparison reliability, while the latter detect information that remains available to downstream users.

The capacity threshold sharpens this distinction. Extra output dimensions can form private subspaces that store modality-specific variation without affecting cross-modal scores. When the output width equals the shared-signal rank, by contrast, a full-signal clean score leaves no room for private directions. Output capacity therefore controls whether reliable similarity also forces encoder denoising.

Limitations. The theory assumes linear encoders, Gaussian data, and expected-norm population InfoNCE; general score-decay rates and nonlinear convergence remain open, and empirical-flow stability is conditional on Appendix E.4. Generic initialization prevents exact finite-time denoising, but positive limiting nuisance is guaranteed only on the open basins of Proposition E.2. Nonlinear resampling supports finite-checkpoint separation; the frozen-CLIP study neither identifies latent nuisance nor establishes general downstream benefits.

AI USE STATEMENT

In this work, we used generative AI tools as an interactive aid in the theoretical development of the paper. The authors formulated the main research questions and mathematical results. Generative AI was used to refine mathematical claims and check arguments for possible gaps or errors. The proofs and final theoretical results were developed and verified by the authors, who critically assessed and, where appropriate, modified AI-assisted suggestions. Generative AI was also used during the writing process to help organize, articulate, and polish mathematical arguments. The final presentation and text were revised and approved by the authors.

For the experiments, the authors determined the research questions, experimental goals, and overall experimental directions. Generative AI was used interactively to suggest possible synthetic and real-data experimental designs and variants. The authors evaluated these suggestions, modified them as needed, and selected the experiments reported in the paper. Generative AI also assisted with portions of the code implementation. All AI-assisted code was inspected, tested, and verified by the authors, and the resulting experimental outputs were independently checked and validated by the authors.

We did not use generative AI for translation. The literature review was led by the authors, with generative AI used interactively to help identify additional potentially relevant literature. All cited works were checked by the authors against the original sources. Generative AI was also used for drafting, editing, and polishing portions of the manuscript for clarity and readability. All AI-assisted work was critically reviewed by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Appendices A–F provide the mathematical assumptions, differential identities, proofs, model consequences, and compression propositions. Appendices G and H specify the experimental protocols, data preprocessing, relevance criteria, numerical indicators, validation-selected linear baselines, and held-out evaluations. The source package includes the LaTeX manuscript, figure assets, indicator implementation, and tests of its conditional-moment identities, normalization, weighting, and numerical floors.

REFERENCES

Antonio Almudévar, José Miguel Hernández-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. Aligning multimodal representations through an information bottleneck. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 1250–1270, 2025.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Zhidong Bai and Jianfeng Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177, 2012.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Raf Cluckers and Daniel J. Miller. Stability under integration of sums of products of real globally subanalytic functions and their logarithms. *Duke Mathematical Journal*, 156(2):311–348, 2011.

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025.

Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, 2016.

Yu Gui, Cong Ma, and Zongming Ma. Multi-modal contrastive learning adapts to intrinsic dimensions of shared latent variables. In *Advances in Neural Information Processing Systems*, 2025.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, 2021.

Philip Hartman. *Ordinary Differential Equations*, volume 38 of *Classics in Applied Mathematics*. SIAM, 2 edition, 2002.

Morris W. Hirsch, Charles C. Pugh, and Michael Shub. Invariant manifolds. *Bulletin of the American Mathematical Society*, 76(5):1015–1019, 1970.

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

Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *Journal of Machine Learning Research*, 24, 2023.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

Sham M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysis. In *Conference on Learning Theory*, 2007.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48(3):769–783, 1998.

Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176:311–337, 2019.

Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. In *Advances in Neural Information Processing Systems*, 2021.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden (eds.), *Handbook of Econometrics*, volume 4, pp. 2111–2245. Elsevier, 1994.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

James B. Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J. Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Leon Simon. Asymptotics for a class of nonlinear evolution equations, with applications to geometric problems. *Annals of Mathematics*, 118(3):525–571, 1983.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Advances in Neural Information Processing Systems*, 2021.

Jiawei Sun, Shuai Zhang, Hongkang Li, and Meng Wang. Contrastive learning with data misalignment: Feature purity, training dynamics and theoretical generalization guarantees. In *Advances in Neural Information Processing Systems*, 2025.

Yuandong Tian. Understanding deep contrastive learning via coordinate-wise optimization. In *Advances in Neural Information Processing Systems*, 2022.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, 2021.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

Lou van den Dries, Angus Macintyre, and David Marker. The elementary theory of restricted analytic fields with exponentiation. *Annals of Mathematics*, 140(1):183–205, 1994.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Yihao Xue, Eric Gan, Jiayi Ni, Siddharth Joshi, and Baharan Mirzasoleiman. Investigating the benefits of projection head for representation learning. In *International Conference on Learning Representations*, 2024.

Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. BiCro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. In *Second Conference on Parsimony and Learning*, 2025.

APPENDIX CONTENTS

The proofs of the trajectory results are in Appendices A–C. The theoretical development continues through Appendix F; the experimental protocols and results are grouped in Appendices G–H. All section and page references below are linked.

Section	Contents	Page
A	**Loss, gradients, and regularity.** Gaussian reduction; expected-norm calculus; existence and uniqueness of the flow.	15
B	**Main trajectory proofs.** Conserved norms; finite-time non-entry; temperature and signal bounds; nuisance-score annihilation.	21
C	**Generic regular entry.** Temperature control, avoidance of zero-loss limits including rank loss, and almost-sure non-collapse.	30
D	**Encoder weight decay.** Radial contraction, the temperature time scale, and the signal-to-nuisance ratio.	46
E	**Further theory and model consequences.** Covariance support; capacity; ranking; finite-sample stability; temperature; conditional moments.	49
F	**Compression consequences.** Score-preserving dimension, variance-retention counterexample, and a finite-gallery ranking certificate.	59
G	**Empirical measurements and robustness.** Controlled linear sweeps; nonlinear indicators; correlations; depth; nuisance interventions; encoder alignment.	61
H	**Retrieval compression experiments.** Fitting, validation-selected baselines, held-out evaluation, reconstruction, and teacher agreement.	74
I	**Additional related work.** Statistical recovery and representation-learning dynamics.	79

Direct entry points. For the score-block proof, see §B.7; for finite-time encoder persistence, see §B.4. The capacity proof is in §E.2. The nonlinear indicators are interpreted in §E.7 and computed in §G.6; the depth-study protocol and all its results are in §G.8. Controlled nuisance interventions and fresh-draw validation are in §G.10. The direct comparison with explicit encoder-alignment regularization is in §G.11.

A Loss, Gradients, and Regularity

We first derive the loss as a quadratic function of the rescaled score matrix. Differentiating this expression gives the coupled encoder–temperature equations. The expected-norm derivatives remain well behaved at every nonzero encoder, including rank-deficient ones; §A.3 proves the modulus of continuity used in the trajectory arguments.

A.1 Closed-form loss and proof notation

For a fixed state, write $\alpha = (\tau N_A N_B)^{-1}$ and $M = W_B^\top W_A$. The score is $S(X,Y) = \alpha Y^\top M X$. Define

$$\mathcal{A} = \|\Sigma_y^{1/2} M \Sigma_x^{1/2}\|_F^2, \qquad \mathcal{B} = \operatorname{tr}(M^\top \Sigma_{yx}).$$

The coefficient $\mathcal{A}$ collects the marginal second moments, whereas $\mathcal{B}$ is the cross-view alignment term. We will show that

$$\mathcal{L}(W_A, W_B, \tau) = \mathcal{A}\alpha^2 - 2\mathcal{B}\alpha. \tag{14}$$

Matched pairs. The cross-covariance identity $\mathbb{E}[YX^\top] = \Sigma_{yx}$ gives

$$\mathbb{E}_{(X,Y)} S(X,Y) = \alpha \operatorname{tr}\big(M^\top \mathbb{E}[YX^\top]\big) = \alpha\mathcal{B}. \tag{15}$$

Independent candidates. Although the bilinear score of two independent Gaussian vectors is not itself Gaussian, it becomes Gaussian after fixing the anchor. Since Y' is independent of X,

$$S(X,Y') \mid X = x \sim \mathcal{N}\big(0, \alpha^2 x^\top M^\top \Sigma_y M x\big).$$

The moment-generating function of a centered scalar Gaussian with variance v is $\exp(v/2)$ at argument one. Hence

$$\mathbb{E}_{Y'} e^{S(X,Y')} = \exp\big(\tfrac{1}{2}\alpha^2 X^\top M^\top \Sigma_y M X\big).$$

Taking the logarithm before averaging over X yields

$$\begin{aligned} \mathbb{E}_X \log \mathbb{E}_{Y'} e^{S(X,Y')} &= \tfrac{1}{2}\alpha^2 \mathbb{E}_X[X^\top M^\top \Sigma_y M X] \\ &= \tfrac{1}{2}\alpha^2 \operatorname{tr}(M^\top \Sigma_y M \Sigma_x) = \tfrac{1}{2}\alpha^2 \mathcal{A}. \end{aligned} \tag{16}$$

For the other anchor direction,

$$S(X',Y) \mid Y = y \sim \mathcal{N}\big(0, \alpha^2 y^\top M \Sigma_x M^\top y\big),$$

so

$$\mathbb{E}_Y \log \mathbb{E}_{X'} e^{S(X',Y)} = \tfrac{1}{2}\alpha^2 \operatorname{tr}(M \Sigma_x M^\top \Sigma_y) = \tfrac{1}{2}\alpha^2 \mathcal{A}.$$

Combining the two candidate terms and the matched-pair term in Eq. (4) proves Eq. (14).

Covariances and score coordinates. The input bases diagonalize the marginal covariances:

$$\Sigma_x = A D_x A^\top, \qquad \Sigma_y = B D_y B^\top,$$

where

$$\begin{aligned} D_x &= \operatorname{diag}\big((1+\sigma^2) I_r, (\sigma^2+\nu^2) I_{d_1 - r}\big), \\ D_y &= \operatorname{diag}\big((1+\sigma^2) I_r, (\sigma^2+\nu^2) I_{d_2 - r}\big). \end{aligned}$$

Thus both covariances have eigenvalues at least $\lambda_0 = \min\{1+\sigma^2, \sigma^2+\nu^2\} > 0$. For $Z \in \mathbb{R}^{d_2 \times d_1}$, define

$$F(Z) = \|\Sigma_y^{1/2} Z \Sigma_x^{1/2}\|_F^2, \qquad G(Z) = \operatorname{tr}(Z^\top \Sigma_{yx}).$$

In particular,

$$\mathcal{L} = F(\widehat{M}) - 2G(\widehat{M}), \qquad \widehat{M} = \alpha M, \qquad \Delta = 2\alpha(\mathcal{A}\alpha - \mathcal{B}) = 2\big(F(\widehat{M}) - G(\widehat{M})\big).$$

We use $\eta = 1/\tau$ when convenient. The temperature equation below gives $\dot\tau = \Delta/\tau$, so this definition of Δ agrees with $\tau\dot\tau$ in the main text.

Encoder quantities. Write $G_A = W_A^\top W_A \in \mathbb{R}^{d_1 \times d_1}$ and $G_B = W_B^\top W_B \in \mathbb{R}^{d_2 \times d_2}$. In output space, use

$$Q_A = W_A W_A^\top, \quad Q_B = W_B W_B^\top, \qquad K_A = W_A \Sigma_x W_A^\top, \quad K_B = W_B \Sigma_y W_B^\top.$$

Thus G_A, G_B act on input coordinates, whereas Q_A, Q_B and K_A, K_B act on output coordinates. The product of the prescribed Frobenius spheres is

$$\mathcal{M} = \{(W_A, W_B) \in \mathbb{R}^{d \times d_1} \times \mathbb{R}^{d \times d_2} : \|W_A\|_F = C_A,\ \|W_B\|_F = C_B\}.$$

The expected-norm gradients are

$$R_A(W_A) = \nabla_{W_A} N_A(W_A) = \mathbb{E}\left[\frac{W_A X X^\top}{\|W_A X\|_2}\right] \in \mathbb{R}^{d \times d_1},$$
$$R_B(W_B) = \nabla_{W_B} N_B(W_B) = \mathbb{E}\left[\frac{W_B Y Y^\top}{\|W_B Y\|_2}\right] \in \mathbb{R}^{d \times d_2}.$$

Lemma A.2 justifies these derivatives. We abbreviate them to R_A, R_B along a trajectory. Conservation of $\mathcal{M}$ and the associated normalizer bounds are established together in §B.1.

Table 2: Quantities used in the trajectory proofs. The input bases A, B are distinct from the scalar loss coefficients $\mathcal{A}, \mathcal{B}$.

Quantity	Role
$M = W_B^\top W_A$	Unscaled bilinear score matrix; $\widetilde{M} = B^\top M A$ contains its signal/nuisance blocks.
$\widehat{M} = \alpha M$	Rescaled score matrix appearing in the loss $F(\widehat{M}) - 2G(\widehat{M})$.
$N_A, N_B;\ R_A, R_B$	Expected encoder norms and their matrix gradients.
$\mathcal{A}, \mathcal{B};\ \Delta$	Quadratic loss coefficient, alignment coefficient, and temperature drift.
$G_A, G_B;\ Q_A, Q_B$	Input-space and output-space encoder Gram matrices.
K_A, K_B	Output covariances, including the marginal input covariances.
$\eta = 1/\tau;\ \Phi, \Psi$	Inverse temperature; signal and nuisance energies.
$\mathcal{M}$	Product of the two fixed-norm encoder spheres.
$\mathcal{B}_{\mathrm{bnd}}, \mathcal{B}_{\mathrm{reg}}$	Trajectory conditions from Definitions 2 and 3.

A.2 Gradient-flow identities

The gradient identities include contributions from both the score product and its expected-norm normalizers.

Lemma A.1 (Gradient-flow identities)**.** *Let* $\Sigma_{xy} = \Sigma_{yx}^\top$ *and*

$$P = \Sigma_{yx} - \Sigma_y \widehat{M} \Sigma_x, \qquad P^\top = \Sigma_{xy} - \Sigma_x \widehat{M}^\top \Sigma_y, \qquad \widehat{M} = \alpha M.$$

Along the gradient flow on its maximal interval,

$$\dot{W}_A = \frac{2}{\tau N_A N_B} W_B P + \frac{\Delta}{N_A} R_A, \qquad \dot{W}_B = \frac{2}{\tau N_A N_B} W_A P^\top + \frac{\Delta}{N_B} R_B,$$

and

$$\dot{\tau} = \frac{\Delta}{\tau}, \qquad \Delta = 2\alpha(\mathcal{A}\alpha - \mathcal{B}).$$

Proof. For fixed α, the product derivatives are

$$\nabla_{W_A} \mathcal{A} = 2 W_B \Sigma_y M \Sigma_x, \qquad \nabla_{W_A} \mathcal{B} = W_B \Sigma_{yx},$$

and symmetrically

$$\nabla_{W_B} \mathcal{A} = 2 W_A \Sigma_x M^\top \Sigma_y, \qquad \nabla_{W_B} \mathcal{B} = W_A \Sigma_{xy}.$$

Since $\alpha = (\tau N_A N_B)^{-1}$,

$$\nabla_{W_A}\alpha = -\frac{\alpha}{N_A}R_A, \qquad \nabla_{W_B}\alpha = -\frac{\alpha}{N_B}R_B, \qquad \partial_\tau \alpha = -\frac{\alpha}{\tau}.$$

Using $\mathcal{L} = \mathcal{A}\alpha^2 - 2\mathcal{B}\alpha$, we obtain

$$\nabla_{W_A}\mathcal{L} = 2\alpha W_B(\alpha\Sigma_y M\Sigma_x - \Sigma_{yx}) - \frac{2\alpha(\mathcal{A}\alpha - \mathcal{B})}{N_A}R_A.$$

Negating the gradient and substituting $\widehat{M} = \alpha M$ and $\Delta = 2\alpha(\mathcal{A}\alpha - \mathcal{B})$ gives the displayed formula for $\dot{W}_A$. The $\dot{W}_B$ formula is identical with $A \leftrightarrow B$ and $x \leftrightarrow y$. Finally,

$$\partial_\tau \mathcal{L} = (2\mathcal{A}\alpha - 2\mathcal{B})\partial_\tau\alpha = -\frac{2\alpha(\mathcal{A}\alpha - \mathcal{B})}{\tau} = -\frac{\Delta}{\tau},$$

so $\dot{\tau} = -\partial_\tau\mathcal{L} = \Delta/\tau$. □

A.3 ODE WELL-POSEDNESS AND EXPECTED-NORM REGULARITY

We use Euclidean norms and inner products, with $\|(H_A, H_B, h)\|^2 = \|H_A\|_F^2 + \|H_B\|_F^2 + h^2$, and induced operator norms for derivatives. The maximal forward time is

$$T_+ = \sup\big\{T \geq 0 : \text{a unique solution to (7) exists on } [0, T)\big\}. \tag{17}$$

The expected-norm normalizer is differentiable at every nonzero encoder, including rank-deficient matrices. Its gradient satisfies the Osgood modulus used for uniqueness and continuation.

EXPECTED-NORM CALCULUS

We first justify differentiating the expected norm. The only exceptional inputs satisfy $WX = 0$; for a nonzero encoder these have probability zero.

Lemma A.2 (Regularity of expected-norm normalization)**.** *Let $X \sim \mathcal{N}(0, \Sigma_x)$ with $\Sigma_x \succ 0$, and define $N(W) = \mathbb{E}[\|WX\|_2]$ for $W \in \mathbb{R}^{d\times d_1}$. Then on the open set $\{W \in \mathbb{R}^{d\times d_1} : W \neq 0\}$:*

(i) $N(W) > 0$;
(ii) N is continuously differentiable, with $\nabla_W N(W) = \mathbb{E}\big[WXX^\top/\|WX\|_2\big]$;
(iii) N is locally Lipschitz on $\{W \neq 0\}$.

The analogous statement holds for $N_B(W) = \mathbb{E}[\|WY\|_2]$ with $Y \sim \mathcal{N}(0, \Sigma_y)$.

Proof. For $W \neq 0$, positive definiteness of Σ_x makes WX a nonzero Gaussian vector, so $\Pr(WX = 0) = 0$ and $N(W) > 0$. For any direction H, the difference quotient of $\|(W + \epsilon H)X\|_2$ is bounded by $\|H\|_F\|X\|_2$. Dominated convergence therefore gives

$$DN(W)[H] = \mathbb{E}\frac{\langle WX, HX\rangle}{\|WX\|_2} = \left\langle \mathbb{E}\frac{WXX^\top}{\|WX\|_2}, H\right\rangle_F.$$

The matrix integrand has norm at most $\|X\|_2$, and converges almost surely when $W_n \to W \neq 0$. A second application of dominated convergence makes the displayed gradient continuous. Thus the partial derivatives are continuous, proving $N \in C^1$ on nonzero matrices. Finally, the triangle inequality gives the stronger global estimate

$$|N(W) - N(V)| \leq \mathbb{E}\|(W - V)X\|_2 \leq \mathbb{E}\|X\|_2\,\|W - V\|_F,$$

which implies the stated local Lipschitz property. □

UNIFORM CONTROL NEAR SINGULAR DIRECTIONS

A derivative of a normalized vector can be large when the vector is small. The next bound controls how often WX is small relative to X, uniformly over a compact set of nonzero encoders.

Lemma A.3 (Relative small-ball estimate). *Let $X \sim \mathcal{N}(0, \Sigma_x)$ with $\Sigma_x \succ 0$, and let $K \subset \{W \neq 0\}$ be compact. There exists $C_K < \infty$ such that, for every $W \in K$ and every $0 < \epsilon \leq 1$,*

$$\mathbb{E}\big[\|X\|_2 \, \mathbf{1}\{\|WX\|_2 \leq \epsilon \|X\|_2\}\big] \leq C_K \epsilon.$$

The same statement holds with X, Σ_x replaced by Y, Σ_y.

Proof.

Step 1 (reduce to a spherical small ball). Write $X = \Sigma_x^{1/2} Z$ and decompose $Z = RU$, where $R = \|Z\|_2$ and $U = Z/\|Z\|_2$ are independent, with U uniform on the unit sphere $S^{d_1 - 1}$. Since K is compact and does not contain zero, there is $s_K > 0$ such that $\sigma_{\max}(W) \geq s_K$ for every $W \in K$. Hence

$$\sigma_{\max}(W\Sigma_x^{1/2}) \geq \sigma_{\max}(W) \, \lambda_{\min}^{1/2}(\Sigma_x) \geq s_K \lambda_{\min}^{1/2}(\Sigma_x) =: a_K > 0.$$

For each $W \in K$, choose a right singular vector $q(W)$ associated with $\sigma_{\max}(W\Sigma_x^{1/2})$. If $\|WX\|_2 \leq \epsilon \|X\|_2$, then, after cancelling the radial factor R,

$$a_K |\langle q(W), U \rangle| \leq \|W\Sigma_x^{1/2} U\|_2 \leq \epsilon \|\Sigma_x^{1/2} U\|_2 \leq \epsilon \|\Sigma_x^{1/2}\|_{\mathrm{op}}.$$

Step 2 (apply the one-dimensional spherical bound). For a uniform spherical vector U, the one-dimensional marginal satisfies

$$\sup_{\|q\|_2 = 1} \Pr\{|\langle q, U \rangle| \leq u\} \leq C_{d_1} u, \qquad 0 < u \leq 1,$$

with the trivial bound used when the right-hand side exceeds one. Therefore the event above has probability at most $C_K \epsilon$, uniformly in $W \in K$. Since R is independent of U and $\|X\|_2 \leq R\|\Sigma_x^{1/2}\|_{\mathrm{op}}$,

$$\mathbb{E}\big[\|X\|_2 \mathbf{1}\{\|WX\|_2 \leq \epsilon \|X\|_2\}\big] \leq \|\Sigma_x^{1/2}\|_{\mathrm{op}} \mathbb{E} R \, \Pr\{\|WX\|_2 \leq \epsilon \|X\|_2\} \leq C_K \epsilon.$$

The proof for Y is identical. □

NORMALIZATION-GRADIENT MODULUS

The small-ball bound compensates for the reciprocal norm in the derivative. Integrating the small-ball bound gives a logarithmic factor, which is sufficient for ODE uniqueness.

Lemma A.4 (Osgood modulus of the normalization gradient). *Let $K \subset \{W \neq 0\}$ be compact. There exists $C_K < \infty$ such that for all $W, V \in K$ with $0 < \|W - V\|_F \leq e^{-1}$,*

$$\|\nabla N(W) - \nabla N(V)\|_F \leq C_K \, \|W - V\|_F \big(1 + |\log \|W - V\|_F|\big).$$

Consequently the gradient-flow vector field on $\mathcal{U}$ satisfies the local Osgood modulus required by Theorem A.6. On every compact regular set

$$\Omega = \{(W_A, W_B, \tau) : \|W_A\|_F = C_A, \ \|W_B\|_F = C_B, \ \tau \in [\tau_{\min}, \tau_{\max}]\} \subset \mathcal{U},$$

the vector field is bounded and the maps $\nabla_{W_A} \mathcal{L}$, $\nabla_{W_B} \mathcal{L}$, and Δ/τ are uniformly continuous.

Proof. Define $h(u) = u/\|u\|_2$ for $u \neq 0$, with $h(0) = 0$. The elementary inequality

$$\|h(u) - h(v)\|_2 \leq 2 \min\left\{1, \frac{\|u - v\|_2}{\|u\|_2 + \|v\|_2}\right\}$$

uses the value one for the fraction at $u = v = 0$. For $0 < \delta = \|W - V\|_F \leq e^{-1}$, put $Q = (\|WX\|_2 + \|VX\|_2)/\|X\|_2$. Since $\nabla N(W) = \mathbb{E}[h(WX)X^\top]$,

$$\|\nabla N(W) - \nabla N(V)\|_F \leq 2\mathbb{E}\left[\|X\|_2 \min\{1, \delta/Q\}\right].$$

Lemma A.3 bounds $\mathbb{E}[\|X\|_2 \mathbf{1}\{Q \leq \epsilon\}] \leq C_K \epsilon$ for $0 < \epsilon \leq 1$. Integrating the indicators of $\{Q \leq \delta/u\}$ and applying Tonelli's theorem gives

$$\begin{aligned}
\mathbb{E}\left[\|X\|_2 \min\{1, \delta/Q\}\right] &= \int_0^1 \mathbb{E}[\|X\|_2 \mathbf{1}\{Q \leq \delta/u\}] \, du \\
&\leq \delta \mathbb{E}\|X\|_2 + C_K \delta \int_\delta^1 \frac{du}{u} \leq C_K' \delta (1 + |\log \delta|).
\end{aligned}$$

The case $\delta = 0$ is immediate. On compact subsets of $\mathcal{U}$, the remaining factors in the flow formula are locally Lipschitz and bounded, with denominators bounded away from zero. Applying this estimate to both normalizers gives the stated Osgood modulus for the full vector field. Its continuity also gives boundedness and uniform continuity on the compact regular set Ω. □

Lemma A.5 (Second derivatives away from rank one)**.** *Let $X \sim \mathcal{N}(0, \Sigma)$, with $\Sigma \succ 0$, and $N(W) = \mathbb{E}\|WX\|$. The function N is C^2 wherever* $\operatorname{rank} W \geq 2$*. On every compact set of nonzero matrices, there is a constant C such that*

$$\|D^2 N(W)\| \leq C\big(1 + |\log \sigma_2(W)|\big) \qquad \text{when } \sigma_2(W) > 0.$$

Proof. The Hessian integrand in directions H, K is bounded by $\|HX\|\|KX\|/\|WX\|$. Near a matrix of rank at least two, two singular values stay positive. They bound the denominator below by a nondegenerate two-dimensional Gaussian radius, whose reciprocal has moments of every order below two. Hölder's inequality and Gaussian numerator moments give local uniform integrability, justifying differentiation and continuity of the Hessian.

On a compact set of nonzero matrices, the leading singular value of $W\Sigma^{1/2}$ is uniformly positive. In Gaussian singular-value coordinates, let g be a standard Gaussian vector and, for a fixed constant $c > 0$, put $\varepsilon = c \min\{\sigma_2(W), 1\}$. Then

$$\|D^2 N(W)\| \leq C\mathbb{E}\frac{\|g\|^2}{\sqrt{g_1^2 + \varepsilon^2 g_2^2}} \leq C(1 + |\log \varepsilon|).$$

For the last bound, condition on g_2 and split the g_1 integral at $\varepsilon|g_2|$. The term with numerator g_1^2 is integrable directly; the other terms use the finite Gaussian polynomial and absolute-log moments. This proves the estimate. □

Osgood continuation

The gradient flow of Section 3 is an autonomous ODE of the form $\dot{\theta} = \mathsf{F}(\theta)$ on the state space $\theta = (W_A, W_B, \tau) \in \mathbb{R}^{d\times d_1} \times \mathbb{R}^{d\times d_2} \times \mathbb{R}$. The following continuation principle combines Peano existence, Osgood uniqueness, and maximal extension (Hartman, 2002, Ch. II, Theorems 2.1 and 3.1; Ch. III, Corollary 6.2).

Theorem A.6 (Osgood uniqueness and continuation)**.** *Let $\mathcal{U} \subseteq \mathbb{R}^N$ be open and let $\mathsf{F} : \mathcal{U} \to \mathbb{R}^N$ be continuous. Suppose that for every compact $K \subset \mathcal{U}$ there are constants $C_K < \infty$ and $0 < \rho_K \leq e^{-1}$ such that*

$$\|\mathsf{F}(x) - \mathsf{F}(y)\| \leq C_K\, \omega(\|x - y\|), \qquad x, y \in K, \qquad \|x - y\| \leq \rho_K,$$

where $\omega(0) = 0$ and

$$\omega(r) = r\big(1 + |\log r|\big), \qquad 0 < r \leq e^{-1}, \qquad \int_{0^+} \frac{dr}{\omega(r)} = \infty.$$

These existence, uniqueness, and continuation results give:

- *(i) uniqueness of local solutions, both forward and backward in time;*
- *(ii) a unique maximal solution $\theta : (T_-, T_+) \to \mathcal{U}$ for every initial condition $\theta_0 \in \mathcal{U}$;*
- *(iii) the usual compact-continuation alternative: if $T_+ < \infty$, then the trajectory eventually leaves every compact subset of $\mathcal{U}$.*

Equivalently, if $\theta([0, T_+))$ is contained in a compact subset of $\mathcal{U}$, then $T_+ = +\infty$.

In our setting, $N = d\,d_1 + d\,d_2 + 1, \mathcal{U} = \{(W_A, W_B, \tau) : W_A \neq 0, W_B \neq 0, \tau > 0\}$, and the vector field F is given by the right-hand sides of the gradient flow defined in Section 3. Lemma A.4 verifies the local Osgood modulus required by Theorem A.6. We use two consequences repeatedly:

- *Forward–backward uniqueness.* If $\theta(t^*) = \bar{\theta}$ for some t^*, then the entire trajectory through $\bar{\theta}$ is uniquely determined. For a relatively closed smooth embedded submanifold $\mathcal{V} \subset \mathcal{U}$ without boundary to which the vector field is tangent, any trajectory reaching $\mathcal{V}$ in finite time must have started in $\mathcal{V}$. Indeed, the times spent in $\mathcal{V}$ form a closed set by continuity and an open set by restricted local existence and forward–backward uniqueness.
- *Global existence via compactness.* Once the trajectory is shown to remain in a compact subset of $\mathcal{U}$, the compact-continuation alternative gives $T_+ = +\infty$. This is the strategy of Proposition B.3.

A.4 EXPECTED-NORM AND SAMPLE-LEVEL NORMALIZATION

The variation of the embedding norm is determined by the spectrum of its covariance. For $X \sim \mathcal{N}(0, \Sigma_x)$, set $C = W\Sigma_x W^\top \neq 0$. Diagonalizing C writes $\|WX\|_2^2$ as $\sum_i \lambda_i G_i^2$ with independent standard Gaussian G_i. Hence

$$\mathbb{E}\|WX\|_2^2 = \operatorname{tr} C, \qquad \frac{\operatorname{Var}(\|WX\|_2^2)}{(\mathbb{E}\|WX\|_2^2)^2} = \frac{2\operatorname{tr}(C^2)}{(\operatorname{tr} C)^2} = \frac{2}{r_{\text{eff}}(C)}, \qquad r_{\text{eff}}(C) = \frac{(\operatorname{tr} C)^2}{\operatorname{tr}(C^2)}.$$

A large effective rank therefore gives concentration of squared norms by Chebyshev's inequality. Increasing the ambient dimension alone gives no such guarantee, since C may remain low rank. The flow can change this spectrum during training.

The two normalizations also have different derivatives. At a fixed pair (x, y), write $u = W_A x \neq 0$, $v = W_B y \neq 0$, $\widehat{u} = u/\|u\|_2$, $\widehat{v} = v/\|v\|_2$, and $S_{\cos} = \widehat{u}^\top \widehat{v}/\tau$. Holding W_B, τ fixed,

$$\nabla_{W_A} S_{\cos} = \frac{(I - \widehat{u}\widehat{u}^\top)\widehat{v}\, x^\top}{\tau\|u\|_2}, \qquad \nabla_{W_A} S(x,y) = \frac{vx^\top}{\tau N_A N_B} - \frac{S(x,y)}{N_A} R_A.$$

Cosine normalization projects each sample's gradient onto its embedding tangent space. Expected-norm normalization uses the population derivative $R_A = \nabla_{W_A} N_A$, shared across samples. Both depend on the encoder parameters. Norm concentration concerns score values; a uniform comparison of the gradient fields would also require control of these derivatives and small embedding norms.

For a bias-free ReLU encoder, the active gates determine a matrix $P_A(x)$ such that $f(x) = P_A(x)x$ on each activation region. With $g(y) = P_B(y)y$, the unnormalized comparison is $y^\top P_B(y)^\top P_A(x)x$. Its effective matrix depends on both inputs and changes when gates switch during training. Consequently, the nonlinear expected-norm objective generally cannot be reduced to the quadratic loss of one fixed score matrix used in the linear analysis. The conditional measurements in Appendix G.10 evaluate the resulting nonlinear score directly.

A.5 LOSS AT AN INTERIOR CRITICAL POINT

At a critical point, the temperature equation relates the two terms in the quadratic loss. This identity will exclude signal-degenerate limits on the negative-loss regime.

A bar denotes evaluation at a fixed state, such as $\bar{N}_A = N_A(\bar{W}_A)$. An ω-limit point is a limit of the joint state along a sequence $t_k \to \infty$. A subsequential matrix limit is defined analogously.

Lemma A.7 (Critical points of $\mathcal{L}$ on $\mathcal{U}$)**.** *Let $\bar{\theta} = (\bar{W}_A, \bar{W}_B, \bar{\tau}) \in \mathcal{U}$ be a critical point of $\mathcal{L}$, i.e., $\nabla\mathcal{L}(\bar{\theta}) = 0$. Then:*

(i) $\bar{\Delta} = 2\bar{\alpha}(\bar{\mathcal{A}}\bar{\alpha} - \bar{\mathcal{B}}) = 0$*, where* $\bar{\alpha} = 1/(\bar{\tau}\bar{N}_A\bar{N}_B)$*;*
(ii) $\bar{\mathcal{B}} \geq 0$*;*
(iii) $\mathcal{L}(\bar{\theta}) = -\bar{\alpha}\bar{\mathcal{B}} \leq 0$*, with equality iff* $\bar{\mathcal{B}} = 0$*.*

Proof. At a critical point, $\partial_\tau \mathcal{L} = -\Delta/\tau = 0$. Since $\bar{\tau} > 0$, this proves $\bar{\Delta} = 0$. The identity defining Δ then gives

$$\bar{\mathcal{B}} = \bar{\mathcal{A}}\bar{\alpha} \geq 0,$$

because $\bar{\mathcal{A}} \geq 0$ and $\bar{\alpha} > 0$. Substitution in the loss yields

$$\mathcal{L}(\bar{\theta}) = \bar{\mathcal{A}}\bar{\alpha}^2 - 2\bar{\mathcal{B}}\bar{\alpha} = -\bar{\alpha}\bar{\mathcal{B}} \leq 0.$$

The last expression is zero precisely when $\bar{\mathcal{B}} = 0$. □

Remark A.8 (Negative loss and signal persistence). At a critical point with $W_{A,s} = W_{B,s} = 0$, one has $\mathcal{B} = 0$, so Lemma A.7 gives zero loss. Lemma B.16 gives a quantitative signal-energy bound directly along the trajectory. The separate upper-temperature bound follows directly from the loss at $\widehat{M} = 0$ in Lemma B.15.

B PROOFS OF THE MAIN TRAJECTORY RESULTS

Two mechanisms establish the trajectory results. Invariance and uniqueness give finite-time encoder non-entry. Finite-length convergence reduces asymptotic score denoising to stationary algebra, from which the projector structure and rank refinement follow. Appendix C proves generic entry into the required regime. The table identifies the result supporting each claim of Corollary 3.

Conclusion	Argument
(i) Positive signal energy at finite times	Invariance of the zero-signal set and Proposition B.11.
(ii) Nonzero nuisance in both encoders	Lemma B.13 and Remark B.14.
(iii) Bounded temperature above	Negative limiting loss excludes the high-temperature tail (Lemma B.15).
(iv) Signal energy bounded away from zero	The loss bound $\mathcal{L} \geq -\alpha\Phi$ (Lemma B.16).
(v) Vanishing nuisance-score blocks	The nuisance equation and asymptotic stationarity (Lemma B.19).

B.1 CONSERVED QUANTITIES AND REGULAR CONTINUATION

The score is unchanged by positive rescaling of either encoder. This fixes the encoder norms; the resulting expected-norm bounds then give a continuation criterion for the joint flow.

NORM CONSERVATION

Proposition B.1 (Norm conservation)**.** *Along every maximal solution of the gradient flow,*

$$\|W_A(t)\|_F = \|W_A(0)\|_F =: C_A, \qquad \|W_B(t)\|_F = \|W_B(0)\|_F =: C_B$$

for every $t \in [0, T_+)$.

Proof. The score $S(X,Y) = X^\top W_A^\top W_B Y/(\tau N_A N_B)$ is invariant under the positive rescaling $W_A \mapsto cW_A$. Indeed, $M \mapsto cM$ and $N_A \mapsto cN_A$, so S and hence $\mathcal{L}$ are unchanged. Differentiating $\mathcal{L}(cW_A, W_B, \tau) = \mathcal{L}(W_A, W_B, \tau)$ at $c = 1$ gives

$$\langle \nabla_{W_A}\mathcal{L}, W_A\rangle_F = 0.$$

Therefore

$$\frac{d}{dt}\frac{1}{2}\|W_A(t)\|_F^2 = \langle W_A, \dot{W}_A\rangle_F = -\langle W_A, \nabla_{W_A}\mathcal{L}\rangle_F = 0.$$

The same argument with $W_B \mapsto cW_B$ gives conservation of $\|W_B\|_F$. □

In particular, the encoder pair remains in $\mathcal{M}$, $\operatorname{tr} G_A = C_A^2$, and $\operatorname{tr} G_B = C_B^2$. The following bounds turn this conservation law into control of the score normalizers.

EXPECTED-NORM BOUNDS

Positive-definite marginal covariances prevent a nonzero fixed-norm encoder from having an arbitrarily small expected output norm.

Proposition B.2 (Expected-norm bounds)**.** *Under the standing assumption of Section 3, along every maximal solution,*

$$0 < N_{A,\min} \leq N_A(t) \leq N_{A,\max} < \infty, \qquad 0 < N_{B,\min} \leq N_B(t) \leq N_{B,\max} < \infty$$

for every $t \in [0, T_+)$. *In particular, on global trajectories the same bounds hold for all* $t \geq 0$. *One can take*

$$\lambda_0 = \min\{1+\sigma^2,\, \sigma^2+\nu^2\}, \qquad N_{A,\max} \;=\; \|\Sigma_x\|_{\mathrm{op}}^{1/2}\, C_A, \qquad N_{A,\min} \;=\; C_A\sqrt{\tfrac{2\lambda_0}{\pi d}},$$

and analogous expressions for B.

Proof. We give the argument for N_A; replacing X, W_A, Σ_x by Y, W_B, Σ_y gives the other bounds.

Upper bound. Jensen's inequality and norm conservation imply

$$\begin{aligned} N_A^2 &\le \mathbb{E}\|W_A X\|_2^2 = \operatorname{tr}(W_A \Sigma_x W_A^\top) \\ &\le \|\Sigma_x\|_{\mathrm{op}} \|W_A\|_F^2 = \|\Sigma_x\|_{\mathrm{op}} C_A^2. \end{aligned}$$

Lower bound. Choose a unit left singular vector u for the largest singular value of W_A. Because $\Sigma_x \succeq \lambda_0 I_{d_1}$, the scalar Gaussian $u^\top W_A X$ has variance at least

$$u^\top W_A \Sigma_x W_A^\top u \ge \lambda_0 \|W_A^\top u\|_2^2 = \lambda_0 \sigma_{\max}(W_A)^2.$$

There are at most d nonzero singular values, so

$$C_A^2 = \sum_i \sigma_i(W_A)^2 \le d\, \sigma_{\max}(W_A)^2.$$

A centered scalar Gaussian with standard deviation s has expected absolute value $s\sqrt{2/\pi}$. Therefore

$$N_A \ge \mathbb{E}|u^\top W_A X| \ge \sigma_{\max}(W_A)\sqrt{2\lambda_0/\pi} \ge C_A \sqrt{\frac{2\lambda_0}{\pi d}}.$$

These estimates hold throughout the maximal interval because C_A is conserved. □

We will use the constants

$$c_{\min} = N_{A,\min} N_{B,\min} > 0, \qquad c_{\max} = N_{A,\max} N_{B,\max} < \infty.$$

They satisfy $c_{\min} \le N_A(t) N_B(t) \le c_{\max}$ for every $t < T_+$.

Continuation away from zero temperature

With the encoder variables confined to $\mathcal{M}$, a finite-time obstruction can only come from the temperature. The next proposition excludes finite-time escape to $+\infty$.

Proposition B.3 (Regular continuation criterion)**.** *Let $(W_A(t), W_B(t), \tau(t))$ be the unique maximal solution on $[0, T_+)$. If $T_+ < \infty$, then necessarily*

$$\liminf_{t \uparrow T_+} \tau(t) = 0.$$

Equivalently, if the whole trajectory $\{(W_A(t), W_B(t), \tau(t)) : 0 \le t < T_+\}$ is contained in a compact subset of $\mathcal{U}$, and in particular if $\inf_{0 \le t < T_+} \tau(t) > 0$, then $T_+ = \infty$.

Proof. Suppose $T_+ < \infty$ and $\liminf_{t \uparrow T_+} \tau(t) > 0$. Continuity gives $\tau \ge \varepsilon > 0$ throughout $[0, T_+)$. Norm conservation and Proposition B.2 bound $\mathcal{A}, \mathcal{B}$ and bound $N_A N_B$ away from zero. The temperature equation therefore gives

$$|\dot\tau| = \left| \frac{2\mathcal{A}}{\tau^3 (N_A N_B)^2} - \frac{2\mathcal{B}}{\tau^2 N_A N_B} \right| \le C(\varepsilon^{-3} + \varepsilon^{-2}) =: C_\varepsilon.$$

Hence $\varepsilon \le \tau(t) \le \tau(0) + C_\varepsilon T_+$, while both encoder norms remain fixed and positive. The trajectory lies in a compact subset of $\mathcal{U}$, contradicting Theorem A.6. Thus finite maximal time requires $\liminf \tau = 0$, and a positive temperature lower bound implies global existence. The compact-continuation assertion is exactly the alternative in Theorem A.6. □

B.2 Scalar loss bounds and asymptotic tools

The loss has a finite lower bound and decreases along the flow, giving an integrable squared gradient. We then establish finite-length convergence on $\mathcal{B}_{\mathrm{bnd}}$. Barbalat's lemma is retained for the compactified infinite-temperature argument in Appendix C.5.

LOSS LOWER BOUND

The decreasing loss has a finite floor determined only by the signal dimension and signal variance. This prevents the gradient flow from using the quadratic loss expression to run off to $-\infty$ and gives a well-defined limiting value.

Lemma B.4 (Loss lower bound)**.** *For all* $(W_A, W_B, \tau) \in \mathcal{U}$*,*

$$\mathcal{L}(W_A, W_B, \tau) \;\geq\; -\frac{r}{(1+\sigma^2)^2}. \tag{18}$$

Proof. Let $a = 1 + \sigma^2$ and $Z_\star = a^{-2} B_s A_s^\top$. Since $\Sigma_y Z_\star \Sigma_x = \Sigma_{yx}$, completing the square in the score matrix gives

$$\mathcal{L} = \|\Sigma_y^{1/2}(\widehat{M} - Z_\star)\Sigma_x^{1/2}\|_F^2 - \frac{r}{a^2} \geq -\frac{r}{a^2}.$$

□

LOSS DECREASE AND GRADIENT INTEGRABILITY

The gradient flow makes $\mathcal{L}(t)$ non-increasing and produces an L^2-integrable gradient. These estimates control finite-time endpoints in Theorem C.3.

Lemma B.5 (Loss decrease and integrability)**.** *Along the gradient flow on its maximal interval,*

$$\frac{d}{dt}\mathcal{L}(t) = -\|\nabla_{W_A}\mathcal{L}(t)\|_F^2 - \|\nabla_{W_B}\mathcal{L}(t)\|_F^2 - \frac{\Delta(t)^2}{\tau(t)^2} \leq 0.$$

Consequently, for every $T < T_+$*,*

$$\int_0^T \|\nabla_{W_A}\mathcal{L}(t)\|_F^2 + \|\nabla_{W_B}\mathcal{L}(t)\|_F^2 + \frac{\Delta(t)^2}{\tau(t)^2}\, dt \leq \mathcal{L}(0) - \mathcal{L}(T).$$

If $T_+ = \infty$*, then* $\mathcal{L}^* = \lim_{t\to\infty} \mathcal{L}(t)$ *exists and*

$$\int_0^\infty \|\nabla_{W_A}\mathcal{L}(t)\|_F^2 + \|\nabla_{W_B}\mathcal{L}(t)\|_F^2 + \frac{\Delta(t)^2}{\tau(t)^2}\, dt < \infty. \tag{19}$$

Proof. The identity follows from the chain rule and the definition of Euclidean gradient flow:

$$\frac{d}{dt}\mathcal{L} = \langle \nabla_{W_A}\mathcal{L}, \dot{W}_A\rangle_F + \langle \nabla_{W_B}\mathcal{L}, \dot{W}_B\rangle_F + \partial_\tau\mathcal{L}\,\dot{\tau} = -\|\nabla_{W_A}\mathcal{L}\|_F^2 - \|\nabla_{W_B}\mathcal{L}\|_F^2 - (\partial_\tau\mathcal{L})^2.$$

Since $\partial_\tau\mathcal{L} = -\Delta/\tau$, the displayed dissipation identity follows. Integrating over $[0, T]$ gives the finite-horizon estimate. If $T_+ = \infty$, Lemma B.4 gives a finite lower bound for $\mathcal{L}$, so the monotone limit $\mathcal{L}^*$ exists and the infinite-horizon integrability follows by monotone convergence. □

Lemma B.6 (Barbalat lemma)**.** *Let* $f : [0, \infty) \to \mathbb{R}_{\geq 0}$ *be uniformly continuous and integrable. Then* $f(t) \to 0$ *as* $t \to \infty$*.*

Proof. If the conclusion failed, there would exist $\varepsilon > 0$ and a sequence $t_n \to \infty$ such that $f(t_n) \geq \varepsilon$. Uniform continuity gives a radius $\delta > 0$ for which $|f(t) - f(s)| \leq \varepsilon/2$ whenever $|t - s| \leq \delta$. Passing to a subsequence with pairwise disjoint intervals $I_n = [t_n - \delta/2, t_n + \delta/2] \cap [0, \infty)$, we get $f(t) \geq \varepsilon/2$ on each I_n, contradicting $f \in L^1([0, \infty))$. □

NONPOSITIVE LIMITING LOSS

The temperature equation and the fixed-scale bounds force the limiting loss of every global trajectory to be nonpositive.

Lemma B.7 (Nonpositive limiting loss on global trajectories)**.** *On every global trajectory of the gradient flow,*

$$\mathcal{L}^* = \lim_{t\to\infty} \mathcal{L}(t) \leq 0.$$

Proof. Lemmas B.4 and B.5 give a finite monotone limit $\mathcal{L}^* $. If $\mathcal{L}^* > 0$, then

$$(\tau^2)' = 2\Delta = 2\{\mathcal{L} + F(\widehat{M})\} \geq 2\mathcal{L}^* > 0,$$

so $\tau \to \infty$. Norm conservation and Proposition B.2 imply $\|\widehat{M}\|_F \leq C_A C_B/(\tau c_{\min}) \to 0$. Continuity of $F - 2G$ then gives $\mathcal{L} \to 0$, a contradiction. □

B.3 Finite length and point convergence

The Gaussian expected norm admits a definable integral representation. This gives point convergence even when an encoder loses rank.

Lemma B.8 (Definability of the population loss)**.** *The expected-norm loss is C^1 and definable in the o-minimal structure $\mathbb{R}_{\mathrm{an,exp}}$ on the original state space $\mathcal{U}$.*

Proof. For either modality, let $(p, \Sigma) = (d_1, \Sigma_x)$ or (d_2, Σ_y). With the input represented as $\Sigma^{1/2}Z$, where $Z \sim \mathcal{N}(0, I_p)$, Gaussian polar coordinates and homogeneity give

$$N(W) = c_p \int_{\|u\|\leq 1} \|W\Sigma^{1/2}u\|\, du, \qquad c_p = \frac{(p+1)\mathbb{E}\|Z\|}{|S^{p-1}|} > 0.$$

Here $|S^{p-1}|$ is the surface area of the unit sphere in $\mathbb{R}^p$. Indeed, the sphere integral equals $(p+1)$ times the ball integral, and the Gaussian radius is independent of its uniform direction. The integrand extended by zero outside the ball is semialgebraic and integrable for every W. The parameter-integration theorem of Cluckers & Miller (2011, Theorem 1.3) makes N constructible, hence definable in $\mathbb{R}_{\mathrm{an,exp}}$, which is o-minimal (van den Dries et al., 1994). Lemma A.2 gives C^1 regularity for every nonzero W. The loss is obtained from these positive normalizers, positive τ, and polynomial score terms by field operations, so it has both stated properties on $\mathcal{U}$. □

Theorem B.9 (Finite length on the bounded-temperature regime)**.** *Every trajectory in $\mathcal{B}_{\mathrm{bnd}}$ has finite arclength and converges to a single stationary state in $\mathcal{U}$. This includes rank-deficient limiting encoders.*

Proof. Norm conservation and the temperature bounds place the trajectory closure in a bounded semialgebraic open subset compactly contained in $\mathcal{U}$. By Lemma B.8, Kurdyka (1998, Theorem 2) applies to the loss restricted to that set and gives finite arclength. The flow is the ambient Euclidean gradient flow; its encoder components are already tangent to the norm spheres. Finite length gives a single limit, which lies in $\mathcal{U}$. Continuity of the vector field makes this limit stationary: a nonzero limiting velocity would contradict convergence. □

We also use the local length estimate from the same theorem. For the loss restricted to any bounded open ball contained in $\mathcal{U}$, there is a continuous increasing function ω_{len}, with $\omega_{\mathrm{len}}(0) = 0$, such that every gradient-trajectory segment remaining in that ball satisfies

$$\int_s^t \|\dot{\theta}(u)\|\, du \leq \omega_{\mathrm{len}}\big(\mathcal{L}(\theta(s)) - \mathcal{L}(\theta(t))\big). \tag{20}$$

This estimate provides a neighborhood of initial states with persistent nuisance in Proposition E.2.

B.4 Finite-time non-entry of signal/nuisance-degenerate sets

The zero-signal set and each nuisance-free encoder set are invariant. Uniqueness in both time directions will show that a trajectory starting outside any one of them cannot reach it at finite time. Generic initialization then gives the two finite-time claims of Corollary 3. The critical-point analysis in Appendix C is separate from this non-entry argument.

Throughout, write $\zeta_s = A_s^\top X$ for the signal coordinates and $\zeta_n = A_\perp^\top X$ for the nuisance coordinates of X (and analogously for Y); these are mutually independent by the data model of Section 3.

THE SIGNAL-DEGENERATE SET

Definition B.1 (Signal-degenerate set)**.** The *signal-degenerate set* is

$$\mathcal{S} \;=\; \{(W_A, W_B, \tau) \in \mathcal{M} \times \mathbb{R}_{>0} : W_{A,s} = 0 \text{ and } W_{B,s} = 0\}.$$

Lemma B.10 (Invariance of $\mathcal{S}$)**.** *$\mathcal{S}$ is positively invariant under the gradient flow.*

Proof. At a point of $\mathcal{S}$, right-multiply the encoder equation by A_s:

$$\dot{W}_{A,s} = \frac{2}{\tau N_A N_B} W_B P A_s + \frac{\Delta}{N_A} R_A A_s.$$

Both terms vanish, for different reasons.

First, $MA_s = W_B^\top W_{A,s} = 0$ and $\Sigma_x A_s = (1+\sigma^2)A_s$. Hence

$$PA_s = \Sigma_{yx} A_s - \alpha \Sigma_y M \Sigma_x A_s = B_s, \qquad W_B P A_s = W_{B,s} = 0.$$

Second, $W_A X = W_{A,\perp}\zeta_n$ on $\mathcal{S}$. Independence of the signal and nuisance coordinates, together with $\mathbb{E}\zeta_s = 0$, gives

$$R_A A_s = \mathbb{E}\left[\frac{W_{A,\perp}\zeta_n \zeta_s^\top}{\|W_{A,\perp}\zeta_n\|_2}\right] = \mathbb{E}\left[\frac{W_{A,\perp}\zeta_n}{\|W_{A,\perp}\zeta_n\|_2}\right] \mathbb{E}[\zeta_s]^\top = 0.$$

Thus $\dot{W}_{A,s} = 0$. The calculation for $W_{B,s}$ is symmetric, proving invariance of $\mathcal{S}$. □

Proposition B.11 (Non-entry into $\mathcal{S}$)**.** *Trajectories starting outside $\mathcal{S}$ cannot reach $\mathcal{S}$ in finite time: if $(W_A(0), W_B(0), \tau(0)) \notin \mathcal{S}$, then $(W_A(t), W_B(t), \tau(t)) \notin \mathcal{S}$ for every finite $t \in [0, T_+)$.*

Proof. Lemma B.10 proves tangency to $\mathcal{S}$ for both the forward field and its negative. Theorem A.6 gives local uniqueness in both time directions. Consequently, if a trajectory meets $\mathcal{S}$, its local forward and backward continuations agree with solutions restricted to $\mathcal{S}$.

If it reached $\mathcal{S}$ at a finite time t^*, the set $\{t \in [0, t^*] : \theta(t) \in \mathcal{S}\}$ would be nonempty, relatively open by this local uniqueness, and closed by continuity. Connectedness of $[0, t^*]$ makes it the whole interval, contradicting $\theta(0) \notin \mathcal{S}$. □

Remark B.12 (Mechanism). The non-entry argument rests on two ingredients: (i) tangency of the vector field to $\mathcal{S}$ (Lemma B.10, $\dot{W}_{A,s}|_{\mathcal{S}} = 0$, which makes $\mathcal{S}$ invariant under *both* the forward and the backward flows); and (ii) ODE uniqueness in both directions (Theorem A.6 (i)). Note that this argument applies in finite time only — it does not preclude $\Phi(t) \to 0$ asymptotically. The asymptotic non-degeneracy $\liminf_{t\to\infty} \Phi(t) > 0$ is the content of Lemma B.16.

NUISANCE-FREE SETS

Definition B.2 (Nuisance-free sets)**.** For each encoder, define

$$\mathcal{N}_A = \{(W_A, W_B, \tau) \in \mathcal{M} \times \mathbb{R}_{>0} : W_{A,\perp} = 0\},$$
$$\mathcal{N}_B = \{(W_A, W_B, \tau) \in \mathcal{M} \times \mathbb{R}_{>0} : W_{B,\perp} = 0\}.$$

These sets are the nuisance-coordinate analogue of the signal-degenerate set $\mathcal{S}$. Their invariance gives the finite-time non-entry statement behind claim (ii) of Corollary 3.

Lemma B.13 (Nuisance-free invariance)**.** *$\mathcal{N}_A$ and $\mathcal{N}_B$ are positively invariant under the gradient flow.*

Proof. Consider a point of $\mathcal{N}_A$, where $W_{A,\perp} = 0$. The nuisance equation is

$$\dot{W}_{A,\perp} = \frac{2}{\tau N_A N_B} W_B P A_\perp + \frac{\Delta}{N_A} R_A A_\perp.$$

Since $MA_\perp = W_B^\top W_{A,\perp} = 0$, the covariance identities imply

$$PA_\perp = \Sigma_{yx} A_\perp - \alpha \Sigma_y M \Sigma_x A_\perp = -(\sigma^2 + \nu^2)\alpha \Sigma_y M A_\perp = 0.$$

For the normalizer term, $W_A X = W_{A,s}\zeta_s$. The nuisance coordinates are independent of ζ_s and centered, so

$$R_A A_\perp = \mathbb{E}\left[\frac{W_{A,s}\zeta_s\zeta_n^\top}{\|W_{A,s}\zeta_s\|_2}\right] = \mathbb{E}\left[\frac{W_{A,s}\zeta_s}{\|W_{A,s}\zeta_s\|_2}\right]\mathbb{E}[\zeta_n]^\top = 0.$$

Therefore $\dot{W}_{A,\perp} = 0$ on $\mathcal{N}_A$. Interchanging the two modalities proves the result for $\mathcal{N}_B$. □

Remark B.14 (Almost-sure non-entry). By the same forward-backward uniqueness argument as Proposition B.11, trajectories starting outside $\mathcal{N}_A$ (resp. $\mathcal{N}_B$) cannot enter it in finite time. The set $\{(W_A, W_B) \in \mathcal{M} : W_A A_\perp = 0\}$ is the product of the nuisance-free section of the A-encoder sphere and the entire B-encoder sphere. It has codimension $d(d_1 - r) > 0$ in $\mathcal{M}$, hence zero induced surface measure. By the absolute-continuity assumption on the initialization, $W_{A,\perp}(0) \neq 0$ almost surely, and by invariance, $W_{A,\perp}(t) \neq 0$ for all finite $t \in [0, T_+)$ almost surely. The argument for W_B is symmetric, and the same codimension argument applied to the encoder section of $\mathcal{S}$ (codimension $2dr > 0$ on $\mathcal{M}$) gives $\Phi(0) > 0$ a.s., hence $\Phi(t) > 0$ for all finite $t \in [0, T_+)$ a.s. via Proposition B.11.

B.5 Temperature boundedness on the negative-loss regime

Negative limiting loss bounds temperature above. If temperature diverged along a subsequence, the fixed encoder norms would force $\widehat{M} \to 0$ and hence $\mathcal{L} \to 0$, contradicting $\mathcal{L}^* < 0$. Combined with a positive temperature lower bound, this estimate supplies the compact regime used below.

Lemma B.15 (Temperature boundedness)**.** *Under the model* (1) *and the gradient flow* (7) *with a generic initialization (Definition 1), suppose the trajectory is global and satisfies* $\mathcal{L}^* < 0$*. Then* $\sup_{t\geq 0}\tau(t) < \infty$.

Proof. If temperature were unbounded, continuity would give times $t_n \to \infty$ with $\tau(t_n) \to \infty$. Propositions B.1 and B.2 yield

$$\|\widehat{M}(t_n)\|_F \leq \frac{C_A C_B}{c_{\min}\tau(t_n)} \longrightarrow 0.$$

Continuity of the quadratic score loss $F - 2G$ gives $\mathcal{L}(t_n) \to 0$, contradicting $\mathcal{L}(t_n) \to \mathcal{L}^* < 0$. Hence $\sup_{t\geq 0}\tau(t) < \infty$. □

B.6 Signal persistence on the regular regime

Negative loss requires positive signal alignment. The normalizer and temperature lower bounds turn this observation into a direct lower bound on signal energy.

Lemma B.16 (Signal persistence)**.** *Under the model* (1) *and the gradient flow* (7) *with a generic initialization (Definition 1), on the regular non-collapsing regime* $\mathcal{B}_{\text{reg}}$*,*

$$\liminf_{t\to\infty}\Phi(t) > 0.$$

Proof. The alignment coefficient satisfies $\mathcal{B} = \langle W_{B,s}, W_{A,s}\rangle_F \leq \Phi/2$. Since $F \geq 0$,

$$\mathcal{L} = F(\widehat{M}) - 2\alpha\mathcal{B} \geq -\alpha\Phi.$$

On $\mathcal{B}_{\text{reg}}$, choose $\tau_{\min} > 0$ with $\tau(t) \geq \tau_{\min}$ on a tail. Proposition B.2 gives $\alpha(t) \leq (\tau_{\min}c_{\min})^{-1}$, so on that tail

$$\Phi(t) \geq -\mathcal{L}(t)\tau(t)N_A(t)N_B(t) \geq -\mathcal{L}(t)\tau_{\min}c_{\min}$$

whenever $\mathcal{L}(t) < 0$. Taking lower limits gives $\liminf_t \Phi(t) \geq -\mathcal{L}^*\tau_{\min}c_{\min} > 0$. □

B.7 NUISANCE-INTERACTION ANNIHILATION ON THE REGULAR REGIME

NUISANCE-BLOCK EVOLUTION

The nuisance equation separates a matrix damping term from the normalization-gradient contribution. Both terms evolve with the other encoder and temperature; their stationary form will imply score-block annihilation.

Right-multiplying the $\dot{W}_A$ identity of Lemma A.1 by $A_\perp$ gives

$$\dot{W}_{A,\perp} \;=\; \frac{2}{\tau N_A N_B}\, W_B P A_\perp \;+\; \frac{\Delta}{N_A}\, R_A A_\perp.$$

Using $\Sigma_{yx} A_\perp = B\left(\begin{smallmatrix} I_r & 0 \\ 0 & 0 \end{smallmatrix}\right) A^\top A_\perp = 0$ (since $A^\top A_\perp$ selects the nuisance columns) and $\Sigma_x A_\perp = A D_x A^\top A_\perp = (\sigma^2+\nu^2) A_\perp$,

$$P A_\perp \;=\; -\,\frac{\sigma^2+\nu^2}{\tau N_A N_B}\, \Sigma_y M A_\perp.$$

Combining with $M A_\perp = W_B^\top W_A A_\perp = W_B^\top W_{A,\perp}$ and defining the damping coefficient

$$c(t) \;=\; \frac{2(\sigma^2+\nu^2)}{(\tau(t)\, N_A(t)\, N_B(t))^2} \;>\; 0,$$

we obtain the nuisance-block ODE

$$\dot{W}_{A,\perp} \;=\; -\,c(t)\left(W_B \Sigma_y W_B^\top\right) W_{A,\perp} \;+\; \frac{\Delta}{N_A}\, R_A A_\perp. \tag{21}$$

The normalization term vanishes on a nuisance-free encoder by Lemma B.13. The symmetric ODE holds with $A \leftrightarrow B$ and $\Sigma_x \leftrightarrow \Sigma_y$.

NUISANCE ANNIHILATION UNDER A TWO-SIDED TEMPERATURE BOUND

Under a two-sided temperature bound, the nuisance-block ODE (21) forces the nuisance-interaction blocks of $\widetilde{M}$ to vanish. On the regular non-collapsing regime, the lower temperature bound is part of the regime definition and the upper bound follows from Lemma B.15.

Proposition B.17 (Nuisance annihilation under bounded temperature)**.** *Suppose the trajectory is global and satisfies* $0 < \tau_{\min} \le \tau(t) \le \tau_{\max}$ *for all* $t \ge 0$*. Then* $\widetilde{M}_{sn}(t) \to 0$*,* $\widetilde{M}_{ns}(t) \to 0$*, and* $\widetilde{M}_{nn}(t) \to 0$*.*

Proof. Theorem B.9 gives a stationary limit $(\bar{W}_A, \bar{W}_B, \bar{\tau}) \in \mathcal{U}$, with $\bar{\Delta} = 0$. At that limit, the nuisance equation (21) gives $\bar{W}_B \Sigma_y \bar{W}_B^\top \bar{W}_{A,\perp} = 0$. Therefore

$$0 = \operatorname{tr}(\bar{W}_{A,\perp}^\top \bar{W}_B \Sigma_y \bar{W}_B^\top \bar{W}_{A,\perp}) = \|\Sigma_y^{1/2} \bar{W}_B^\top \bar{W}_{A,\perp}\|_F^2.$$

Positive definiteness of Σ_y implies $\bar{M} A_\perp = \bar{W}_B^\top \bar{W}_{A,\perp} = 0$, which removes the signal–nuisance and nuisance–nuisance blocks. The symmetric argument gives $B_\perp^\top \bar{M} = 0$, removing the nuisance–signal block. Continuity transfers these identities to the three stated trajectory limits. □

Remark B.18. The temperature bound is the only non-trivial assumption in Proposition B.17. On the regular non-collapsing regime, the lower bound is part of the regime definition and Lemma B.15 supplies the upper bound from $\mathcal{L}^* < 0$.

On the regular non-collapsing regime, the late-time temperature is bounded both below and above. The conditional nuisance-annihilation proposition therefore applies directly after discarding a finite initial interval.

Lemma B.19 (Nuisance block annihilation)**.** *Under the model* (1) *and the gradient flow* (7) *with a generic initialization (Definition 1), on the regular non-collapsing regime* $\mathcal{B}_{\mathrm{reg}}$*,*

$$\widetilde{M}_{sn}(t) \to 0, \qquad \widetilde{M}_{ns}(t) \to 0, \qquad \widetilde{M}_{nn}(t) \to 0$$

as $t \to \infty$*.*

Proof. Lemma B.15 bounds temperature above on $\mathcal{B}_{\mathrm{reg}}$. Its eventual positive lower bound extends over every finite initial interval by continuity. Proposition B.17 therefore gives all three limits. □

B.8 Completion of the main proofs and the rank bound

Theorem 1. On $\mathcal{B}_{\mathrm{bnd}}$, the temperature is eventually bounded below and is bounded above for all time. Continuity and positivity also give a lower bound on every finite initial interval. Proposition B.17 therefore proves the three block limits of Theorem 1.

We next combine finite-time non-entry with the negative-loss consequences, then characterize the limiting shared score.

Proof of Corollary 3. Generic initialization avoids the zero-signal and nuisance-free sets almost surely, as shown in Remark B.14. Their invariance and forward–backward uniqueness give claims (i)–(ii) by Proposition B.11 and Lemma B.13. On $\mathcal{B}_{\mathrm{reg}}$, negative limiting loss gives the upper temperature bound in Lemma B.15, proving (iii). Lemmas B.16 and B.19 give (iv) and (v), respectively. □

Remark B.20 (Interpretation). Claims (ii) and (v) distinguish exact finite-time encoder denoising from asymptotic score denoising. The dissipative term in Eq. (21) acts through $W_B^\top W_{A,\perp}$. For a fixed W_B, a column of $W_{A,\perp}$ in $\ker(W_B^\top)$ is invisible to this score term. Along the coupled flow, both this kernel and the normalization contribution can change. Norm conservation constrains total encoder energy but does not determine its asymptotic division between signal and nuisance. Consequently, the finite-time positivity in claim (ii) is compatible with either a positive limiting nuisance component or asymptotic encoder denoising, as the capacity analysis in Appendix E.2 illustrates.

Proof of Corollary 4. Let $\widehat{M}(t_n) \to \widehat{M}^*$. Continuity of the quadratic loss gives $\mathcal{L}^* = F(\widehat{M}^*) - 2G(\widehat{M}^*) < 0$. Since $F \geq 0$, Cauchy–Schwarz yields

$$0 < -\frac{\mathcal{L}^*}{2} \leq G(\widehat{M}^*) = \langle \widehat{M}^*, \Sigma_{yx} \rangle_F \leq \|\widehat{M}^*\|_F \|\Sigma_{yx}\|_F.$$

Dividing by $\|\Sigma_{yx}\|_F > 0$ proves the stated lower bound. □

Proposition B.21 (Structure of limiting shared scores)**.** *Let $a = 1 + \sigma^2$. On every trajectory in $\mathcal{B}_{\mathrm{bnd}}$, the rescaled score converges to a single limit of the form*

$$\widehat{M}(t) \longrightarrow \widehat{M}^* = a^{-2} B_s \Pi A_s^\top, \qquad \Pi^\top = \Pi, \quad \Pi^2 = \Pi, \quad \operatorname{rank} \Pi = k.$$

Here $k \in \{0, \ldots, \min(d, r)\}$ and $\mathcal{L}^* = -k/a^2$. On $\mathcal{B}_{\mathrm{reg}}$, $k \geq 1$. If $-\mathcal{L}^* > (r-1)/a^2$, then $k = r$ and

$$\widehat{M}(t) \longrightarrow a^{-2} B_s A_s^\top = \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1}.$$

Proof. By Theorem B.9, the joint state converges to $(\bar{W}_A, \bar{W}_B, \bar{\tau}) \in \mathcal{U}$, which is stationary. Theorem 1 makes the limiting off-signal score blocks zero. Write

$$U = \bar{W}_A A_s, \quad V = \bar{W}_B B_s, \quad \bar{\alpha} = (\bar{\tau} \bar{N}_A \bar{N}_B)^{-1}, \quad Z = \bar{\alpha} V^\top U.$$

At the limit, $\bar{\Delta} = 0$ and $\bar{P} A_s = B_s (I_r - a^2 Z)$. The stationary encoder equation gives $V(I_r - a^2 Z) = 0$. Multiplying by $\bar{\alpha} U^\top$ yields

$$Z^\top = a^2 Z^\top Z.$$

The right-hand side is symmetric positive semidefinite, so $Z = Z^\top = a^2 Z^2$. Hence $\Pi = a^2 Z$ is an orthogonal projector of rank $k \leq \min(d, r)$, and

$$\mathcal{L}^* = a^2 \|Z\|_F^2 - 2 \operatorname{tr} Z = -k/a^2.$$

Joint-state convergence gives $\widehat{M}(t) \to a^{-2} B_s \Pi A_s^\top$, proving the asserted limit. Negative limiting loss forces $k \geq 1$. Under the stated threshold, $k > r - 1$, so $k = r$ and $\Pi = I_r$, giving the unique optimal score. □

The limiting shared score thus selects a signal subspace whose nonzero singular values equal a^{-2}. Theorem B.9 gives one limiting projector for each trajectory. Full shared rank makes it the unique optimal score.

Corollary B.22 (Rank refinement on the regular non-collapsing regime)**.** *Under the model* (1) *and the gradient flow* (7) *with a generic initialization (Definition 1), on the regular non-collapsing regime* $\mathcal{B}_{\mathrm{reg}}$ *of Corollary 3, the rescaled score product* $\widehat{M}(t) = \alpha(t)M(t)$ *converges to a single limit* $\widehat{M}^*$ *satisfying*

$$\operatorname{rank}(\widehat{\widetilde{M}}^*_{ss}) \;=\; -\mathcal{L}^* \cdot (1+\sigma^2)^2.$$

In particular, if $-\mathcal{L}^* > (r-1)/(1+\sigma^2)^2$*, then* $\operatorname{rank}(\widehat{\widetilde{M}}^*_{ss}) = r$ *exactly.*

Proof of Corollary B.22. Lemma B.15 places every trajectory in $\mathcal{B}_{\mathrm{reg}}$ inside $\mathcal{B}_{\mathrm{bnd}}$. By Proposition B.21, with $a = 1+\sigma^2$, $\widehat{\widetilde{M}}^*_{ss} = a^{-2}\Pi$ for a rank-k orthogonal projector and $\mathcal{L}^* = -k/a^2$. Thus

$$\operatorname{rank}(\widehat{\widetilde{M}}^*_{ss}) = k = -\mathcal{L}^* a^2,$$

which is the stated identity. The same representation gives

$$\operatorname{tr}(\widehat{\widetilde{M}}^*_{ss}) = -\mathcal{L}^*, \qquad \|\widehat{\widetilde{M}}^*_{ss}\|_F^2 = \frac{-\mathcal{L}^*}{a^2}. \tag{22}$$

Under the displayed threshold, the integer $k \le r$ exceeds $r-1$, hence $k = r$. □

Remark B.23 (Empirical verification of the rank refinement). The unconditional global minimum of $\mathcal{L}$ on the cross-block-vanishing stratum is $\mathcal{L}_{\min} = -r/(1+\sigma^2)^2$, attained at $\widehat{\widetilde{M}}_{ss} = I_r/(1+\sigma^2)^2$. In the default setting of Figure 1 ($r = 3$, $\sigma^2 = 0.5$), the threshold is $2/2.25 \approx 0.889$ and the analytic floor is $-3/2.25 \approx -1.333$, so the terminal loss can be compared with this threshold as a finite-horizon diagnostic of full signal-block rank. The same cell-specific comparison is used throughout the (d, σ^2) sweep.

Remark B.24 (Nonzero limiting signal). On $\mathcal{B}_{\mathrm{reg}}$, Corollary 4 bounds every limiting rescaled score away from zero. Since the three off-signal blocks vanish, the remaining signal–signal block is nonzero. Corollary 2 proves almost-sure entry into this regime under its dimension conditions.

C GENERIC ENTRY INTO THE NON-COLLAPSING REGIME

We prove Corollary 2 in three stages. First, finite-length estimates give global existence and a positive temperature lower bound. Next, a trajectory that never reaches negative loss must lose all shared signal. We then show that the initial states producing this behavior have zero surface measure, both when temperature is bounded and when it diverges.

We retain the notation of Appendix A.1. In this appendix,

$$a = 1 + \sigma^2, \qquad b = \sigma^2 + \nu^2, \qquad W = (W_A, W_B), \qquad \phi = \sqrt{\Phi}.$$

For the inverse-temperature analysis, normalize the loss coefficients by the expected encoder norms:

$$\mathcal{A}_N = \frac{\mathcal{A}}{(N_A N_B)^2}, \qquad \mathcal{B}_N = \frac{\mathcal{B}}{N_A N_B}, \qquad \mathcal{L} = \eta^2 \mathcal{A}_N - 2\eta \mathcal{B}_N. \tag{23}$$

Here $\eta = 1/\tau$, as in Appendix A.1. The coefficients $\mathcal{A}_N, \mathcal{B}_N$ depend only on the encoder pair.

C.1 TEMPERATURE-BOUNDARY ALTERNATIVES

Lemma C.1 (Asymptotic behavior at the τ-boundary)**.** *Let $\theta : [0, \infty) \to \mathcal{U}$ be the gradient flow trajectory.*

(i) If $\tau(t) \to \infty$ along a subsequence, then $\mathcal{L}(t) \to 0$ along the same subsequence, and $\mathcal{L}^* = \lim_t \mathcal{L}(t) = 0$ (using $\mathcal{L}$ non-increasing).

(ii) If $\tau(t) \to 0$ along a subsequence, then $\alpha(t) \to \infty$ on that subsequence; the loss is bounded if and only if αM is bounded.

Proof. Norm conservation gives $\|M\|_F \le C_A C_B$. Consequently, $\mathcal{A}$ and $\mathcal{B}$ are bounded along the trajectory. Proposition B.2 also bounds $N_A N_B$ above and away from zero.

(i) Temperature tending to infinity. Along any sequence with $\tau(t_n) \to \infty$, the normalizer lower bound gives $\alpha(t_n) \to 0$. Using the bounded loss coefficients,

$$\mathcal{L}(t_n) = \mathcal{A}(t_n)\alpha(t_n)^2 - 2\mathcal{B}(t_n)\alpha(t_n) \longrightarrow 0.$$

The loss decreases and has a finite lower bound (Lemmas B.4 and B.5), so its limit exists. A subsequence tending to zero therefore implies $\mathcal{L}^* = 0$.

(ii) Temperature tending to zero. The normalizer upper bound now gives $\alpha(t_n) \to \infty$. For any score matrix Z,

$$F(Z) \ge \lambda_0^2 \|Z\|_F^2, \qquad |G(Z)| \le \|\Sigma_{yx}\|_F \|Z\|_F.$$

Thus $F(Z) - 2G(Z)$ tends to $+\infty$ whenever $\|Z\|_F \to \infty$. Applied to $Z = \alpha M$, this proves that bounded loss requires bounded αM. Conversely, bounded αM gives bounded loss by continuity of the quadratic function $F - 2G$. Along the flow the loss is bounded by monotonicity and its lower bound, so $\alpha(t_n)M(t_n)$ is bounded. Therefore

$$M(t_n) = \frac{\alpha(t_n)M(t_n)}{\alpha(t_n)} \longrightarrow 0.$$

□

Loss coercivity and the temperature boundary. The quadratic loss controls the rescaled product $\widehat{M}$. It permits algebraic families in which $\tau \to 0$ and $M \to 0$ together while their rescaled score remains optimal (Proposition E.9). The following dynamical argument excludes approach to this boundary along a trajectory.

C.2 GLOBAL EXISTENCE AND POSITIVE TEMPERATURE

As temperature approaches zero, bounded loss forces the encoder product toward zero. The output-Gram equations and the logarithmic normalization-gradient bound control these two limits jointly.

Lemma C.2 (Exclusion of a zero-temperature endpoint)**.** *Suppose a trajectory has a finite joint-state limit as $t \uparrow T$, where $T \le \infty$, and $\int_{t_0}^{T} |(\tau^2)'(t)|\, dt < \infty$. Then its limiting temperature is positive.*

Proof. Loss coercivity bounds $\widehat{M}$ throughout the trajectory. If $\tau \to 0$, norm conservation and the normalizer bounds give $M = \tau N_A N_B \widehat{M} \to 0$. The nonzero limiting encoders therefore have orthogonal output ranges E_A, E_B. Let P_{E_A}, P_{E_B} be their orthogonal projectors and $S_0 = P_{E_A} - P_{E_B}$, so $-I \preceq S_0 \preceq I$.

Write $Q_A = W_A W_A^\top$, $Q_B = W_B W_B^\top$, and consider their difference $Q_A - Q_B$. The encoder equations give

$$\frac{d}{dt}(Q_A - Q_B) = 2\Delta(\mathcal{Q}_A - \mathcal{Q}_B), \qquad \mathcal{Q}_A = \frac{R_A W_A^\top}{N_A}, \quad \mathcal{Q}_B = \frac{R_B W_B^\top}{N_B}, \qquad (\tau^2)' = 2\Delta. \tag{24}$$

The terms containing P cancel in the derivative above. For the remaining terms,

$$R_A W_A^\top = \mathbb{E}\left[\frac{(W_A X)(W_A X)^\top}{\|W_A X\|}\right].$$

Thus $\mathcal{Q}_A$ is positive semidefinite with trace one; likewise for $\mathcal{Q}_B$. Define

$$g_A = \operatorname{tr}((I - S_0)Q_A), \qquad g_B = \operatorname{tr}((I + S_0)Q_B),$$
$$g = g_A + g_B = C_A^2 + C_B^2 - \operatorname{tr}(S_0(Q_A - Q_B)), \qquad q = g + 2\tau^2.$$

Under the contrary assumption, $g \ge 0$, $q > 0$, and $q \to 0$. Equation (24) gives the derivative directly:

$$q' = (\tau^2)' \left[\operatorname{tr}((I - S_0)\mathcal{Q}_A) + \operatorname{tr}((I + S_0)\mathcal{Q}_B)\right]. \tag{25}$$

Both trace terms are nonnegative.

Let $\delta_A = \|(I - P_{E_A})W_A\|_F$ measure the distance from the limiting output range. Then $\delta_A^2 \le g_A \le 2\delta_A^2$, and both $P_{E_A}W_A, W_A$ approach the same nonzero limit. The range of $R_A(P_{E_A}W_A)$ is contained in E_A, so Lemma A.4 gives, on a sufficiently late tail,

$$\begin{aligned} N_A \operatorname{tr}((I - S_0)\mathcal{Q}_A) &= \langle (I - S_0)[R_A(W_A) - R_A(P_{E_A}W_A)], W_A - P_{E_A}W_A \rangle_F \\ &\le C\delta_A^2(1 + |\log \delta_A|). \end{aligned}$$

At $\delta_A = 0$, the expression is zero. The normalizer has a positive lower bound. Applying the same argument to the projection $P_{E_B}W_B$, and using $g_A, g_B \le g \le q$, gives

$$0 \le \operatorname{tr}((I - S_0)\mathcal{Q}_A) + \operatorname{tr}((I + S_0)\mathcal{Q}_B) \le Cg(1 + |\log g|) \le Cq(1 + |\log q|)$$

for sufficiently small q, with value zero at $g = 0$. Thus $|q'| \le C|(\tau^2)'|q(1 + |\log q|)$. On a tail where $q < 1$, Gronwall's inequality bounds $1 - \log q$, since $\int_{t_0}^{T} |(\tau^2)'|dt < \infty$. This contradicts $q \to 0$. □

Theorem C.3 (Global existence and a positive temperature floor)**.** *For every initial state in $\mathcal{U}$, the original Euclidean gradient flow is global and*

$$\inf_{t \ge 0} \tau(t) > 0.$$

If $\mathcal{L}(t_0) < 0$ at any finite time, temperature is also bounded above, the joint trajectory has finite length, and it converges to a negative-loss stationary state. Consequently, $\mathcal{B}_{\text{reg}} = \{\mathcal{L}^* < 0\} = \{\exists t_0 < \infty : \mathcal{L}(t_0) < 0\}$. The lower bound may depend on the initial state.

Proof. Suppose first that $T_+ < \infty$. The loss floor and dissipation identity bound $\int_0^{T_+} \|\dot{\theta}\|^2 dt$. Cauchy–Schwarz gives finite length and a finite joint-state limit at T_+. In particular τ is bounded and $\int_0^{T_+} |(\tau^2)'|dt < \infty$. Lemma C.2 makes the limiting temperature positive. Norm conservation makes

the limiting encoders nonzero, so continuation in $\mathcal{U}$ contradicts maximality. Hence every trajectory is global.

If the loss is always nonnegative, the identity $\Delta = \mathcal{L} + F(\widehat{M})$, with $F(\widehat{M}) \geq 0$, gives $\dot{\tau} \geq 0$. Therefore $\tau(t) \geq \tau(0) > 0$. Otherwise, monotonicity gives $\mathcal{L}^* < 0$. The loss estimate used in Lemma B.15, which does not use genericity, bounds τ above. Norm conservation now places the trajectory in a bounded semialgebraic open subset of $\mathcal{U}$; its closure need not yet avoid $\tau = 0$. Lemma B.8 and Kurdyka (1998, Theorem 2) give finite length on this bounded open domain. That theorem does not require an extension of the loss to the domain boundary. Thus the state has a limit and $\int_0^\infty |(\tau^2)'|dt < \infty$. Lemma C.2 excludes a zero limiting temperature. Positivity on finite initial intervals gives the claimed global lower bound. The limit is interior and stationary by continuity of the vector field, and its loss equals $\mathcal{L}^* < 0$. □

The temperature lower bound depends on initialization: the equilibrium family in Proposition E.9 has no uniform positive lower bound. Negative-loss entry is a separate dynamical property. For example, exactly anti-aligned signal-only encoders can remain fixed while temperature diverges and loss decreases to zero from above.

Remark C.4 (Negative loss and non-collapse). Negative loss distinguishes selective score denoising from collapse of the entire score product. Lemma C.5 identifies the bounded-temperature zero-loss alternative, in which $M(t) \to 0$. Theorem C.3 gives a positive temperature lower bound for every trajectory and shows that any finite-time negative loss implies membership in $\mathcal{B}_{\mathrm{reg}}$. Proposition C.8 characterizes failure of entry by complete signal extinction, and Corollary 2 excludes that alternative almost surely under its dimension conditions.

C.3 Zero-loss limits and score collapse

For global trajectories with positive signal energy and an upper temperature bound, the following trichotomy separates negative limiting loss from zero-loss product collapse.

Lemma C.5 (Negative loss limit, structural form)**.** *Under the model* (1) *and the gradient flow* (7) *with a generic initialization (Definition 1), suppose the trajectory is global,* $\Phi(t) > 0$ *for all* $t \geq 0$*, and* $\tau_{\max} = \sup_{t \geq 0} \tau(t) < \infty$*. Then exactly one of the following alternatives occurs:*

- *(A)* $\Delta(t_1) < 0$ *for some* $t_1 \geq 0$*. Then* $\mathcal{L}^* < 0$*.*
- *(B)* $\Delta(t) \geq 0$ *for all* $t \geq 0$*,* $\tau(t) \to \tau_\infty \in [\tau(0), \tau_{\max}]$*, and one of:*
  - *(B.1) Some* ω*-limit has* $\bar{M} \neq 0$*. Then* $\mathcal{L}^* < 0$*.*
  - *(B.2) Every* ω*-limit has* $\bar{M} = 0$*. Then* $\mathcal{L}^* = 0$*,* $M(t) \to 0$*, and nuisance-interaction annihilation holds vacuously.*

Thus $\mathcal{L}^* < 0$ *is equivalent, within this trichotomy, to excluding the residual zero-loss collapse in case (B.2).*

Proof. If $\Delta(t_1) < 0$, then $\mathcal{L}(t_1) = \Delta(t_1) - F(\widehat{M}(t_1)) < 0$, and loss monotonicity gives case (A). Otherwise $\Delta \geq 0$, so τ is nondecreasing and converges to a positive limit. The trajectory lies in $\mathcal{B}_{\mathrm{bnd}}$, and Theorem B.9 gives a single stationary limit $\bar{\theta}$. At that limit,

$$\bar{\mathcal{B}} = \bar{\alpha}\bar{\mathcal{A}}, \qquad \mathcal{L}^* = -\bar{\alpha}^2\bar{\mathcal{A}}, \qquad \bar{\mathcal{A}} = \|\Sigma_y^{1/2}\bar{M}\Sigma_x^{1/2}\|_F^2.$$

Positive-definite covariances imply $\bar{\mathcal{A}} > 0$ exactly when $\bar{M} \neq 0$. This gives case (B.1). If $\bar{M} = 0$, then $\mathcal{L}^* = 0$ and $M(t) \to 0$, giving case (B.2). The cases are mutually exclusive and exhaust the stated alternatives. □

C.4 Signal-degenerate critical points and driving force

At $M = 0$, the first-order driving force in the loss alignment term is proportional to the signal energy Φ.

Lemma C.6 (Critical points on $\mathcal{S}$)**.** *The critical points of the gradient flow on* $\mathcal{S}$ *are precisely the points where* $M = 0$*.*

Proof. On $\mathcal{S}$, both signal factors vanish, so $\mathcal{B} = 0$ and $\Delta = 2\alpha^2\mathcal{A}$. Temperature stationarity therefore forces $\mathcal{A} = 0$, hence $M = 0$. Conversely, on $\mathcal{S} \cap \{M = 0\}$, one has $P = \Sigma_{yx}$ and $\Delta = 0$. The encoder equations reduce to

$$\dot{W}_A = 2\alpha W_{B,s} A_s^\top = 0, \qquad \dot{W}_B = 2\alpha W_{A,s} B_s^\top = 0.$$

The temperature equation also vanishes, so the state is critical. □

Proposition C.7 (Signal driving force)**.** *At every point with $M = 0$,*

$$\left.\frac{d\mathcal{B}}{dt}\right|_{M=0} = \frac{2\Phi}{\tau N_A N_B}.$$

Proof. At $M = 0$, the flow gives $\dot{W}_{A,s} = 2\alpha W_{B,s}$ and $\dot{W}_{B,s} = 2\alpha W_{A,s}$. Differentiating the alignment coefficient directly yields

$$\dot{\mathcal{B}} = \frac{d}{dt}\langle W_{B,s}, W_{A,s}\rangle_F = 2\alpha\big(\|W_{A,s}\|_F^2 + \|W_{B,s}\|_F^2\big) = \frac{2\Phi}{\tau N_A N_B}.$$

□

C.5 Negative-loss entry and signal extinction

Proposition C.8 (Negative-loss entry or complete signal extinction)**.** *Suppose at least one encoder nuisance factor is initially nonzero. Then exactly one of the following occurs:*

(i) The loss becomes negative at a finite time. The trajectory belongs to $\mathcal{B}_{\mathrm{reg}}$, and $\Phi(t)$ converges to a strictly positive limit.

(ii) The loss stays nonnegative and $\Phi(t) \to 0$.

Thus negative-loss entry is equivalent to $\limsup_{t\to\infty} \Phi(t) > 0$. The nuisance hypothesis holds almost surely under generic initialization.

Proof. Positive constants c, C below are independent of time and may change between estimates. Theorem C.3 gives global existence and a positive temperature lower bound. If the loss becomes negative, it also gives joint-state convergence in $\mathcal{B}_{\mathrm{reg}}$; the nonzero limiting shared score forces a positive limiting signal energy.

Suppose instead that the loss is always nonnegative. Since $\Delta = \mathcal{L} + F(\widehat{M}) \geq 0$, temperature is nondecreasing. If it is bounded, Theorem B.9 gives a stationary limit. Temperature stationarity gives $\mathcal{L} + F(\widehat{M}) = 0$; both terms are nonnegative, so the limiting loss and score product are zero. The stationary encoder equations then force both signal blocks to zero.

For $\tau \to \infty$, use the normalized coefficients in Eq. (23) and rescale time by $ds/dt = \eta$. The functions $\mathcal{B}_N, \mathcal{A}_N$ and their gradients are bounded on the encoder spheres. The temperature equation gives $\dot{\tau} \leq C/\tau^2$ for large τ, so $s(t) \to \infty$. Primes in this proof denote derivatives with respect to s. The equations are

$$W' = 2\nabla\mathcal{B}_N - \eta\nabla\mathcal{A}_N, \qquad \eta' = -2\eta^3(\eta\mathcal{A}_N - \mathcal{B}_N) \leq 0.$$

The nonnegative rescaled loss $\mathcal{L}/\eta = -2\mathcal{B}_N + \eta\mathcal{A}_N$ satisfies

$$(\mathcal{L}/\eta)' = -\|W'\|^2 + \mathcal{A}_N\eta' \leq -\|W'\|^2. \tag{26}$$

Thus $\mathcal{L}/\eta$ converges, $\mathcal{B}_N$ has a nonpositive limit, and $\int_0^\infty \|W'\|^2 ds < \infty$. The compactified field is bounded and continuous on the encoder spheres times $[0, \eta(0)]$, so the trajectory is Lipschitz and W' is uniformly continuous along it. Barbalat's lemma gives $W' \to 0$, hence $\nabla\mathcal{B}_N \to 0$.

A nonzero nuisance factor rules out a negative limit of $\mathcal{B}_N$. On the A side, the output $W_A X$ is the sum of independent signal and nuisance contributions. Its covariance is K_A, and its nuisance covariance is $bW_{A,\perp}W_{A,\perp}^\top \preceq K_A$. Gaussian regression gives

$$\langle R_A A_\perp, W_{A,\perp}\rangle_F = b\,\mathrm{tr}\left(W_{A,\perp}W_{A,\perp}^\top K_A^\dagger \mathbb{E}\frac{(W_A X)(W_A X)^\top}{\|W_A X\|}\right) \geq \frac{2b}{\pi N_A}\|W_{A,\perp}\|_F^2.$$

To verify the inequality, diagonalize K_A on its positive-eigenvalue range. The corresponding eigenvalues of $K_A^\dagger \mathbb{E}[(W_A X)(W_A X)^\top / \|W_A X\|]$ are $\mathbb{E}[Z_i^2/(\sum_j \lambda_j Z_j^2)^{1/2}]$. Cauchy–Schwarz bounds each eigenvalue below by $(\mathbb{E}|Z_i|)^2/N_A = 2/(\pi N_A)$. The nuisance covariance is supported on the same range, so the bound holds at every encoder rank.

For $\Psi_A = \|W_{A,\perp}\|_F^2$, the exact nuisance equation in s-time is

$$\Psi_A' = -\frac{4b\eta}{(N_A N_B)^2}\,\mathrm{tr}(W_{A,\perp}^\top K_B W_{A,\perp}) + \frac{4(\eta\mathcal{A}_N - \mathcal{B}_N)}{N_A}\langle R_A A_\perp, W_{A,\perp}\rangle_F.$$

Uniform normalizer and encoder bounds therefore imply $\Psi_A' \geq [-C\eta + c(\eta\mathcal{A}_N - \mathcal{B}_N)]\Psi_A$. Finite-time non-entry makes $\Psi_A > 0$ if it is initially nonzero. A negative limiting $\mathcal{B}_N$ would give $\Psi_A' \geq c\Psi_A$ on a tail, contradicting the conserved encoder norm. The B-side argument is identical. Hence $\mathcal{B}_N \to 0$.

Finally,

$$\nabla_{W_A}\mathcal{B}_N = \frac{W_B\Sigma_{yx}}{N_A N_B} - \frac{\mathcal{B}_N}{N_A}R_A, \qquad \nabla_{W_B}\mathcal{B}_N = \frac{W_A\Sigma_{xy}}{N_A N_B} - \frac{\mathcal{B}_N}{N_B}R_B.$$

Since $\mathcal{B}_N, \nabla\mathcal{B}_N \to 0$, these identities force both signal factors to zero. This proves the second alternative also when temperature is unbounded. □

The dichotomy identifies signal survival as sufficient for regular entry. The following results characterize the exceptional complete-extinction trajectories.

C.6 Local instability of signal-degenerate equilibria

Signal perturbations with independently chosen amplitudes give negative curvature at every zero-loss equilibrium.

Proposition C.9 (Strict saddle property at zero-loss equilibria)**.** *Every zero-loss equilibrium in $\mathcal{U}$ is a strict saddle on $\mathcal{M} \times \mathbb{R}_{>0}$. Its $2rd$-dimensional signal tangent subspace has exactly rd negative and rd positive Hessian eigenvalues, at every positive temperature.*

Proof. At an equilibrium, $\Delta = 0$ gives $\mathcal{L} = -\alpha^2\mathcal{A}$. Zero loss therefore forces $M = 0$, and the stationary encoder equations force $W_{A,s} = W_{B,s} = 0$. Let $a = 1+\sigma^2$, $b = \sigma^2+\nu^2$, $Q_A = \bar{W}_A\bar{W}_A^\top$, $Q_B = \bar{W}_B\bar{W}_B^\top$, and $\alpha_0 = (\bar{\tau}\bar{N}_A\bar{N}_B)^{-1}$. The zero product gives $Q_A Q_B = 0$.

For signal perturbations $H_A A_s^\top, H_B B_s^\top$, with $H_A, H_B \in \mathbb{R}^{d\times r}$, radial normalization to the original encoder spheres leaves the quadratic loss coefficient unchanged:

$$\mathcal{L}(\epsilon) = \epsilon^2\left[\alpha_0^2 ab\big(\mathrm{tr}(H_A^\top Q_B H_A) + \mathrm{tr}(H_B^\top Q_A H_B)\big) - 2\alpha_0\,\mathrm{tr}(H_B^\top H_A)\right] + o(\epsilon^2).$$

The normalizers are positive and continuous, while both score coefficients vanish to second order. For each shared coordinate, the signal Hessian is

$$2\begin{pmatrix} \alpha_0^2 ab Q_B & -\alpha_0 I_d \\ -\alpha_0 I_d & \alpha_0^2 ab Q_A \end{pmatrix}. \tag{27}$$

The symmetric matrices Q_A, Q_B commute. In a common eigenbasis, their eigenvalues satisfy $q_A q_B = 0$, so each two-dimensional block has determinant $-4\alpha_0^2 < 0$. Each block has one negative and one positive eigenvalue, proving the stated index. First-order nuisance-score variations occupy disjoint blocks and have no quadratic cross term with signal variations. Temperature also has no quadratic coupling here. □

Equal-amplitude certificate. For the restricted perturbation $H_A = H_B = ue_1^\top$, the same calculation reads

$$\left.\frac{d^2\mathcal{L}}{d\epsilon^2}\right|_0 = 2(\alpha_0^2\mathcal{A}_2 - 2\alpha_0\|u\|^2), \tag{28}$$

where

$$\mathcal{A}_2 = ab\,u^\top(Q_A + Q_B)u. \tag{29}$$

Choosing u as a unit eigenvector for $\lambda_{\min}(Q_A + Q_B)$, this direction is descending when

$$\bar{\tau} > \tau_c(\bar{W}_A, \bar{W}_B) = \frac{ab}{2\bar{N}_A\bar{N}_B}\lambda_{\min}(Q_A + Q_B). \tag{30}$$

The threshold applies to equal-amplitude perturbations; Proposition C.9 supplies negative curvature at every positive temperature by allowing the two amplitudes to differ.

Proposition C.10 (Finite-time entry from positive loss)**.** *Every zero-loss equilibrium has arbitrarily close nonempty relative open sets of positive-loss initial states whose trajectories reach negative loss in finite time and hence belong to $\mathcal{B}_{\mathrm{reg}}$.*

Proof. Use orthonormal local coordinates x on the encoder spheres and temperature line at the equilibrium. Both smooth score coefficients $\mathcal{A}, \mathcal{B}$ vanish to second order, while α is C^1. Differentiating $\mathcal{L} = \alpha^2\mathcal{A} - 2\alpha\mathcal{B}$ therefore gives $\nabla\mathcal{L}(x) = Hx + o(\|x\|)$, where H is the Hessian at this point. The coordinate metric contributes only higher-order terms. Thus $\dot{x} = -Hx + o(\|x\|)$.

For a fixed $T < \infty$, the trajectory from $x(0) = \epsilon v$ satisfies

$$\sup_{0\le t\le T} \|\epsilon^{-1}x(t) - e^{-tH}v\| \longrightarrow 0.$$

Indeed, $\|\dot{x}\| \le C\|x\|$ keeps it in an $O(\epsilon)$ neighborhood over this fixed interval, and variation of constants controls the remaining $o(\epsilon)$ term. This argument uses differentiability at the equilibrium, not C^1 regularity of the vector field on an entire neighborhood.

Choose orthonormal signal eigenvectors e_+, e_- of H with eigenvalues $\lambda > 0, -\mu < 0$. For $v = e_+ + ce_-$, where $0 < c^2 < \lambda/\mu$, the initial quadratic loss is positive, whereas along the linearized flow it is

$$\tfrac{1}{2}\big(\lambda e^{-2\lambda t} - \mu c^2 e^{2\mu t}\big) < 0 \quad \text{if} \quad t > \frac{\log(\lambda/(\mu c^2))}{2(\lambda+\mu)}.$$

Fix such a time. The second-order loss expansion gives the same strict signs for sufficiently small $\epsilon > 0$ in the actual flow. Continuous dependence makes these signs persist on an open neighborhood of each initial state. Taking ϵ arbitrarily small proves the entry claim. Theorem C.3 then gives membership in $\mathcal{B}_{\mathrm{reg}}$. □

C.7 Almost-sure exclusion of smooth zero-loss limits

Proposition C.11 (Fixed-temperature exclusion of smooth zero-loss limits)**.** *Fix $C_A, C_B > 0$ and any initial temperature $\tau_0 > 0$. Under generic encoder initialization, the event that the loss stays nonnegative, temperature is bounded, and both limiting encoder ranks are at least two has probability zero.*

Proof. First, the vector field has a Lipschitz inverse flow on finite segments with nonnegative loss, including at rank-one encoders. On a compact regular set, write

$$\mathsf{F}(\theta) = V(\theta) + c_A(\theta)(R_A, 0, 0) + c_B(\theta)(0, R_B, 0), \qquad c_A = \Delta/N_A, \quad c_B = \Delta/N_B.$$

Here V, c_A, c_B are locally C^1, hence Lipschitz on a compact enlargement, while R_A, R_B are bounded. Convexity of each expected norm makes its gradient monotone. For two states x, y with $\Delta \ge 0$, split

$$c_A(x)R_A(x) - c_A(y)R_A(y) = c_A(x)(R_A(x) - R_A(y)) + (c_A(x) - c_A(y))R_A(y).$$

The inner product of the first term with the encoder difference is nonnegative; the second is bounded below by $-C\|x - y\|^2$. Including the B side and V gives

$$\langle \mathsf{F}(x) - \mathsf{F}(y), x - y\rangle \ge -C\|x - y\|^2.$$

Gronwall's inequality therefore bounds the inverse time-T map by a Lipschitz constant e^{CT} on such segments. Nonnegative loss implies $\Delta = \mathcal{L} + F(\widehat{M}) \ge 0$, as required.

Second, Lemma A.5 makes the vector field C^1 near any limit at which both encoders have rank at least two.

A bounded nonnegative-loss trajectory converges to a zero-loss equilibrium by Theorem B.9. Proposition C.9 supplies rd expanding directions of the linearized flow there. Applying the local center-stable theorem to its time-one map gives a manifold of dimension at most $D - rd$, where $D = dd_1 + dd_2 - 1$ is the dimension of $\mathcal{M} \times \mathbb{R}_{>0}$; see the smooth avoidance framework in Lee et al. (2019). Countably many such neighborhoods and integer entry times cover all converging trajectories of this type. A countable compact exhaustion of their finite initial segments, followed by the inverse-Lipschitz bound above, preserves this Hausdorff-dimension bound under pullback.

If a zero-product equilibrium exists, its two nonzero output ranges are orthogonal, so $d \geq 2$ and $rd \geq 2$. Thus $D - rd < D - 1$, the dimension of every fixed-temperature initialization slice. The exceptional set has zero surface measure on each such slice. When $d = 1$, no zero-product equilibrium with nonzero encoders exists. This proves the assertion for every prescribed τ_0. □

Lemma C.12 (Signal decay on bounded zero-loss trajectories)**.** *Suppose the loss stays nonnegative and temperature is bounded above. There are constants $C, \beta > 0$, depending on the trajectory, such that*

$$\|W_{A,s}(t)\|_F + \|W_{B,s}(t)\|_F \leq Ce^{-\beta t}, \qquad t\Delta(t) \longrightarrow 0.$$

No limiting encoder-rank condition is needed.

Proof. Theorem B.9 gives a stationary limit. Nonnegative loss and temperature stationarity force zero loss and $M = 0$; the stationary encoder equations then force both signal factors to zero. Since $\Delta = \mathcal{L} + F(\widehat{M}) \geq \mathcal{L} \geq 0$ and $(\tau^2)' = 2\Delta$,

$$\int_0^\infty \mathcal{L}(t)\, dt \leq \int_0^\infty \Delta(t)\, dt = \tfrac{1}{2}\big(\tau_\infty^2 - \tau(0)^2\big) < \infty.$$

The loss is nonincreasing, so

$$t\mathcal{L}(t) \leq 2\int_{t/2}^t \mathcal{L}(u)\, du \longrightarrow 0. \tag{31}$$

Let $y = (W_{A,s}, W_{B,s})$, so $\|y\| = \phi = \sqrt{\Phi}$. The smooth score part of the signal equations is linear in y, with coefficients converging to a matrix A_*. By Eq. (27), A_* has rd positive and rd negative eigenvalues. The normalizer gradient has zero signal component at zero signal; Lemma A.4 therefore gives

$$\dot{y} = A_* y + R(t), \qquad \|R(t)\| \leq \epsilon_0(t)\phi + C\Delta\phi(1 + |\log\phi|), \qquad \epsilon_0(t) \to 0,$$

with the logarithmic term defined as zero at $\phi = 0$. All constants are uniform on the compact regular trajectory closure. The identity $\Delta = 2\mathcal{L} + 2\alpha\mathcal{B}$, where $\mathcal{B} = \operatorname{tr}(W_{B,s}^\top W_{A,s})$, implies $0 \leq \Delta \leq 2\mathcal{L} + C\phi^2$. The term $\phi^3(1 + |\log\phi|)$ is $o(\phi)$. Fix $\gamma > 0$. If $\phi \geq e^{-\gamma t}$, then $\mathcal{L}\phi(1 + |\log\phi|) \leq (1 + \gamma t)\mathcal{L}\phi = o(\phi)$ by Eq. (31). If $\phi < e^{-\gamma t}$, monotonicity of $x(1 + |\log x|)$ for small $x > 0$ gives $\mathcal{L}\phi(1 + |\log\phi|) \leq (1 + \gamma t)\mathcal{L}e^{-\gamma t} \leq Ce^{-\gamma t}$. Consequently

$$\|R(t)\| \leq \epsilon(t)\phi(t) + Ce^{-\gamma t}, \qquad \epsilon(t) \to 0.$$

Choose $\lambda > 0$ below the absolute values of the eigenvalues of A_*. Variation of constants on its stable subspace from a sufficiently late time T, and on its unstable subspace from infinity using $y(t) \to 0$, yields

$$\phi(t) \leq Ce^{-\lambda(t-T)} + C\int_T^\infty e^{-\lambda|t-u|}\big(\epsilon(u)\phi(u) + e^{-\gamma u}\big)\, du.$$

For $0 < \beta < \min\{\lambda, \gamma\}$, the positive integral operator multiplying ϕ has norm at most

$$C \sup_{u \geq T} \epsilon(u) \left(\frac{1}{\lambda - \beta} + \frac{1}{\lambda + \beta}\right)$$

in the norm $\sup_{t \geq T} e^{\beta(t-T)}|\phi(t)|$. Increasing T makes this bound, and the corresponding unweighted norm, less than one. Iterating the inequality first eliminates the remainder in the unweighted norm because ϕ is bounded. The resulting Neumann series converges in the weighted norm, proving exponential signal decay. Finally, $0 \leq t\Delta(t) \leq 2t\mathcal{L}(t) + Ct\phi(t)^2 \to 0$. □

Lemma C.13 (Finite-time preservation of nuisance rank)**.** *Suppose the loss stays nonnegative and* $\operatorname{rank} W_{A,\perp}(0) \geq 2$*. There is a constant* $C > 0$ *such that*

$$\sigma_2(W_A(t)) \geq \sigma_2(W_{A,\perp}(t)) \geq \sigma_2(W_{A,\perp}(0))e^{-Ct} > 0 \qquad (t \geq 0).$$

The same conclusion holds for B.

Proof. At rank at least two, Gaussian regression gives

$$R_A A_\perp = b\mathsf{T}(K_A)W_{A,\perp}, \qquad \mathsf{T}(K_A) = \mathbb{E}\frac{gg^\top}{(g^\top K_A g)^{1/2}}, \quad g \sim \mathcal{N}(0, I_d).$$

Thus the nuisance equation is

$$\dot{W}_{A,\perp} = \left[-2\alpha^2 b K_B + \frac{b\Delta}{N_A}\mathsf{T}(K_A)\right] W_{A,\perp}.$$

Nonnegative loss gives $\Delta \geq 0$ and $\tau(t) \geq \tau(0)$. The second matrix term is positive semidefinite, and the first is uniformly bounded by norm conservation and the normalizer lower bounds. Therefore

$$\frac{d}{dt}(W_{A,\perp}^\top W_{A,\perp}) \succeq -2CW_{A,\perp}^\top W_{A,\perp}.$$

Integration gives the exponential lower bound and excludes a finite first loss of rank two. Finally, $W_A W_A^\top = W_{A,\perp}W_{A,\perp}^\top + W_{A,s}W_{A,s}^\top$ gives the first inequality. The B-side proof is identical. □

C.8 Exclusion of infinite-temperature escape

Theorem C.14 (Almost-sure boundedness of temperature)**.** *Suppose* $r \geq 1$*,* $d \geq 2$*, and* $d_1 - r, d_2 - r \geq 2$*. Fix* $C_A, C_B > 0$ *and any initial temperature* $\tau_0 > 0$*. Under generic encoder initialization, temperature is bounded above almost surely. More precisely, among trajectories whose two initial nuisance blocks have rank at least two, the set with unbounded temperature has Hausdorff dimension at most*

$$D - rd, \qquad D = d(d_1 + d_2) - 1,$$

in the state space consisting of the two prescribed encoder spheres and the positive temperature line.

We study the boundary $\eta = 1/\tau \to 0$ using the normalized loss coefficients from Eq. (23). Norm conservation bounds $\mathcal{B}_N, \mathcal{A}_N$ and their gradients uniformly. An unbounded-temperature trajectory cannot enter negative loss, by the temperature estimate in Lemma B.15. Consequently $\Delta = \mathcal{L} + F(\widehat{M}) \geq 0$, so τ is nondecreasing and tends to infinity.

Define s by $ds/dt = \eta$, and use primes for derivatives in s. The equations become

$$W' = 2\nabla\mathcal{B}_N - \eta\nabla\mathcal{A}_N, \qquad \eta' = 2\mathcal{B}_N\eta^3 - 2\mathcal{A}_N\eta^4 = -2\eta^3(\eta\mathcal{A}_N - \mathcal{B}_N), \tag{32}$$

Since $\mathcal{L} \geq 0$,

$$\mathcal{B}_N \leq \eta\mathcal{A}_N/2, \qquad -\eta' \geq \eta^4\mathcal{A}_N. \tag{33}$$

Both $s(t)$ and $u(t) = \int_0^t \tau(v)^{-2}\,dv$ tend to infinity. Indeed, boundedness of $\mathcal{B}_N, \mathcal{A}_N$ gives $\dot{\tau} = 2\mathcal{A}_N/\tau^3 - 2\mathcal{B}_N/\tau^2 \leq C/\tau^2$ once $\tau \geq 1$, and hence $\tau(t) \leq C'(1+t)^{1/3}$.

The stable-manifold argument at $\eta = 0$ requires vanishing signal and smoothness near the limiting nuisance encoders. The following lemmas establish integrable signal decay, a positive limit of $\mathcal{A}_N$, and a lower bound on the second nuisance singular value.

Proposition C.8 gives $\mathcal{B}_N \to 0$ and $\Phi \to 0$ whenever at least one initial nuisance block is nonzero and $\tau \to \infty$.

Lemma C.15 (Integrable signal amplitude)**.** *Suppose* $\tau(t) \to \infty$ *and at least one initial nuisance block is nonzero. Then*

$$\int_0^\infty \sqrt{\Phi(s)}\,ds < \infty, \qquad \mathcal{A}_N(s) \longrightarrow \mathcal{A}_{N,*}, \qquad \int_0^\infty \eta\|\nabla\mathcal{A}_N\|^2 ds < \infty.$$

Proof. Write $y = (W_{A,s}, W_{B,s})$, $\phi = \|y\| = \sqrt{\Phi}$, and $p = \eta\mathcal{A}_N$. Let $T(y_A, y_B) = (y_B, y_A)$. The signal equations have the form

$$y' = c_{\mathrm{sig}}(s)Ty + R(s), \qquad c_{\mathrm{sig}}(s) = 2/(N_A N_B) \geq c_0 > 0.$$

Since $\mathcal{B}_N$ has a bilinear signal numerator, $|\mathcal{B}_N| \leq C\phi^2$. Projecting each encoder onto its nuisance columns leaves a matrix bounded away from zero on a tail. Its normalizer gradient has zero signal columns. Lemma A.4 therefore gives

$$\|R_A A_s\|_F + \|R_B B_s\|_F \leq C\phi(1 + |\log\phi|).$$

The polynomial signal derivative of $\mathcal{A}_N$ is $O(\phi)$, and its normalization contribution is bounded by $C\mathcal{A}_N\phi(1 + |\log\phi|)$. Consequently

$$\|R(s)\| \leq C\{\eta\phi + (p + \phi^2)\phi(1 + |\log\phi|)\}.$$

Splitting at $\phi = \exp(-1/\sqrt{p})$ gives, for small p, ϕ,

$$p\phi(1 + |\log\phi|) \leq C\sqrt{p}\,\phi + C\sqrt{p}\,e^{-1/\sqrt{p}}.$$

The expression on the right is defined by continuity at $p = 0$. Hence

$$\|R(s)\| \leq \epsilon(s)\phi(s) + f(s), \qquad \epsilon = C\{\eta + \sqrt{p} + \phi^2(1 + |\log\phi|)\} \to 0, \quad f = C\sqrt{p}\,e^{-1/\sqrt{p}}.$$

Boundedness of $\mathcal{A}_N$ and Eq. (33) give $f \leq Cp^4 \leq C\eta^4\mathcal{A}_N \leq C(-\eta')$ on a tail, so $f \in L^1(ds)$.

The projections onto the fixed ± 1 eigenspaces of T have scalar evolution rates $\pm c_{\mathrm{sig}}(s)$. Variation of constants forward from a late s_0 for the stable component, and backward from infinity for the unstable component using $y \to 0$, yields

$$\phi(s) \leq C\phi(s_0)e^{-c_0(s-s_0)} + C\int_{s_0}^{\infty} e^{-c_0|s-v|}\{\epsilon(v)\phi(v) + f(v)\}\,dv.$$

The positive integral operator multiplying ϕ has norm at most $2C\sup_{v\geq s_0}\epsilon(v)/c_0$ on both L^1 and L^∞. Choose s_0 so that this bound is below one. Iterating the inequality, the remainder tends to zero in L^∞, since ϕ is bounded, and the resulting Neumann series has finite L^1 norm, since its forcing term is integrable. Thus $\phi \in L^1(ds)$.

On nuisance columns $\nabla\mathcal{B}_N = O(\phi^2)$, whereas on signal columns $\nabla\mathcal{B}_N = O(\phi)$ and $\nabla\mathcal{A}_N = O(\phi(1 + |\log\phi|))$. Therefore

$$\mathcal{A}_N' = -\eta\|\nabla\mathcal{A}_N\|^2 + e(s), \qquad |e(s)| = 2|\langle\nabla\mathcal{A}_N, \nabla\mathcal{B}_N\rangle| \leq C\phi^2(1 + |\log\phi|) \leq C\phi.$$

The error is integrable. Subtracting its primitive from $\mathcal{A}_N$ gives a nonincreasing function bounded below. This proves convergence of $\mathcal{A}_N$ and the stated dissipation bound. □

Lemma C.16 (Positive nuisance-score limit)**.** *Suppose $\tau(t) \to \infty$ and at least one initial nuisance block is nonzero. Then the limit in Lemma C.15 satisfies $\mathcal{A}_{N,*} > 0$.*

Proof. We show that nondecreasing, unbounded temperature is incompatible with $M \to 0$. Define

$$Q_A = W_A W_A^\top, \quad Q_B = W_B W_B^\top, \quad Q = Q_A - Q_B,$$
$$\mathcal{Q}_A = R_A W_A^\top/N_A, \qquad \mathcal{Q}_B = R_B W_B^\top/N_B.$$

The matrices $\mathcal{Q}_A, \mathcal{Q}_B$ are positive semidefinite with trace one, and the exact Gram identity gives

$$Q' = (\tau^2)'(\mathcal{Q}_A - \mathcal{Q}_B).$$

The nuclear norm satisfies $\|Q\|_* \leq C_A^2 + C_B^2$. Let $\mathrm{sgn}(Q)$ be the matrix sign, equal to zero on the kernel. The convex chain rule for the nuclear norm gives, almost everywhere,

$$\frac{d}{ds}\|Q\|_* = (\tau^2)'\,\mathrm{tr}\{\mathrm{sgn}(Q)(\mathcal{Q}_A - \mathcal{Q}_B)\}.$$

If $M \to 0$, any convergent encoder subsequence has nonzero orthogonal limiting output ranges E_A, E_B. Passing to a further subsequence, the symmetric contractions $\mathrm{sgn}(Q)$ converge. Their limit acts as $+I$ on E_A and $-I$ on E_B, because the limiting Q is strictly positive on E_A and strictly

negative on E_B. Continuity of the normalizer gradients gives limits of $\mathcal{Q}_A, \mathcal{Q}_B$ supported in E_A, E_B, respectively, each with trace one. Thus

$$\mathrm{tr}\{\mathrm{sgn}(Q)(\mathcal{Q}_A - \mathcal{Q}_B)\} \longrightarrow 2.$$

The subsequence argument proves this limit despite possible zero eigenvalues of Q. Since $(\tau^2)' \geq 0$, eventually $\frac{d}{ds}\|Q\|_* \geq (\tau^2)'$, contradicting bounded $\|Q\|_*$. Finally, $\mathcal{A}_{N,*} = 0$ would imply $M \to 0$ by positive covariance and normalizer bounds. Hence $\mathcal{A}_{N,*} > 0$. □

Lemma C.17 (A rank barrier for nuisance factors)**.** *If $\tau(t) \to \infty$ and $\operatorname{rank} W_{A,\perp}(0) \geq 2$, then*

$$\liminf_{t\to\infty} \sigma_2(W_{A,\perp}(t)) > 0, \qquad \liminf_{t\to\infty} \sigma_2(W_A(t)) > 0.$$

The analogous statement holds for B.

Proof. The output covariance decomposes as

$$K_A = bW_{A,\perp}W_{A,\perp}^\top + aW_{A,s}W_{A,s}^\top.$$

The Gaussian regression identity from Lemma C.13 is $R_A A_\perp = b\mathsf{T}(K_A)W_{A,\perp}$, where $\mathsf{T}(K_A)$ is finite and positive semidefinite when the encoder has rank at least two. Thus

$$W'_{A,\perp} = \left[-\frac{2b\eta}{(N_A N_B)^2}K_B + \frac{2b(\eta\mathcal{A}_N - \mathcal{B}_N)}{N_A}\mathsf{T}(K_A)\right] W_{A,\perp}. \tag{34}$$

Lemma C.13 keeps the nuisance factor at rank at least two at every finite time, so this equation holds throughout the trajectory.

We need a uniform transverse bound. If $\sigma_2(W_{A,\perp}) \to 0$, $W_{A,s} \to 0$, and $\|W_{A,\perp}\|_F$ stays bounded above and away from zero, then every unit second left singular vector v of $W_{A,\perp}$ satisfies

$$v^\top \mathsf{T}(K_A)v \longrightarrow \infty. \tag{35}$$

Along any convergent subsequence, K_A approaches a positive rank-one covariance, and v approaches a vector perpendicular to its leading direction. In a covariance eigenbasis, the transverse diagonal entries of $\mathsf{T}(K_A)$ are $\mathbb{E}[g_i^2/\sqrt{\sum_j \lambda_j g_j^2}]$, $i \geq 2$. The leading eigenvalue is bounded. If $\max_{i\geq 2}\lambda_i \leq \delta^2$, restrict $1 \leq |g_i| \leq 2$, the other transverse coordinates to a fixed bounded box, and $\delta \leq |g_1| \leq 1$. Gaussian densities are bounded below there, and the denominator is at most a constant times $|g_1|$. Each transverse entry is therefore bounded below by

$$c\int_\delta^1 \frac{dx}{x} = c\log(1/\delta) \longrightarrow \infty.$$

The squared transverse projection of v tends to one. Positivity of $\mathsf{T}(K_A)$ proves Eq. (35), uniformly by compactness.

In the same rescaled time s, factor the positive inverse temperature η out of the coefficient matrix in Eq. (34). The remaining matrix is

$$-\frac{2b}{(N_A N_B)^2}K_B + \frac{2b(\mathcal{A}_N - \mathcal{B}_N/\eta)}{N_A}\mathsf{T}(K_A).$$

Lemma C.16 and $\mathcal{B}_N/\eta \leq \mathcal{A}_N/2$ bound the second coefficient below by a positive constant on a tail; the first matrix is bounded. Also $W_{A,s} \to 0$ and $\|W_{A,\perp}\|_F \to C_A$. Equation (35) therefore supplies s_0 and $\delta_0 > 0$ such that this matrix has Rayleigh quotient at least one on every second left singular vector whenever $s \geq s_0$ and $\sigma_2(W_{A,\perp}) \leq \delta_0$. Choose δ_0 small enough to separate the leading singular value. For repeated second singular values the bound holds on the whole corresponding eigenspace. The directional eigenvalue formula gives, almost everywhere in this region,

$$\frac{d}{ds}\sigma_2(W_{A,\perp})^2 \geq 2\eta\sigma_2(W_{A,\perp})^2.$$

A downward crossing of the level δ_0 is impossible, so the second singular value is bounded below by $\min\{\delta_0, \sigma_2(W_{A,\perp}(s_0))\} > 0$. Finally $W_A W_A^\top = W_{A,\perp}W_{A,\perp}^\top + W_{A,s}W_{A,s}^\top$ gives $\sigma_2(W_A) \geq \sigma_2(W_{A,\perp})$. □

Proof of Theorem C.14. Let $\mathcal{M}$ be the product of the two prescribed encoder spheres. We first construct countably many smooth auxiliary fields that agree with the relevant trajectory tails.

Lemma A.5 gives C^2 regularity away from rank one. For a single nonzero encoder matrix W, define the smooth, degree-zero rank statistic

$$\rho_2(W) = \frac{\{\mathrm{tr}(WW^\top)\}^2 - \mathrm{tr}\{(WW^\top)^2\}}{2\|W\|_F^4}. \tag{36}$$

It vanishes exactly at rank one. Let χ_m be a smooth cutoff equal to zero for $\rho_2 \le 1/(2m)$ and one for $\rho_2 \ge 1/m$, and set

$$\widetilde{N}_m(W) = \chi_m(\rho_2(W))N(W) + [1 - \chi_m(\rho_2(W))]\|W\|_F.$$

This positive, degree-one homogeneous function is C^2 on nonzero matrices and retains signal sign symmetry. It equals the original expected norm when $\rho_2 \ge 1/m$. Apply this construction to N_A and N_B separately, writing $\widetilde{N}_A, \widetilde{N}_B$ with the fixed cutoff index m suppressed. Substitution in Eq. (23) defines $\widetilde{\mathcal{B}}_N, \widetilde{\mathcal{A}}_N$. By Lemma C.17, every trajectory under consideration eventually lies in $\rho_2(W_A), \rho_2(W_B) > 1/m$ for some integer m.

Add a circle coordinate ϑ, and choose a smooth real function ρ on the circle equal to ϑ near zero. On the compact manifold $\mathcal{M} \times S^1$, consider

$$W' = 2\nabla\widetilde{\mathcal{B}}_N - \delta\rho(\vartheta)\nabla\widetilde{\mathcal{A}}_N, \qquad \vartheta' = 2\delta^2\widetilde{\mathcal{B}}_N\rho(\vartheta)^3 - 2\delta^3\widetilde{\mathcal{A}}_N\rho(\vartheta)^4.$$

Homogeneity makes the encoder vector fields tangent to their spheres. This C^1 field converges in C^1 to $(2\nabla\widetilde{\mathcal{B}}_N, 0)$ as $\delta \to 0$. The compact boundaryless manifold

$$\mathcal{C} = \{W_{A,s} = W_{B,s} = 0\} \times S^1$$

is invariant for every δ. At $\delta = 0$ it consists of equilibria, its tangent linearization is zero, and its normal signal linearization is

$$(W_{A,s}, W_{B,s})' = \frac{2}{\widetilde{N}_A\widetilde{N}_B}(W_{B,s}, W_{A,s}).$$

It therefore has stable and unstable normal bundles of dimension rd, with rates uniformly separated from zero. The compact normally hyperbolic manifold theorem applied to the time-one map supplies a C^1 local center-stable manifold of codimension rd for all sufficiently small $\delta > 0$; it contains all forward orbits remaining in a sufficiently small neighborhood of $\mathcal{C}$ (Hirsch et al., 1970, Theorems 2.1–2.2). The persisted manifold is $\mathcal{C}$ itself: its invariance places it in the intersection of the local stable and unstable sets, and the two compact connected manifolds have the same dimension.

Fix such a δ for each m. Where $\rho(\vartheta) = \vartheta$, the change $\eta = \delta\vartheta$ recovers Eq. (32). Each actual infinite-temperature trajectory with initial nuisance ranks at least two eventually agrees with one of these auxiliary fields and stays in its chosen neighborhood of $\mathcal{C}$, by Proposition C.8 and Lemma C.17. Its tail belongs to a center-stable manifold of dimension $D - rd$. Changing from η to τ preserves this dimension on every compact positive-temperature chart.

To pull back to initialization, use integer times in the original physical clock. Lemma C.13 keeps both nuisance ranks at least two at every finite time. Thus both full encoders have rank at least two along each relevant finite segment, and the original vector field is C^1 on a neighborhood of that compact segment. Its local inverse physical-time maps are C^1, hence locally Lipschitz. Countably many compact charts cover the finite segments and the positive-temperature coordinate changes.

Every trajectory under consideration enters one of the countably many center-stable neighborhoods before some integer physical time. Pullback from those integer times preserves Hausdorff dimension, so the full exceptional set has dimension at most $D - rd$. Since $rd \ge 2$, this is strictly below $D - 1$, the dimension of every fixed-temperature encoder slice. Its intersection with each such slice has zero surface measure. Finally, when $d, d_1 - r, d_2 - r \ge 2$, initial nuisance rank below two is a proper algebraic condition and has zero surface measure. Generic initialization avoids these sets, proving almost-sure boundedness of temperature. □

C.9 Excluding zero-loss limits at any encoder rank

Near a rank-one encoder, the normalization gradient is log-Lipschitz. We construct a smooth auxiliary field that agrees with each relevant trajectory on a tail. A center-stable graph for this field then bounds the dimension of initial states that approach zero loss.

Theorem C.18 (Exclusion of bounded zero-loss convergence)**.** *Fix $C_A, C_B > 0$. Among initial states with $\operatorname{rank} W_{A,\perp}(0), \operatorname{rank} W_{B,\perp}(0) \geq 2$, let $\mathcal{E}_{\mathrm{b}}$ consist of those whose loss stays nonnegative and whose temperature is bounded above. In the state manifold $\mathcal{M} \times \mathbb{R}_{>0}$,*

$$\dim_H \mathcal{E}_{\mathrm{b}} \leq D - rd, \qquad D = d(d_1 + d_2) - 1.$$

In particular, $\mathcal{E}_{\mathrm{b}}$ has zero surface measure on every fixed-temperature initialization slice. No limiting encoder-rank condition is imposed.

Proof. We cover the exceptional trajectories by countably many Lipschitz graphs and pull those graphs back to initialization.

Every trajectory in $\mathcal{E}_{\mathrm{b}}$ converges by Theorem B.9. At its stationary limit, $\Delta = \mathcal{L} + F(\widehat{M}) = 0$; nonnegative loss therefore forces $\mathcal{L} = 0$ and $M = 0$. The limit is a zero-loss equilibrium.

Lemma C.19 gives a neighborhood of each such equilibrium and a countable family of smooth auxiliary flows. Their linear part is the negative loss Hessian, with exactly rd positive eigenvalues by Proposition C.9; the other eigenvalues are nonpositive. Choose the Lipschitz remainder small enough for Lemma C.20. For each auxiliary flow and each integer starting time n, all bounded forward trajectories then lie on a Lipschitz graph of dimension $D - rd$.

Take a countable subcover of the equilibrium neighborhoods. Every original trajectory under consideration eventually stays in one selected neighborhood and agrees there with one auxiliary flow. At some integer time n, its state therefore lies on the corresponding graph. The chart, auxiliary flow, and entry time have only countably many choices.

Lemma C.13 keeps both encoder ranks at least two at every finite time. The original vector field is consequently C^1 near every compact finite orbit segment, so its inverse time-n map is locally Lipschitz. A countable cover by such inverse-flow neighborhoods pulls the graphs back without increasing Hausdorff dimension. Hence $\dim_H \mathcal{E}_{\mathrm{b}} \leq D - rd$. Finally, the initial rank assumption gives $d \geq 2$, so $rd \geq 2$ and $D - rd < D - 1$, the dimension of each fixed-temperature slice. □

The following two lemmas separate the approximation estimates from the geometric argument. The first supplies smooth flows that agree exactly with the original trajectory on a tail; the second controls bounded trajectories of each auxiliary flow.

Lemma C.19 (Smooth auxiliary flows with exact tail agreement)**.** *Fix a zero-loss equilibrium p. For any prescribed positive Lipschitz tolerance, there is a coordinate neighborhood U_p and a countable family of globally defined time-dependent vector fields*

$$\widehat{\mathsf{F}}_t(x) = \mathsf{A}x + r_t(x), \qquad \mathsf{A} = -\operatorname{Hess} \mathcal{L}(p), \quad r_t(0) = 0,$$

where $x(p) = 0$. The linear part A has exactly rd positive eigenvalues and all other eigenvalues are nonpositive. The fields are continuous in time and C^1 in space, and the Lipschitz constants of r_t are below the prescribed tolerance, uniformly in time and across the family. Every nonnegative-loss, bounded-temperature trajectory with initial nuisance ranks at least two and limit in U_p agrees in these coordinates with one field on a tail. The schedules use the original physical time.

Proof.

Smoothing the normalization gradient. Fix a zero-loss equilibrium p, and use orthonormal product coordinates x on the encoder spheres and temperature line, with $x(p) = 0$. Put $c_i = \Delta/N_i$ for $i \in \{A, B\}$. Write the coordinate vector field as

$$\mathsf{F}(x) = V(x) + c_A(x)\mathcal{R}_A(x) + c_B(x)\mathcal{R}_B(x).$$

Here V is C^1, and $\mathcal{R}_i$ is the corresponding normalization-gradient contribution after applying the smooth coordinate differential. This differential can be extended smoothly to ambient vectors, so the

decomposition does not require its individual summands to be tangent before the coordinate change. At the origin,

$$V(0) = 0, \quad DV(0) = \mathsf{A}, \quad c_i(0) = 0, \quad Dc_i(0) = 0. \tag{37}$$

The coefficient identities follow because both smooth score coefficients vanish to second order at p, while the normalizers are positive and C^1. The pointwise linearization $\mathsf{A} = -\operatorname{Hess}\mathcal{L}(p)$ is symmetric in these coordinates. Proposition C.9 supplies exactly rd positive eigenvalues on the signal subspace. The remaining Hessian terms are the nonnegative squared first-order nuisance scores, with no quadratic coupling to signal or temperature. Thus all remaining eigenvalues of A are nonpositive.

On a compact coordinate neighborhood, $\mathcal{R}_i$ is bounded and log-Lipschitz by Lemma A.4. It is C^1 wherever encoder i has rank at least two, and

$$\|D\mathcal{R}_i\| \le C\big(1 + |\log \sigma_2(W_i)|\big). \tag{38}$$

This follows from Lemma A.5 and the smooth coordinate change. On the fixed positive-norm neighborhood,

$$c\sigma_2(W_i)^2 \le \rho_2(W_i) \le C\sigma_2(W_i)^2, \qquad \|D\rho_2(W_i)\| \le C\sqrt{\rho_2(W_i)}, \tag{39}$$

for positive constants c, C, independent of the smoothing radius below. The numerator is twice the sum of products of pairs of squared singular values, proving comparability. It is also twice the sum of squared two-by-two minors; differentiation and Cauchy–Schwarz give the derivative bound.

We first smooth $\mathcal{R}_i$, then make the approximation equal to the original field away from rank one. Extend it from a larger coordinate ball with a smooth cutoff and convolve with a mollifier of radius δ. Denote the result by $\mathcal{R}_i^\delta$. On a smaller fixed ball, for sufficiently small $\delta > 0$, the log-Lipschitz modulus gives

$$\|\mathcal{R}_i^\delta - \mathcal{R}_i\| \le C\delta(1 + |\log \delta|), \qquad \|D\mathcal{R}_i^\delta\| \le C(1 + |\log \delta|).$$

The second estimate uses the zero integral of the mollifier derivative. Choose a smooth scalar cutoff χ, equal to zero on $(-\infty, 1]$ and one on $[4, \infty)$, and set

$$\widehat{\mathcal{R}}_{i,\delta} = \mathcal{R}_i^\delta + \chi(\rho_2(W_i)/\delta^2)(\mathcal{R}_i - \mathcal{R}_i^\delta).$$

This is C^1 on the whole ball, agrees with $\mathcal{R}_i$ when $\rho_2(W_i) \ge 4\delta^2$, and satisfies

$$\|\widehat{\mathcal{R}}_{i,\delta}\| \le C, \qquad \|D\widehat{\mathcal{R}}_{i,\delta}\| \le C(1 + |\log \delta|). \tag{40}$$

Near rank one the cutoff vanishes identically. Wherever it is nonzero, Eq. (39) bounds σ_2 below by a constant times δ, so Eq. (38) applies. On the transition region, $\|D\chi(\rho_2(W_i)/\delta^2)\| \le C/\delta$; multiplying by the mollification error gives the same logarithmic bound. All constants are independent of δ.

Clipping its coefficient. The derivative bound grows as $|\log \delta|$. We control this growth by truncating the coefficient multiplying the smoothed gradient. Choose a smooth bounded function κ with bounded derivative, such that $\kappa(u) = u$ for $|u| \le 1$ and $|\kappa(u)| \le C\min\{|u|, 1\}$. For integers $j, K \ge 1$, physical time $t \ge 0$, and a constant $\epsilon > 0$, define

$$\delta_t = e^{-j-Kt}, \qquad a_t = \frac{\epsilon}{1 + j + Kt}, \qquad \widehat{c}_{i,t} = a_t\kappa(c_i/a_t).$$

Restrict j to be sufficiently large for all mollification radii to fit within the coordinate margin. Since

$$D\widehat{c}_{i,t} = \kappa'(c_i/a_t)Dc_i, \qquad |\widehat{c}_{i,t}| \le Ca_t,$$

the auxiliary field

$$\widehat{\mathsf{F}}_t(x) = V(x) + \sum_{i=A,B} \widehat{c}_{i,t}(x)\widehat{\mathcal{R}}_{i,\delta_t}(x)$$

fixes the origin and obeys, on a fixed ball,

$$\sup_x \|D\widehat{\mathsf{F}}_t(x) - \mathsf{A}\| \le \sup_x \|DV(x) - \mathsf{A}\| + C\sum_i \sup_x \|Dc_i(x)\| + C\epsilon. \tag{41}$$

The two schedules are chosen so that $a_t(1 + |\log \delta_t|) = \epsilon$. Thus the derivative estimate is uniform in t, j, K. Equation (37) controls the first two terms by shrinking the ball. Choosing ϵ small then makes the full bound arbitrarily small on that fixed ball.

Let the coordinate ball have radius ρ. Multiply $\widehat{\mathsf{F}}_t(x) - \mathsf{A}x$ by a smooth spatial cutoff equal to one on $B_{\rho/2}$ and zero outside B_ρ. This defines the global remainder r_t in the statement. If the derivative bound in Eq. (41) is small, its Lipschitz constant is small as well: the cutoff derivative has size $O(1/\rho)$, whereas the uncut remainder is bounded by that derivative bound times $\|x\|$. Choosing the ball and ϵ accordingly meets the prescribed tolerance. Set $U_p = B_{\rho/4}$.

Exact agreement on each relevant tail. Every trajectory in the statement converges to a zero-loss equilibrium by Theorem B.9: at its stationary limit, $\Delta = \mathcal{L} + F(\widehat{M}) = 0$, so nonnegative loss forces $\mathcal{L} = 0$ and $M = 0$. Lemma C.12 gives

$$tc_A(t) \to 0, \qquad tc_B(t) \to 0, \qquad c_A = \Delta/N_A, \quad c_B = \Delta/N_B. \tag{42}$$

Both normalizers have positive lower bounds on the prescribed encoder spheres.

For each fixed trajectory, Lemma C.13 gives positive constants c, C, independent of time, such that

$$\sigma_2(W_i(t)) \geq ce^{-Ct} > 0, \qquad i \in \{A, B\}.$$

Thus the original vector field is C^1 near every finite trajectory segment.

For a trajectory in the statement, the finite-time rank estimate and Eq. (39) give $\rho_2(W_i(t)) \geq ce^{-Ct}$, after adjusting $c, C > 0$. Choose an integer $K > C$, then choose an integer j large enough that

$$\rho_2(W_i(t)) \geq 4e^{-2j-2Kt} \qquad (t \geq 0).$$

The rank smoothing is then inactive wherever the trajectory lies in the selected chart. Equation (42) supplies an integer n such that

$$|c_i(t)| \leq \frac{\epsilon}{1 + j + Kt} \qquad (t \geq n).$$

Thus coefficient clipping is also inactive on the entire tail. Once the tail stays in $B_{\rho/2}$, it therefore satisfies the auxiliary equation exactly. All schedules use global physical time, so j, K can be fixed before increasing n. The integers j, K give a countable family. A trajectory with limit anywhere in U_p eventually stays in $B_{\rho/2}$, so the spatial cutoff is also inactive on its tail. Agreement does not require convergence to the chart center. □

Lemma C.20 (Center-stable graphs for a time-dependent perturbation)**.** *Let* A *be symmetric with at least one positive eigenvalue, and let its orthogonal spectral subspaces be* E_{cs} *and* E_u *for its nonpositive and positive eigenvalues. Let* $\lambda > 0$ *be its smallest positive eigenvalue. Suppose* $r_t(x)$ *is continuous in* (t, x)*, vanishes at* $x = 0$*, and is globally Lipschitz in* x *with a uniform constant* $\ell < \lambda/4$*. At each starting time* $t_0 \geq 0$*, the initial states of bounded forward solutions of*

$$\dot{x} = \mathsf{A}x + r_t(x)$$

lie on a Lipschitz graph over E_{cs}*, of dimension* $\dim E_{cs}$*.*

Proof. Write P_{cs}, P_u for the orthogonal projections and $\mathsf{A}_{cs}, \mathsf{A}_u$ for the restrictions of A. For a prescribed stable component $v \in E_{cs}$, write $s = t - t_0$ for elapsed physical time and consider

$$x_{cs}(s) = e^{\mathsf{A}_{cs}s}v + \int_0^s e^{\mathsf{A}_{cs}(s-u)}P_{cs}r_{t_0+u}(x(u))\,du,$$

$$x_u(s) = -\int_s^\infty e^{\mathsf{A}_u(s-u)}P_u r_{t_0+u}(x(u))\,du, \qquad s \geq 0.$$

On the complete space of continuous paths with finite norm $\sup_{s\geq 0} e^{-\lambda s/2}\|x(s)\|$, the two integral operators have Lipschitz constants at most $2\ell/\lambda$ each. Their sum is a contraction since $4\ell/\lambda < 1$. The fixed point depends Lipschitz-continuously on v; evaluating its unstable component at zero gives the claimed graph.

Every bounded forward solution satisfies these equations by variation of constants. In the unstable component, the terminal term vanishes when its terminal time tends to infinity. Such a solution is therefore the unique fixed point for its initial stable component, and lies on the graph. The graph has dimension $\dim E_{cs}$, since its parameterization is Lipschitz and its inverse is the orthogonal projection. □

Proof of Corollary 2. Initial nuisance rank below two is a proper algebraic condition on the encoder spheres and has zero surface measure under the stated dimensions. For the remaining initializations, Theorem C.14 excludes unbounded temperature outside a null set on the prescribed τ_0 slice. Theorem C.18 excludes bounded trajectories whose loss never becomes negative, again on every fixed-temperature slice. These two alternatives exhaust failure of finite-time negative-loss entry. Theorem C.3 then gives membership in $\mathcal{B}_{\mathrm{reg}}$. □

Proposition C.21 (Generic entry for scalar output)**.** *For $d = 1$, generic encoder initialization reaches $\mathcal{B}_{\mathrm{reg}}$ almost surely at every prescribed $\tau(0) > 0$, including when either modality has only one nuisance coordinate.*

Proof. For nonzero row encoders, $N_A = \sqrt{2/\pi}\sqrt{K_A}$ and $N_B = \sqrt{2/\pi}\sqrt{K_B}$, where the output covariances K_A, K_B are positive scalars. These normalizers are smooth. The rank-one product identity gives

$$\mathcal{A} = (W_A \Sigma_x W_A^\top)(W_B \Sigma_y W_B^\top), \qquad \mathcal{A}_N = \frac{\mathcal{A}}{(N_A N_B)^2} = \frac{\pi^2}{4}.$$

Thus the compactified encoder equation (32) is the autonomous smooth flow $W' = 2\nabla \mathcal{B}_N$, independent of temperature. A bounded trajectory with nonnegative loss would converge to a zero-product equilibrium. This is impossible: the outer product of two nonzero rows is nonzero. Every failure of negative-loss entry therefore has $\tau \to \infty$, with compactified time $s \to \infty$ as established in Appendix C.8.

Generic initialization has nonzero nuisance factors, so Proposition C.8 makes both signal factors tend to zero on such a trajectory. The compact zero-signal submanifold of the encoder spheres consists of equilibria of $2\nabla \mathcal{B}_N$, with zero tangent linearization and normal signal equations

$$(W_{A,s}, W_{B,s})' = \frac{2}{N_A N_B}(W_{B,s}, W_{A,s}).$$

It has r expanding and r contracting directions, uniformly separated from zero. The normally hyperbolic manifold theorem gives a local stable manifold of codimension r in the encoder state space. Its countably many inverse integer-time images have zero encoder surface measure. This argument also applies when a nuisance sphere has dimension zero: it is a finite compact smooth manifold, and its components are treated separately. Since the compactified encoder flow is independent of $\tau(0)$, the null set applies to every fixed initial temperature. Theorem C.3 converts finite-time negative loss into entry into $\mathcal{B}_{\mathrm{reg}}$. □

C.10 Analytic refinement at full column rank

On the full-column-rank stratum, analyticity also gives convergence rates through the classical Łojasiewicz exponent.

Remark C.22 (Real-analytic refinement on the full-column-rank stratum). Write $X = \Sigma_x^{1/2} Z$, with $Z \sim \mathcal{N}(0, I_{d_1})$, and set

$$\widetilde{G}(W) = \Sigma_x^{1/2} W^\top W \Sigma_x^{1/2} \in \mathbb{R}^{d_1 \times d_1}, \qquad N(W) = \mathbb{E}\sqrt{Z^\top \widetilde{G}(W) Z}.$$

At a full-column-rank matrix W_0, $\widetilde{G}_0 = \widetilde{G}(W_0) \succ 0$. On a sufficiently small neighborhood of W_0, $\|\widetilde{G}_0^{-1/2}(\widetilde{G}(W) - \widetilde{G}_0)\widetilde{G}_0^{-1/2}\|_{\mathrm{op}} \le q < 1$. For $Z \neq 0$, expand

$$\sqrt{Z^\top \widetilde{G}(W) Z} = \sqrt{Z^\top \widetilde{G}_0 Z}\left(1 + \frac{Z^\top(\widetilde{G}(W) - \widetilde{G}_0)Z}{Z^\top \widetilde{G}_0 Z}\right)^{1/2}.$$

The binomial series is uniformly absolutely convergent in the ratio, whose absolute value is at most q, and is dominated by a constant times the integrable quantity $\sqrt{Z^\top \widetilde{G}_0 Z}$. Termwise expectation gives a convergent power series in $\widetilde{G}(W) - \widetilde{G}_0$. Hence N is real-analytic on $\{W \in \mathbb{R}^{d\times d_1} : \operatorname{rank}(W) = d_1\}$.

When $d \geq \max\{d_1, d_2\}$, both full-column-rank conditions define an open dense subset of $\mathcal{U}$, and $\mathcal{L}$ is analytic there. The main trajectory results use the C^1 regularity of Lemma A.2 and the Osgood modulus of Lemma A.4, which also cover rank-deficient nonzero encoders. The analytic theorem below refines the description for trajectories precompact within that region.

Theorem C.23 (Łojasiewicz–Simon convergence; Simon, 1983)**.** *Let $L : \mathcal{U} \to \mathbb{R}$ be real-analytic on the open set $\mathcal{U} \subseteq \mathbb{R}^N$, and let $\theta : [0,\infty) \to \mathcal{U}$ be a maximal gradient flow trajectory $\dot\theta = -\nabla L(\theta)$. Suppose $\theta([0,\infty))$ is precompact in $\mathcal{U}$, i.e., $\overline{\theta([0,\infty))} \subset \mathcal{U}$ is compact. Then there exists a critical point $\bar\theta \in \mathcal{U}$ ($\nabla L(\bar\theta) = 0$) such that $\theta(t) \to \bar\theta$ as $t \to \infty$, and the trajectory has finite arclength: $\int_0^\infty \|\dot\theta(t)\|\, dt < \infty$. The convergence rate is governed by the Łojasiewicz exponent at $\bar\theta$.*

Proof. Precompactness and energy dissipation give a critical accumulation point $\bar\theta$. Put $e(t) = L(\theta(t)) - L(\bar\theta)$, so $e(t) \downarrow 0$. Near $\bar\theta$, the analytic gradient inequality gives $\|\nabla L\| \geq c|L - L(\bar\theta)|^{1-\beta}$ for some $c > 0$ and $\beta \in (0, 1/2)$ (Simon, 1983, Eq. (2.4), pp. 537–538). Wherever $e > 0$, energy dissipation yields

$$-\frac{d}{dt} e(t)^\beta = \beta e(t)^{\beta-1} \|\nabla L(\theta(t))\|^2 \geq c\beta \|\dot\theta(t)\|.$$

Choose a sufficiently late visit to a smaller ball about $\bar\theta$. The remaining length bound $e(t)^\beta/(c\beta)$ is smaller than the distance to the boundary of the gradient-inequality neighborhood, so a first exit is impossible. Integrating gives finite tail length and hence convergence to $\bar\theta$; the initial finite interval also has finite length. If e reaches zero, energy dissipation makes the subsequent trajectory stationary. This is the length argument of Simon (1983, Section 3, Lemma 1). Combining the gradient inequality with $e' = -\|\nabla L\|^2$ gives the rate bound. □

For the original state space in Eq. (8), norm conservation and two-sided temperature bounds give precompactness. Applying Theorem C.23 within the full-column-rank analytic region additionally requires that the closure of the trajectory remain inside that region, away from rank loss. By Sylvester's rank inequality and Proposition B.21, a stationary limit with both encoders of full column rank requires $d_1 + d_2 - d \leq r$. Thus the default setting $d = 25$, $d_1 = d_2 = 20$, $r = 3$ has a rank-deficient encoder limit. Theorem B.9 covers these limits and supplies point convergence on the original state space.

D ℓ_2 WEIGHT DECAY: RADIAL ACTION AND TEMPERATURE COUPLING

Positive-rescaling invariance separates isotropic encoder weight decay into a common radial contraction and a change in the relative speed of temperature learning. We derive the two effects and their consequences for signal-to-nuisance ratios.

Setup. Let $\lambda \geq 0$ and consider the weight-decay flow (weight decay applied only to the encoders, not to the temperature):

$$\begin{aligned} \dot{W}_A &= -\nabla_{W_A}\mathcal{L}(W_A, W_B, \tau) - \lambda W_A, \\ \dot{W}_B &= -\nabla_{W_B}\mathcal{L}(W_A, W_B, \tau) - \lambda W_B, \\ \dot{\tau} &= -\nabla_{\tau}\mathcal{L}(W_A, W_B, \tau). \end{aligned} \tag{43}$$

We write $W_A^{(\lambda)}$ etc. when emphasis on the regularization level is required. Define the encoder *nuisance energy*

$$\Psi(t) \;=\; \|W_{A,\perp}(t)\|_F^2 + \|W_{B,\perp}(t)\|_F^2,$$

in parallel with the signal energy $\Phi(t) = \|W_{A,s}(t)\|_F^2 + \|W_{B,s}(t)\|_F^2$ introduced in Section 3.3.

Theorem D.1 (ℓ_2 weight decay as exact radial rescaling)**.** *Define the scale-corrected trajectory*

$$\widetilde{W}_A(t) = e^{\lambda t}\, W_A^{(\lambda)}(t), \qquad \widetilde{W}_B(t) = e^{\lambda t}\, W_B^{(\lambda)}(t),$$

with corresponding rescaled signal and nuisance energies

$$\widetilde{\Phi}(t) = \|\widetilde{W}_{A,s}\|_F^2 + \|\widetilde{W}_{B,s}\|_F^2, \qquad \widetilde{\Psi}(t) = \|\widetilde{W}_{A,\perp}\|_F^2 + \|\widetilde{W}_{B,\perp}\|_F^2.$$

Then, for every t in the maximal interval of existence,

$$W_A^{(\lambda)}(t) = e^{-\lambda t}\widetilde{W}_A(t), \qquad W_B^{(\lambda)}(t) = e^{-\lambda t}\widetilde{W}_B(t),$$

and hence

$$\Phi^{(\lambda)}(t) = e^{-2\lambda t}\widetilde{\Phi}(t), \qquad \Psi^{(\lambda)}(t) = e^{-2\lambda t}\widetilde{\Psi}(t), \qquad \frac{\Phi^{(\lambda)}(t)}{\Psi^{(\lambda)}(t)} = \frac{\widetilde{\Phi}(t)}{\widetilde{\Psi}(t)}$$

whenever the ratio is defined. In the rescaled time

$$s(t) = \int_0^t e^{2\lambda u}\, du = \begin{cases} \dfrac{e^{2\lambda t} - 1}{2\lambda}, & \lambda > 0, \\ t, & \lambda = 0. \end{cases}$$

the scale-corrected encoders satisfy the unregularized encoder-gradient equations driven by a non-autonomous temperature:

$$\frac{d\widetilde{W}_A}{ds} = -\nabla_{W_A}\mathcal{L}(\widetilde{W}_A, \widetilde{W}_B, \tau^{(\lambda)}), \qquad \frac{d\widetilde{W}_B}{ds} = -\nabla_{W_B}\mathcal{L}(\widetilde{W}_A, \widetilde{W}_B, \tau^{(\lambda)}),$$

while

$$\frac{d\tau^{(\lambda)}}{ds} = e^{-2\lambda t(s)}\big[-\nabla_{\tau}\mathcal{L}(\widetilde{W}_A, \widetilde{W}_B, \tau^{(\lambda)})\big].$$

Thus isotropic ℓ_2 weight decay has no direct angular preference between signal and nuisance components; it only introduces a common radial contraction plus a non-autonomous slowing of the temperature equation.

Proof. The separate positive-rescaling invariance used in Proposition B.1 gives

$$\mathcal{L}(cW_A, W_B, \tau) = \mathcal{L}(W_A, W_B, \tau) = \mathcal{L}(W_A, cW_B, \tau), \qquad c > 0. \tag{44}$$

Differentiation gives the gradient homogeneity

$$\nabla_{W_A}\mathcal{L}(cW_A, W_B, \tau) = c^{-1}\nabla_{W_A}\mathcal{L}(W_A, W_B, \tau), \tag{45}$$

and its B-side counterpart. Substituting the weight-decay equation into the derivative of $\widetilde{W}_A = e^{\lambda t} W_A^{(\lambda)}$ cancels the radial term:

$$\begin{aligned}\dot{\widetilde{W}}_A &= -e^{\lambda t} \nabla_{W_A} \mathcal{L}(e^{-\lambda t}\widetilde{W}_A, e^{-\lambda t}\widetilde{W}_B, \tau^{(\lambda)}) \\ &= -e^{2\lambda t} \nabla_{W_A} \mathcal{L}(\widetilde{W}_A, \widetilde{W}_B, \tau^{(\lambda)}).\end{aligned}$$

The same identity holds for B, while scale invariance leaves $\dot{\tau}^{(\lambda)} = -\nabla_\tau \mathcal{L}(\widetilde{W}_A, \widetilde{W}_B, \tau^{(\lambda)})$. Dividing by $ds/dt = e^{2\lambda t}$ proves the stated equations. Each signal or nuisance block inherits the common factor $e^{-\lambda t}$; its squared norm inherits $e^{-2\lambda t}$, which cancels in the ratio. □

Remark D.2 (Norm decay). The rescaling-invariance identity $\langle W_A, \nabla_{W_A} \mathcal{L} \rangle_F = 0$ from Eq. (44) also gives the exact full-norm equation $\frac{d}{dt}\|W_A^{(\lambda)}\|_F^2 = -2\lambda \|W_A^{(\lambda)}\|_F^2$, whose integration yields $\|W_A^{(\lambda)}(t)\|_F = e^{-\lambda t} \|W_A^{(\lambda)}(0)\|_F$. In particular, $\|W_{A,\perp}(t)\|_F, \|W_{A,s}(t)\|_F \le \|W_A^{(\lambda)}(t)\|_F = O(e^{-\lambda t})$: this is a common upper envelope, not equality of their individual decay rates. On a global trajectory with $\lambda > 0$, all unscaled encoder components tend to zero. Their relative proportions depend on the scale-corrected dynamics.

Corollary D.3 (Cancellation of the direct decay term)**.** *At a fixed state, write a and b for the contributions of the contrastive vector field to $\dot{\Phi}$ and $\dot{\Psi}$. The regularized equations give*

$$\dot{\Phi} = a - 2\lambda\Phi, \qquad \dot{\Psi} = b - 2\lambda\Psi, \qquad \frac{d}{dt}\frac{\Phi}{\Psi} = \frac{a\Psi - \Phi b}{\Psi^2}$$

whenever $\Psi > 0$. Thus there is no direct selective decay contribution to this derivative. Equality of ratios between two separately trained runs does not follow, since their states and temperature trajectories may differ. If temperature is fixed, the scale-corrected encoder flow is exactly the unregularized flow at the rescaled time; the obstruction above is specific to joint temperature evolution.

Proof. Differentiate the quotient and cancel the two $2\lambda\Phi\Psi$ terms. The fixed-temperature assertion follows directly from Theorem D.1, since there is then no temperature equation to reweight. □

On the ambient state space $\mathcal{U}$, the sets defined by $W_{A,\perp} = 0$ and $W_{B,\perp} = 0$ are relatively closed smooth submanifolds. Their intersections with $\mathcal{M} \times \mathbb{R}_{>0}$ are $\mathcal{N}_A$ and $\mathcal{N}_B$. The ambient sets remain invariant under weight decay: Lemma B.13 gives tangency of the contrastive field, and the added nuisance-block term $-\lambda W_{A,\perp}$ vanishes on the A-side set, with the same argument for B. Adding a linear field preserves the local Osgood modulus, so Theorem A.6 gives forward–backward uniqueness and finite-time non-entry while the solution remains in $\mathcal{U}$. The encoder norms follow Remark D.2; the radial factor sends every unscaled component to zero on a global trajectory when $\lambda > 0$. Figure 3 compares the relative signal and nuisance energies.

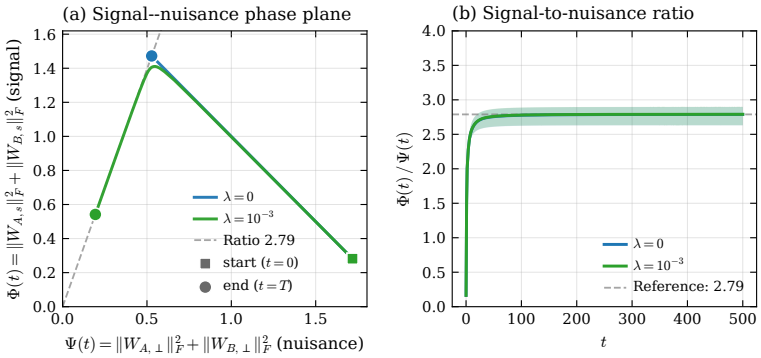

Figure 3: **Signal and nuisance under encoder weight decay.** Population-linear flows with $\lambda \in \{0, 10^{-3}\}$, summarized over 10 seeds. (a) Nuisance energy $\Psi(t)$ versus signal energy $\Phi(t)$. (b) Signal-to-nuisance ratio $\Phi(t)/\Psi(t)$. The dashed ray and horizontal line mark the numerical reference ratio 2.79. The final ratios are close in these runs. Theorem D.1 describes the common radial contraction and the distinct relative speed of temperature learning.

E Further Theory and Model Consequences

This appendix develops consequences that complement the main trajectory proofs. We characterize covariance support and output-capacity geometry, separate free score optimization from factorized dynamics, establish fixed-horizon stability and ranking statements, analyze temperature boundaries, and derive the conditional-moment interpretation of the nonlinear indicators.

E.1 Equivalent covariance descriptions and where independence enters

The latent decomposition in Eq. (1) identifies shared and modality-specific directions. Its marginal covariances are

$$\Sigma_x = (1+\sigma^2)A_s A_s^\top + (\sigma^2+\nu^2)A_\perp A_\perp^\top,$$
$$\Sigma_y = (1+\sigma^2)B_s B_s^\top + (\sigma^2+\nu^2)B_\perp B_\perp^\top.$$

Thus $A_\perp \xi_x + \varepsilon_x$ can equivalently be viewed as an anisotropic Gaussian residual, and analogously for Y. We retain the latent decomposition because it records which directions carry the shared latent variable and which variation is restricted to one modality.

Replacing the residual by isotropic noise of variance $\sigma^2+\nu^2$ is a special case rather than an equivalent representation of the full family. The isotropic replacement has shared-direction variance $1+\sigma^2+\nu^2$ and nuisance-direction variance $\sigma^2+\nu^2$, so the shared variance is always larger by one. In the stated model, the corresponding variances are $1+\sigma^2$ and $\sigma^2+\nu^2$, whose difference is $1-\nu^2$; the modality-specific variance may therefore be smaller than, equal to, or larger than the shared-direction variance. The qualitative population score-side conclusion holds in the isotropic special case, but the original family covers a wider range of variance geometries.

Where independence is used. The cross-covariance satisfies

$$\Sigma_{yx} A_\perp = 0, \qquad \Sigma_{yx}^\top B_\perp = 0. \tag{46}$$

These identities put the cross-covariance entirely in the signal–signal block. They remove the direct cross-view source from the nuisance-factor equation.

The marginal covariances satisfy

$$\Sigma_x A_\perp = (\sigma^2+\nu^2)A_\perp, \qquad \Sigma_y B_\perp = (\sigma^2+\nu^2)B_\perp. \tag{47}$$

The marginal covariances therefore preserve the signal and nuisance subspaces, with positive variance along the latter. These identities determine the damping coefficient.

Finally, within-modality independence makes the normalization-gradient term vanish on each nuisance-free set, as shown in Lemma B.13. Correlations can invalidate these properties; §G.7 tests two specified departures from independence.

Alignment versus score denoising. Independence and centering give the exact identity

$$\mathbb{E}\|W_A X - W_B Y\|_2^2 = \|W_A A_s - W_B B_s\|_F^2 + \nu^2\big(\|W_A A_\perp\|_F^2 + \|W_B B_\perp\|_F^2\big) + \sigma^2\big(\|W_A\|_F^2 + \|W_B\|_F^2\big).$$

Thus $\sigma^2 > 0$ precludes exact nonzero representation alignment. When $\sigma^2 = 0$, matching signal-only encoders can align exactly. Off-signal score denoising isolates the other three coordinate blocks, while measurement noise in the shared coordinates contributes through the remaining signal block.

E.2 Embedding dimension and feasible limiting geometry

A full-rank clean score and two nonzero nuisance factors must fit within the same output space. The relevant constraint is their combined dimension. Let

$$U_s = \mathrm{col}(W_A A_s), \quad U_n = \mathrm{col}(W_A A_\perp), \qquad V_s = \mathrm{col}(W_B B_s), \quad V_n = \mathrm{col}(W_B B_\perp).$$

Proof of Proposition 5. Full rank of $\widetilde{M}_{ss}$ gives $\dim U_s = r$ and $U_s \cap V_s^\perp = \{0\}$. The three vanishing off-signal blocks give

$$U_n \subset V_s^\perp, \qquad V_n \perp U_s, \qquad V_n \perp U_n.$$

Thus $U_s \cap U_n = \{0\}$ and $V_n \perp (U_s \oplus U_n)$. Counting dimensions yields

$$d \geq \dim(U_s \oplus U_n) + \dim V_n = r + \dim U_n + \dim V_n.$$

At $d = r$, both nuisance spaces are zero; at $d = r + 1$, at least one is zero. Two nonzero nuisance factors therefore require $d \geq r + 2$.

For attainment, choose a common r-dimensional signal subspace and two mutually orthogonal nuisance directions in its orthogonal complement. Full-rank signal maps and nonzero nuisance maps on these directions give exactly the required blocks. The construction fits every $d \geq r + 2$. □

Proposition E.1 (Optimal scores with nuisance-bearing encoders)**.** *For every $d \geq r + 2$ and prescribed positive encoder norms C_A, C_B, the model admits a global-minimizer equilibrium in $\mathcal{B}_{\mathrm{reg}}$ with both nuisance factors nonzero and*

$$\widehat{M} = \frac{B_s A_s^\top}{(1+\sigma^2)^2}.$$

The dimension threshold is necessary for a full-rank clean score with both nuisance factors nonzero.

Proof. Let $a = 1 + \sigma^2$. Choose orthonormal output vectors $e_1, \ldots, e_r, u, v$ and positive s_A, s_B with $rs_A^2 < C_A^2$, $rs_B^2 < C_B^2$. In the input bases, set the signal columns of W_A, W_B to $s_A[e_1, \ldots, e_r]$ and $s_B[e_1, \ldots, e_r]$. Set their first nuisance columns to $n_A u, n_B v$, where

$$n_A = \sqrt{C_A^2 - rs_A^2} > 0, \qquad n_B = \sqrt{C_B^2 - rs_B^2} > 0,$$

and set all other nuisance columns to zero. The standing assumptions $d_1, d_2 > r$ provide these columns. The encoder norms are C_A, C_B, the signal score block is $s_A s_B I_r$, and the off-signal score blocks vanish. Since $N_A, N_B > 0$, choose

$$\tau = \frac{a^2 s_A s_B}{N_A N_B} > 0.$$

Then $\widehat{M} = B_s A_s^\top / a^2$, the unique minimizer in the completed-square identity of Lemma B.4. At this score, $P = 0$ and $\Delta = 0$, so the encoder and temperature equations all vanish. The resulting constant trajectory has loss $-r/a^2 < 0$ and positive temperature, and hence belongs to $\mathcal{B}_{\mathrm{reg}}$. Both nuisance norms remain positive. Necessity follows from Proposition 5. □

Proposition E.2 (An open basin with persistent encoder nuisance)**.** *Fix positive encoder norms and $d \geq r + 2$. Every global-minimizer equilibrium $\theta_* = (W_{A,*}, W_{B,*}, \tau_*)$ with both nuisance factors nonzero has a relative open neighborhood V in $\mathcal{M} \times \mathbb{R}_{>0}$ such that every initialization in V generates a global finite-length trajectory converging to a global minimizer. Both nuisance norms remain uniformly positive, and temperature stays bounded above and below by positive constants. In particular, all these trajectories belong to $\mathcal{B}_{\mathrm{reg}}$.*

Proof. Write $a = 1 + \sigma^2$, $\mathcal{L}_{\min} = -r/a^2$, and $n_{A,*}, n_{B,*} > 0$ for the nuisance norms at θ_*. Choose a Euclidean ball $B_\rho(\theta_*)$ whose closure lies in $\mathcal{U}$, on which

$$\tau > \tau_*/2, \qquad \|W_{A,\perp}\|_F > n_{A,*}/2, \qquad \|W_{B,\perp}\|_F > n_{B,*}/2, \qquad \mathcal{L} < \mathcal{L}_{\min} + \frac{1}{2a^2}.$$

Let ω_{len} be the length modulus in Eq. (20) for that ball, and define

$$V = \left\{\theta_0 \in \mathcal{M} \times \mathbb{R}_{>0} : \ \|\theta_0 - \theta_*\| + \omega_{\mathrm{len}}\big(\mathcal{L}(\theta_0) - \mathcal{L}_{\min}\big) < \rho/2\right\}.$$

This is a relative open neighborhood of θ_*. Before any possible exit from the ball, loss monotonicity and its global lower bound give

$$\|\theta(t) - \theta_*\| \leq \|\theta_0 - \theta_*\| + \int_0^t \|\dot{\theta}(u)\|\, du \leq \|\theta_0 - \theta_*\| + \omega_{\mathrm{len}}\big(\mathcal{L}(\theta_0) - \mathcal{L}_{\min}\big) < \rho/2.$$

Thus exit is impossible. Compact continuation gives global existence, and the finite length estimate gives a stationary limit in the ball. Every stationary state is a constant trajectory in $\mathcal{B}_{\mathrm{bnd}}$, so Proposition B.21 makes its loss $-k/a^2$ for an integer $0 \leq k \leq r$. The ball's loss bound forces $k = r$. The limit is a global minimizer, and the ball's bounds give the temperature and nuisance conclusions throughout the trajectory. □

Proposition E.1 supplies such equilibria at the tight threshold $d = r + 2$. Their basins contain nonstationary initial states: changing only temperature gives $\mathcal{L}(\tau) - \mathcal{L}_{\min} = ra^{-2}(\tau_*/\tau - 1)^2 > 0$ when $\tau \neq \tau_*$. The neighborhood is in the joint encoder–temperature space; the proposition does not specify its size or its intersection with a prescribed initial-temperature slice.

Proposition E.3 (Narrow-output trajectory consequence)**.** *Suppose a trajectory belongs to $\mathcal{B}_{\mathrm{bnd}}$ and $\liminf_{t\to\infty} \sigma_r(\widetilde{M}_{ss}(t)) > 0$, where σ_r is the smallest singular value of the $r \times r$ signal block. If $d = r$, both $\|W_{A,\perp}(t)\|_F$ and $\|W_{B,\perp}(t)\|_F$ tend to zero. If $d = r + 1$, then*

$$\min\{\|W_{A,\perp}(t)\|_F, \|W_{B,\perp}(t)\|_F\} \longrightarrow 0.$$

At least one limiting nuisance factor is zero at $d = r + 1$. On $\mathcal{B}_{\mathrm{reg}}$, the singular-value condition holds whenever $-\mathcal{L}^* > (r-1)/(1+\sigma^2)^2$.

Proof. Theorem B.9 gives a joint stationary limit. Its off-signal score blocks vanish by Theorem 1, and its signal block has rank r by the singular-value hypothesis. Proposition 5 therefore forces both limiting nuisance factors to vanish at $d = r$, and at least one to vanish at $d = r + 1$. Continuity gives the claimed norm limits.

On $\mathcal{B}_{\mathrm{reg}}$, the stated loss threshold and Proposition B.21 give $\widehat{M}(t) \to a^{-2}B_sA_s^\top$, where $a = 1 + \sigma^2$. Since the joint limit has $\bar{\alpha} > 0$, the unscaled signal block converges to $(\bar{\alpha}a^2)^{-1}I_r$, which gives the singular-value lower bound. □

Uniform full rank is an additional condition: a full-rank signal block at each finite time can still lose rank in the limit. Point convergence selects at least one denoised modality at $d = r + 1$.

At a regular rank-k limit, Corollary E.7 gives the exact limiting pairwise success probability from this projector structure.

E.3 Score-matrix optimization and factorized dynamics

The closed-form population loss depends on the rescaled product $Z = \widehat{M} = \alpha M \in \mathbb{R}^{d_2 \times d_1}$ through

$$\ell(Z) = \|\Sigma_y^{1/2} Z \Sigma_x^{1/2}\|_F^2 - 2\,\mathrm{tr}(Z^\top \Sigma_{yx}).$$

Because $\Sigma_x, \Sigma_y \succ 0$, this is strictly convex in Z.

Proposition E.4 (Independently optimized score matrix)**.** *If Z were optimized as a free matrix, the unique minimizer would be*

$$Z_\star = \Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}.$$

Its three nuisance blocks vanish.

Proof. The completed-square identity in Lemma B.4 is valid for every free score matrix Z. Positive-definite covariances make its squared term vanish only at $Z_\star = \Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1} = B_sA_s^\top/(1+\sigma^2)^2$, which has no nuisance blocks. □

The learning trajectory evolves W_A, W_B, τ jointly through both the score product and its expected-norm normalizers. Multiplying the encoder equation by $A_\perp$ gives

$$\dot{W}_{A,\perp} = -\frac{2(\sigma^2+\nu^2)}{(\tau N_A N_B)^2}\big(W_B\Sigma_y W_B^\top\big)W_{A,\perp} + \frac{\Delta}{N_A}R_A A_\perp. \tag{48}$$

The identity $\Sigma_{yx}A_\perp = 0$ removes the direct cross-view source. The remaining dynamics depend on $K_B(t) = W_B\Sigma_y W_B^\top \in \mathbb{R}^{d\times d}$, the normalization gradient, and temperature. The matrix K_B can be singular and changes during training; nuisance therefore affects both the quadratic score loss and its normalizers.

On $\mathcal{B}_{\mathrm{bnd}}$, point convergence reduces this coupled equation to its stationary form, which forces all three off-signal score blocks to vanish (Appendix B.7). Corollary 3 gives finite-time encoder persistence, and Proposition E.2 establishes an open basin with positive limiting nuisance in both encoders.

E.4 FINITE-TIME ATTENUATION AND FIXED-HORIZON FINITE-SAMPLE APPROXIMATION

The finite-time encoder conclusion is Corollary 3(ii); it is distinct from asymptotic score-block annihilation. Noise variances enter the nuisance-factor equation through

$$c(t) = \frac{2(\sigma^2 + \nu^2)}{(\tau(t) N_A(t) N_B(t))^2}.$$

The coefficient changes with the noise variances, normalizers, and temperature. Its matrix multiplier and the normalization-gradient contribution evolve as well; the variance sweep therefore measures their combined finite-time effect.

Full-batch empirical objective. For paired observations $(X_i, Y_i)_{i=1}^n$, one natural full-batch version of Eq. (4) is

$$\begin{aligned}\mathcal{L}_n(\theta) = &-\frac{2}{n}\sum_i S_n(X_i, Y_i) + \frac{1}{n}\sum_i \log\left(\frac{1}{n}\sum_j e^{S_n(X_i, Y_j)}\right) \\ &+ \frac{1}{n}\sum_j \log\left(\frac{1}{n}\sum_i e^{S_n(X_i, Y_j)}\right), \\ S_n(x, y) = \frac{y^\top W_B^\top W_A x}{\tau N_{A,n} N_{B,n}}, &\quad N_{A,n} = \frac{1}{n}\sum_i \|W_A X_i\|_2, \quad N_{B,n} = \frac{1}{n}\sum_i \|W_B Y_i\|_2.\end{aligned}$$

This definition uses all n candidates in each logarithm and includes the paired index. Other negative-sampling conventions, stochastic minibatches, detached normalizers, log-temperature parameterization, and Adam require separate statements. In particular, the real-data label-masked objective is not $\mathcal{L}_n$.

All random quantities in the probability statements below are assumed measurable.

Proposition E.5 (Conditional fixed-horizon stability)**.** *Let $\dot\theta = \mathsf{F}(\theta) = -\nabla\mathcal{L}(\theta)$, fix $T < \infty$, and put $K = \theta([0, T])$. Assume:*

(i) K has a compact neighborhood $K' \subset \mathcal{U}$, with positive distance from K to the boundary of K'.

(ii) For each empirical field F_n, solutions from the same initial state can be continued until they exit K'.

(iii) The empirical fields converge uniformly in probability on K':

$$\varepsilon_n = \sup_{\vartheta \in K'} \|\mathsf{F}_n(\vartheta) - \mathsf{F}(\vartheta)\| \xrightarrow{p} 0. \tag{49}$$

Here probability refers to the sampled dataset. Then the empirical trajectories exist on $[0, T]$ with probability tending to one, and

$$\sup_{0 \le t \le T} \|\theta_n(t) - \theta(t)\| \xrightarrow{p} 0.$$

The conclusion also holds pathwise whenever $\varepsilon_n \to 0$.

Proof.

Control before exit. Appendix A.3 gives an Osgood modulus for the population field on K':

$$\|\mathsf{F}(x) - \mathsf{F}(y)\| \le C\omega(\|x - y\|), \qquad \omega(r) = r(1 + |\log r|) \quad \text{near zero.}$$

Set $\omega(0) = 0$, and extend it continuously as an increasing function with at most linear growth away from zero. Up to the first exit of θ_n from K', the error $e_n(t) = \|\theta_n(t) - \theta(t)\|$ satisfies

$$D^+ e_n(t) \le \varepsilon_n + C\omega(e_n(t)), \qquad e_n(0) = 0,$$

where $D^+e_n(t) = \limsup_{h\downarrow 0}(e_n(t+h) - e_n(t))/h$. The first term bounds the difference between the two vector fields; the second controls the population field at two nearby states.

A scalar comparison. For $\varepsilon > 0$, define y_ε by

$$\int_0^{y_\varepsilon(t)} \frac{dr}{\varepsilon + C\omega(r)} = t,$$

and set $y_0(t) = 0$. Scalar comparison gives $e_n(t) \le y_{\varepsilon_n}(t)$, including the case $\varepsilon_n = 0$ by Osgood uniqueness. For any sufficiently small $\delta > 0$, monotone convergence and the Osgood condition imply

$$\int_0^{\delta} \frac{dr}{\varepsilon + C\omega(r)} \longrightarrow \infty \qquad (\varepsilon \downarrow 0).$$

Thus $y_\varepsilon(T) < \delta$ for sufficiently small ε.

Excluding exit. Choose δ smaller than the distance from K to the boundary of K'. At a first exit before T, the empirical trajectory would be at least this boundary distance from the population trajectory, contradicting the comparison bound. The continuation assumption therefore gives existence through T, and $\sup_{t\le T} e_n(t) \to 0$ follows when $\varepsilon_n \to 0$. Finally, apply this deterministic conclusion on the events $\{\varepsilon_n \le \varepsilon\}$. Their probabilities tend to one for every fixed $\varepsilon > 0$, which proves convergence in probability. □

Statistical and regularity hypotheses. Equation (49) and empirical continuation are explicit hypotheses of Proposition E.5. For i.i.d. paired observations, ordinary sample averages satisfy a uniform law of large numbers on a compact parameter set under almost-sure parameter continuity and an integrable dominating envelope (Newey & McFadden, 1994, Lemma 2.4 and footnote 17), with the measurability convention above. For the displayed full-batch objective, gradients also involve ratios of exponential sample averages, so uniform denominator and integrability control are required.

The empirical norm is nondifferentiable when $W_A X_i = 0$ or $W_B Y_i = 0$. For $d < d_1$ and i.i.d. nondegenerate Gaussian inputs, any fixed neighborhood of a nonzero W_A contains such matrices with probability tending to one as $n \to \infty$: the perturbation $-(W_A X_i)X_i^\top / \|X_i\|^2$ annihilates X_i and has norm $\|W_A X_i\|/\|X_i\|$, which can be arbitrarily small near a kernel direction. The same argument applies to W_B when $d < d_2$. Pointwise almost-sure differentiability therefore does not establish a classical empirical gradient field on the fixed neighborhood K'. The proposition applies to empirical fields defined throughout K' that satisfy the stated continuation and uniform-convergence hypotheses. Its conclusion holds on each fixed horizon $[0, T]$; growing horizons or discrete optimizers require additional quantitative control.

E.5 A PAIRWISE-RANKING CERTIFICATE

Let

$$D = S(X_i, Y_i) - S(X_i, Y_j),$$

where (X_i, Y_i) is a matched pair and Y_j is an independent distractor.

Proposition E.6 (Pairwise-ranking certificate)**.** *If* $\operatorname{tr}(\widetilde{M}_{ss}) > 0$*, then*

$$\mathbb{E}[D] = \alpha \operatorname{tr}(\widetilde{M}_{ss}), \qquad \operatorname{Var}(D) = \alpha^2(V_{\mathrm{sig}} + V_{\mathrm{nui}}),$$

where

$$V_{\mathrm{sig}} = 2(1+\sigma^2)^2 \|\widetilde{M}_{ss}\|_F^2 + \operatorname{tr}(\widetilde{M}_{ss}^2),$$

and

$$V_{\mathrm{nui}} = 2(1+\sigma^2)(\sigma^2+\nu^2)\left(\|\widetilde{M}_{sn}\|_F^2 + \|\widetilde{M}_{ns}\|_F^2\right) + 2(\sigma^2+\nu^2)^2 \|\widetilde{M}_{nn}\|_F^2.$$

Consequently,

$$\Pr(D > 0) \ge \frac{\operatorname{tr}(\widetilde{M}_{ss})^2}{\operatorname{tr}(\widetilde{M}_{ss})^2 + V_{\mathrm{sig}} + V_{\mathrm{nui}}}.$$

For fixed $\widetilde{M}_{ss}$, the bound strictly improves when V_{nui} decreases. In particular, decreasing a nonzero nuisance-block norm while holding the other blocks fixed improves the bound; decreasing one block does not suffice if another increases enough to offset it.

Proof. Set $Q = Y_i^\top M X_i$, $Q' = Y_j^\top M X_i$, $a = 1 + \sigma^2$, and $b = \sigma^2 + \nu^2$. Then $D = \alpha(Q - Q')$.

Means and variances. The cross-covariance identity gives $\mathbb{E}Q = \mathrm{tr}(\widetilde{M}_{ss})$, while independence gives $\mathbb{E}Q' = 0$. The centered Gaussian fourth-moment formula gives

$$\mathrm{Var}(Q) = \|\Sigma_y^{1/2} M \Sigma_x^{1/2}\|_F^2 + \mathrm{tr}(\widetilde{M}_{ss}^2),$$
$$\mathrm{Var}(Q') = \|\Sigma_y^{1/2} M \Sigma_x^{1/2}\|_F^2.$$

Although both scores contain X_i, their covariance is zero: $\mathbb{E}[Q' \mid X_i, Y_i] = 0$ implies $\mathrm{Cov}(Q, Q') = 0$. Expanding in signal and nuisance coordinates gives

$$\|\Sigma_y^{1/2} M \Sigma_x^{1/2}\|_F^2 = a^2 \|\widetilde{M}_{ss}\|_F^2 + ab\big(\|\widetilde{M}_{sn}\|_F^2 + \|\widetilde{M}_{ns}\|_F^2\big) + b^2 \|\widetilde{M}_{nn}\|_F^2.$$

Combining these identities yields the stated $\mathbb{E}D$ and $\mathrm{Var}(D)$.

The ranking bound. Write $\mu = \mathbb{E}D > 0$. Cantelli's inequality gives

$$\Pr(D \le 0) \le \frac{\mathrm{Var}(D)}{\mathrm{Var}(D) + \mu^2}.$$

Subtract this bound from one and cancel the common factor α^2. The term $\mathrm{tr}(\widetilde{M}_{ss}^2)$ is the trace of a matrix square, not a squared Frobenius norm. It can be negative; the complete variance is nonnegative. □

For a fixed signal block, nuisance enters the ranking certificate through the variance term V_{nui}. The signal block also evolves during training, so this fixed-state comparison gives neither monotonic improvement along the trajectory nor a top-K retrieval guarantee.

Corollary E.7 (Exact limiting pairwise reliability)**.** *On $\mathcal{B}_{\mathrm{reg}}$, let $k \ge 1$ be the rank of the limiting projector in Proposition B.21. For the expected-norm score difference $D(t) = S_t(X_i, Y_i) - S_t(X_i, Y_j)$, with a matched pair and an independent marginal distractor,*

$$\lim_{t\to\infty} \Pr(D(t) > 0) = F_{t_k}\left(\sqrt{\frac{k}{2(1+\sigma^2)^2 - 1}}\right) > \frac{1}{2}, \tag{50}$$

where F_{t_k} is the Student-t distribution function with k degrees of freedom.

Proof. Set $a = 1 + \sigma^2$. Rotate the shared coordinates into the range of the limiting projector and denote the retained query, matched, and distractor coordinates by $x, y, y' \in \mathbb{R}^k$. Their marginal covariances are aI_k, the matched cross-covariance is I_k, and y' is independent of (x, y). Gaussian conditioning gives

$$x^\top (y - y') \mid x \sim \mathcal{N}\left(\frac{\|x\|^2}{a}, \left(2a - \frac{1}{a}\right)\|x\|^2\right).$$

The limiting score difference is $a^{-2} x^\top (y - y')$. Writing $x = \sqrt{a}\, g$ and taking an independent $Z \sim \mathcal{N}(0,1)$, its probability of being positive is

$$\Pr\left(Z < \frac{\|g\|}{\sqrt{2a^2 - 1}}\right) = F_{t_k}\left(\sqrt{\frac{k}{2a^2 - 1}}\right), \qquad g \sim \mathcal{N}(0, I_k),$$

because $Z/(\|g\|/\sqrt{k})$ has the Student-t_k law. The argument is positive, so the CDF exceeds one half. Convergence of the rescaled score matrix gives $D(t) \to D^*$ almost surely. Since $x \neq 0$ almost surely and the conditional variance above is positive, $\Pr(D^* = 0) = 0$; bounded convergence of the sign indicators proves Eq. (50). □

E.6 TEMPERATURE BOUNDS AND SCORE-PRODUCT COLLAPSE

Theorem C.3 gives global existence and a positive temperature lower bound for every initial state. Negative limiting loss excludes zero-loss collapse and supplies the upper temperature bound. Thus reaching negative loss suffices for all temperature conditions in Corollary 3.

Coercivity also relates temperature to the unscaled score product along a trajectory. The following comparison is obtained directly from the loss, before imposing a temperature lower bound. Write

$$\lambda_0 = \min\{1+\sigma^2, \sigma^2+\nu^2\}, \quad K_\Sigma = \|\Sigma_{yx}\|_F, \quad R_0 = \frac{K_\Sigma + \sqrt{K_\Sigma^2 + \lambda_0^2 \mathcal{L}(0)_+}}{\lambda_0^2}.$$

Here $x_+ = \max\{x, 0\}$, so $\mathcal{L}(0)_+$ is the positive part of the initial loss. Let $c_{\min}, c_{\max} > 0$ be the normalizer-product bounds following Proposition B.2.

Proposition E.8 (Temperature/product reduction)**.** *On the maximal interval, $\|\widehat{M}(t)\|_F \le R_0$, so*

$$\tau(t) \ge \frac{\|M(t)\|_F}{c_{\max} R_0}.$$

On any global trajectory with $\mathcal{L}^* < 0$, there are $t_0 < \infty$ and $\rho_* > 0$ such that, for $t \ge t_0$,

$$c_{\min}\rho_* \tau(t) \le \|M(t)\|_F \le c_{\max} R_0 \tau(t).$$

Consequently $\liminf_t \tau(t) > 0$ if and only if $\liminf_t \|M(t)\|_F > 0$. Along every sequence $t_k \to \infty$, $\tau(t_k) \to 0$ if and only if $\|M(t_k)\|_F \to 0$.

Proof.

Upper bound on the rescaled score. The covariance lower bound and loss monotonicity give

$$\lambda_0^2 \|\widehat{M}(t)\|_F^2 - 2K_\Sigma \|\widehat{M}(t)\|_F \le \mathcal{L}(t) \le \mathcal{L}(0).$$

Solving this quadratic inequality, with $\mathcal{L}(0)_+$ in place of $\mathcal{L}(0)$ for a uniform expression, gives $\|\widehat{M}(t)\|_F \le R_0$. Since $M = \tau N_A N_B \widehat{M}$, it follows that $\|M\|_F \le c_{\max} R_0 \tau$.

Lower bound under negative limiting loss. On a global trajectory with $\mathcal{L}^* < 0$, choose t_0 such that $\mathcal{L}(t) \le \mathcal{L}^*/2$ for $t \ge t_0$. Using $\mathcal{L} = F(\widehat{M}) - 2G(\widehat{M})$ and $F \ge 0$, we get

$$G(\widehat{M}(t)) \ge -\mathcal{L}^*/4, \qquad \|\widehat{M}(t)\|_F \ge \rho_* = \frac{-\mathcal{L}^*}{4K_\Sigma} > 0.$$

Here the second inequality is Cauchy–Schwarz. Multiplying by $\tau N_A N_B$ yields

$$c_{\min}\rho_* \tau(t) \le \|M(t)\|_F \le c_{\max} R_0 \tau(t) \qquad (t \ge t_0).$$

The two-sided comparison proves the equivalence of positive lower limits and the equivalence of vanishing along a subsequence. No temperature lower bound was assumed in obtaining it. □

The objective value alone does not give a uniform lower temperature bound, as the following family of equilibria shows.

Proposition E.9 (Global minimizers can approach the temperature boundary)**.** *Let*

$$Z_\star = \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1}.$$

Every state with $\widehat{M} = Z_\star$ is a global minimizer of $\mathcal{L}$ and an equilibrium of the gradient flow. For any prescribed positive encoder norms C_A, C_B, such equilibria exist with arbitrarily small $\tau > 0$ whenever $d \ge r + 1$.

Proof. Proposition E.4 gives the unique optimal score $Z_\star = a^{-2}\Sigma_{yx}$, where $a = 1 + \sigma^2$. At this score, $P = 0$ and $\Delta = 0$, so Lemma A.1 makes every factorization an equilibrium.

Choose orthonormal output vectors $e_1, \dots, e_r, u$. In the input bases, set the signal columns of W_A, W_B to $s_A[e_1, \dots, e_r]$ and $s_B[e_1, \dots, e_r]$, where

$$0 < s_A < C_A/\sqrt{r}, \qquad s_B = C_B/\sqrt{r}.$$

Set the first nuisance column of W_A to $\sqrt{C_A^2 - rs_A^2}\,u$, and all remaining nuisance columns of both encoders to zero. The standing assumption $d_1 > r$ supplies this column. Both encoder norms have their prescribed values, the signal score block is $s_A s_B I_r$, and all off-signal score blocks vanish. Choosing

$$\tau = \frac{a^2 s_A s_B}{N_A N_B}$$

gives $\widehat{M} = Z_\star$. Proposition B.2 bounds $N_A N_B$ away from zero uniformly in this family. With s_B fixed, letting $s_A \downarrow 0$ therefore gives optimal equilibria with $\tau \downarrow 0$. □

All states in this family have the same minimal objective value, while their temperatures approach zero across the family. Each individual equilibrium has a positive constant temperature. Thus the objective value alone provides no positive lower bound uniform over global minimizers. Theorem C.3 gives a positive bound for each individual trajectory.

Practical temperature constraints can also be imposed algorithmically. The CLIP paper reports capping the inverse-temperature logit scale at 100, corresponding to $\tau \geq 0.01$ (Radford et al., 2021). Clipping and log-temperature optimization define a different algorithm from the unconstrained Euclidean flow analyzed here. Appendices G.4, G.3, and G.7 report the corresponding finite-horizon temperature diagnostics.

E.7 Nonlinear indicators: interpretation and conditional moments

For encoders f, g with outputs in $\mathbb{R}^d$, set

$$u = f(X), \quad v = g(Y), \qquad \widehat{u} = \frac{u}{\|u\|_2}, \quad \widehat{v} = \frac{v}{\|v\|_2}, \quad s = \widehat{u}^\top \widehat{v}.$$

For nonzero embeddings, s is the cosine score used for retrieval; the numerical norm floor is specified below. At fixed shared signal z, define

$$\mu_{u,z} = \mathbb{E}[\widehat{u} \mid z] \in \mathbb{R}^d, \qquad \mu_{v,z} = \mathbb{E}[\widehat{v} \mid z] \in \mathbb{R}^d, \qquad \mu_{s,z} = \mathbb{E}[s \mid z] \in \mathbb{R}.$$

The score- and encoder-level indicators are

$$\begin{aligned} C_{\mathrm{nl}} &= \frac{\mathbb{E}_z \mathbb{E}[(s - \mu_{s,z})^2 \mid z]}{\mathbb{E}_z \mu_{s,z}^2 + \epsilon}, \\ I_{\mathrm{nl}} &= \frac{1}{2} \sum_{\phi \in \{u,v\}} \frac{\mathbb{E}_z \mathbb{E}[\|\widehat{\phi} - \mu_{\phi,z}\|_2^2 \mid z]}{\mathbb{E}_z \|\mu_{\phi,z}\|_2^2 + \epsilon}, \qquad \epsilon = 10^{-30}. \end{aligned} \tag{51}$$

Each indicator measures conditional variance relative to squared conditional-mean magnitude. Discrimination between candidates is evaluated separately by retrieval; a small ratio alone does not ensure an informative score. The encoder indicator captures variation around a vector mean; the score indicator captures variation in the scalar comparison. Samples are normalized before computing means, the two encoder ratios are averaged separately, and the paired-score mean is measured directly. On real data, z is a class label and eligible classes receive equal weight, so within-class variation includes fine-grained biology or semantics.

Equation (51) compares the conditional variability of a cosine score with that of its two normalized representations. To make the roles of the numerators and denominators explicit, write

$$\begin{aligned} N_s &= \mathbb{E}_z \operatorname{Var}(s \mid z), & D_s &= \mathbb{E}_z \mu_{s,z}^2, \\ N_\phi &= \mathbb{E}_z \mathbb{E}[\|\widehat{\phi} - \mu_{\phi,z}\|_2^2 \mid z], & D_\phi &= \mathbb{E}_z \|\mu_{\phi,z}\|_2^2, \quad \phi \in \{u, v\}. \end{aligned}$$

Then $C_{\mathrm{nl}} = N_s/(D_s + \epsilon)$ and

$$I_{\mathrm{nl}} = \frac{1}{2}\left(\frac{N_u}{D_u + \epsilon} + \frac{N_v}{D_v + \epsilon}\right).$$

The six components separate changes in conditional variability from changes in conditional-mean magnitude. The definition averages the two encoder-specific ratios; Tables 6 and 7 also show the score and encoder indicators individually.

Proposition E.10 (Conditional-moment interpretation)**.** *Suppose the embeddings are nonzero almost surely and use exact unit normalization. Define*

$$T_{u,z} = \mathbb{E}[\widehat{u}\widehat{u}^\top \mid z] \in \mathbb{R}^{d\times d}, \qquad T_{v,z} = \mathbb{E}[\widehat{v}\widehat{v}^\top \mid z] \in \mathbb{R}^{d\times d}.$$

Then $N_u = 1 - D_u$ and $N_v = 1 - D_v$. If the two embeddings are conditionally independent given z, then

$$\mu_{s,z} = \mu_{u,z}^\top \mu_{v,z},$$
$$\mathrm{Var}(s \mid z) = \mathrm{tr}(T_{u,z}T_{v,z}) - (\mu_{u,z}^\top \mu_{v,z})^2.$$

Neither conditional independence nor equality of the two expressions for the conditional score mean is required by Eq. (51) *itself.*

Proof. For $\phi \in \{u, v\}$, expanding the square and using the definition of the conditional mean gives

$$\mathbb{E}[\|\widehat{\phi} - \mu_{\phi,z}\|_2^2 \mid z] = \mathbb{E}[\|\widehat{\phi}\|_2^2 \mid z] - \|\mu_{\phi,z}\|_2^2 = 1 - \|\mu_{\phi,z}\|_2^2.$$

Average over z. Under conditional independence, expectation of the scalar product factors, giving the stated mean. Also

$$s^2 = \mathrm{tr}(\widehat{u}\widehat{u}^\top \widehat{v}\widehat{v}^\top).$$

Taking the conditional expectation and factoring the two second moments gives $\mathbb{E}[s^2 \mid z] = \mathrm{tr}(T_{u,z}T_{v,z})$. Subtract the square of the conditional mean. □

The distinction can be expressed directly in terms of conditional covariances. Write

$$\Gamma_{\phi,z} = T_{\phi,z} - \mu_{\phi,z}\mu_{\phi,z}^\top, \qquad \phi \in \{u, v\}.$$

Under conditional independence, expansion of the second-moment identity gives

$$\mathrm{Var}(s \mid z) = \mu_{v,z}^\top \Gamma_{u,z}\mu_{v,z} + \mu_{u,z}^\top \Gamma_{v,z}\mu_{u,z} + \mathrm{tr}(\Gamma_{u,z}\Gamma_{v,z}). \tag{52}$$

In contrast, the encoder variances are $\mathrm{tr}\,\Gamma_{u,z}$ and $\mathrm{tr}\,\Gamma_{v,z}$. The score therefore weights each representation's variation through the other representation's mean and covariance. Large encoder variation can have little effect on comparison when these directions are orthogonal. This is the conditional-moment counterpart of the interaction geometry in the linear score, with sample normalization included in the random representations. For dependent pairs the directly measured score moments remain the definitions used in Eq. (51).

The score variance can be zero even when both representations vary. For example, let z, ξ_x, ξ_y be independent uniform random signs in $\{-1, 1\}$, let $0 < a < 1$, and put $b = \sqrt{1 - a^2}$. The already normalized embeddings

$$\widehat{u} = (az, b\xi_x, 0)^\top, \qquad \widehat{v} = (az, 0, b\xi_y)^\top$$

have $s = a^2$, $D_s = a^4$, $N_s = 0$, and $D_u = D_v = a^2$, $N_u = N_v = b^2$. Hence $C_{\mathrm{nl}} = 0$ while $I_{\mathrm{nl}} = b^2/(a^2 + \epsilon) > 0$. Replacing az by the constant a leaves these indicators unchanged but makes every matched and unmatched score equal. Thus the same indicator values are compatible with either signal-dependent or uninformative comparisons.

More generally, adding a common coordinate gives unit embeddings

$$\widehat{\phi}_\gamma = (\sqrt{1-\gamma}, \sqrt{\gamma}\,\widehat{\phi})^\top, \qquad \phi \in \{u, v\}, \quad 0 < \gamma < 1.$$

Their score is $s_\gamma = 1 - \gamma + \gamma s$, so all candidate rankings are preserved and the original embeddings can be recovered. Yet

$$N_{s,\gamma} = \gamma^2 N_s, \quad D_{s,\gamma} = \mathbb{E}_z(1 - \gamma + \gamma\mu_{s,z})^2, \quad N_{\phi,\gamma} = \gamma N_\phi, \quad D_{\phi,\gamma} = 1 - \gamma + \gamma D_\phi.$$

If $N_u + N_v > 0$, these identities give $C_{\mathrm{nl},\gamma}/I_{\mathrm{nl},\gamma} = \gamma(2N_s/(N_u + N_v) + o(1)) \to 0$. The ratio describes normalized comparison geometry; candidate discrimination and within-class pair retrieval are measured separately in Appendix G.9.

PAIRED OBSERVATIONS AND CONDITIONING

For general paired observations define

$$K_{uv,z} = \mathbb{E}[(\widehat{u} - \mu_{u,z})(\widehat{v} - \mu_{v,z})^\top \mid z] \in \mathbb{R}^{d\times d}.$$

Then $\mu_{s,z} = \mu_{u,z}^\top \mu_{v,z} + \operatorname{tr} K_{uv,z}$, so the paired-score mean includes the within-group cross-covariance. The estimator computes this mean directly from paired scores and computes representation means after sample normalization. A coarse label can leave residual dependence and meaningful within-class variability.

WHY COSINE VARIATION DIFFERS FROM BLOCK NUISANCE

The block indicators in Eq. (13) measure the matrices in the population theorem. The nonlinear indicators in Eq. (51) measure sample-normalized embeddings, whose denominators can vary with nuisance even when the bilinear product is clean. For example, at fixed $z \neq 0$, let

$$u = (z, \xi_x, 0)^\top, \qquad v = (z, 0, \xi_y)^\top,$$

with independent nondegenerate Gaussian ξ_x, ξ_y. The bilinear product is $u^\top v = z^2$, whereas

$$s = \frac{z^2}{\sqrt{z^2 + \xi_x^2}\sqrt{z^2 + \xi_y^2}}$$

varies under nuisance resampling. This variability can reverse a ranking: with query $(1, 0, 0)^\top$, matched output $(1, 0, 2)^\top$, and distractor $(1/2, 0, 1/2)^\top$, the bilinear scores are $1 > 1/2$, while the cosine scores are $1/\sqrt{5} < 1/\sqrt{2}$. Changing only the matched nuisance coordinate from 2 to 0 restores the cosine ranking. Sample normalization therefore introduces a source of conditional score variation beyond the bilinear blocks. The nonlinear experiments measure the relative score–encoder variability of this cosine comparison; exact block annihilation is tested separately in the linear experiments.

F Compression Consequences of Score–Encoder Separation

F.1 Exact score preservation and encoder reconstruction

Proof of Proposition 6. Write $U_s = \mathrm{Im}(W_{A,s})$, $U_n = \mathrm{Im}(W_{A,\perp})$, and define V_s, V_n analogously. Full rank of $W_{B,s}^\top W_{A,s}$ gives $\dim U_s = \dim V_s = r$ and $U_s \cap V_s^\perp = \{0\}$. The clean-score conditions imply $U_n \subset V_s^\perp$, hence $U_s \cap U_n = \{0\}$. The symmetric argument gives $V_s \cap V_n = \{0\}$. Therefore

$$\mathrm{rank}(W_A) = r + q_A, \qquad \mathrm{rank}(W_B) = r + q_B.$$

In the input bases, $M = W_B^\top W_A$ has only its full-rank r-by-r signal block, so $\mathrm{rank}(M) = r$.

Let $M = L\Lambda R^\top$ be a compact SVD. Its right and left singular spaces lie in the row spaces of W_A and W_B, respectively. With the Moore–Penrose inverse denoted by $\dagger$, define

$$T_A = \Lambda^{1/2} R^\top W_A^\dagger, \qquad T_B = \Lambda^{1/2} L^\top W_B^\dagger.$$

Then $T_A W_A = \Lambda^{1/2} R^\top$ and $T_B W_B = \Lambda^{1/2} L^\top$, giving

$$(T_B W_B y)^\top (T_A W_A x) = y^\top M x \quad \text{for all } x, y.$$

The fixed positive score multiplier is retained separately or absorbed into the two maps. Conversely, any exact bilinear representation through k coordinates has score rank at most k, so $k \geq r$.

Exact linear reconstruction of $W_A x$ through k_A coordinates requires $\mathrm{rank}(W_A) \leq k_A$. Equality is attained by recording coordinates in a basis of $\mathrm{Im}(W_A)$. The same argument applies to W_B. Since both input covariance matrices are positive definite, almost-sure exact linear reconstruction also implies the corresponding matrix identity. Thus the same lower bounds hold for reconstruction on the data distribution. □

This is a statement about postprocessing a fixed full-rank clean score. It does not require the latent rank as an input to a practical compression algorithm. Propositions E.1 and E.2 establish global optima and open basins with strictly positive encoder nuisance. Generic full-rank convergence is not an assumption of this compression statement.

F.2 Feature-variance retention does not certify score preservation

Proposition F.1 (Arbitrarily high variance retention with zero compressed score)**.** *For every $\varepsilon \in (0, 1)$ and equal prescribed positive encoder norms, the paper's model admits a global-minimizer equilibrium with $r = 1$, $d = 3$ and nuisance in both encoders such that pooled PCA retains at least $1 - \varepsilon$ of the normalized feature variance in two coordinates but makes every cross-modal score zero. A one-coordinate linear postprocessing of each encoder preserves the original score exactly.*

Proof. Put $a = 1 + \sigma^2$, $b = \nu^2 + \sigma^2 > 0$, and take input dimensions $d_1 = d_2 = 3$. In the input bases let

$$W_A = \begin{pmatrix} s & 0 & 0 \\ 0 & n & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad W_B = \begin{pmatrix} s & 0 & 0 \\ 0 & 0 & 0 \\ 0 & n & 0 \end{pmatrix}, \qquad s, n > 0, \quad s^2 + n^2 = C^2.$$

The expected norms agree; denote them by $N > 0$. Choosing $\tau = a^2 s^2 / N^2$ gives $\widehat{M} = \mathrm{diag}(1, 0, 0)/a^2$, the full-signal global minimum. The population gradient is zero there, as in Proposition E.1. Both nuisance factors are nonzero.

The outputs have zero mean, and their equally weighted normalized covariance is

$$\frac{1}{N^2} \mathrm{diag}\left(as^2, \frac{bn^2}{2}, \frac{bn^2}{2} \right).$$

Choose s sufficiently small that

$$\frac{bn^2}{as^2} > \max\left\{ 2, \frac{1 - \varepsilon}{\varepsilon} \right\}.$$

The leading two-dimensional PCA space is then $\mathrm{span}(e_2, e_3)$. Its retained-variance fraction is $bn^2/(as^2 + bn^2) > 1 - \varepsilon$. The projected outputs lie on the orthogonal axes e_2 and e_3, so every compressed cross-modal score is zero. Keeping the e_1 coordinate instead preserves the original inner product and therefore the score with its fixed multiplier. □

The proposition concerns an exact configuration in the population model. It is not a claim that every pretrained encoder has this geometry. The real-data experiments compare the two compression objectives on unmodified features.

F.3 A DIRECT CERTIFICATE FOR PRESERVING RETRIEVAL COMPARISONS

For one finite-gallery query, let p be the teacher softmax distribution and q the compressed model's softmax distribution. All their coordinates are positive. The following elementary certificate makes the connection between distribution matching and retrieval explicit, without assumptions on the encoder.

Proposition F.2 (KL threshold for preserving a top-ranked candidate)**.** *Suppose p has a unique maximum p_a, and let $p_b = \max_{j \neq a} p_j$. Define*

$$c(p) = p_a \log \frac{2p_a}{p_a + p_b} + p_b \log \frac{2p_b}{p_a + p_b}.$$

If $\mathrm{KL}(p\|q) < c(p)$, then $q_a > q_j$ for every $j \neq a$. The threshold is the infimum KL needed to change the top-ranked candidate. Moreover, on a full rectangular score matrix, zero KL in both retrieval directions is equivalent to preserving all logits up to one common additive constant.

Proof. Fix a competitor $j \neq a$. Minimize $\mathrm{KL}(p\|q)$ over the simplex subject to $q_j \geq q_a$. The unconstrained minimum $q = p$ violates this condition. The constrained minimum has

$$q_a = q_j = \frac{p_a + p_j}{2}, \qquad q_\ell = p_\ell \quad (\ell \notin \{a, j\}).$$

These coordinates satisfy the simplex constraint and the first-order optimality conditions of the strictly convex problem. Its value is

$$f(p_a, p_j) = p_a \log \frac{2p_a}{p_a + p_j} + p_j \log \frac{2p_j}{p_a + p_j}.$$

For $0 < p_j < p_a$, its derivative with respect to p_j is $\log(2p_j/(p_a + p_j)) < 0$. The least expensive competitor therefore has probability p_b, and every top-rank change has KL at least $c(p)$. The displayed constrained distribution attains a tie at this value; an arbitrarily small further perturbation reverses the order, proving sharpness of the infimum.

For the last statement, equality of each row's softmax implies $\widehat{S}_{ij} - S_{ij} = u_i$, while equality of each column's softmax implies $\widehat{S}_{ij} - S_{ij} = v_j$. Since all row–column pairs occur, $u_i = v_j$ is a single common constant. The converse follows from softmax invariance to such a shift. □

If the teacher's top-ranked candidate is relevant, every certified query remains correct after compression. The certificate depends on the teacher's observable probabilities, not latent nuisance labels. It is query-specific: a small average KL alone does not certify every query, especially those with a small teacher margin. Exact score preservation from Proposition 6 gives zero KL in both directions. In the retrieval-compression experiment, we minimize KL within a normalized linear-map family and evaluate retrieval directly.

G Empirical Measurements and Robustness Studies

This appendix collects the empirical evidence for score–encoder separation. Controlled linear experiments measure the exact score blocks, and the subsequent studies test correlation robustness, nonlinear conditional variability, nuisance-resampled rankings, and explicit encoder alignment. Each experiment states the comparison property it evaluates; none treats a low variability indicator alone as a retrieval guarantee.

G.1 Synthetic experimental setup for Figure 2

Both panels of Figure 2 use $r = 3$, $d_1 = d_2 = 20$, $d = 25$, $\sigma^2 = 0.5$, and $\nu^2 = 0.3$. Panel (a) integrates the expected-norm population flow with learned temperature. Panel (b) measures linear encoder blocks after stochastic cosine training, using the following protocol.

Finite-data setting	Value
Objective; normalization	One-way paired InfoNCE; sample-wise cosine
Optimizer; learning rate	Adam; 10^{-3}
Batch size; displayed update budgets	256; 10,000 and 50,000
Temperature	Fixed 0.5
Principal sample size	20,000
Sample sizes in the inset	1,000, 5,000, 20,000
Runs per setting; summaries	10; medians and $[10\%, 90\%]$ intervals

Both panels evaluate $C_{\mathrm{lin}}, I_{\mathrm{lin}}$ from Eq. (13). The finite-data panel reports separate budget measurements; the stored records do not establish a paired trajectory between the two budgets. Its loss, normalization, optimizer, and temperature rule differ from the population flow, so this panel measures block separation under stochastic training. The inset reports the 50,000-update measurements for $n \geq 1000$. The stored sweep also includes $n = 500$, where two of ten runs have $C_{\mathrm{lin}} \geq I_{\mathrm{lin}}$.

Numerical population integration. The population simulations use exact Gaussian covariance matrices in the loss coefficients and fixed Monte Carlo samples for the expected norms and their gradients. The weak-signal and initial-temperature sweeps use 50,000 samples per modality (Monte Carlo seed 9999), relative/absolute solver tolerances $10^{-8}/10^{-10}$, and 2000 equally spaced recorded times on $[0, T]$, with $T = 50$ and 5000, respectively. The dimension/noise-grid driver uses 5000 samples, tolerances $10^{-6}/10^{-8}$, and 500 recorded times through $T = 1000$. The reported grid extrema refer to these recorded times.

G.2 Growth of signal from near-signal-degenerate initialization

Proposition C.9 identifies descending directions at trivial critical points on the signal-degenerate set $\mathcal{S}$. To examine the growth of weak signal in the full dynamics, we initialize the signal blocks at a small prescribed norm while retaining the nuisance blocks at their usual scale.

Setup. Generate Gaussian encoder matrices normalized to unit Frobenius norm. Replace each signal block by a matrix in the same direction with Frobenius norm $\varepsilon = 10^{-2}$, leaving its nuisance block unchanged; do not renormalize the resulting encoder. Thus $\|W_{A,s}(0)\|_F = \|W_{B,s}(0)\|_F = \varepsilon$, $\Phi(0) = 2\varepsilon^2 = 2 \times 10^{-4}$, and the alignment trace $\mathcal{B}(0) = \mathrm{tr}(M(0)^\top \Sigma_{yx})$ is small. The initialization is close to $\mathcal{S}$; the retained nuisance blocks need not give an exactly critical state. We integrate the full population gradient flow for 20 seeds to $T = 50$.

Plot coordinates. Using the alignment trace $\mathcal{B}(t) = \mathrm{tr}(M(t)^\top \Sigma_{yx})$, define

$$s(t) = \int_0^t \frac{du}{\tau(u)}, \qquad w_+(t) = \frac{\Phi(t)}{4} + \frac{\mathcal{B}(t)}{2}, \qquad w_-(t) = \frac{\Phi(t)}{4} - \frac{\mathcal{B}(t)}{2}.$$

Thus $(\Phi, \mathcal{B}) = w_+(2, 1) + w_-(2, -1)$. These are coordinates of the full coupled trajectory, with $s(t)$ measuring inverse-temperature-weighted time.

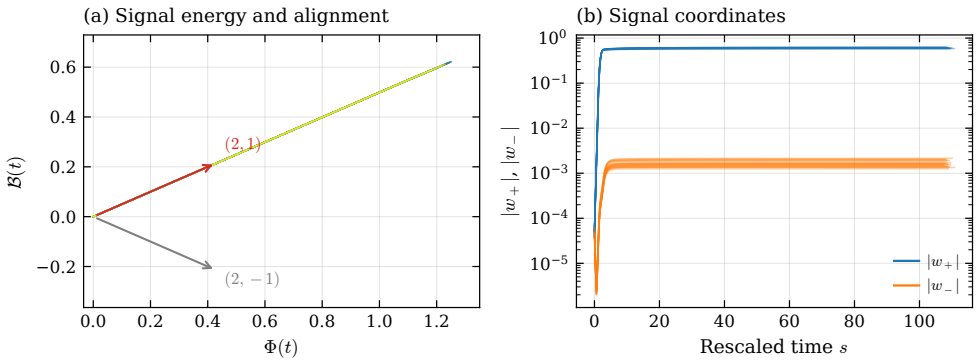

Figure 4: Signal growth from weak shared-signal initialization, over all 20 trajectories. **(a)** Full trajectories in the $(\Phi, \mathcal{B})$-plane, with coordinate directions $(2, 1)$ and $(2, -1)$. **(b)** The coordinates w_+, w_- against $s(t) = \int_0^t du/\tau(u)$. The w_+ coordinate becomes dominant.

Findings. Both coordinates end above their initial values in all 20 runs. At $T = 50$, w_+ is more than two orders of magnitude larger than w_-. The trajectories consequently approach the direction $(2, 1)$ as the signal energy increases. Figure 4 shows growth and relative alignment from a weak shared-signal initialization. Since the initial nuisance blocks need not form a critical state, these trajectories do not estimate the local eigenvalues in Proposition C.9.

G.3 Initial-temperature dependence

Initial temperature can affect the terminal allocation of encoder energy between signal and nuisance. We sweep $\tau(0)$ over five decades under standard random initialization and measure signal energy $\Phi(t)$ and nuisance energy $\Psi(t) = \|W_{A,\perp}(t)\|_F^2 + \|W_{B,\perp}(t)\|_F^2$ at a common integration horizon.

Setup. We sweep $\tau(0) \in \{10^{-4}, 3\times10^{-4}, 10^{-3}, 1.7\times10^{-3}, 3\times10^{-3}, 6\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}, 3\times10^{-1}, 1, 3, 10\}$ — a 13-point grid spanning five decades — with 5 seeds per value and the standard random initialization $\|W_A(0)\|_F = \|W_B(0)\|_F = 1$. For each $(\tau(0), \text{seed})$ pair we run the population gradient flow to $T = 5000$ and record both energies. Norm conservation gives $\Phi(t) + \Psi(t) = 2$ up to numerical error.

Findings. For $\tau(0) \le 0.03$, median terminal signal energy is approximately 0.92. It reaches 0.99 at $\tau(0) = 0.3$ and 1.47 at $\tau(0) = 1$, then approaches 2 at higher initial temperatures. Median terminal nuisance energy is 0.528, 1.27×10^{-3}, and 2.16×10^{-18} at $\tau(0) = 1, 3, 10$, respectively. Figure 5 shows both energies, including the strong nuisance attenuation at high initial temperatures.

Interpretation. Output capacity permits nuisance retention but does not require it for every initialization. The observed allocation of encoder energy depends on the full trajectory and is compatible with both finite-time non-entry and the open basin of nuisance-bearing optima. This finite-horizon sweep concerns random initialization; the equal-amplitude criterion in Eq. (30) concerns perturbations of a zero-loss equilibrium.

G.4 Temperature and non-collapse across dimension and noise

Let T be a run's terminal integration time. Figure 6 sweeps $d \in \{5, 10, 15, 20, 25, 30, 40\}$ and $\sigma^2 \in \{0.01, 0.05, 0.1, 0.3, 0.5, 1, 2\}$, with 20 seeds per cell. All 980 runs have negative terminal loss, a minimum recorded temperature above 10^{-2}, and relative encoder-norm drift below 10^{-6}. Across the grid, the median reduction in $\|\widetilde{M}_{nn}\|_F$ over 20 seeds ranges from 3.1 to 10.5 orders of magnitude. These finite-horizon observations are consistent with the regular regime. Appendix G.3 separately varies initial temperature over five decades, and Appendix E.6 relates temperature nondegeneracy to the unscaled encoder product.

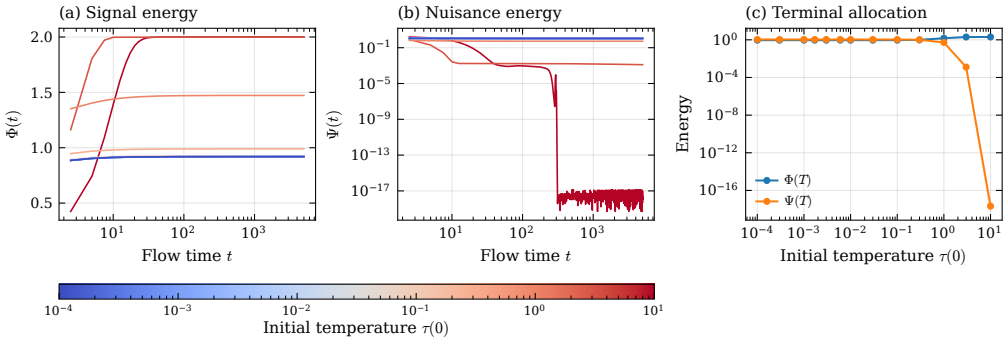

Figure 5: Initial-temperature dependence of signal and nuisance energy. (a,b) Median trajectories over five seeds at each of the 13 initial temperatures. (c) Terminal energies at $T = 5000$, with medians and $[10\%, 90\%]$ envelopes. Nuisance energy is measured directly from the encoder blocks.

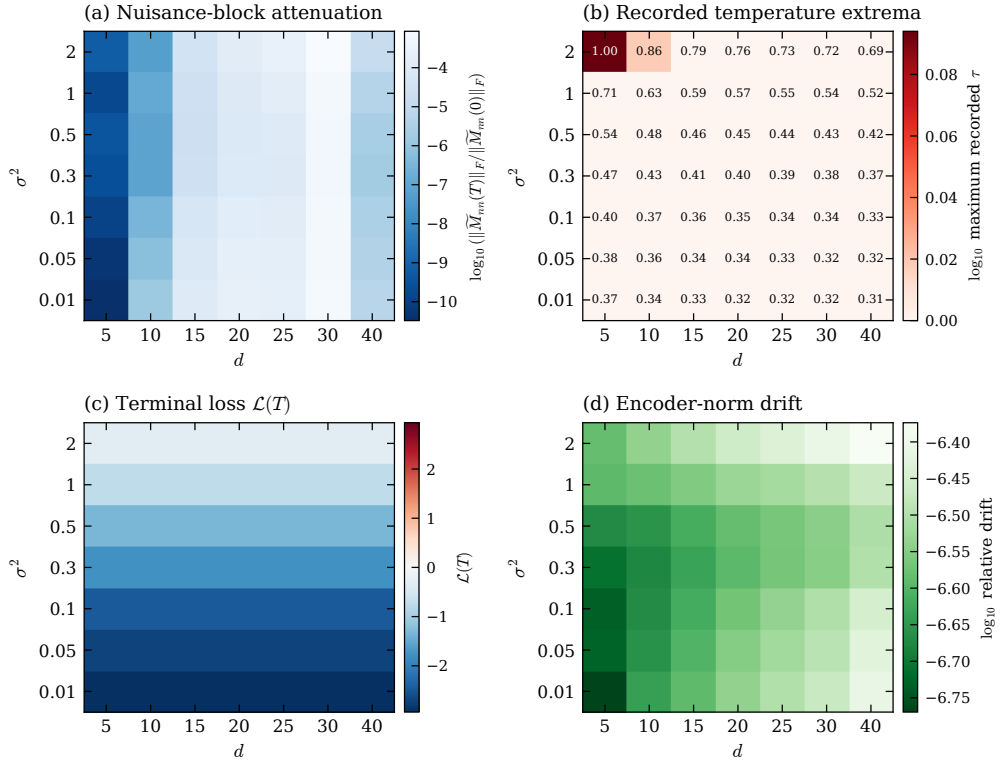

Figure 6: **Finite-horizon diagnostics on the (d, σ^2) grid.** Each cell summarizes 20 seeds. (a) Median $\log_{10}(\|\widetilde{M}_{nn}(T)\|_F/\|\widetilde{M}_{nn}(0)\|_F)$. (b) Maximum recorded temperature, annotated with the minimum recorded temperature; extrema are computed within each trajectory before taking seed medians. (c) Median terminal loss. (d) Median log relative numerical drift in the conserved encoder norms.

G.5 OUTPUT-CAPACITY SWEEP

The population experiment uses three seeds per embedding dimension and extends the dimension sweep down to $d = 1$. Table 3 gives seed medians for $d = 25, 5, 4, 3$, which straddle the threshold. The combined nuisance-score norm is $\left(\|\widetilde{M}_{sn}\|_F^2 + \|\widetilde{M}_{ns}\|_F^2 + \|\widetilde{M}_{nn}\|_F^2\right)^{1/2}$. The $d < r$ runs are outside the full-signal-capacity standing convention but inside the stated any-d extension of the score theorem; no full-rank signal block is asserted there.

Table 3: Bottleneck sweep with $r = 3$. Each entry is a seed median. The combined score norm is the square root of the sum of the three nuisance-block squared Frobenius norms.

d	Combined score norm	$\|W_{A,s}\|_F$	$\|W_{B,s}\|_F$	$\|W_{A,\perp}\|_F$	$\|W_{B,\perp}\|_F$
25	1.8×10^{-6}	.8549	.8627	.52	.51
$5 = r + 2$	1.3×10^{-9}	.9862	.9496	.17	.31
$4 = r + 1$	3.0×10^{-10}	.9819	1.0000	.19	2.1×10^{-10}
$3 = r$	1.2×10^{-10}	1.0000	1.0000	1.5×10^{-10}	1.6×10^{-10}

At the terminal integration time, both nuisance norms are substantial in the displayed $d \geq r + 2$ cells, one is near numerical zero at $d = r + 1$, and both are near numerical zero at $d = r$. The finite-horizon measurements follow the capacity pattern of Proposition 5. Corollary 3(ii) concerns exact non-entry at finite time and permits the near-zero terminal norms observed at the bottleneck.

G.6 FINITE-SAMPLE COMPUTATION OF NONLINEAR INDICATORS

For synthetic diagnostics, hold z fixed and resample both modality-specific variables and measurement noises; average within each such group and then across sampled signals. For real data, write ℓ_i for the class label of evaluation observation i, let $J_c = \{i : \ell_i = c\}$, and put $n_c = |J_c|$. Let $\mathcal{C} = \{c : n_c \geq 2\}$, assumed nonempty.

1. Normalize samples and form paired scores. Normalize each embedding with

$$\widehat{\phi}_i = \frac{\phi_i}{\max\{\|\phi_i\|_2, \delta\}}, \qquad \delta = 10^{-12}, \qquad s_i = \widehat{u}_i^\top \widehat{v}_i.$$

2. Compute group means. Form $\widehat{\mu}_{\phi,c} = n_c^{-1} \sum_{i \in J_c} \widehat{\phi}_i$ and $\widehat{\mu}_{s,c} = n_c^{-1} \sum_{i \in J_c} s_i$.

3. Average within-group residuals and mean magnitudes. The empirical components are

$$\widehat{N}_s = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{n_c} \sum_{i \in J_c} (s_i - \widehat{\mu}_{s,c})^2, \qquad \widehat{D}_s = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \widehat{\mu}_{s,c}^2,$$

$$\widehat{N}_\phi = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{n_c} \sum_{i \in J_c} \|\widehat{\phi}_i - \widehat{\mu}_{\phi,c}\|_2^2, \qquad \widehat{D}_\phi = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \|\widehat{\mu}_{\phi,c}\|_2^2.$$

4. Form the ratios. Substitute these six components into the ratios above, using $\epsilon = 10^{-30}$. The plug-in within-class variance uses divisor n_c. Each eligible class receives weight $1/|\mathcal{C}|$, independently of its sample count.

Numerical floors. The norm floor δ stabilizes sample normalization; ϵ stabilizes the final ratios. A floor-affected embedding has normalized norm below one, giving $N_\phi = \mathbb{E}\|\widehat{\phi}\|_2^2 - D_\phi$. Denominator magnitudes and the frequency of floor-affected embeddings are therefore relevant diagnostics. Away from the norm floor, the indicators are invariant to positive sample-wise rescaling. Temperature affects them through the learned embeddings, even though it does not appear explicitly in the evaluation score.

G.7 Effect of cross-modal and signal–nuisance correlations

We test two zero-mean jointly Gaussian departures from independence, retaining the original latent marginals and independent measurement noises. In the nuisance–nuisance setting, z is independent of (ξ_x, ξ_y), and

$$\mathrm{Cov}(\xi_x, \xi_y) = \rho\nu^2 I_{17},$$

which violates Eq. (46). In the nuisance–signal setting,

$$\mathrm{Cov}(z, \xi_x) = \mathrm{Cov}(z, \xi_y) = \rho\nu[I_r\ 0],$$

which violates both Eqs. (46) and (47). The nuisance vectors are independent conditional on z, which also induces

$$\mathrm{Cov}(\xi_y, \xi_x) = \rho^2\nu^2\,\mathrm{diag}(I_r, 0_{17-r}).$$

Here $r = 3$, $d_1 = d_2 = 20$, $d = 25$, $\sigma^2 = .5$, and $\nu^2 = .3$. The population protocol integrates to $T = 2000$ and summarizes three seeds per setting. The constructed joint covariance is checked for positive semidefiniteness before integration. The selector $[I_r\ 0] \in \mathbb{R}^{3\times 17}$ acts on the first r nuisance coordinates. Table 4 reports seed medians in the original signal–nuisance coordinates.

Table 4: Score and encoder indicators with correlated variables. Each row is a separate correlation setting, summarized by the median over three seeds.

Correlation	ρ	C_{lin}	I_{lin}
Nuisance–nuisance	0.0	2×10^{-11}	.359
	0.1	.063	.78
	0.3	.57	1.9
	0.5	1.6	3.1
Nuisance–signal	0.0	2×10^{-11}	.359
	0.1	.0024	.36
	0.3	.023	.38
	0.5	.073	.41

Correlations make the original off-signal coordinates informative across views, and their score blocks remain nonzero. The ordering $C_{\mathrm{lin}} < I_{\mathrm{lin}}$ holds in every tested cell. At $\rho = .5$, the off-signal cross-covariance norm $\|\Sigma_{yx}A_\perp\|_F$ is 0.62 and 0.49 in the nuisance–nuisance and nuisance–signal settings, respectively, compared with the baseline signal-block norm $\sqrt{r} = 1.73$. The shared-signal score norm $\|\widehat{\widetilde{M}}_{ss}\|_F$ is 0.770 in the first setting and changes from 0.770 to 0.719 in the second.

G.8 Nonlinear retrieval and conditional variability across depths

This study evaluates pair retrieval and relative conditional variability on CITE-seq and COCO. The indicators $C_{\mathrm{nl}}, I_{\mathrm{nl}}$ from Eq. (51) describe score and representation variability within classes. Five depths, two datasets, and two training recipes give twenty settings, each run with seeds $0, 1, 2, 3, 4$. Both recipes are evaluated with the same normalized embeddings and paired cosine score.

Retrieval criterion. Candidates are ranked by cosine similarity. Class Recall@5 counts a query as successful when a same-class item appears among its five highest-ranked candidates; pair Recall@5 requires the original matched observation. Each experiment specifies its cohort and candidate pool.

Architecture and optimization. Each modality has a bias-free multilayer perceptron with $2, 4, 6, 8, 10$ hidden ReLU layers, width 256 throughout, and a linear output layer into $\mathbb{R}^{32}$. We use Kaiming-uniform weight initialization. Both recipes use AdamW with learning rate 10^{-3}, batch size 512, and zero weight decay. The log-temperature is initialized at zero, giving $\tau_0 = 1$; the temperature used by the objective is constrained below by 0.08. CITE-seq models are trained for 600 epochs and COCO models for 1000 epochs. We evaluate before training, after the first epoch, every 25 epochs thereafter, every 100 epochs, and at the final epoch. Tables report the final checkpoint.

Expected-norm training estimates each normalizer on at most 4096 training observations, retains gradients through the normalizer, and uses symmetric paired InfoNCE with same-label off-diagonal negatives masked. The normalizer is bounded below by 10^{-8}, and the total gradient norm is clipped at 10. Cosine training uses one-way paired InfoNCE without a label mask or gradient-norm clipping. The comparison therefore evaluates two complete training recipes with a common architecture, optimizer, horizon, and evaluation protocol.

Table 5: Nonlinear depth-study protocol. Depth counts hidden ReLU layers; each encoder ends with a separate linear output layer. Cohort sizes and splits are fixed across the models in this study.

Setting	CITE-seq	COCO
Paired observations	8000	2048
Training / evaluation	$6400/1600$	$1638/410$
Training epochs	600	1000
Hidden layers	$2, 4, 6, 8, 10$	$2, 4, 6, 8, 10$
Hidden width / output dimension	$256/32$	$256/32$
Optimizer / learning rate	AdamW / 10^{-3}	AdamW / 10^{-3}
Batch size / weight decay	$512/0$	$512/0$
Initial temperature / lower bound	$1/0.08$	$1/0.08$
Seeds / reported checkpoint	$0, 1, 2, 3, 4$ / final	$0, 1, 2, 3, 4$ / final

Preprocessing and evaluation. CITE-seq uses spleen/lymph-node RNA and ADT measurements accessed through `scvi-tools`.[1] COCO uses frozen CLIP ViT-B/32 image and caption features.[2] A fixed random permutation, generated with seed 0, splits each loaded cohort at $\lfloor 0.8n \rfloor$, where n is the cohort size in Table 5. For CITE-seq, each cell is library-size normalized to 10^4 and log-transformed. The 3000 genes with largest training-set variance are retained, followed by training-set standardization and 200-component PCA. ADT measurements undergo the same cell-wise normalization and log transform, followed by training-set standardization. Final feature standardizers for both datasets also use only training rows. Every fitted transformation is then applied to the evaluation rows. Cosine retrieval is evaluated from RNA to ADT and from images to captions, using the entire evaluation cohort as the candidate pool. Class labels define retrieval relevance and the conditioning groups in Eq. (51). The indicators average equally over classes with at least two observations; Recall@5 averages over queries. Labels also mask negatives during expected-norm training, but do not mask negatives in the cosine recipe. These conventions and the temperature constraint distinguish the experiment from the unconstrained population flow.

Table 6: **Nonlinear indicators and pair retrieval under the two training recipes.** Entries use five seeds. $C_{\mathrm{nl}}, I_{\mathrm{nl}}$ are means; R and Recall@5 are means $\pm$ sample standard deviations. Pair retrieval requires the original matched observation. Ratios are formed within runs before aggregation.

Dataset	Recipe	Depth	C_{nl}	I_{nl}	R	Pair R@5
CITE-seq	Expected-norm	2	0.173	0.638	0.271 ± 0.035	0.066 ± 0.004
CITE-seq	Expected-norm	10	0.316	0.561	0.565 ± 0.055	0.053 ± 0.002
CITE-seq	Cosine	2	0.458	2.753	0.167 ± 0.021	0.032 ± 0.002
CITE-seq	Cosine	10	1.403	4.407	0.321 ± 0.078	0.073 ± 0.009
COCO	Expected-norm	2	0.153	0.728	0.211 ± 0.015	0.345 ± 0.018
COCO	Expected-norm	10	0.447	0.852	0.523 ± 0.042	0.258 ± 0.020
COCO	Cosine	2	0.356	1.804	0.197 ± 0.021	0.171 ± 0.020
COCO	Cosine	10	1.162	2.156	0.537 ± 0.092	0.225 ± 0.010

[1] `https://scvi-tools.org`; built-in `scvi.data.spleen_lymph_cite_seq`.

[2] MS-COCO: `https://cocodataset.org`. Annotations use CC BY 4.0; image content is subject to the Flickr Terms of Use. CLIP ViT-B/32: `https://huggingface.co/openai/clip-vit-base-patch32` (MIT license).

Aggregation. For a fixed dataset, recipe, and depth, let $j \in \{0, 1, 2, 3, 4\}$ index the seed. Each run gives $R_j = C_{\mathrm{nl}}^{(j)} / I_{\mathrm{nl}}^{(j)}$. We report the mean and sample standard deviation across seeds for each indicator, ratio, and retrieval value. The ratio is formed within each run before aggregation.

Let $\bar{R}_{\mathrm{first}}$ and $\bar{R}_T$ denote the mean ratios after one epoch and at the final checkpoint. The training factor is $A_{\mathrm{train}} = \bar{R}_{\mathrm{first}} / \bar{R}_T$; a value above one indicates that the ratio decreased between these checkpoints. The untrained checkpoint is reported separately. Table 7 gives the complete final grid. Table 8 gives initial values and the initial-to-final score-variance numerator and squared-mean denominator. The per-run records contain all six components of the indicators at both checkpoints.

Interpretation. All 100 final models have $R < 1$, giving the same ordering of the normalized conditional indicators across datasets, recipes, and depths. Across the 20 settings, the mean ratio decreases by a factor of 1.3–18.2 from the first epoch to the final checkpoint, so every setting has $A_{\mathrm{train}} > 1$. Between the untrained and final checkpoints, R decreases in 98 runs and increases in two: COCO depth 4, seed 3, under each recipe. Absolute score variance N_s increases in all 100 runs, while C_{nl} decreases in 69. Growth of the squared conditional mean therefore contributes to the reduction in normalized variability. This ordering describes relative conditional variability; pair discrimination and robustness under controlled nuisance resampling are evaluated in Appendices G.9 and G.10.

Table 7: Complete depth grid. Indicators and class Recall@5 are five-seed means ± sample standard deviations. $A_{\mathrm{train}} = \bar{R}_{\mathrm{first}} / \bar{R}_T$ compares the mean ratio after one epoch with its final value; values above one indicate a decrease.

Depth	C_{nl}	I_{nl}	R	Class R@5	A_{train}
CITE-seq / Expected-norm					
2	0.173 ± 0.029	0.638 ± 0.037	0.271 ± 0.035	0.892 ± 0.009	1.759
4	0.204 ± 0.023	0.565 ± 0.036	0.362 ± 0.032	0.882 ± 0.008	1.489
6	0.225 ± 0.023	0.539 ± 0.040	0.418 ± 0.034	0.887 ± 0.013	2.419
8	0.315 ± 0.046	0.538 ± 0.055	0.594 ± 0.131	0.884 ± 0.011	4.218
10	0.316 ± 0.020	0.561 ± 0.036	0.565 ± 0.055	0.861 ± 0.015	5.653
CITE-seq / Cosine					
2	0.458 ± 0.052	2.753 ± 0.077	0.167 ± 0.021	0.800 ± 0.008	1.760
4	0.689 ± 0.093	4.201 ± 0.086	0.164 ± 0.019	0.834 ± 0.012	2.429
6	0.977 ± 0.181	4.592 ± 0.203	0.212 ± 0.030	0.835 ± 0.008	3.211
8	0.939 ± 0.360	4.345 ± 0.873	0.212 ± 0.066	0.856 ± 0.020	9.860
10	1.403 ± 0.303	4.407 ± 0.236	0.321 ± 0.078	0.862 ± 0.004	18.239
COCO / Expected-norm					
2	0.153 ± 0.010	0.728 ± 0.033	0.211 ± 0.015	0.820 ± 0.007	3.988
4	0.198 ± 0.019	0.645 ± 0.022	0.306 ± 0.023	0.813 ± 0.014	3.051
6	0.327 ± 0.045	0.789 ± 0.048	0.413 ± 0.046	0.792 ± 0.023	2.414
8	0.391 ± 0.061	0.838 ± 0.081	0.466 ± 0.059	0.773 ± 0.012	4.578
10	0.447 ± 0.058	0.852 ± 0.049	0.523 ± 0.042	0.763 ± 0.012	4.033
COCO / Cosine					
2	0.356 ± 0.041	1.804 ± 0.059	0.197 ± 0.021	0.601 ± 0.022	3.100
4	0.465 ± 0.076	1.774 ± 0.069	0.261 ± 0.035	0.628 ± 0.012	1.999
6	0.670 ± 0.066	1.946 ± 0.071	0.344 ± 0.025	0.651 ± 0.022	1.342
8	0.853 ± 0.181	2.086 ± 0.053	0.408 ± 0.079	0.696 ± 0.021	2.849
10	1.162 ± 0.226	2.156 ± 0.076	0.537 ± 0.092	0.662 ± 0.022	3.176

Table 8: Indicators at the untrained checkpoint, followed by the initial-to-final score-variance numerator and squared-mean denominator. Initial indicators use mean ± sample standard deviation across five seeds; the last two columns are means. All six numerator/denominator components at both checkpoints are provided in the per-run measurements.

Depth	$C_{\mathrm{nl}}(0)$	$I_{\mathrm{nl}}(0)$	$R(0)$	$N_s(0) \to N_s(T)$	$D_s(0) \to D_s(T)$
CITE-seq / Expected-norm					
2	3.660 ± 1.580	0.666 ± 0.067	5.724 ± 2.960	$0.0203 \to 0.0711$	$0.0065 \to 0.4139$
4	1.347 ± 1.305	0.347 ± 0.072	3.614 ± 3.192	$0.0136 \to 0.0942$	$0.0193 \to 0.4622$
6	0.928 ± 0.853	0.251 ± 0.039	3.771 ± 3.460	$0.0106 \to 0.1056$	$0.0280 \to 0.4707$
8	1.059 ± 0.556	0.178 ± 0.045	6.379 ± 3.249	$0.0079 \to 0.1200$	$0.0110 \to 0.3901$
10	0.960 ± 1.329	0.136 ± 0.040	6.481 ± 8.117	$0.0070 \to 0.1316$	$0.0246 \to 0.4181$
CITE-seq / Cosine					
2	3.660 ± 1.580	0.666 ± 0.067	5.724 ± 2.960	$0.0203 \to 0.0380$	$0.0065 \to 0.0833$
4	1.347 ± 1.305	0.347 ± 0.072	3.614 ± 3.192	$0.0136 \to 0.0443$	$0.0193 \to 0.0650$
6	0.928 ± 0.853	0.251 ± 0.039	3.771 ± 3.460	$0.0106 \to 0.0568$	$0.0280 \to 0.0595$
8	1.059 ± 0.556	0.178 ± 0.045	6.379 ± 3.249	$0.0079 \to 0.0693$	$0.0110 \to 0.0866$
10	0.960 ± 1.329	0.136 ± 0.040	6.481 ± 8.117	$0.0070 \to 0.0939$	$0.0246 \to 0.0690$
COCO / Expected-norm					
2	1.123 ± 0.692	0.450 ± 0.049	2.559 ± 1.728	$0.0152 \to 0.0522$	$0.0205 \to 0.3404$
4	0.690 ± 0.506	0.281 ± 0.048	2.741 ± 2.590	$0.0105 \to 0.0790$	$0.0406 \to 0.4001$
6	0.940 ± 0.939	0.169 ± 0.020	5.773 ± 5.821	$0.0092 \to 0.1096$	$0.0252 \to 0.3384$
8	1.298 ± 1.306	0.135 ± 0.038	9.721 ± 10.019	$0.0068 \to 0.1194$	$0.0124 \to 0.3094$
10	0.601 ± 0.634	0.121 ± 0.013	5.037 ± 5.334	$0.0050 \to 0.1385$	$0.0142 \to 0.3118$
COCO / Cosine					
2	1.123 ± 0.692	0.450 ± 0.049	2.559 ± 1.728	$0.0152 \to 0.0237$	$0.0205 \to 0.0671$
4	0.690 ± 0.506	0.281 ± 0.048	2.741 ± 2.590	$0.0105 \to 0.0305$	$0.0406 \to 0.0665$
6	0.940 ± 0.939	0.169 ± 0.020	5.773 ± 5.821	$0.0092 \to 0.0410$	$0.0252 \to 0.0611$
8	1.298 ± 1.306	0.135 ± 0.038	9.721 ± 10.019	$0.0068 \to 0.0601$	$0.0124 \to 0.0729$
10	0.601 ± 0.634	0.121 ± 0.013	5.037 ± 5.334	$0.0050 \to 0.0719$	$0.0142 \to 0.0641$

G.9 Pair discrimination within a class

To test whether the learned comparisons distinguish observations within the conditioning groups, we restrict each query's candidates to observations with its class label and measure top-1 recovery of the original pair. The encoders and evaluation split are those of Appendix G.8. A query is eligible when its class contains at least two observations. For a class of size n_c, a class-only score with uniform tie breaking succeeds with probability $1/n_c$. Averaging over eligible queries gives a chance level of $26/1600 = 0.01625$ on CITE-seq and $47/399 \simeq 0.11779$ on COCO. The latter excludes eleven singleton queries from the 410-pair evaluation cohort.

Table 9 reports full-pool pair Recall@5 and same-class top-1 accuracy. Top-1 recovery requires the paired score to strictly exceed every competing score; tied maxima count as failures. All 100 final models exceed their dataset's class-only chance level. CITE-seq accuracies range from 0.02188 to 0.05375, and COCO accuracies from 0.15539 to 0.28822. No evaluated maximum is tied. Thus the learned scores retain pair discrimination within the groups used to measure conditional variability.

Within-class pairing control. We keep each final score matrix fixed and permute the pair assignment within each class $9{,}999$ times, preserving one-to-one pairing. All models of a dataset use the same permutations. For each model, the one-sided Monte Carlo p-value is $(1 + b)/10{,}000$, where b counts permutations with at least as many correct pairs as observed. The values range from 0.0001 to 0.0256 on CITE-seq and from 0.0001 to 0.0044 on COCO. After Holm correction across all 100 models, the largest adjusted value is 0.039. This control tests random pair assignment within classes, conditional on the fitted models and evaluation cohorts.

Table 9: Pair retrieval for the complete depth grid. Entries are five-seed means ± sample standard deviations. Pair Recall@5 uses all evaluation candidates; within-class top-1 uses only same-class candidates and excludes singleton queries.

Depth	Initial pair R@5	Final pair R@5	Within-class top-1
CITE-seq / Expected-norm			
2	0.003 ± 0.001	0.066 ± 0.004	0.033 ± 0.002
4	0.004 ± 0.002	0.064 ± 0.003	0.028 ± 0.006
6	0.003 ± 0.001	0.057 ± 0.004	0.030 ± 0.003
8	0.004 ± 0.002	0.053 ± 0.008	0.027 ± 0.003
10	0.005 ± 0.002	0.053 ± 0.002	0.026 ± 0.001
CITE-seq / Cosine			
2	0.003 ± 0.001	0.032 ± 0.002	0.028 ± 0.003
4	0.004 ± 0.002	0.055 ± 0.007	0.041 ± 0.003
6	0.003 ± 0.001	0.055 ± 0.009	0.041 ± 0.008
8	0.004 ± 0.002	0.071 ± 0.012	0.044 ± 0.007
10	0.005 ± 0.002	0.073 ± 0.009	0.037 ± 0.005
COCO / Expected-norm			
2	0.015 ± 0.002	0.345 ± 0.018	0.225 ± 0.015
4	0.011 ± 0.003	0.345 ± 0.019	0.221 ± 0.011
6	0.012 ± 0.004	0.303 ± 0.020	0.203 ± 0.024
8	0.012 ± 0.003	0.280 ± 0.019	0.192 ± 0.024
10	0.012 ± 0.005	0.258 ± 0.020	0.173 ± 0.016
COCO / Cosine			
2	0.015 ± 0.002	0.171 ± 0.020	0.239 ± 0.011
4	0.011 ± 0.003	0.176 ± 0.008	0.216 ± 0.020
6	0.012 ± 0.004	0.199 ± 0.021	0.230 ± 0.019
8	0.012 ± 0.003	0.235 ± 0.031	0.249 ± 0.026
10	0.012 ± 0.005	0.225 ± 0.010	0.214 ± 0.009

G.10 Pairwise reliability under nuisance resampling

This experiment measures pairwise accuracy, ranking stability, and encoder sensitivity under controlled nuisance interventions with a known continuous shared signal.

Model and training. We use Eq. (1) with $r = 3$, $d_1 = d_2 = 20$, $\sigma^2 = 0.5$, and $\nu^2 = 0.3$; orthogonal input bases come from Gaussian QR decompositions. Each bias-free encoder has $1, 2, 4, 6, 8$, or 10 hidden ReLU layers of width 64, a linear output into $\mathbb{R}^{25}$, and Kaiming-uniform initialization. Three seeds $0, 1, 2$ per depth specify the input bases, minibatch stream, and encoder initialization, giving eighteen runs.

Training uses one-way paired cosine InfoNCE with fixed $\tau = 0.5$, Adam with learning rate 10^{-3}, batch size 256, and $20{,}000$ fresh Gaussian minibatches. There is no gradient clipping, weight decay, or label mask. Initial and final checkpoints are evaluated on identical draws, with evaluation seed $1000j + L$ for seed j and depth L. Both checkpoints are retained for the additional evaluations below.

Interventions and measurements. Each of 50 groups fixes independent shared vectors z_i, z_j. The query and positive candidate share z_i, and the distractor has z_j. We draw 100 independent noise realizations per group. In the first intervention, we resample only ξ_x, ξ_y for the three observations, holding each observation's measurement noise ε fixed. In the second, we also resample the measurement noise. The query is shared between its positive and negative comparisons within each draw.

We measure how often $s(X_i, Y_i) > s(X_i, Y_j)$, and how often two independent resamples reverse this ordering while keeping the shared vectors fixed. If n_+ and n_- are the numbers of positive and negative score differences among the 100 draws, the latter probability is estimated by

$$\frac{2n_+n_-}{100 \cdot 99}.$$

This averages the reversal indicator over all distinct pairs of draws; we then average over the 50 groups. There are no ties in the evaluated comparisons. At a fixed group success probability p, the population reversal probability is $2p(1-p)$. We report accuracy alongside reversal to distinguish stable correct and stable incorrect rankings. Both statistics are invariant to a common positive affine transformation of the score, so adding a common embedding coordinate cannot explain their improvement.

Cosine comparisons. Table 10 gives the complete grid. Under nuisance-only resampling, the mean success probability across eighteen runs rises from 0.502 to 0.801, while the mean reversal probability falls from 0.383 to 0.046. When measurement noise is also resampled, the corresponding changes are $0.502 \to 0.806$ and $0.488 \to 0.271$. Every run improves both statistics in both interventions. Thus training produces more reliable pairwise rankings under noise resampling.

Bilinear comparisons on the same encoders. At a fixed checkpoint, expected-norm normalization divides every bilinear score by the same positive constant, so its rankings equal those of $f(X)^\top g(Y)$. We evaluate this score on the same cosine-trained encoders and the same draws; the training objective is unchanged. Table 10 shows improvement in all eighteen runs for both interventions. The nonlinear observations complement the limiting expected-norm result in Corollary E.7.

Table 10: Pairwise reliability under nuisance resampling, using cosine and bilinear scores on the same cosine-trained encoders. Entries are means $\pm$ sample standard deviations across three seeds. Initial values precede training; final values follow $20{,}000$ updates. Success requires the matched score to exceed the distractor score. Reversal compares two resamples at fixed shared signals.

Depth	Initial success	Final success	Initial reversal	Final reversal
Cosine / modality nuisance only				
1	0.484 ± 0.030	0.828 ± 0.047	0.358 ± 0.003	0.052 ± 0.023
2	0.561 ± 0.031	0.782 ± 0.030	0.361 ± 0.022	0.040 ± 0.013
4	0.478 ± 0.017	0.814 ± 0.052	0.384 ± 0.010	0.048 ± 0.015
6	0.488 ± 0.004	0.792 ± 0.031	0.409 ± 0.003	0.041 ± 0.025
8	0.492 ± 0.031	0.809 ± 0.031	0.411 ± 0.016	0.056 ± 0.039
10	0.509 ± 0.015	0.780 ± 0.046	0.373 ± 0.031	0.037 ± 0.020
Cosine / modality nuisance and measurement noise				
1	0.500 ± 0.016	0.802 ± 0.047	0.489 ± 0.009	0.274 ± 0.068
2	0.513 ± 0.003	0.800 ± 0.028	0.484 ± 0.010	0.274 ± 0.034
4	0.503 ± 0.003	0.820 ± 0.004	0.492 ± 0.005	0.261 ± 0.006
6	0.499 ± 0.005	0.813 ± 0.007	0.493 ± 0.001	0.257 ± 0.009
8	0.495 ± 0.015	0.810 ± 0.031	0.491 ± 0.008	0.268 ± 0.036
10	0.502 ± 0.007	0.790 ± 0.031	0.482 ± 0.014	0.290 ± 0.032
Bilinear / modality nuisance only				
1	0.476 ± 0.022	0.828 ± 0.050	0.345 ± 0.016	0.048 ± 0.012
2	0.561 ± 0.029	0.768 ± 0.047	0.357 ± 0.029	0.058 ± 0.020
4	0.475 ± 0.029	0.805 ± 0.046	0.373 ± 0.017	0.055 ± 0.009
6	0.491 ± 0.013	0.766 ± 0.017	0.395 ± 0.007	0.040 ± 0.008
8	0.493 ± 0.033	0.832 ± 0.038	0.378 ± 0.025	0.042 ± 0.020
10	0.505 ± 0.023	0.775 ± 0.031	0.344 ± 0.010	0.037 ± 0.010
Bilinear / modality nuisance and measurement noise				
1	0.497 ± 0.016	0.788 ± 0.055	0.486 ± 0.010	0.276 ± 0.073
2	0.514 ± 0.004	0.783 ± 0.022	0.485 ± 0.009	0.277 ± 0.032
4	0.506 ± 0.008	0.808 ± 0.006	0.486 ± 0.004	0.266 ± 0.007
6	0.501 ± 0.004	0.798 ± 0.010	0.487 ± 0.006	0.261 ± 0.013
8	0.502 ± 0.032	0.801 ± 0.037	0.484 ± 0.003	0.269 ± 0.037
10	0.505 ± 0.007	0.780 ± 0.034	0.477 ± 0.010	0.286 ± 0.029

Encoder variability under the same nuisance intervention. We evaluate Eq. (51) on the query–positive pairs from the nuisance-only intervention, conditioning on $(z, \varepsilon_x, \varepsilon_y)$ in place of z. Denote these indicators by C_ξ and I_ξ. They measure variation caused by the same modality-nuisance draws used for the ranking statistics. Table 11 reports the six-depth results. Across eighteen runs, mean I_ξ decreases from 0.2103 to 0.0125; every run decreases and retains nonzero final variability. Final

I_ξ ranges from 0.00208 to 0.01414, and C_ξ/I_ξ ranges from 0.390 to 0.926. Thus improved ranking robustness coexists with residual encoder sensitivity to modality nuisance and a lower normalized score indicator in every run.

Table 11: Conditional indicators under modality-nuisance-only resampling. Shared signal and measurement noise are fixed within each group. Entries are means $\pm$ sample standard deviations over three seeds; ratios are formed within runs.

Depth	$I_\xi(0)$	$I_\xi(T)$	$C_\xi(T)$	$C_\xi(T)/I_\xi(T)$
1	0.3204 ± 0.0111	0.0131 ± 0.0008	0.0099 ± 0.0035	0.752 ± 0.230
2	0.3134 ± 0.0524	0.0137 ± 0.0004	0.0094 ± 0.0014	0.686 ± 0.106
4	0.2123 ± 0.0304	0.0135 ± 0.0006	0.0089 ± 0.0007	0.661 ± 0.031
6	0.1475 ± 0.0185	0.0124 ± 0.0011	0.0074 ± 0.0016	0.596 ± 0.094
8	0.1671 ± 0.0405	0.0133 ± 0.0008	0.0086 ± 0.0035	0.647 ± 0.246
10	0.1009 ± 0.0558	0.0092 ± 0.0062	0.0068 ± 0.0043	0.797 ± 0.131

With both modality nuisance and measurement noise resampled at fixed z, final I_{nl} ranges from 0.599 to 0.796, while $C_{\mathrm{nl}}/I_{\mathrm{nl}}$ ranges from 0.318 to 0.468. These groups fix the continuous shared signal exactly, so the indicator ordering also holds without class-based conditioning. This measurement includes noise along the shared coordinates. To distinguish its contribution, consider the signal-only projections $f(X) = (A_s^\top X, 0)$ and $g(Y) = (B_s^\top Y, 0)$, padded to the same output dimension. Both encoders have zero off-signal blocks. On the same evaluation draws, their mean I_{nl} is 0.652 and mean $C_{\mathrm{nl}}/I_{\mathrm{nl}}$ is 0.503, whereas nuisance-only encoder variability is at numerical precision. The nuisance-only measurements therefore isolate the residual modality-nuisance sensitivity from conditional variability due to measurement noise.

Validation on fresh intervention draws. We reevaluate the same initial and final checkpoints on 200 new signal groups with 100 draws per group, using seed $400000 + 1000j + L$. Each checkpoint pair receives identical draws. Table 12 reports means over all eighteen runs. Accuracy increases and reversal decreases in every run under both interventions. Nuisance-only encoder variability also decreases in every run and remains positive in both encoders separately. Final I_ξ ranges from 0.00270 to 0.01368, and C_ξ/I_ξ ranges from 0.517 to 0.843.

Within each run, we resample the 200 groups with replacement $2{,}000$ times, preserving initial/final pairing and recomputing the ratios from their components. Each run's percentile 95% interval for the accuracy gain and reversal reduction lies above zero in both interventions. Each final nuisance-only indicator-ratio interval lies below one. These are pointwise intervals for evaluation-group sampling at fixed checkpoints; they do not quantify variation over new training runs.

Table 12: Fresh-draw evaluation of the same eighteen checkpoint pairs. Entries are means across runs; all groups are separate from those in the preceding tables. Conditional encoder variability uses nuisance-only draws for I_ξ and both noise sources for I_{nl}.

Resampled variables	Measurement	Initial	Final
ξ only	Pairwise accuracy	0.4963	0.8223
ξ only	Reversal probability	0.3774	0.0403
ξ only	I_ξ	0.2063	0.0125
ξ, ε	Pairwise accuracy	0.4996	0.8132
ξ, ε	Reversal probability	0.4886	0.2646
ξ, ε	I_{nl}	0.5826	0.7050

The full-noise encoder indicator increases on average, while nuisance-only sensitivity decreases. Conditioning and intervention therefore determine what the indicator measures; the residual modality-nuisance conclusion uses the nuisance-only arm.

G.11 Direct comparison with explicit encoder-alignment regularization

Question and protocol. The preceding experiments keep the contrastive objective fixed. We now test directly whether an objective designed to remove modality-specific information changes the distinction between score reliability and encoder minimality. We use the positive-pair alignment term in the official implementation of Almudévar et al. (2025),

$$\mathcal{L}_\beta = \mathcal{L}_{\mathrm{InfoNCE}} + \beta \left(1 - \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)^\top g(y_i)}{\|f(x_i)\|_2 \|g(y_i)\|_2} \right). \tag{53}$$

This is a controlled comparison in our continuous-latent model, not a reproduction of every architecture and dataset in that work.

We use Eq. (1) with $r = 3$, $d_1 = d_2 = 20$, $\sigma^2 = 0.5$, and $\nu^2 = 0.3$. Each modality has a separate bias-free MLP with one width-64 ReLU hidden layer and a 25-dimensional linear output. Training uses one-way cosine InfoNCE, fixed $\tau = 0.5$, Adam with learning rate 10^{-3}, batch size 128, and 2,000 fresh Gaussian minibatches. We sweep $\beta \in \{0, 0.1, 0.3, 1, 3\}$ over ten seeds. Within each seed, every value of β uses the same input bases, initialization, minibatch stream, and evaluation draws. Explicit NumPy backpropagation was checked against centered finite differences, with maximum relative error 1.92×10^{-11} on the checked coordinates.

Evaluation follows Appendix G.10, but uses 100 groups and 100 draws per group: shared signals and measurement noises are fixed, while modality-specific nuisance is resampled. We report pairwise accuracy, reversal probability, positive-pair cosine, and the indicators I_ξ, C_ξ. We also fit held-out linear ridge probes for the modality nuisance and the shared signal; the probes split by intervention group, so different nuisance draws of one fixed shared signal cannot cross the train–test boundary.

Results. Table 13 reports means and sample standard deviations over the ten paired seeds. Unregularized InfoNCE already has substantially lower normalized score than encoder sensitivity (mean $C_\xi / I_\xi = 0.217$) while nuisance remains decodable from individual encoder outputs. Increasing β then removes progressively more encoder nuisance without a corresponding improvement in pairwise accuracy.

Table 13: **Direct controlled comparison with explicit positive-pair alignment.** Values are mean ± sample standard deviation over ten paired seeds. Acc. is matched-versus-distractor pairwise accuracy; Rev. is the nuisance-induced reversal probability; Cos. is positive-pair cosine. The final two columns are held-out linear-probe R^2 for modality nuisance and shared signal, respectively.

β	Acc.	Rev.	Cos.	I_ξ	C_ξ	Nuis. R^2	Signal R^2
0	0.8083±0.0285	0.0471±0.0139	0.5947±0.0423	0.03846±0.00205	0.008388±0.001472	0.0751±0.0092	0.5256±0.0453
0.1	0.8087±0.0283	0.0472±0.0136	0.6121±0.0409	0.03077±0.00142	0.008108±0.001497	0.0526±0.0087	0.5313±0.0445
0.3	0.8092±0.0286	0.0483±0.0150	0.6424±0.0385	0.02188±0.00098	0.007045±0.001411	0.0269±0.0075	0.5402±0.0429
1	0.8111±0.0313	0.0482±0.0130	0.7463±0.0275	0.01035±0.00046	0.003206±0.000699	0.0093±0.0047	0.5612±0.0425
3	0.7919±0.0283	0.1124±0.0169	0.9899±0.0022	0.00174±0.00013	0.000010±0.000004	−0.0008±0.0055	0.5477±0.0349

For $\beta = 1$ relative to $\beta = 0$, the paired mean accuracy difference is 0.00282, with Student t 95% interval $[-0.00395, 0.00959]$, and the reversal difference is 0.00109, with interval $[-0.00288, 0.00506]$. In contrast, I_ξ decreases by 73.1% (paired difference -0.02811, interval $[-0.02930, -0.02692]$) and nuisance-probe R^2 decreases by 87.6% (difference -0.06573, interval $[-0.07262, -0.05884]$); both decrease in all ten seeds. Positive-pair cosine increases by 0.15155, and signal-probe R^2 increases by 0.03565, also in every seed. Thus a moderate alignment penalty makes the encoders substantially more minimal without a detectable change in comparison reliability in this experiment.

At $\beta = 3$, positive-pair cosine reaches 0.9899 and both nuisance measurements approach zero, but reversal probability increases by 0.06528 with paired 95% interval $[0.05569, 0.07488]$, increasing in every seed. The accuracy difference is -0.01635, although its interval $[-0.03727, 0.00457]$ remains inconclusive. The reversal result does not contradict the small absolute C_ξ: the alignment term does not directly maximize matched-versus-distractor margins, so small score changes can reverse an ordering when the margin is also small.

These observations sharpen the distinction with Almudévar et al. (2025). Their regularizer targets encoder minimality; our separation concerns the weaker condition needed for a fixed joint score to be

relatively nuisance-insensitive. Moderate regularization can improve the former without changing the latter, while near-perfect positive-pair alignment does not by itself certify stable rankings. Because this experiment is synthetic and uses a fixed temperature, it is a controlled mechanism test and does not establish the same numerical tradeoff for pretrained CLIP.

H Retrieval Compression: Protocol and Additional Results

H.1 Representations, fitting data, and comparison objective

We use the public `openai/clip-vit-large-patch14` checkpoint at revision `32bd64288804d66eefd0ccbe215aa642df71cc41`. The backbone and its original 768-dimensional projection outputs are frozen. Feature extraction uses FP16 autocast, followed by FP32 unit normalization; TF32 is disabled. The original logit scale is $\gamma = 100.00000762939453$. Each dataset supplies 10,000 calibration images and all their captions. The validation cohorts contain 5,000 COCO images and 1,014 Flickr30k images. Neither fit uses relevance labels: image and caption queries are sampled independently, and the targets are the original model's comparison distributions. Relevance labels are used for validation and evaluation.

For unit feature vectors x_i, y_j, let

$$S_{ij} = \gamma x_i^\top y_j, \qquad \widehat{S}_{ij} = \widehat{\gamma}\frac{(P_A^\top x_i)^\top (P_B^\top y_j)}{\|P_A^\top x_i\|\,\|P_B^\top y_j\|}, \qquad P_A, P_B \in \mathbb{R}^{768\times k}.$$

The fitting objective is

$$\frac{1}{2n_I}\sum_{i=1}^{n_I} \mathrm{KL}\Big(\mathrm{softmax}(S_{i,:}) \,\|\, \mathrm{softmax}(\widehat{S}_{i,:})\Big) + \frac{1}{2n_T}\sum_{j=1}^{n_T} \mathrm{KL}\Big(\mathrm{softmax}(S_{:,j}) \,\|\, \mathrm{softmax}(\widehat{S}_{:,j})\Big). \tag{54}$$

Each sampled query uses the full opposite-modality calibration gallery. The scalar $\widehat{\gamma} > 0$ is learned in log coordinates. It calibrates the fitting distributions and does not change the final retrieval order. Both maps are fitted jointly, and the compressed features are normalized during fitting and retrieval. This finite-gallery objective implements the comparison-preserving choice; the population linear-model theorem does not prescribe this particular optimizer or guarantee its empirical approximation error.

Initialization and optimization. Let $C_A = n_I^{-1}\sum_i x_i x_i^\top$ and $C_B = n_T^{-1}\sum_j y_j y_j^\top$. From the SVD $C_A^{1/2}C_B^{1/2} = U\Lambda V^\top$, initialize $P_A = C_A^{-1/2}U_k\Lambda_k^{1/2}$ and $P_B = C_B^{-1/2}V_k\Lambda_k^{1/2}$, using pseudoinverses below a relative eigenvalue threshold of 10^{-12}. This minimizes the unregularized two-sided weighted squared score error. We first take 1,000 Adam steps on the symmetric KL within independent batches of 1,024 images and 2,048 captions, using unnormalized projected scores and the original scale. The learning rate is $10^{-4} + 9\cdot 10^{-4}(1+\cos(\pi t/999))/2$, for $t = 0, \ldots, 999$.

We then take 250 Adam steps on Eq. (54), sampling 256 image queries and 512 text queries per step with replacement. At step t, the map learning rate is $2\cdot 10^{-4}h(t)$ and the log-scale learning rate is $2\cdot 10^{-3}h(t)$, where $h(t) = 0.1 + 0.9(1+\cos(\pi t/1999))/2$. This uses the first 250 steps of the 2,000-step schedule selected during development; the schedule is not rescaled to 250 steps. The scale is initialized as γ times the product of the two modalities' mean projected norms. Adam uses its default momentum parameters and no weight decay. The same procedure and hyperparameters are used on both datasets. Optimizer seeds are 20260926, 20260927, and 20260928; the first is the primary seed.

Linear baselines and storage. The pooled second moment gives equal weight to the two modalities, $(C_A + C_B)/2$. We consider its eigenvectors and those of the pooled centered covariance, with either projected dot products or projected cosine similarity. All projections act on the original feature vectors; "centered" refers to the basis estimation. Validation selects among these four candidates at each width. Cosine retrieval is selected on both datasets, with an uncentered basis on COCO and a centered basis on Flickr30k. An extension also fits independent PCA, Procrustes-aligned independent PCA, cross-covariance SVD, and regularized CCA. Their centering, singular-value weighting, and ridge choices are selected using validation only; Table 14 records the selected configurations. All methods store one k-coordinate vector per item. The two comparison-preserving maps add $2\cdot 768k$ shared parameters; storage-reduction factors refer to the vectors, not backbone memory or end-to-end encoding latency.

Table 14: **Validation-only baseline selection.** Widths remain 120 for COCO and 96 for Flickr30k. The strongest baseline is identified by validation mean R@1, without using either held-out cohort. The entry none denotes no singular-value weighting; CCA entries give (λ_x, λ_y).

Dataset	Family	Selected configuration	Validation mean R@1 (%)
COCO	Pooled PCA	uncentered; cosine	32.10
COCO	Independent PCA (raw)	centered; cosine	0.03
COCO	PCA + Procrustes	centered; cosine	23.41
COCO	Cross-covariance SVD	centered; cosine; none	34.69
COCO	Regularized CCA	uncentered; cosine; none; $(0.1, 10)$	**42.96**
Flickr30k	Pooled PCA	centered; cosine	61.73
Flickr30k	Independent PCA (raw)	centered; cosine	0.17
Flickr30k	PCA + Procrustes	centered; cosine	50.44
Flickr30k	Cross-covariance SVD	centered; cosine; none	66.62
Flickr30k	Regularized CCA	centered; cosine; square-root; $(0.1, 10)$	**76.48**

Table 15: **Compression on fresh held-out cohorts.** Bidirectional mean R@1 (%); mean $\pm$ sample standard deviation over three optimizer seeds for the comparison-preserving maps. PCA and the original model are deterministic. Regularized CCA is the strongest validation-selected linear baseline. The gain is relative to same-width PCA. Storage counts coordinates per stored vector.

Dataset	Width	Storage gain	Original	PCA	CCA	Comparisons	Comp.–PCA (pp)
COCO	120	6.4×	46.10	31.95	**43.28**	42.25 ± 0.05	+10.30
Flickr30k	96	8×	76.31	60.64	**76.60**	74.00 ± 0.05	+13.36

H.2 Development history and evaluation cohorts

The first squared-score-error compression failed the target retrieval gain. Matching softmax distributions improved it, but minibatch-only fitting still gave a small COCO advantage. Development then evaluated full-gallery fitting and normalization of the projected outputs. The final procedure uses both; all earlier outcomes remain in the experiment record. Width and checkpoint selection used validation retrieval. For COCO, the selection rule was amended before its new test evaluation to favor joint slack in the gain and retention constraints; this selected width 120 at step 250 over a width-112 candidate closer to the retention boundary. Flickr30k's width was fixed at 96. The final COCO procedure was then applied unchanged to Flickr30k. Standard test splits had been evaluated during this development, so we treat their results as exploratory.

Before inspecting the new cohorts' features or scores, we froze all estimator choices and sampled 5,000 COCO and 1,000 Flickr30k images uniformly from the remaining Karpathy train/restval pools, with seed 2026092601. We excluded the current fitting, validation, and standard test IDs, as well as image IDs in saved older compression-pilot bundles. Every sampled image and its original captions was retained, without filtering by category, model score, or availability. The resulting galleries have 25,018 COCO captions and 5,000 Flickr30k captions. Image IDs and file-byte hashes have zero overlap with the current fitting, validation, and standard test cohorts. These are additional held-out cohorts, not the standard benchmark test splits or a new data domain. No choices were revised using their outcomes.

For image-to-text retrieval, the rank is the best rank among that image's captions; for text-to-image retrieval, it is the rank of the caption's source image. The reported mean R@1 equally weights the two directions. We use the full gallery, including 5,000 images for COCO, rather than averaging 1,000-image folds. For uncertainty, 5,000 paired bootstrap replicates resample image-query clusters, including their captions, with the ranked gallery fixed. Thus intervals describe query sampling conditional on a gallery and fitted maps, not uncertainty over new galleries or training datasets. Seed standard deviations separately measure optimizer variation.

Table 16: **Standard test results (exploratory).** Same notation as Table 15. COCO uses 5,000 images and 25,010 captions; Flickr30k uses 1,000 images and 5,000 captions.

Dataset	Width	Storage gain	Original	PCA	CCA	Comparisons	Comp.–PCA (pp)
COCO	120	6.4×	46.42	31.67	**42.78**	42.41 ± 0.13	+10.74
Flickr30k	96	8×	74.97	60.00	**74.97**	71.98 ± 0.03	+11.98

Table 17: **Held-out objective tradeoffs.** Mean R@1 and exact teacher top-1 agreement are percentages. R^2 is image/text affine reconstruction. Comparisons reports the mean over the three fixed maps. Bold marks the highest compressed label R@1, not a universally preferred method.

Dataset	Method	Mean R@1	R^2 I/T	Teacher top-1 agreement
COCO	Original 768D	46.10	1.000/1.000	100.00
COCO	Pooled PCA	31.95	0.643/0.768	45.94
COCO	Independent PCA (raw)	0.04	0.681/0.790	0.01
COCO	PCA + Procrustes	23.72	0.681/0.790	16.61
COCO	Cross-covariance SVD	35.58	0.645/0.772	25.30
COCO	Regularized CCA	**43.28**	0.542/0.769	30.45
COCO	Comparisons	42.25	0.602/0.746	60.06
Flickr30k	Original 768D	76.31	1.000/1.000	100.00
Flickr30k	Pooled PCA	60.64	0.583/0.711	59.93
Flickr30k	Independent PCA (raw)	0.19	0.626/0.733	0.10
Flickr30k	PCA + Procrustes	48.89	0.626/0.733	31.78
Flickr30k	Cross-covariance SVD	66.53	0.585/0.709	45.33
Flickr30k	Regularized CCA	**76.60**	0.469/0.690	52.18
Flickr30k	Comparisons	74.00	0.541/0.680	72.06

H.3 Extended linear baselines and three evaluation objectives

The baseline extension keeps the data splits, widths, comparison maps, and evaluation code fixed. Independent PCA fits each modality separately. Its raw coordinates need not share an axis system; PCA + Procrustes therefore uses paired calibration items to fit an orthogonal alignment. Cross-covariance SVD uses the leading left and right singular vectors of the paired cross-covariance. Regularized CCA whitens both views with validation-selected ridges and optionally weights canonical coordinates. Centering, cosine versus dot-product scoring, singular-value weighting, and CCA ridges are selected only on validation (Table 14); held-out results do not select families or hyperparameters.

Table 17 reports three different quantities on the additional held-out cohorts. Reconstruction R^2 measures linearly recoverable encoder information. Label mean R@1 measures retrieval against dataset relevance. Teacher top-1 agreement measures whether compression returns exactly the original model's top candidate. For image-to-text agreement, selecting a different caption of the same image is a disagreement even if that caption is also relevant; agreement is therefore a descriptive comparison-rule diagnostic, not another relevance metric or a claim of full score preservation.

The independently fitted PCA maps make the distinction especially sharp: they reconstruct each encoder better than the comparison maps but, before cross-modal alignment, retrieve at approximately chance. Procrustes alignment repairs part of this coordinate mismatch without matching the comparison maps' retrieval. Cross-covariance SVD also reconstructs both modalities better while remaining 6.66 and 7.47 points below the comparison maps on held-out COCO and Flickr30k. Conversely, regularized CCA has the highest label R@1 but preserves the teacher's exact top choice much less often. The paired teacher-agreement advantages of the comparison maps over CCA are 29.61 points $[28.72, 30.48]$ on COCO and 19.88 points $[17.87, 21.93]$ on Flickr30k.

The exploratory standard test gives the same qualification: comparisons minus CCA is -0.37 points $[-1.17, 0.46]$ on COCO and -2.99 points $[-4.57, -1.37]$ on Flickr30k. Thus the evidence rejects a universal same-width linear-baseline superiority claim. Its positive conclusion is the objective separation: high encoder reconstruction does not ensure retrieval or teacher-score preservation, and high label retrieval does not ensure fidelity to the original comparison rule.

The prespecified confirmation criteria are width at most 192, at least 90% retention of original R@1 in each direction, a mean R@1 gain of at least ten percentage points over same-width validation-selected

Table 18: **Paired held-out mean-R@1 differences.** Comparison maps minus each validation-selected baseline, averaging per-query success over the three fixed maps before 5,000 image-group bootstrap replicates. Intervals condition on the fitted maps and fixed gallery.

Dataset	Baseline	Difference (pp)	Paired 95% interval
COCO	Pooled PCA	10.30	[9.60, 11.00]
COCO	PCA + Procrustes	18.52	[17.61, 19.44]
COCO	Cross-covariance SVD	6.66	[5.83, 7.51]
COCO	Regularized CCA	-1.03	[-1.86, -0.24]
Flickr30k	Pooled PCA	13.36	[11.87, 14.85]
Flickr30k	PCA + Procrustes	25.11	[23.04, 27.14]
Flickr30k	Cross-covariance SVD	7.47	[5.63, 9.31]
Flickr30k	Regularized CCA	-2.60	[-4.13, -1.10]

PCA, and a positive lower endpoint of the paired 95% interval. All three seeds pass on each dataset, on both the standard and additional held-out cohorts. On the latter, the minimum directional retention across seeds is 90.18% for COCO and 96.51% for Flickr30k.

H.4 Encoder reconstruction and individual retrieval failures

For each modality and compression map, we fit an affine least-squares decoder on the calibration features. Its held-out reconstruction statistic is

$$R^2 = 1 - \frac{\sum_i \|z_i - \widehat{z}_i\|^2}{\sum_i \|z_i - \overline{z}_{\text{train}}\|^2},$$

where z_i is the original 768-dimensional unit feature. The decoder uses the raw projected coordinates, which can be stored and normalized when computing cosine retrieval without increasing the width. PCA obtains image/text $R^2 = 0.643/0.768$ on the fresh COCO cohort, versus $0.602/0.746$ for the comparison-preserving maps. On Flickr30k the corresponding values are $0.583/0.711$ and $0.541/0.680$. PCA therefore retains more linearly recoverable encoder information in every comparison, despite its lower retrieval accuracy. These measurements do not assign true nuisance labels to CLIP features.

We also enumerate image queries for which the original model ranks a relevant caption first, PCA moves every relevant caption below rank 100, and the primary compressed model recovers rank one. At the operating widths, there is one such COCO image in the standard test and one in the fresh cohort, with PCA ranks 541 and 167 respectively. There are no qualifying Flickr30k images and no reverse cases under the same criterion. For COCO image 576463, the original and compressed models retrieve the caption "the cover of a magazine with a soccer player on the front," while PCA retrieves "A man adjusting his tie while having a blurry hand." The full enumeration and source captions are saved with the results.

An earlier, distinct severe-failure criterion required at least three cases while PCA retained at least 95% of pooled feature variance. It found zero qualifying cases and remains unsuccessful. The two operating-width cases do not satisfy that earlier requirement. Proposition F.1 is an exact possibility result in the theoretical model, separately from these real-data observations.

Finally, the query-wise certificate in Proposition F.2 certifies 13.98% of COCO image queries and 27.56% of its text queries on the fresh cohort, with zero observed violations. Flickr30k's corresponding coverage is 29.80% and 54.50%, also with zero violations. Coverage is lower than empirical top-candidate agreement, consistent with the certificate being sufficient. We independently recomputed 9,216 sampled full-gallery ranks across the baseline extension using double-precision scoring and stable sorting; all matched the recorded ranks exactly. Five unit tests passed, and repeated fitting, validation selection, and held-out evaluation reproduced the saved deterministic maps and metrics.

Data integrity and evaluation scope. Cache and manifest digests, split disjointness, finite features, caption mappings, and unique confirmation IDs were checked. COCO has one train/test byte-duplicate image group; excluding its single standard-test query changes every reported mean R@1 by at most 0.01 point, and the confirmation cohort has no byte overlap with any other split. The confirmation

cohort is independent of fitting and validation. Its comparison-map and pooled-PCA outcomes were available when the baseline study was designed; baseline selection uses only validation outcomes. The standard test is exploratory. Teacher top-1 agreement is a parameter-free post-evaluation diagnostic and was not used for method or hyperparameter selection.

I Additional Related Work

Alignment/uniformity and information-theoretic analyses describe what contrastive objectives encourage at the representation level (van den Oord et al., 2018; Tschannen et al., 2020; Wang & Isola, 2020). Multi-view theories study canonical correlation analysis and related recovery guarantees, augmentation or latent-class assumptions, and CLIP-style models (Hotelling, 1936; Kakade & Foster, 2007; Tosh et al., 2021; Saunshi et al., 2019; Lee et al., 2021; HaoChen et al., 2021). Spiked-covariance models separate low-rank signal from higher-dimensional variation (Johnstone, 2001; Baik et al., 2005; Bai & Yao, 2012).

Almudévar et al. (2025) provide the closest representation-level treatment of modality-specific nuisance. They posit an "essence" variable shared by the modalities and separate nuisance variables, define a representation as sufficient when it retains the essence and minimal when it discards modality-specific information, and relate perfect population representation alignment to this minimality. Their variational information-bottleneck term compares stochastic representations of positive pairs and explicitly encourages nuisance removal. Their controlled experiments measure retained nuisance with classifiers and alignment with CKA, study depth and temperature, and their real-data experiments examine the captioning–retrieval tradeoff after adding the regularizer. The present work does not seek perfect representation alignment or propose a nuisance-removal penalty. It keeps the contrastive objective fixed, uses a noisy linear–Gaussian model with explicit score blocks, and analyzes the long-time flow of the two factors and learned temperature. The central possibility is consequently different: modality-specific variation may remain in encoder outputs while lying in directions that are invisible to the other encoder, so the joint bilinear score is nuisance-insensitive. Theorem 1 and Corollary 3 establish this score-side behavior along the flow, while Propositions 5–E.2 give the output-capacity threshold and an open basin with positive limiting nuisance. Thus their work motivates explicit encoder minimality for alignment; ours distinguishes encoder minimality from reliability of a fixed cross-modal comparison.

Dufumier et al. (2025) address a complementary limitation of pairwise multimodal contrastive learning: cross-modal alignment emphasizes redundant information and may omit unique or synergistic interactions. Their CoMM objective constructs a fused multimodal representation intended to retain all three information types. Our two-encoder setting instead asks which modality-specific directions affect a particular pairwise score. Retaining unique information for general multimodal tasks and making a fixed cross-modal comparison insensitive to irrelevant variation are therefore distinct goals.

Representation-learning dynamics have been studied through deep linear networks, self-supervised objectives, dimensional collapse, coordinate-wise optimization, and stepwise feature learning (Saxe et al., 2014; Arora et al., 2018; Ji & Telgarsky, 2019; Tian et al., 2021; Jing et al., 2022; Tian, 2022; Simon et al., 2023). Closest statistically, Ji et al. (2023) analyze linear feature recovery, and Nakada et al. (2023) relate multimodal contrastive learning to cross-covariance and singular-vector structure. Their gradient identity uses a contrastive cross-covariance determined by the current parameters; it does not by itself describe the long-time trajectory of the factors and learned temperature. Our trajectory analysis treats temperature boundaries and rank-deficient encoder limits, and establishes an open basin of nuisance-bearing global optima. Our bounded-temperature analysis obtains score-block decay without requiring the factors to converge to a global minimizer. The fixed-horizon empirical stability result in Appendix E.4 states its vector-field hypotheses separately.

Yaras et al. (2025) study temperature and modality separation through gradient flow for finite paired data with sample-normalized embeddings. Their parallel-embedding parameterization introduces an offset between modality centers, and their slow-gap-closure result assumes a positive matching margin. The present model tracks signal and nuisance blocks of linear encoders under population expected-norm InfoNCE. Since $\mathbb{E}X = \mathbb{E}Y = 0$, both encoder population means are zero at every time, even when their nuisance blocks and off-signal score interactions are nonzero. Center separation and score-block denoising therefore describe different properties of the learned representations.

Projection-head analyses characterize how feature weighting changes across layers and why representations before the projection can benefit downstream tasks (Xue et al., 2024). Here the objects are the two encoder outputs and the scalar comparison formed from them. Nuisance can remain in the representations used by the contrastive objective itself while its interactions vanish from their score. This distinguishes the geometry of a cross-modal comparison from feature selection across successive layers.

Noisy-correspondence methods modify learning when observed positive pairs may be incorrect. BiCro (Yang et al., 2023) estimates soft correspondence labels using bidirectional cross-modal similarity consistency. In Eq. (1), each pair shares the same latent z; nuisance and measurement noise enter these correctly paired observations. Our results describe the implicit dynamics of this objective, while correspondence rectification changes the training pairs' contribution to the objective.

Gui et al. (2025) study intrinsic-dimension adaptation assuming aligned representations, whereas independent positive-variance measurement noise in our model prevents exact nonzero alignment (Appendix E.1). Sun et al. (2025) study individual feature-purity dynamics. Our score-level question permits nonzero encoder nuisance at every finite time. Strict-saddle and matrix-factorization results provide the background for Appendix C (Ge et al., 2016; Lee et al., 2019; Stöger & Soltanolkotabi, 2021).