EXAMINING WHAT SELF-SUPERVISED MODELS LEARN FROM LONGITUDINAL 3D MEDICAL IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

In longitudinal medical studies, representations of three-dimensional (3D) images are expected to capture meaningful changes within a subject over time. However, these representations may encode noise factors or omit information about meaningful temporal changes. We introduce the `Longitudinal Representation Audit (LRA)` framework for assessing what information is gained through the training process. LRA compares each trained encoder with its matched initialization and, where possible, controls for measured noise factors to isolate the information of interest. We apply LRA across several self-supervised learning (SSL) models on longitudinal knee magnetic resonance imaging (MRI) data using two training budgets. The experiments reveal that features from a randomly initialized 3D backbone in one configuration already achieve 0.630 temporal-order prediction accuracy before SSL training. Across all five models, training produces no statistically supported improvement in this accuracy over matched initialization. Training nevertheless benefits phenotype prediction: Static SSL improves mean AUROC by 0.118 over initialization, despite showing no detectable improvement in temporal-order accuracy. Favorable representation geometry also does not guarantee improved temporal-order performance: after 100 epochs, IET achieves the highest participation ratio (33.76), yet its temporal-order accuracy is 0.109 below initialization. Together, these findings show that decodability, non-collapse, and favorable representation geometry alone cannot establish that longitudinal information was acquired through training. LRA therefore shifts evaluation from *what can be decoded?* to *what did training actually add?*

1 INTRODUCTION

In recent years, computer vision (CV) research has grown rapidly across a wide range of downstream tasks, particularly in medical imaging (disease or lesion detection, diagnosis, and segmentation). Among these tasks, longitudinal (follow-up) medical imaging is receiving increasing attention for assessing disease progression and treatment response using images acquired from the same patient across multiple visits (Rivail et al., 2019; Zhao et al., 2021; Foo et al., 2022). Many clinical questions are fundamentally questions of change: *what has progressed, what has remained stable, and whether treatment has been effective*. For example, the Osteoarthritis Initiative (OAI) provides repeated knee magnetic resonance imaging (MRI) scans acquired at various time points (Nevitt et al., 2006; Peterfy et al., 2008), allowing structural changes such as cartilage damage and bone marrow lesions to be tracked over time (Hunter et al., 2011). However, existing studies mainly evaluate longitudinal CV models using downstream metrics (e.g., accuracy or AUROC), and whether their representations capture meaningful temporal changes or incidental differences between visits remains a black box. This raises a central question: *when a model is evaluated on a longitudinal task, what information did the training process actually add to its representation?*

At the same time, obtaining high-quality annotations for longitudinal medical images requires expert knowledge and is time-consuming and costly (Huang et al., 2023), particularly when multiple scans are collected across visits. Self-supervised learning (SSL) is increasingly used to learn useful representations directly from images while reducing reliance on detailed annotations (Huang et al., 2023). Several studies have reported successful applications of SSL across different types of medical images, including X-ray (Tiu et al., 2022), histopathology (Ghesu et al., 2022), MRI and computed

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.

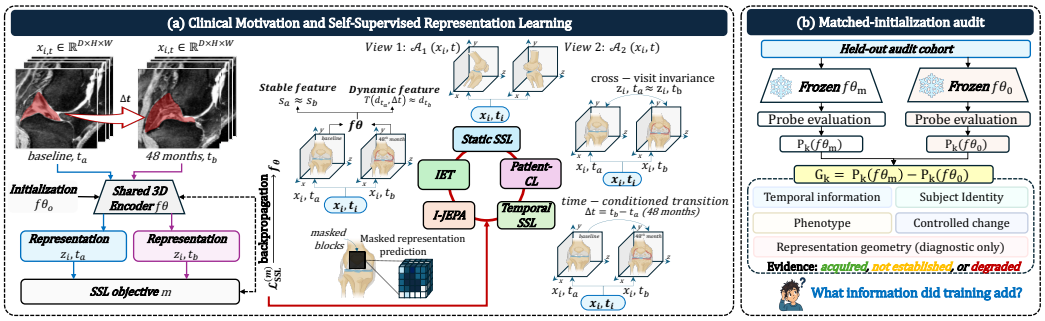

Figure 1: **Overview of LRA**. (a) Longitudinal representation learning under five self-supervised objectives, spanning static, temporal, and controlled-change learning signals. (b) Matched-initialization audit across four information axes, with representation geometry assessed separately.

tomography (CT) (Li et al., 2026). These results have encouraged the use of SSL in longitudinal imaging, where relationships between scans acquired from the same patient across multiple visits can be used to learn information that remains stable or changes over time (Rivail et al., 2019; Zhao et al., 2021). However, the relationships among these scans can reflect multiple sources of variation, including changes related to disease progression or treatment and batch effects caused by differences in scanners, acquisition protocols, patient positioning, field of view and image intensity (Hu et al., 2023). These factors make it difficult to determine which changes are clinically meaningful and which are caused by image acquisition. A probe score, in other words, does not say where the information came from: the training objective, the architecture, or the scanner.

To address the concerns raised above, we introduce the *Longitudinal Representation Audit* (LRA), a framework for examining what information SSL training adds to representations learned from longitudinal three-dimensional (3D) medical images (Figure 1). In this study, we apply LRA to longitudinal 3D knee MRI from the OAI (Nevitt et al., 2006; Peterfy et al., 2008). Using a common 3D encoder architecture, we train separate encoders for 10 and 100 epochs under five SSL objectives: Static Self-Supervised Learning (Static SSL), Patient-Level Contrastive Learning (Patient-CL), Temporal Self-Supervised Learning (Temporal SSL), Image-based Joint-Embedding Predictive Architecture (I-JEPA)(Assran et al., 2023), and Invariant–Equivariant Temporal Representation Learning (IET). For each objective, LRA freezes the trained encoder and its matched initialization and evaluates both under the same protocol. We perform this comparison across temporal order, knee identity, phenotype, and controlled change, with controls for measured nuisance factors where possible. The paired difference in probe performance measures what training added, while a reference built from low-level image statistics tests whether the learned representation offers anything beyond non-learned image cues. To the best of our knowledge, this study is among the first to systematically examine longitudinal SSL representations through comparisons with their matched initializations across multiple information axes. Before SSL training, a randomly initialized 3D backbone already achieves a temporal-order accuracy of 0.630. A fixed set of 32 low-level image statistics, without representation learning, similarly reaches 0.636. After training, none of the objectives shows a statistically supported improvement in temporal-order accuracy over its matched initialization, while Static SSL improves mean phenotype AUROC by 0.118. These results show that information decodable after training is not necessarily information acquired through training.

2 Related Work

2.1 Longitudinal Learning in Medical Imaging

Capturing temporal changes through visual learning from longitudinal imaging data is a significant research objective in medical computer vision, which is essential for tracking lesion evolution over time or assessing treatment efficacy. Approaches to this problem have used longitudinal imaging data to identify subtle changes that may support early disease detection (Sabanayagam et al., 2020; Rivail et al., 2019; Zhao et al., 2021). Image-based progression models further exploit repeated observations to capture changes across visits, for example by modeling fine-grained biological regions and their relationships over time for disease progression prediction (Foo et al., 2022). Related invariant–equivariant approaches separate stable from transformation-sensitive features (Dangovski et al., 2022; Garrido et al., 2023b), while predictive methods such as I-JEPA learn structure by pre-

dicting target representations from context (Assran et al., 2023). Despite their different objectives, these approaches are typically evaluated primarily using diagnostic labels, making it difficult to tell how much of the observed structure was learned, already present at initialization, or associated with acquisition variation. Our LRA focuses on this evaluation problem rather than introducing another longitudinal learning objective.

2.2 Self-Supervised Representation Evaluation

Effective supervised learning in medical imaging relies on high-quality annotations, but obtaining expert labels is costly and time-consuming. In recent years, SSL has gained increasing attention for its ability to learn useful representations from unannotated images, reducing reliance on limited expert annotations (Chen et al., 2020; Bardes et al., 2022; Assran et al., 2023). Representations of SSL are typically evaluated through downstream tasks, but longitudinal data raise a different question: whether training actually captures within-subject change over time. Existing tools only partly address this. Linear probes measure what frozen features make decodable (Alain & Bengio, 2016), geometric diagnostics characterize rank and collapse (Jing et al., 2022; Garrido et al., 2023a; Agrawal et al., 2022), and CKA measures representation similarity (Kornblith et al., 2019). Probe outcomes also depend on the task and probe itself (Hewitt & Liang, 2019; Belinkov, 2022). These evaluations characterize accessibility or geometry, but do not establish whether the information was acquired during training, especially when temporal structure is already decodable at initialization.

2.3 From Model Interpretation to Training-Acquired Information

Interpretability in medical imaging has largely focused on explaining predictions, using tools such as saliency and class-activation maps to identify image regions that influence model outputs. These methods can show whether a prediction relies on plausible anatomy, but they do not reveal what information training added to the representation. This distinction matters because learned features may encode anatomy, texture, or acquisition characteristics, and medical-imaging models can exploit hospital, scanner, and other acquisition signatures as shortcuts (Zech et al., 2018; DeGrave et al., 2021; Geirhos et al., 2020). Moreover, useful image structure may already be present before training: randomly weighted networks can provide non-trivial visual features, while untrained architectures impose strong image priors (Jarrett et al., 2009; Saxe et al., 2011; Ulyanov et al., 2020). Motivated by these observations, we treat representation analysis as an attribution problem. We compare each trained encoder with its matched initialization, examine multiple information axes, and control measured nuisance structure where possible. This shifts the question from *what can be decoded?* to *what information can actually be attributed to training?*

3 Problem Formulation

3.1 Longitudinal Representation Learning Setting

We observe subjects $i \in \mathcal{I}$ at visits $t \in \mathcal{T}_i$, with scans $\mathbf{x}_{i,t} \in \mathcal{X}$, measured nuisance descriptors $\mathbf{c}_{i,t}$, and, where available, annotations $\mathbf{y}_{i,t}$. An encoder $f_\theta : \mathcal{X} \to \mathbb{R}^p$ maps each scan to $\mathbf{z}_{i,t} = f_\theta(\mathbf{x}_{i,t})$, with each representation branch audited separately. Training by an algorithm $\mathcal{A}$ starts from an architecture- and seed-matched initialization θ_0 and yields $\theta = \mathcal{A}(\theta_0, \mathcal{D}_{\text{train}})$. For each information axis $k \in \mathcal{K}$, such as temporal order, subject identity, phenotype, or controlled change, $P_k(f)$ denotes held-out probe performance under a fixed evaluation protocol, not an intrinsic information-theoretic quantity. Within each axis, encoders share the same probe family, hyperparameter search, and patient-disjoint partition, so differences in P_k reflect representation differences under a common evaluation procedure.

3.2 Decodable, Learned, and Useful Information

A high probe score can mean several things, so we ask three separate questions. Can the information be decoded? Did training add it? And does what training added survive the relevant controls? These questions are routinely conflated, and we define each in turn.

Decodable. Information k is decodable from f_θ if $P_k(f_\theta)$ exceeds the chance level π_k of the axis with a confidence interval excluding it. Decodability answers whether information is accessible, but it is silent on where the information came from.

Learned. Information k is learned if training increased its decodability relative to the encoder that training started from. The natural reference is the matched initialization f_{θ_0}, which shares architecture, seed, input pipeline, and probe with f_θ and differs only in whether $\mathcal{A}$ was applied:

$$G_k(f_\theta) \;=\; P_k(f_\theta) - P_k(f_{\theta_0}). \tag{1}$$

G_k answers what training added. It is the central quantity of this paper. Throughout this work, learned information means an increase in probe-decodable information relative to matched initialization; Appendix D.1 explains how far this measure can be interpreted. An encoder with $P_k(f_\theta) = 0.70$ whose matched initialization achieves $P_k(f_{\theta_0}) = 0.72$ has not learned k, however high 0.70 may look in isolation.

Controlled and beyond reference. Information k is controlled-learned if the gain survives the removal of measured nuisance structure, and useful if it is relevant beyond a low-level reference that requires no representation learning at all. Let R_c denote the axis-specific residualization operator that removes from the probe input of that axis the component linearly predictable from the measured nuisance covariates, with coefficients estimated on the training partition only (Section 4.4); we suppress the axis index when unambiguous. The nuisance-controlled gain is

$$G_k^{\perp}(f_\theta) \;=\; P_k\big(R_c[f_\theta]\big) - P_k\big(R_c[f_{\theta_0}]\big). \tag{2}$$

Let $\mathbf{b} : \mathcal{X} \to \mathbb{R}^m$ be a fixed low-level reference, in our case a battery of $m = 32$ intensity, texture, and geometry statistics computed directly from the image. The reference margin is

$$U_k(f_\theta) \;=\; P_k(f_\theta) - P_k(\mathbf{b}), \qquad U_k^{\perp}(f_\theta) \;=\; P_k\big(R_c[f_\theta]\big) - P_k\big(R_c[\mathbf{b}]\big). \tag{3}$$

These quantities answer different questions, and none can be read off another without the matching control:

$$\text{Decodable} \not\Rightarrow \text{Learned}, \qquad \text{Learned} \not\Rightarrow \text{Beyond-reference utility}. \tag{4}$$

The first non-implication can hold whenever $P_k(f_{\theta_0}) > \pi_k$; in our benchmark, this occurs for temporal order. The second can hold whenever the learned increment is carried by nuisance structure or is already available to the low-level reference.

3.3 Control-Conditioned Information Gain

An exact decomposition of absolute performance. Writing $P_k^{\perp}(f) = P_k(R_c[f])$, the absolute decodability of a trained encoder decomposes identically into four terms:

$$\begin{aligned} P_k(f_\theta) \;=\; & \underbrace{P_k^{\perp}(f_{\theta_0})}_{\text{controlled prior}} \;+\; \underbrace{\big[P_k(f_{\theta_0}) - P_k^{\perp}(f_{\theta_0})\big]}_{\text{prior change under nuisance control}} \\ & +\; \underbrace{G_k^{\perp}(f_\theta)}_{\text{controlled gain}} \;+\; \underbrace{\big[G_k(f_\theta) - G_k^{\perp}(f_\theta)\big]}_{\text{gain change under nuisance control}} . \end{aligned} \tag{5}$$

The identity holds by cancellation: the first two terms sum to $P_k(f_{\theta_0})$ and the last two to $G_k(f_\theta)$, whose sum is $P_k(f_\theta)$ by Equation 1. Equation 5 makes the ambiguity of Section 1 explicit: $P_k(f_\theta)$ combines four terms with distinct scientific meanings, while the controlled gain is the quantity relevant to training-acquired information that survives measured nuisance control. Standard evaluation reports only the total; LRA separates its components.

Absolute performance does not reveal the gain. A high final score does not imply that training improved the representation, since the same $P_k(f_\theta)$ may represent a gain or loss relative to initialization. We therefore measure the gain G_k of Equation 1, whose sign indicates whether training increased or reduced decodable information on axis k. Likewise, a favorable label-free diagnostic $Q(f_\theta)$ does not establish acquisition unless it tracks G_k across encoders (Section 5.3).

Estimation. We estimate gains using a paired, patient-clustered bootstrap with B resamples of the held-out patients, yielding bootstrap gains $G_k^{(\beta)}$, $\beta = 1, \ldots, B$. We report the bootstrap mean

$\bar{G}_k = \frac{1}{B}\sum_{\beta=1}^{B} G_k^{(\beta)}$, the empirical 95% percentile interval, and the two-sided bootstrap p-value

$$p_k = \min\Big\{1,\ \frac{2}{B}\min\Big(\sum_{\beta=1}^{B}\mathbf{1}\big[G_k^{(\beta)} \le 0\big],\ \sum_{\beta=1}^{B}\mathbf{1}\big[G_k^{(\beta)} \ge 0\big]\Big)\Big\}, \tag{6}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Pairing directly quantifies uncertainty in the gain. Where several contrasts form a family (branches or labels), p-values are Holm-corrected (Holm, 1979), and the adjusted value is denoted $\tilde{p}_k$ (Appendix D.6).

Representation quality is a profile, not a scalar. We summarize gains as $\mathbf{G}(f) = (G_k(f))_{k\in\mathcal{K}}$. Encoder f_A dominates f_B only if $G_k(f_A) \ge G_k(f_B)$ on every axis; otherwise, different axes may favour different encoders. A single ranking would therefore require an external preference rule, such as $\mathbf{w}^\top\mathbf{G}$. LRA instead retains the full profile (Section 5.3; Appendix E.4).

4 Method

4.1 Longitudinal Representation Audit

The Longitudinal Representation Audit (LRA) measures what information is added through training by comparing a trained encoder with its matched initialization. For each information axis, the two encoders are evaluated using the same preprocessing, data split, probe family, and hyperparameter search, so differences in performance can be attributed to training under the evaluation protocol. These differences define the training gain G_k and, where nuisance measurements or a low-level reference are available, the controlled quantities $G_k^{\perp}$ and $U_k^{\perp}$. LRA summarizes these measurements in the LIP, with uncertainty estimated using shared patient-clustered bootstrap resamples. Representation geometry, such as participation ratio, is reported separately as a diagnostic and does not determine the evidence verdicts. The complete procedure is given in Figure 1 and Algorithm 1 (Appendix D.2).

4.2 Matched-Initialization Control

The matched initialization f_{θ_0} uses the same architecture, configuration, and random seed as the trained encoder, recovering the state from which training began. Both encoders share the same augmentation-free input pipeline and probe family, with hyperparameters tuned under the same procedure. Thus, G_k measures the change associated with training while holding architecture, preprocessing, and probe capacity fixed. Initialization and normalization checks are reported in Appendix F.2.

4.3 Proposed Multi-Axis Representation Assessment

Longitudinal representations may encode different aspects of the data, so a single measure cannot fully characterize what training has added. We therefore evaluate four complementary axes: temporal order, patient identity, anatomical phenotype, and sensitivity to anatomical versus acquisition-related change. Each axis is evaluated separately using a corresponding probe, with full definitions provided in Appendix D.

Temporal order. We ask whether a representation distinguishes the chronological order of two visits of the same knee. Let $\tilde{\mathbf{z}}$ denote the representation standardized with training-partition statistics. For each ordered pair of visits $a < b$ of the same knee, an ℓ_2-regularised logistic probe receives the antisymmetric feature $\pm(\tilde{\mathbf{z}}_b - \tilde{\mathbf{z}}_a)/\sqrt{p}$, labelled 1 for the $+$ sign and 0 for the $-$ sign, and predicts which visit is later; $A_{\mathrm{ord}}(f)$ denotes its held-out accuracy. The same probe family is fitted to the trained and to the matched-initialization representation, giving the temporal gain

$$G_{\mathrm{temp}} = A_{\mathrm{ord}}(f_\theta) - A_{\mathrm{ord}}(f_{\theta_0}), \tag{7}$$

so that a positive gain indicates temporal information attributable to training rather than already exposed by the architecture. Pair construction, probe fitting, regularisation selection, and the estimator of A_{ord} (Equation 23) are given in Appendix D.4.

Patient identity. We measure identity retention by leave-one-out cosine retrieval, where scans from the same knee are treated as relevant. We report Recall@1 (R@1) and mean average precision

(mAP), together with the change in R@1 from initialization, $\Delta\text{R@1} = \text{R@1}(f_\theta) - \text{R@1}(f_{\theta_0})$ (Appendix D.8).

Phenotype. We probe 12 binary MOAKS labels (Hunter et al., 2011), spanning bone marrow lesions, cartilage damage, osteophytes, meniscal abnormality, and effusion-synovitis, using standardized logistic readouts on patient-disjoint splits. With $\text{AUC}_\ell(f)$ the held-out AUROC of the probe for label ℓ, the phenotype gain is the mean paired AUROC difference over the twelve labels,

$$G_{\text{phen}}(f_\theta) = \frac{1}{12}\sum_{\ell=1}^{12}\left[\text{AUC}_\ell(f_\theta) - \text{AUC}_\ell(f_{\theta_0})\right]. \tag{8}$$

Probe details and Holm correction across labels are given in Appendix D.9.

Controlled change. We compare representation sensitivity to a smooth, localized anatomical deformation τ_A and a nuisance perturbation τ_N matched for in-support RMS intensity change. Let $\cos(\mathbf{u},\mathbf{v}) = \mathbf{u}^\top\mathbf{v}/(\|\mathbf{u}\|_2\|\mathbf{v}\|_2)$ denote cosine similarity and $D_f(\mathbf{x},\tau) = 1 - \cos\big(f(\mathbf{x}), f(\tau(\mathbf{x}))\big)$ the representation distance that a perturbation τ induces under encoder f. Selectivity is the ratio of the mean distances over the N evaluation scans,

$$S(f) = \frac{\sum_{i=1}^N D_f(\mathbf{x}_i,\tau_A)}{\sum_{i=1}^N D_f(\mathbf{x}_i,\tau_N)}, \tag{9}$$

so that $S(f) > 1$ indicates greater sensitivity to anatomical than to matched nuisance change (Appendices D.1 and D.11; Figure 3).

Geometric descriptors. We characterize representation geometry using participation ratio (PR), effective rank r_{eff}, and mean pairwise cosine similarity (cos). These are diagnostic descriptors, not information axes, and are compared with the measured training gains. Full definitions are given in Appendix D.10.

4.4 Nuisance-Controlled Auditing

A measured gain may still reflect nuisance structure rather than the information of interest. We therefore control for measured nuisance factors separately for each axis, with all control models fitted only on training data.

Image-derived descriptors for temporal order. Visits of the same knee may differ in low-level image properties because of coil placement, operator, or scanner software. We capture these differences using 32 statistics describing intensity, gradient, edge, frequency, and support properties (Table 7; Appendix D.7). For a visit pair (a,b), let $\Delta\tilde{\mathbf{z}}_{ab} = \tilde{\mathbf{z}}_b - \tilde{\mathbf{z}}_a$ and $\Delta\mathbf{c}_{ab} = \tilde{\mathbf{c}}_b - \tilde{\mathbf{c}}_a$ denote the within-pair differences of the standardized representation and of the standardized descriptors (descriptors that are constant within a knee vanish in this difference). The descriptor differences are regressed out of the representation differences by a multi-output ridge regression fitted on the training pairs $\mathcal{P}_{\text{tr}}$ only:

$$R_c[\Delta\tilde{\mathbf{z}}_{ab}] = \Delta\tilde{\mathbf{z}}_{ab} - \hat{\mathbf{\Gamma}}\,\Delta\mathbf{c}_{ab}, \qquad \hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \sum_{(a,b)\in\mathcal{P}_{\text{tr}}} \left\|\Delta\tilde{\mathbf{z}}_{ab} - \mathbf{\Gamma}\,\Delta\mathbf{c}_{ab}\right\|_2^2 + \alpha\|\mathbf{\Gamma}\|_F^2, \tag{10}$$

where $\alpha > 0$ is the ridge penalty selected by cross-validation. The temporal probe is then re-evaluated on the residualized differences to obtain $G^{\perp}_{\text{temp}}$. Descriptor filtering, regularization selection, and residualization controls are detailed in Appendix D.5.

Acquisition control and low-level reference. Acquisition metadata alone predicts the MOAKS labels with a mean AUROC of 0.578, indicating that acquisition factors may contribute to phenotype prediction. We therefore residualize the learned representation, its matched initialization, and the 32-statistic reference against the same acquisition design matrix before probing. The residualization model is fitted only on the training split. The 32-statistic battery then serves as the non-learned reference $\mathbf{b}$ in Equation 3, allowing us to test whether the learned representation provides predictive information beyond measured low-level image statistics. Failure to exceed this reference does not imply that the two representations encode the same information, which we examine in Section 5.4. Full details of the acquisition control are provided in Appendix D.9.

4.5 Longitudinal Information Profile

We summarize these measurements in the Longitudinal Information Profile,

$$\mathrm{LIP}(f_\theta) = \Big(\underbrace{G_{\mathrm{temp}},\, G^{\perp}_{\mathrm{temp}}}_{\text{temporal}},\ \underbrace{\Delta\mathrm{R@1}}_{\text{identity}},\ \underbrace{G_{\mathrm{phen}},\, U^{\perp}_{\mathrm{phen}}}_{\text{phenotype}},\ \underbrace{S(f_\theta)}_{\text{controlled change}} \ \Big|\ \underbrace{\mathrm{PR},\, r_{\mathrm{eff}},\, \cos}_{\text{descriptors}} \Big), \tag{11}$$

where the entries to the left of the bar are the audited quantities (contrasts against initialization or reference, and the selectivity ratio), each reported with the absolute scores of both encoders and a bootstrap interval, and the entries to the right of the bar are label-free geometric descriptors reported for diagnosis only (notation in Appendix A). The profile is deliberately not collapsed into a scalar. As shown in Section 3.3, a ranking of encoders requires either dominance on every axis or an external weighting, and the weighting would encode a scientific priority, such as whether temporal ordering matters more than phenotype, that the evaluation itself cannot supply. Table 1 reports the profile of all ten encoders in our benchmark.

From profile to verdicts. Each inferential profile entry is mapped to a common verdict: gains with patient-clustered intervals above zero and Holm-adjusted $p < 0.05$ are ACQUIRED, those with intervals below zero are DEGRADED, and the rest are NOT ESTABLISHED, indicating insufficient evidence rather than zero gain. Controlled change is judged against $S = 1$, while identity and geometric descriptors remain descriptive. These rules were codified after the analyses as a transparent summary, not as pre-registered tests; full criteria are given in Appendix D.3, with pre-registered analyses reported in Table 8 and Appendix D.15.

5 Experiments and Results

5.1 Experimental Setup

Data and evaluation. We use longitudinal sagittal 3D knee MRI from the OAI (Nevitt et al., 2006; Peterfy et al., 2008), acquired at baseline and 12, 24, and 48 months. To reduce geometric and padding cues, scans are brought to a common orientation and resolution, rigidly registered across visits of the same knee, restricted to their common support, and intensity-normalized. We train the encoders on 464 patients (1,268 scans; 1,167 within-knee pairs) and evaluate them on a separate cohort of 690 patients (768 knees; 2,036 scans). A further cohort of 690 patients is kept sealed and is not used in this study. Cohort construction and preprocessing are detailed in Appendix B and Figures 4 and 5.

Objectives. We evaluate five self-supervised objectives that differ in how they use longitudinal information. *Static SSL* applies VICReg (Bardes et al., 2022) to two views of the same scan, while *Patient-CL* applies it across visits of the same knee. *Temporal SSL* additionally models time-conditioned transitions and predicts visit order. *JEPA* predicts masked target representations from visible context using an exponential-moving-average target encoder (Assran et al., 2023). Finally, *IET* separates stable and dynamic latents, combining cross-visit invariance and time-equivariance with order prediction and anti-collapse regularization. IET serves as a longitudinal stress test rather than a proposed method, allowing us to examine whether explicit temporal structure and favorable representation geometry correspond to longitudinal information gained through training (Appendix F). Full objectives and loss weights are provided in Appendix C.2.

Models and representations. All objectives use the same 3D ViT-S backbone and training protocol, with separate 10- and 100-epoch runs. Our main analysis uses the shared trunk embedding $\mathbf{h} \in \mathbb{R}^{384}$, which is optimized by every objective. Results for the projected latents, including IET's dynamic latent $\mathbf{d}$, are reported in Table 10 and Appendix E.2; full architecture and training details are provided in Appendix C.3.

5.2 Temporal Decodability Does Not Imply Temporal Learning

Temporal order is already strongly decodable before training. The untrained trunk achieves 0.630 accuracy $[0.597, 0.662]$, comparable to the non-learned 32-statistic battery at 0.636 $[0.603, 0.668]$; the battery further improves from 0.599 at 12 months to 0.697 at 48 months (Figure 2a,d). Despite this strong baseline, none of the five SSL objectives shows a statistically supported positive

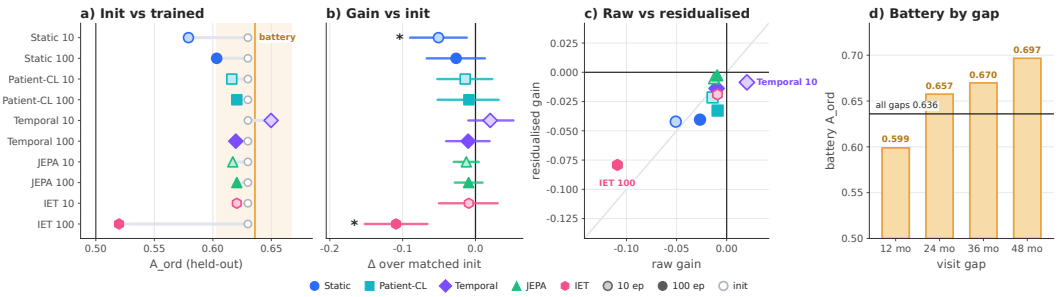

Figure 2: **Temporal-order results for the trunk embedding.** (a) Accuracy before and after training. (b) Gain over matched initialization with 95% patient-clustered intervals. (c) Raw and residualised gains. (d) 32-statistic reference across visit gaps. No encoder shows a significant positive temporal gain.

Table 1: LIP for the ten benchmark encoders using the trunk representation $\mathbf{h}$. G_{temp} and G_{phen} measure gains over matched initialization, $U^{\perp}_{\text{phen}}$ measures the acquisition-controlled margin over the 32-statistic reference, S measures controlled-change selectivity, and PR describes representation rank. Brackets show 95% patient-clustered bootstrap intervals; N.E. denotes NOT ESTABLISHED.

Encoder	G_{temp} [95% CI]	Verdict	G_{phen} (Holm)	$U^{\perp}_{\text{phen}}$ [95% CI]	S [95% CI]	PR
Static SSL 10	-0.051 $[-0.090, -0.012]$	DEG.	$+0.072$ (2/12)	-0.035 $[-0.064, -0.005]$	0.23 $[0.22, 0.25]$	1.49
Static SSL 100	-0.027 $[-0.067, +0.013]$	N.E.	$+0.118$ (6/12)	$+0.016$ $[-0.012, +0.043]$	2.33 $[2.11, 2.56]$	1.32
Patient-CL 10	-0.014 $[-0.053, +0.023]$	N.E.	$+0.068$ (3/12)	-0.042 $[-0.070, -0.014]$	0.31 $[0.29, 0.34]$	1.29
Patient-CL 100	-0.009 $[-0.052, +0.032]$	N.E.	$+0.102$ (4/12)	-0.015 $[-0.043, +0.014]$	2.26 $[2.08, 2.46]$	1.35
Temporal SSL 10	$+0.020$ $[-0.010, +0.052]$	N.E.	$+0.048$ (1/12)	-0.055 $[-0.084, -0.024]$	0.06 $[0.06, 0.07]$	5.13
Temporal SSL 100	-0.010 $[-0.040, +0.019]$	N.E.	$+0.032$ (0/12)	-0.078 $[-0.109, -0.046]$	0.12 $[0.11, 0.13]$	4.72
JEPA 10	-0.013 $[-0.030, +0.004]$	N.E.	-0.009 (0/12)	-0.113 $[-0.152, -0.073]$	0.08 $[0.04, 0.21]$	2.04
JEPA 100	-0.009 $[-0.028, +0.010]$	N.E.	-0.022 (0/12)	-0.121 $[-0.159, -0.082]$	0.11 $[0.07, 0.19]$	2.57
IET 10	-0.009 $[-0.050, +0.031]$	N.E.	-0.018 (0/12)	-0.123 $[-0.161, -0.088]$	1.17 $[1.11, 1.24]$	9.67
IET 100	-0.109 $[-0.152, -0.066]$	DEG.	$+0.010$ (0/12)	-0.096 $[-0.126, -0.065]$	0.67 $[0.65, 0.69]$	33.76

temporal gain over its matched initialization (Table 1; Figure 2b). Eight of ten trunk gains are indistinguishable from zero, while Static SSL at 10 epochs and IET at 100 epochs significantly reduce temporal accuracy by 0.051 and 0.109, respectively. Temporal SSL provides the clearest example of why absolute decodability can be misleading: it reaches the highest 10-epoch accuracy (0.650), yet its gain over initialization is only $+0.020$ $[-0.010, +0.052]$. Even IET's explicitly temporal dynamic latent $\mathbf{d}$ shows no positive gain at either training budget (Table 10). This conclusion remains unchanged after residualizing measured low-level image statistics and across three 10-epoch seeds (Appendix E.10). Thus, the strongest temporal result in this benchmark is not high decoding accuracy itself, but the gap between what is decodable and what can be attributed to training.

5.3 Representation Geometry Does Not Predict Acquired Information

Representation geometry is not a reliable proxy for information acquired through training. At 100 epochs, IET achieves the highest participation ratio (33.76 versus 5.18 at initialization) but also the largest temporal loss (-0.109), whereas Static SSL remains strongly low-rank (PR $=$ 1.32) yet achieves the largest phenotype gain ($+0.118$, six significant labels; Table 1). Across all ten encoders, participation ratio is associated with same-knee retrieval ($\rho = +0.67$, $p = 0.033$) and negatively with phenotype gain ($\rho = -0.66$, $p = 0.038$), but not with temporal gain ($\rho = +0.16$, $p = 0.65$; Figure 10), with other geometric measures showing similar limitations (Appendix E.3). Training budget is similarly axis dependent: for IET, extending training from 10 to 100 epochs reduces temporal gain from -0.009 to -0.109 while phenotype gain increases from -0.018 to $+0.010$. More broadly, 44 of 45 encoder pairs are incomparable by point-estimate dominance across temporal, identity, and phenotype gains, showing that neither favorable geometry nor longer training implies uniformly better representations across information axes; because the 100-epoch models use a single seed, these longer-training patterns are treated as benchmark observations rather than general effects.

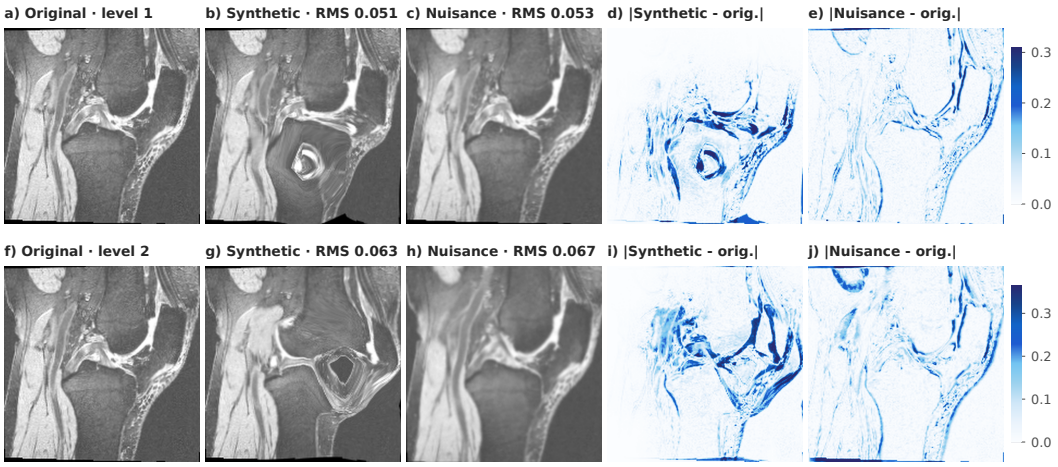

Figure 3: **Controlled-change perturbations at two matched levels:** original MRI, synthetic local deformation τ_A, matched nuisance perturbation τ_N, and absolute difference maps. The deformation probes representation sensitivity, not real disease progression. Selectivity results are reported in Appendix G.1 (Table 19).

5.4 Phenotype Information Is Compact and Label-Family Dependent

Phenotype and controlled-change analyses reveal different aspects of acquired information. After acquisition control, the non-learned 32-statistic battery remains a strong phenotype reference (mean AUROC $= 0.626$). Static SSL at 100 epochs achieves the largest phenotype gain over initialization ($+0.118$, six of twelve labels Holm-significant), yet its margin over the battery is only $+0.016$ $[-0.012, +0.043]$ ($p = 0.27$). Thus, training adds phenotype-predictive information, but performance beyond the low-level reference is not statistically established. The two representations also favor different label families: learned features perform better on several meniscal, effusion-synovitis, bone-marrow-lesion, and patellar-cartilage labels, whereas the battery performs better on osteophytes and femoral and tibial cartilage. Controlled-change selectivity provides a distinct signal: at 10 epochs, IET is the only encoder selective for anatomical deformation over matched nuisance perturbation ($S = 1.17$ $[1.11, 1.24]$) despite no established phenotype gain, whereas at 100 epochs Static SSL and Patient-CL show the strongest selectivity ($S = 2.33$ and 2.26) alongside the largest phenotype gains. Controlled-change sensitivity therefore captures a useful property of the representation, but does not serve as a direct proxy for temporal or phenotype information acquired through training (Figures 13 and 3; Appendices E.8, E.5, G.1, G.4, and G.2).

6 Conclusion

We introduced the Longitudinal Representation Audit (LRA) to measure what information self-supervised training adds to longitudinal representations by comparing trained encoders with matched initialization across multiple information axes. Across five SSL objectives on longitudinal knee MRI, temporal order was already decodable before training, and none showed a statistically supported positive temporal gain. Training nevertheless improved phenotype information in some models, while representation geometry did not consistently reflect these gains. These findings show that longitudinal representations should be evaluated by what training adds, not only by what the final representation makes decodable.

7 Limitations

Our study is limited to one longitudinal knee MRI cohort, one imaging sequence, and one backbone family. The 10-epoch experiments were repeated across three seeds, whereas the 100-epoch experiments used a single seed. Nuisance control is limited to measured factors, and controlled change uses synthetic perturbations rather than observed disease progression. Future work should evaluate LRA across broader datasets, modalities, architectures, and longitudinal tasks.

ETHICS STATEMENT

This work does not involve human participants or personally identifiable information. We are not aware of any significant ethical concerns associated with this study.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the methodology, experimental setup, implementation details, and evaluation procedures to facilitate reproducibility. Additional details are provided in the supplementary material.

AI USE STATEMENT

Generative AI tools were used solely for grammar checking and language editing. All scientific content, analyses, and conclusions were developed and verified by the authors.

REFERENCES

Kumar Krishna Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17626–17638, 2022.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

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

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*, 2022.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. doi: 10.1038/s42256-021-00338-7.

Alex Foo, Wynne Hsu, Mong Li Lee, and Gavin S. W. Tan. DP-GAT: A framework for image-based disease progression prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2903–2912. Association for Computing Machinery, 2022. doi: 10.1145/3534678.3539113.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 2023a.

Quentin Garrido, Laurent Najman, and Yann LeCun. Self-supervised learning of split invariant equivariant representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10975–10996. PMLR, 2023b.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Florin C. Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Dominik Neumann, Pujan Patel, R. S. Vishwanath, James M. Balter, Yue Cao, Sasa Grbic, and Dorin Comaniciu. Contrastive self-supervised learning from 100 million medical images with optional supervision. *Journal of Medical Imaging*, 9(6):064503, 2022. doi: 10.1117/1.JMI.9.6.064503.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Fengling Hu, Andrew A. Chen, Hannah Horng, Vishnu Bashyam, Christos Davatzikos, Aaron Alexander-Bloch, Mingyao Li, Haochang Shou, Theodore D. Satterthwaite, Meichen Yu, and Russell T. Shinohara. Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *NeuroImage*, 274: 120125, 2023. doi: 10.1016/j.neuroimage.2023.120125.

Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *npj Digital Medicine*, 6(1):74, 2023. doi: 10.1038/s41746-023-00811-0.

David J. Hunter, Ali Guermazi, Grace H. Lo, Andrew J. Grainger, Philip G. Conaghan, Robert M. Boudreau, and Frank W. Roemer. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI osteoarthritis knee score). *Osteoarthritis and Cartilage*, 19(8):990–1002, 2011. doi: 10.1016/j.joca.2011.05.004.

Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2146–2153, 2009. doi: 10.1109/ICCV.2009.5459469.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Yuheng Li, Yuan Gao, Haoyu Dong, Yuxiang Lai, Shansong Wang, Mojtaba Safari, James E. Baciak, and Xiaofeng Yang. Universal CT representations from anatomy to disease phenotype through agglomerative pretraining. *arXiv preprint arXiv:2605.21906*, 2026.

Michael C. Nevitt, David T. Felson, and Gayle Lester. The Osteoarthritis Initiative: Protocol for the cohort study. Osteoarthritis Initiative Study Protocol, Version 1.1, 2006. URL https://nda.nih.gov/static/docs/StudyDesignProtocolAndAppendices.pdf.

C. G. Peterfy, E. Schneider, and M. Nevitt. The Osteoarthritis Initiative: Report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and Cartilage*, 16(12):1433–1441, 2008. doi: 10.1016/j.joca.2008.06.016.

Antoine Rivail, Ursula Schmidt-Erfurth, Wolf-Dieter Vogl, Sebastian M. Waldstein, Sophie Riedl, Christoph Grechenig, Zhichao Wu, and Hrvoje Bogunović. Modeling disease progression in retinal OCTs with longitudinal self-supervised learning. In *Predictive Intelligence in Medicine*, volume 11843 of *Lecture Notes in Computer Science*, pp. 44–52. Springer, 2019. doi: 10.1007/978-3-030-32281-6_5.

Charumathi Sabanayagam, Dejiang Xu, Daniel S. W. Ting, Simon Nusinovici, Riswana Banu, Haslina Hamzah, Cynthia Lim, Yih-Chung Tham, Carol Y. Cheung, E. Shyong Tai, Ya Xing Wang, Jost B. Jonas, Ching-Yu Cheng, Mong Li Lee, Wynne Hsu, and Tien Y. Wong. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *The Lancet Digital Health*, 2(6):e295–e302, 2020. doi: 10.1016/S2589-7500(20)30063-7.

Andrew M. Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y. Ng. On random weights and unsupervised feature learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1089–1096, 2011.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest X-Ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022. doi: 10.1038/s41551-022-00936-9.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *International Journal of Computer Vision*, 128(7):1867–1888, 2020. doi: 10.1007/s11263-020-01303-4.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

Qingyu Zhao, Zixuan Liu, Ehsan Adeli, and Kilian M. Pohl. Longitudinal self-supervised learning. *Medical Image Analysis*, 71:102051, 2021. doi: 10.1016/j.media.2021.102051.

APPENDIX

A NOTATION

Table 2: Summary of notation, grouped by where each symbol is introduced.

symbol	meaning
Setup and encoders (§3.1)	
$\mathbf{x}_{i,t}$	scan of subject i at visit t
$\mathbf{c}_{i,t}$	nuisance descriptors of the scan
$\mathbf{y}_{i,t}$	annotation where available
$f_\theta : \mathcal{X} \to \mathbb{R}^p$	encoder; $\mathbf{z} = f_\theta(\mathbf{x})$ its representation
$\mathbf{h}$, $\mathbf{s}$, $\mathbf{d}$	trunk embedding; projected (stable) latent; dynamic latent (§C.1)
$\tilde{\mathbf{z}}$	representation standardized on training
θ_0, f_{θ_0}	matched initialization and its encoder
$\mathcal{A}$, $\mathcal{D}_{\text{train}}$	training algorithm; training data
$\mathcal{K}$, k	set of information axes; axis index
$P_k(f)$	probe performance on axis k
π_k	chance level of axis k
Gains, nuisance control, reference margin (§3.2)	
$G_k(f_\theta)$	training gain $P_k(f_\theta) - P_k(f_{\theta_0})$ (Equation 1)
R_c	nuisance residualization operator (Equation 10, Equation 27)
$\hat{\mathbf{\Gamma}}$, α	nuisance regression coefficients; ridge penalty
$G_k^\perp$	nuisance-controlled gain (Equation 2)
$\mathbf{b}$	low-level reference battery (32 statistics)
U_k, $U_k^\perp$	raw/controlled reference margin over $\mathbf{b}$ (Equation 3)
Estimation, profile, verdicts (§3.3, §4.5)	
$G_k^{(\beta)}$, B	bootstrap replicate β of the gain; number of resamples
p_k, $\tilde{p}_k$	two-sided bootstrap p-value (Equation 6); its Holm-adjusted value
$\mathbf{G}(f)$	full gain profile across axes (not a scalar)
$\text{LIP}(f_\theta)$	Longitudinal Information Profile (Equation 11)
ACQUIRED/DEGRADED/NOT ESTABLISHED	verdicts from bootstrap intervals ($p < 0.05$)
Per-axis probes (§4.3, §4.4)	
A_{ord}	temporal-order accuracy
G_{temp}	temporal-order gain (Equation 7)
R@1, mAP, ΔR@1	same-knee identity retrieval (Equation 24); change in R@1 from initialization
AUC_ℓ	AUROC on MOAKS label ℓ
G_{phen}	phenotype gain over 12 labels (Equation 8)
τ_A, τ_N, D_f	anatomical vs. nuisance perturbation; induced representation distance
$S(f)$	selectivity (Equation 9)
PR, r_{eff}, cos	geometric diagnostics, not axes

B DATASET AND PREPROCESSING DETAILS

B.1 SOURCE DATA AND COHORTS

All images are sagittal three-dimensional double-echo steady-state water-excitation (SAG-3D-DESS-WE) knee MRI from the Osteoarthritis Initiative (Nevitt et al., 2006; Peterfy et al., 2008), acquired at the baseline, 12-, 24-, and 48-month visits (there is no 36-month imaging visit). After

removing 16 same-day repeat series, the image index contains 6,584 scans from 2,446 patients and 2,864 knees (1,892 baseline, 2,174 12-month, 1,425 24-month, and 1,093 48-month scans); 632 knees have one visit, 1,024 two, 928 three, and 280 all four. Raw volumes are $384 \times 384 \times 160$ voxels at $0.3646 \times 0.3646 \times 0.70$ mm on protocol. Laterality was established per scan by joining acquisition barcodes to the OAI imaging metadata, with complete coverage.

Table 3 summarises the cohorts. Encoders are trained on a *pilot* cohort of knees with at least two visits, registered to a common pose; its patients are split 464/59/57 into training, validation, and test by a patient-level assignment stratified by the number of visits. Because the training loader keeps only scans present in the registered cache, the training split contributes 1,268 scans and 1,167 within-knee visit pairs at gaps of 12, 24, 36, and 48 months (543, 335, 213, and 76 pairs). All analyses use a separately registered *diagnostic* cohort with no patient overlap with the pilot cohort. This cohort is then split by patient into a development half, on which every result in this paper is computed, and a holdout half that was sealed before the analyses and remains unused. Patient overlap between any two of the pilot, development, and holdout cohorts is zero. The procedure that drew the development/holdout division was not recorded as a script; only the resulting counts and the absence of overlap are verifiable, and we list this among the limitations in Appendix I.2.

Evaluation subsets. For each downstream evaluation, we split the development cohort by patient so that no patient appears in both the training and evaluation sets. The temporal-order analysis uses 62% of development patients, giving 1,162 held-out within-knee visit pairs. For phenotype evaluation, 40% of patients are reserved for testing, corresponding to 831 of the 2,035 scans with acquisition metadata. Geometry and initialization-similarity analyses use the same fixed subset of 1,024 training scans. This subset is also used for the exploratory Q_4 analysis described in Appendix G.4.

Table 3: Cohorts. Pairs are within-knee visit pairs at gaps of 12–48 months. The full-index count enumerates every within-knee pair of the image index for reference only; no encoder is trained on it. The pilot training row counts the scans and pairs seen by the training loader after restriction to the registered cache.

Cohort	Patients	Knees	Scans	Four-visit knees	Pairs
Full image index	2,446	2,864	6,584	280	5,488
Pilot (registered)	580	599	1,597	75	–
training split (loader)	464	476	1,268	59	1,167
validation split	59	63	168	8	155
Diagnostic (registered)	1,380	1,535	4,078	191	3,742
development half (audited)	690	768	2,036	94	1,862
holdout half (sealed)	690	767	2,042	97	1,880

B.2 Preprocessing Pipeline

Each volume passes through five stages, shown for one scan in Figure 4. (i) *Canonicalisation and resampling*: reorientation to the closest canonical frame, linear resampling to $0.70 \times 0.7292 \times 0.7292$ mm (an in-plane factor of exactly one half), centre crop or pad to $160 \times 192 \times 192$ voxels, and per-volume percentile normalisation to $[0, 1]$ between the 1st and 99th percentiles. (ii) *Registration to a within-knee common pose*: every later visit of a knee is registered directly (never chained) to its earliest visit with a six-degree-of-freedom rigid transform, using Mattes mutual information with 32 histogram bins and 20% random sampling, regular-step gradient descent (learning rate 2.0, at most 300 iterations), and a three-level pyramid with shrink factors $(4, 2, 1)$ and smoothing $(2, 1, 0)$ mm. To avoid privileging the reference visit, rotations are averaged on $\mathrm{SO}(3)$ by the chordal quaternion mean and translations arithmetically, and every visit, including the reference, is resampled exactly once from its original array into this common pose; resampling only the moving visits made interpolation blur alone predict temporal order with accuracy 0.86. Registration succeeded for all 998 pilot and all 2,543 diagnostic moving volumes (median 29 iterations), improving the mean Mattes mutual-information objective, a negated similarity for which lower is better, from -0.319 to -0.494 and from -0.320 to -0.498 respectively; Figure 5 shows a four-year pair before and after alignment, where the mean absolute difference falls from 0.215 to 0.115. (iii) *Common-support masking*: every visit of a knee is multiplied by the intersection of the non-zero supports of all its

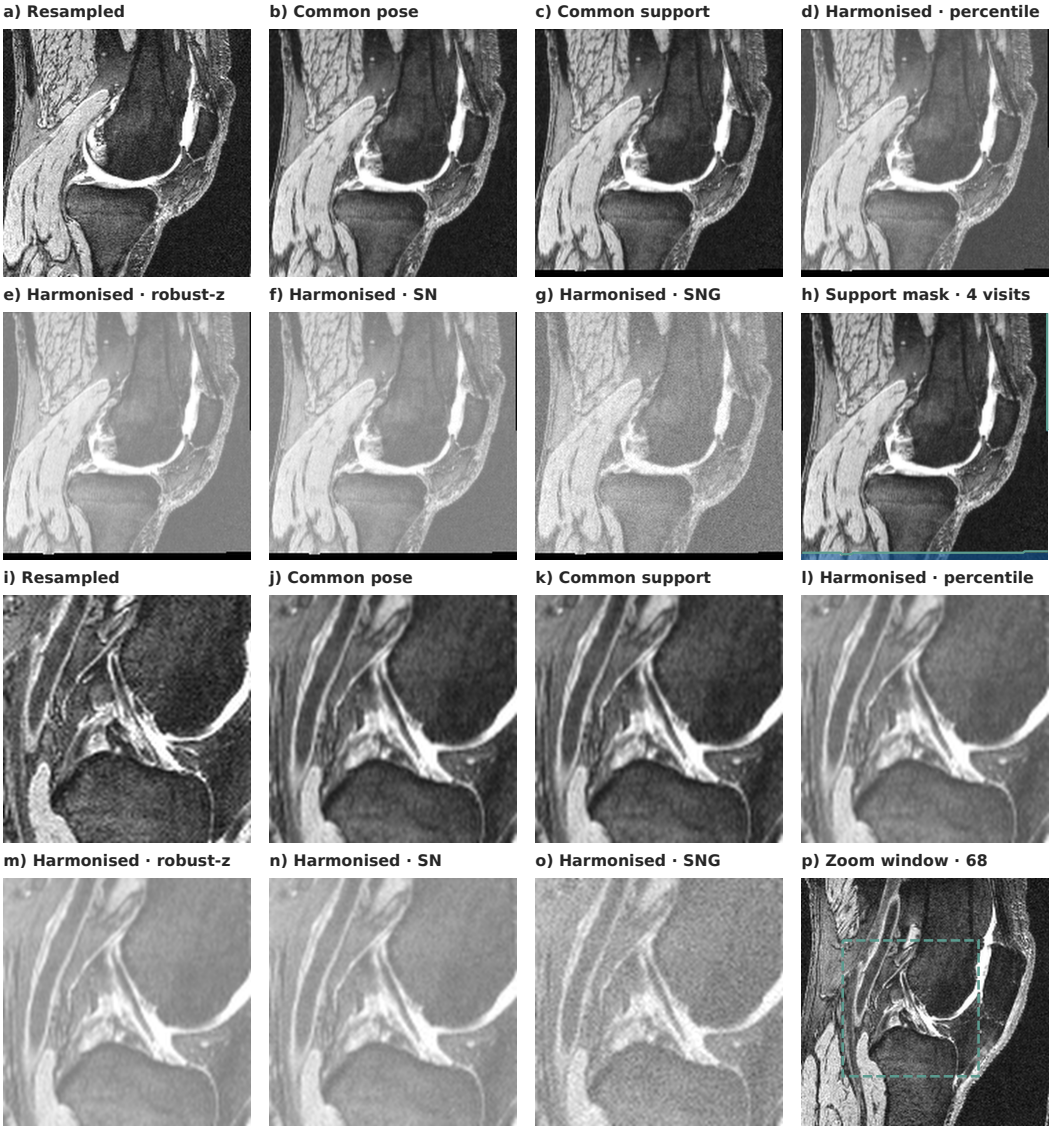

Figure 4: One pilot scan through the preprocessing stages, each panel displayed with its own intensity window: resampled and percentile-normalized volume, common pose, common-support masking, and four harmonisation variants. Rows 1–2: a condylar slice at full field of view; the last panel shades voxels outside the support common to the four visits of this knee. Rows 3–4: the same stages zoomed into the intercondylar notch, where the cruciate ligaments run (slice 68), with the zoom window outlined in the last panel.

visits, because the fraction of zero voxels alone ordered visits with accuracy 0.73 before and 0.83 after registration. (iv) *Intensity normalisation*: inside a tissue mask defined once per knee by Otsu thresholding of the mean volume within the common support, each volume is mapped by a robust z-transform $u = (v - \text{median})/\text{IQR}$, clipped to $[-4, 4]$, and affinely placed in $[0.02, 0.98]$ so that zero continues to denote the outside of the support. (v) *Caching* as 16-bit floating point. Training uses the preprocessed pilot data, while all evaluations use the separately preprocessed diagnostic data; the sharpness- and noise-matched variants in Figure 4 were evaluated during development and are not used by any encoder in the benchmark.

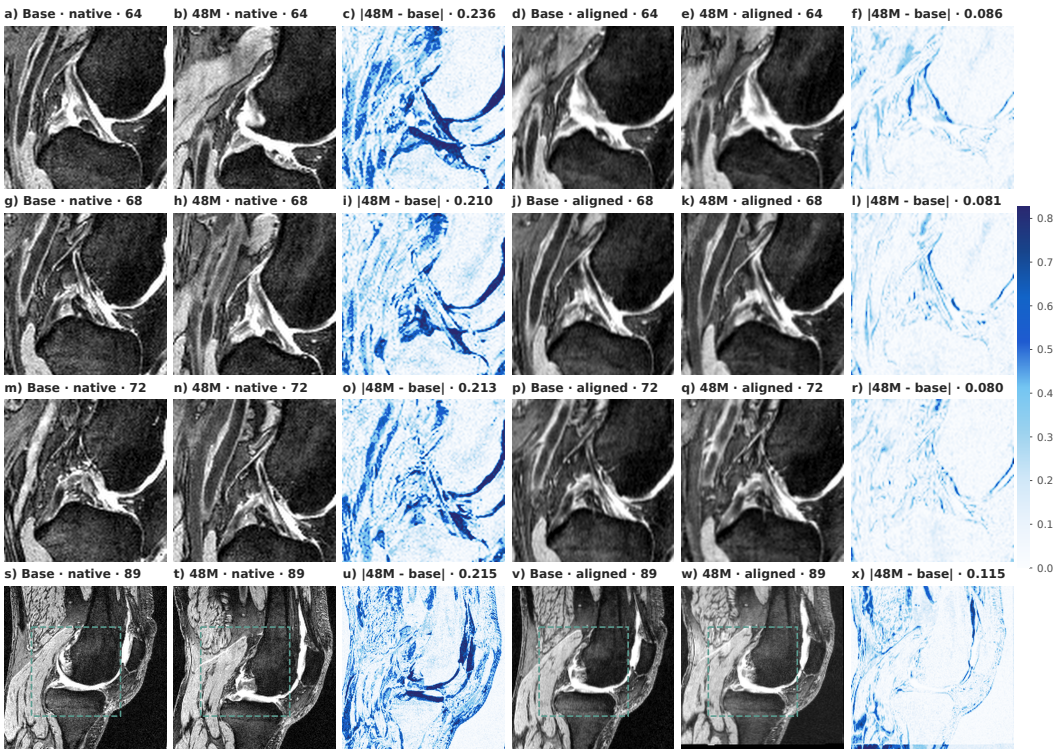

Figure 5: Registration quality control for one pilot knee. Rows 1–3: the intercondylar notch, where the cruciate ligaments run, zoomed on sagittal slices 64, 68, and 72. Row 4: slice 89 at full field of view, with the zoom window outlined. Left three columns: baseline and 48-month scans in their native frames and their absolute difference. Right three columns: the same pair after alignment to the common pose. All difference maps share one colour scale, and each label gives the mean absolute difference over the region shown: it falls from 0.210–0.236 to 0.080–0.087 in the notch and from 0.215 to 0.115 over the full slice.

B.3 MOAKS Phenotype Labels and Acquisition Covariates

Phenotype labels are 14 binary MOAKS (Hunter et al., 2011) indicators per scan: bone marrow lesions, cartilage damage, and osteophytes in the patellar, femoral, and tibial compartments, medial and lateral meniscal abnormality, anterior and posterior cruciate ligament abnormality, and effusion-synovitis. Labels with prevalence outside $[0.06, 0.94]$ on the probe pool are removed, which excludes the two cruciate ligament labels and leaves twelve (Table 4). The probe pool is the development cohort restricted to scans with acquisition metadata (2,035 scans, 690 patients). We could not establish from the available export which MOAKS reading project the labels derive from or whether longitudinal readings were paired, so the labels are used cross-sectionally only.

Acquisition covariates are the technologist identifier and the scanner-upgrade flag of each imaging visit, joined by barcode for 6,583 of 6,584 scans. The technologist group is the first character of the identifier, an inferred grouping that most plausibly encodes site. The residualization design has ten columns: an intercept, seven indicators for the eight technologist groups, and two indicators for the three upgrade levels, each coded against a reference level so that the design has full column rank. Table 4 reports, for each label, the battery AUROC before and after acquisition residualization and the AUROC of the acquisition covariates alone.

B.4 Clinical Endpoints

Three longitudinal clinical endpoints are defined on the development cohort. *Radiographic progression* (dJSM) is an increase in the OARSI medial joint-space-narrowing grade (0–3) between the two visits of a pair, from a single central reading project; the pair set contains 1,822 pairs from 671

Table 4: MOAKS labels used for phenotype evaluation, with prevalence in the probe pool and test-set AUROC for the 32-statistic battery before and after acquisition residualization, together with the acquisition covariates alone. Holm p refers to the change in battery AUROC after residualization.

Label	Prevalence	Battery raw	Battery resid.	Acquisition only	Holm p (drop)
BML patella	0.34	0.653	0.623	0.614	< 0.001
BML femur	0.47	0.645	0.626	0.586	0.152
BML tibia	0.35	0.658	0.643	0.590	1.000
Cartilage patella	0.73	0.631	0.583	0.543	0.490
Cartilage femur	0.71	0.671	0.663	0.580	1.000
Cartilage tibia	0.58	0.666	0.656	0.510	1.000
Osteophyte patella	0.26	0.689	0.652	0.607	0.090
Osteophyte femur	0.31	0.707	0.661	0.616	< 0.001
Osteophyte tibia	0.25	0.674	0.631	0.619	0.020
Meniscus medial	0.83	0.591	0.596	0.534	1.000
Meniscus lateral	0.40	0.506	0.500	0.551	1.000
Effusion-synovitis	0.43	0.695	0.679	0.588	1.000
Mean	–	0.649	0.626	0.578	3/12 significant

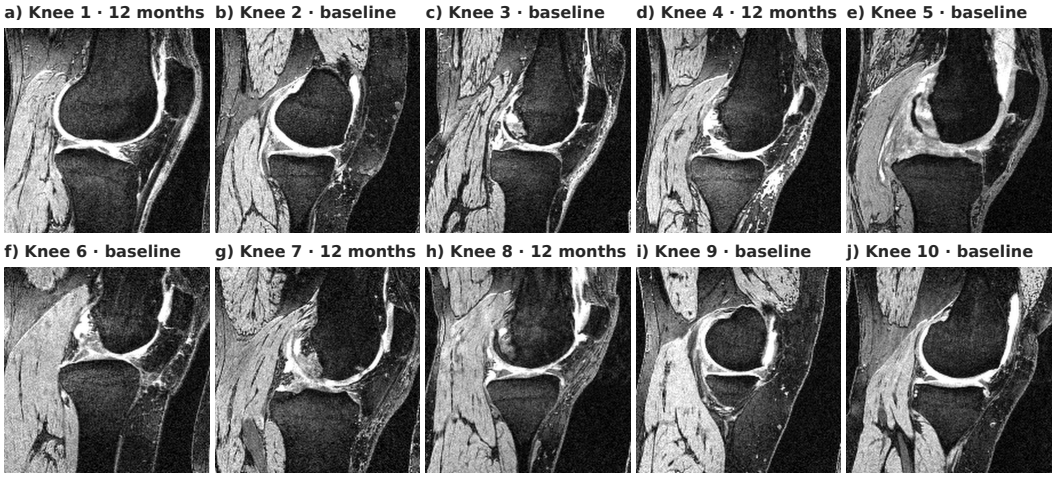

Figure 6: Ten knees from ten patients in the image index, one structured central sagittal slice each, displayed with femur superior and anterior to the right.

patients with 154 progression events. *Knee replacement* (TKR) is an adjudicated, confirmed replacement after the earliest cached MRI of a knee, with knees already replaced at baseline excluded and death treated as a terminating event; the cohort has 768 knees with 106 incident events over follow-up, 66 of which occur within the primary eight-year horizon. *MOAKS cartilage change* (dCART) is an increase in the summed cartilage-lesion grade over 28 MOAKS items between baseline and 24 months; 409 pairs carry both dCART and dJSM, with 145 cartilage worsenings. The cartilage endpoint was pre-specified as exploratory. Probing protocols are given in Appendix H.1.

B.5 Example Volumes

Figure 6 shows the most structured central sagittal slice of ten knees from ten different patients, Figure 7 shows two volumes in all three orthogonal planes, and Figure 8 shows two knees across all four visits in their common pose, where the same slice index corresponds to the same anatomy. The visual similarity of visits in Figure 8 illustrates why longitudinal change is subtle relative to between-knee variation, and why a representation that retains knee identity can decode visit order from small acquisition differences without encoding disease change.

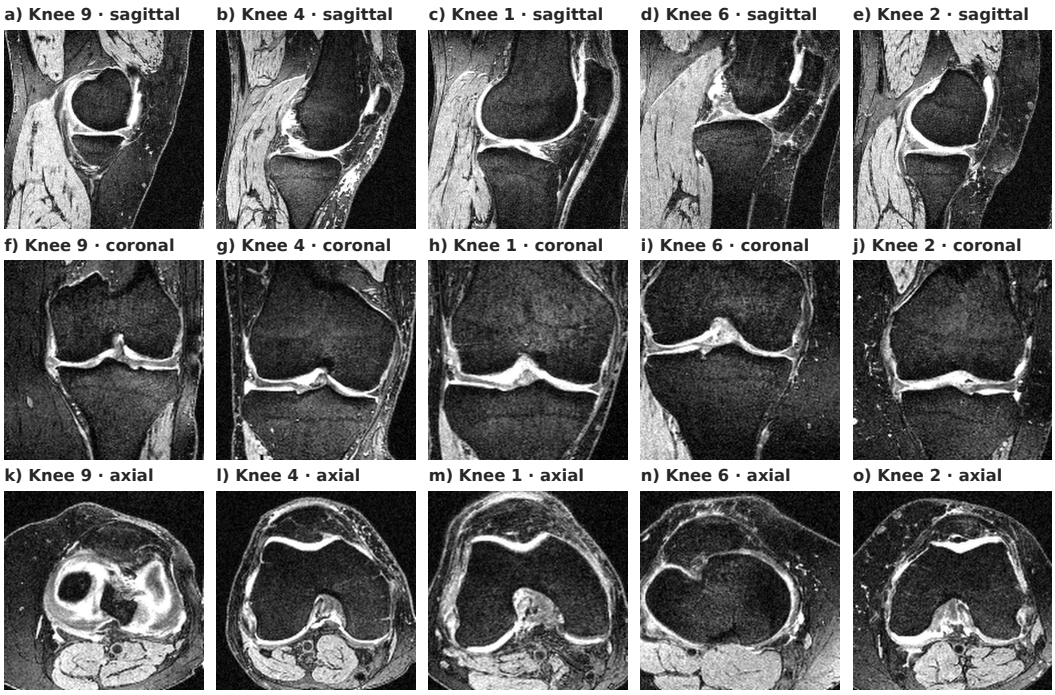

Figure 7: Two DESS volumes in sagittal, coronal, and axial planes after resampling to $160 \times 192 \times 192$ voxels; coronal and axial panels are centre-cropped to squares.

C Implementation and Training Details

C.1 Architecture

The encoder is a three-dimensional ViT-S: patch embedding by a 16^3 convolution with stride 16 produces a $10 \times 12 \times 12$ grid of 1,440 tokens of width 384, processed by 12 pre-norm transformer blocks with six attention heads, MLP ratio 4, GELU activations, LayerScale initialised at 10^{-5}, and stochastic depth rising linearly to 0.1. There is no absolute position embedding and no class token; positions enter through three-dimensional axial rotary embeddings on queries and keys with base 100. A learned-query attention pool followed by LayerNorm produces the trunk embedding $\mathbf{h} \in \mathbb{R}^{384}$ (23.47M parameters including the pool). A projection head (MLP $384 \to 1024 \to 1024 \to 256$ with BatchNorm and GELU) produces the latent $\mathbf{s} \in \mathbb{R}^{256}$; for IET a second head of identical shape produces the dynamic latent $\mathbf{d} \in \mathbb{R}^{256}$, while for single-latent objectives $\mathbf{d}$ and $\mathbf{s}$ are the same tensor. VICReg-style expanders (three-layer MLPs of width 512) map $\mathbf{h}$, $\mathbf{s}$, and $\mathbf{d}$ to the spaces where variance and covariance terms are computed. The time-conditioned transition used by Temporal SSL and IET is $T(\mathbf{d}, \Delta) = \mathbf{d} + \mathbf{W}_{\text{out}}\mathbf{u}$, with $\mathbf{u}$ produced by a two-layer residual network whose input $\text{SiLU}(\mathbf{W}_{\text{in}}\mathbf{d} \odot (1 + \boldsymbol{\gamma}_\Delta) + \boldsymbol{\beta}_\Delta)$ is modulated by FiLM parameters $(\boldsymbol{\gamma}_\Delta, \boldsymbol{\beta}_\Delta)$ computed from a 64-dimensional Fourier embedding of the elapsed time Δ; $\mathbf{W}_{\text{out}}$ is zero-initialised so that T is the identity at initialization. The order head $\psi_{\text{ord}}(\mathbf{d}_a, \mathbf{d}_b)$ is an MLP on $[\mathbf{d}_a, \mathbf{d}_b, \mathbf{d}_b - \mathbf{d}_a]$ that outputs the logit that visit a precedes visit b. JEPA adds a per-token predictor ($384 \to 1024 \to 384$) and an exponential-moving-average copy of the encoder as target. Trainable parameter counts are 30.94M for IET, 29.23M for the single-latent objectives, and 30.02M for JEPA (excluding its frozen target), not counting two adversarial heads that are instantiated but receive zero loss weight in every run.

C.2 Objectives

Each training step draws a batch of 16 within-knee visit pairs $(\mathbf{x}^1, \mathbf{x}^2, \Delta)$ with $\mathbf{x}^1$ the earlier visit and Δ the elapsed time, and three independent augmentations $\mathbf{x}^{1a}$, $\mathbf{x}^{1b}$, $\mathbf{x}^{2a}$ (Appendix C.4); superscripts on $\mathbf{h}$, $\mathbf{s}$, and $\mathbf{d}$ below name the view they are computed from. For a matrix of expander outputs

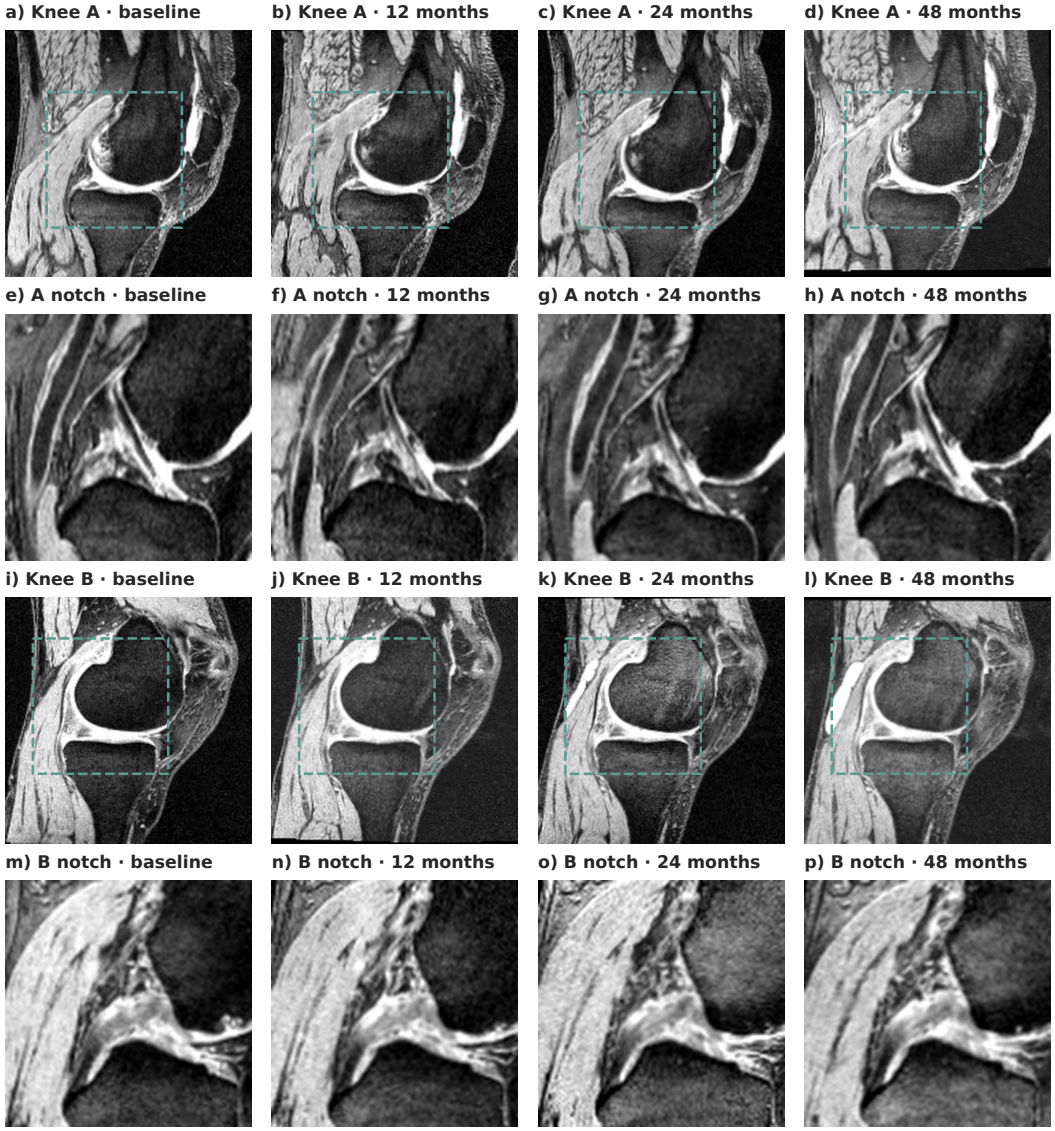

Figure 8: Two knees at baseline, 12, 24, and 48 months after registration to their within-knee common pose (rows 1–2: a pilot knee; rows 3–4: a development-cohort knee). For each knee, the first row shows a condylar slice at full field of view with the zoom window outlined, and the second row zooms into the intercondylar notch, where the cruciate ligaments run (slices 68 and 76). Within each row the slice index is fixed, so differences reflect change and residual acquisition variation rather than pose.

$\mathbf{P} \in \mathbb{R}^{n\times q}$ (n rows, q expander dimensions) define the variance and covariance regularisers

$$V_\gamma(\mathbf{P}) = \frac{1}{q}\sum_{j=1}^{q} \mathrm{ReLU}\Big(\gamma - \sqrt{\mathrm{Var}(\mathbf{P}_{\cdot j}) + 10^{-4}}\Big), \qquad \mathrm{Cov}(\mathbf{P}) = \frac{1}{q}\sum_{i\neq j} \Sigma_{ij}^2, \tag{12}$$

with $\mathbf{\Sigma} = [\Sigma_{ij}]$ the sample covariance matrix of the columns of $\mathbf{P}$ and $[\cdot\,;\cdot]$ denoting row-wise stacking, and the VICReg loss between two views $\mathrm{L}_{\mathrm{VIC}}(\mathbf{P}_1, \mathbf{P}_2) = \lambda_{\mathrm{sim}}\,\mathrm{MSE}(\mathbf{P}_1, \mathbf{P}_2) + \lambda_{\mathrm{std}} V_\gamma([\mathbf{P}_1; \mathbf{P}_2]) + \lambda_{\mathrm{cov}}\,\mathrm{Cov}([\mathbf{P}_1; \mathbf{P}_2])$ with $\lambda_{\mathrm{sim}} = \lambda_{\mathrm{std}} = 25$, $\lambda_{\mathrm{cov}} = 1$, and $\gamma = 1$. Let E_h, E_s, E_d denote the expanders, $\bar{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|_2$ the ℓ_2-normalization of a vector, sg the stop-gradient, SL_1 the smooth-ℓ_1 loss, $\mathrm{CE}(\cdot, y)$ the binary cross-entropy of a logit against label y, and ψ_{ord} the order head of Appendix C.1. Write $\mathbf{Z}_s = E_s([\mathbf{s}^{1a}; \mathbf{s}^{1b}; \mathbf{s}^{2a}])$ and $\mathbf{Z}_d = E_d([\mathbf{d}^{1a}; \mathbf{d}^{1b}; \mathbf{d}^{2a}])$ for the expanded latents of the three views, and $\mathbf{O}_\xi = [\xi^{1a}; \xi^{1b}; \xi^{2a}]$ for the raw (unexpanded)

Table 5: Loss weights of the five objectives. A dash denotes zero. IET's encoder-level variance and covariance weights are listed for the trunk, stable, and dynamic outputs.

Objective	L_{ssl}	L_{inv}	L_{aug}	L_{dyn}	L_{eq}	L_{adyn}	L_{ord}	Split latent	L_{enc} (w^{var}; w^{cov})
Static SSL	1 (VICReg)	–	–	–	–	–	–	no	–
Patient-CL	–	1	0.5	–	–	–	–	no	–
Temporal SSL	–	–	–	1	1	0.5	1.0	no	–
JEPA	1 (JEPA)	–	–	–	–	–	–	no	–
IET	1 (VICReg)	1	0.1	1	1	0.5	0.5	yes	25/2/1; 0.2/0.05/0.05

outputs of branch $\xi \in \{\mathbf{h}, \mathbf{s}, \mathbf{d}\}$. The component losses are

$$L_{ssl} = L_{VIC}\big(E_h(\mathbf{h}^{1a}), E_h(\mathbf{h}^{1b})\big), \tag{13}$$

$$L_{inv} = \lambda_{sim}\,\mathrm{MSE}\left(E_s(\mathbf{s}^{1a}), E_s(\mathbf{s}^{2a})\right) + \lambda_{std} V_\gamma(\mathbf{Z}_s) + \lambda_{cov}\,\mathrm{Cov}(\mathbf{Z}_s), \tag{14}$$

$$L_{aug} = \lambda_{sim}\,\mathrm{MSE}\left(E_s(\mathbf{s}^{1a}), E_s(\mathbf{s}^{1b})\right), \tag{15}$$

$$L_{dyn} = \lambda_{std} V_\gamma(\mathbf{Z}_d) + \lambda_{cov}\,\mathrm{Cov}(\mathbf{Z}_d), \tag{16}$$

$$L_{eq} = \tfrac{1}{2}\Big[\mathrm{SL}_1\left(\overline{T(\mathbf{d}^{1a}, \Delta)}, \mathrm{sg}\ \bar{\mathbf{d}}^{2a}\right) + \mathrm{SL}_1\left(\overline{T(\mathbf{d}^{2a}, -\Delta)}, \mathrm{sg}\ \bar{\mathbf{d}}^{1a}\right)\Big], \tag{17}$$

$$L_{adyn} = \mathrm{SL}_1\left(\bar{\mathbf{d}}^{1a}, \mathrm{sg}\ \bar{\mathbf{d}}^{1b}\right), \tag{18}$$

$$L_{ord} = \tfrac{1}{2}\Big[\mathrm{CE}\left(\psi_{ord}(\mathbf{d}^{1a}, \mathbf{d}^{2a}), 1\right) + \mathrm{CE}\left(\psi_{ord}(\mathbf{d}^{2a}, \mathbf{d}^{1a}), 0\right)\Big], \tag{19}$$

$$L_{enc} = \sum_{\xi \in \{\mathbf{h},\mathbf{s},\mathbf{d}\}} \Big[w_\xi^{var}\, V_\gamma(\mathbf{O}_\xi) + w_\xi^{cov}\,\mathrm{Cov}(\mathbf{O}_\xi)\Big], \tag{20}$$

where L_{enc} acts on the raw encoder outputs $\mathbf{O}_\xi$ rather than on expander outputs, and L_{ord} trains the order head to output label 1 for the chronological pair and 0 for the reversed pair. JEPA replaces L_{ssl} by masked-token prediction: token blocks of size $2\times4\times4$ are masked until at least 60% of tokens are covered, the student encodes $\mathbf{x}^{1a}$ with mask tokens substituted, and $L_{jepa} = \mathrm{SL}_1\left(\overline{g(\mathbf{t}_M^{stu})}, \mathrm{sg}\ \overline{\mathbf{t}_M^{tgt}}\right)$ over masked tokens M, with the target encoder updated by exponential moving average with momentum rising from 0.996 to 1 on a cosine schedule. The five objectives combine these terms with the weights in Table 5. Two implementation facts matter for interpretation. First, Patient-CL uses no negative pairs; it is a cross-visit invariance objective. Second, in Static SSL and JEPA no loss term depends on the projection head, so its parameters remain bit-identical to their initialization and only its BatchNorm running statistics can still change, but the projection-head parameters themselves remain at initialization. We therefore use the shared trunk embedding $\mathbf{h}$ for the main analysis in Section 5.

C.3 Training Protocol and Compute

Table 6 lists the shared protocol. All runs start from the same resolved configuration and differ only in the keys of Table 5, the epoch budget, and, for the additional analyses, the seed or a rank bottleneck. Each training budget follows its full cosine schedule, and all evaluations use the final checkpoint; the checkpoint with the lowest validation loss is saved but never used. Wall-clock time on one NVIDIA RTX PRO 4000 Blackwell (23.4 GiB) was 16.4 minutes for Static SSL and JEPA and about 20.3 minutes for Patient-CL, Temporal SSL, and IET at 10 epochs, and ten times that at 100 epochs. The rank-bottleneck and budget-sweep encoders of Appendix G.4 were trained on an NVIDIA RTX 5090 (5.2–5.6 minutes per 10 epochs), and the seed-replication encoders of Appendix E.10 on the RTX PRO 4000. Seeds control initialization and augmentation draws but not cuDNN algorithm selection, so runs are not bit-reproducible across devices; replaying one configuration and seed on the second GPU shifted Q_4 by about 0.05.

C.4 Augmentation

All objectives use the same nuisance augmentation pipeline, applied on the GPU in the order random resized crop (probability 0.8, isotropic scale $U[0.7, 1.0]$), affine transform (0.8; rotations $U[\pm10^\circ]$

Table 6: Shared training protocol.

Setting	Value
Training data	pilot training split: 464 patients, 1,268 scans, 1,167 visit pairs
Input	harmonised $160 \times 192 \times 192$ volumes, pairs within knee at $\Delta \in \{12, 24, 36, 48\}$ months
Batch	16 pairs (48 augmented volumes per step), gradient clipping at norm 1.0
Optimizer	AdamW, learning rate 3×10^{-4}, $\beta = (0.9, 0.95)$, weight decay 0.05 (none on norms, biases, queries)
Schedule	linear warm-up for 2 epochs, cosine decay to 0 at the end of the run's own budget
Steps	72 per epoch: 720 at 10 epochs, 7,200 at 100 epochs
Precision	bfloat16 autocast
Budgets	10 and 100 epochs (benchmark); 25 and 50 epochs (budget sweep, Static SSL)
Seeds	0 (benchmark); 1 and 2 (replication)
Evaluated checkpoint	final (fully annealed)

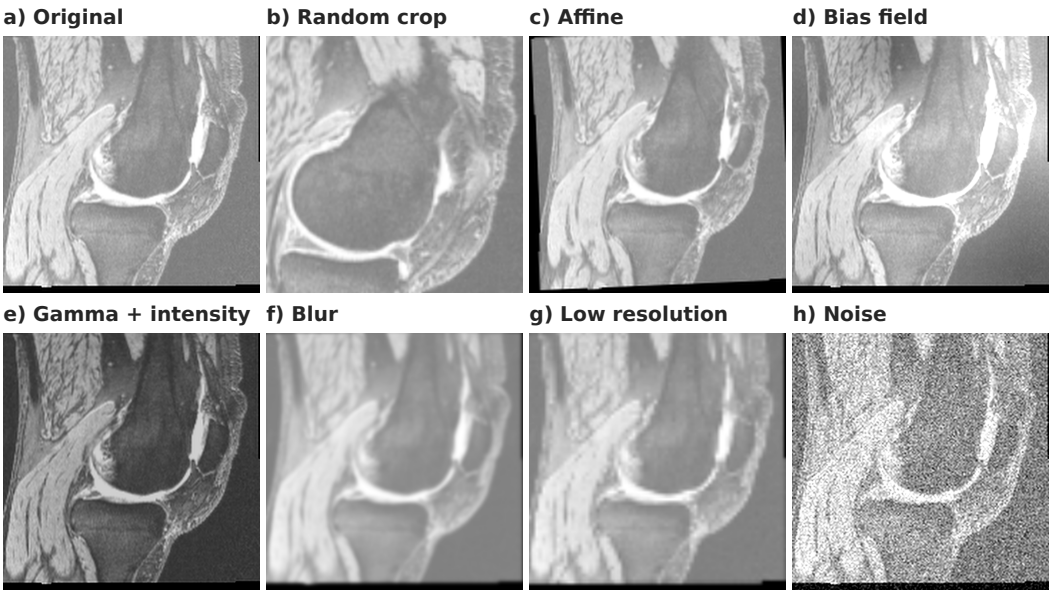

Figure 9: The nuisance augmentation operations applied individually to one scan with fixed seeds, each at the maximum magnitude allowed by the training configuration (during training each operation is applied only with its configured probability and a random magnitude up to this maximum), all panels on one display window: original, random resized crop, affine transform, bias field, gamma and intensity, blur, resolution degradation, and noise.

about each axis, per-axis scale $U[0.9, 1.1]$, shift $U[\pm 0.05]$ of the field of view), multiplicative smooth bias field (0.5; 4^3 control grid, strength $U[0, 0.3]$), gamma (0.5; $U[0.7, 1.4]$), affine intensity (0.8; gain $U[0.85, 1.15]$, offset $U[-0.1, 0.1]$), Gaussian blur (0.3; $\sigma \sim U[0.3, 1.2]$ voxels), resolution degradation (0.3; down- and up-sampling by $U[1.0, 2.5]$), and additive Gaussian noise (0.5; $\sigma \sim U[0, 0.08]$), followed by clamping to $[-3, 5]$. Whether each operation is applied is decided once per batch, and the blur and noise magnitudes are shared within a batch. Figure 9 applies each operation separately at the edge of its range to one scan. The same pipeline, scaled toward identity, provides the nuisance perturbation τ_N of the controlled-change probe (Appendix D.11).

D Evaluation Metrics and Statistical Procedures

D.1 Interpretation of Evaluation Quantities

Operational meaning of information. The information quantities used by LRA are operational rather than information-theoretic. For an axis k, $P_k(f)$ denotes held-out performance under a fixed probe family, fitting protocol, data partition, and evaluation metric. Consequently, G_k measures the change in probe-decodable information relative to matched initialization; it should not be interpreted as a direct estimate of mutual information or of the total information contained in the representation. Probe performance also carries sampling and optimization uncertainty, which is why every contrast is reported with a patient-clustered interval (Appendix D.6).

Axis-specific nuisance control. residualization is defined separately for each applicable information axis. More precisely, $R_{c,k}$ denotes the operator applied to the corresponding probe input using coefficients estimated exclusively from the training partition; the axis index is suppressed in the main text when unambiguous. For temporal order, residualization is applied to representation differences (Equation 10), whereas for phenotype prediction it is applied to scan-level representations (Appendix D.9). Accordingly, differences between raw and residualized performance quantify changes under the measured nuisance control and should not be interpreted as causal effects of the nuisance variables. The same reading applies to the two bracketed terms of Equation 5, which are algebraic residuals of that control rather than attributions of performance to nuisance structure; the calibration controls of Appendix D.5 bound how specific the control is on this cohort.

Beyond-reference utility. The quantity U_k is termed a reference margin and measures predictive performance relative to a fixed low-level image reference. Where the phrase "beyond-reference utility" is used, it refers only to this incremental predictive performance and not to decision-theoretic utility. Exceeding the reference therefore establishes that the representation provides predictive value beyond the measured low-level statistics under the specified probe protocol, not that it is universally more useful.

Interpretation of controlled change. The controlled-change axis measures relative representation sensitivity to a synthetic anatomical deformation and a matched nuisance perturbation. Therefore, $S > 1$ indicates greater sensitivity to the former under this controlled experiment. It does not establish sensitivity to natural disease progression or imply that the corresponding representation encodes clinically meaningful longitudinal change.

D.2 LRA Procedure

Algorithm 1 describes the evaluation for one encoder and one information axis; there, $P_k(\sigma)$ denotes the axis metric computed from held-out probe scores σ on the rows of the current bootstrap resample. Repeating the procedure across all axes gives the measurements used to construct the LIP in Equation 11.

D.3 Evidence Rules and Verdicts

Continuous estimates are necessary but invite selective reading, so LRA maps each inferentially estimated profile entry to a verdict with rules that are fixed and applied mechanically to every encoder. For an axis k with a paired-bootstrap interval $[\ell_k, u_k]$ and Holm-adjusted p-value $\tilde{p}_k$:

$$\text{verdict}_k = \begin{cases} \text{ACQUIRED} & \text{if } \ell_k > 0 \text{ and } \tilde{p}_k < 0.05, \\ \text{DEGRADED} & \text{if } u_k < 0 \text{ and } \tilde{p}_k < 0.05, \\ \text{NOT ESTABLISHED} & \text{otherwise.} \end{cases} \tag{21}$$

A verdict of NOT ESTABLISHED means that the benchmark provides insufficient evidence for a gain, not that the true gain is zero. Phenotype information is ACQUIRED only when $G_{\text{phen}} > 0$ and at least one label has a positive Holm-significant gain, and BEYOND REFERENCE only when the bootstrap interval of $U^{\perp}_{\text{phen}}$ lies above zero. For controlled change, S is compared with 1: an interval above 1 indicates SYNTHETIC-ANATOMY-SELECTIVE behavior, while one below 1 indicates NUISANCE-DOMINATED behavior. We do not compare S with initialization because both initialization distances

Algorithm 1 Longitudinal Representation Audit for one encoder and one information axis

Require: configuration κ with seed ω and trained parameters θ; evaluation pool with patient-disjoint partition $(\mathcal{I}_{\mathrm{tr}}, \mathcal{I}_{\mathrm{te}})$; probe family $\mathcal{H}_k$ with hyperparameter grid Λ_k; optional nuisance descriptors $\mathbf{c}$ for axis k; optional reference features $\mathbf{b}$ for axis k; bootstrap size B

Seed all generators with ω; build the model from κ; record θ_0 and its parameter hash
Load θ into a second copy of the model built from κ
for $f \in \{f_{\theta_0}, f_\theta\}$ **do**
    Embed every pool scan in evaluation mode without augmentation: Z_f
    Standardize with training-partition statistics; form axis-specific rows (pairs for order, scans for phenotype)
    Select $\lambda_f \in \Lambda_k$ by patient-grouped cross-validation on $\mathcal{I}_{\mathrm{tr}}$; fit probe $p_f \in \mathcal{H}_k$
    Compute held-out scores σ_f on $\mathcal{I}_{\mathrm{te}}$
    if nuisance descriptors $\mathbf{c}$ are defined for axis k **then**
        Fit residualization R_c on $\mathcal{I}_{\mathrm{tr}}$ only; refit the probe on residuals; compute held-out scores $\sigma_f^{\perp}$
    end if
end for
if a low-level reference $\mathbf{b}$ is defined for axis k **then**
    Fit and score $\mathbf{b}$ identically, raw and, if $\mathbf{c}$ is defined, residualized
end if
for $\beta = 1, \ldots, B$ **do**
    Resample patients of $\mathcal{I}_{\mathrm{te}}$ with replacement (one index set shared by all conditions)
    $G_k^{(\beta)} \leftarrow P_k(\sigma_{f_\theta}) - P_k(\sigma_{f_{\theta_0}})$
    where defined: $G_k^{\perp(\beta)} \leftarrow P_k(\sigma_{f_\theta}^{\perp}) - P_k(\sigma_{f_{\theta_0}}^{\perp})$ and $U_k^{\perp(\beta)} \leftarrow P_k(\sigma_{f_\theta}^{\perp}) - P_k(\sigma_b^{\perp})$
end for
Report bootstrap means, percentile 95% intervals, and two-sided bootstrap p-values; Holm-adjust within the declared family
Apply the evidence rules of Equation 21 to obtain the verdict for axis k

Ensure: absolute scores and G_k with uncertainty; $G_k^{\perp}$ and $U_k^{\perp}$ where defined; a verdict

are below 10^{-4}, making the ratio unstable; moreover, S measures response to synthetic deformation, not real disease change. Identity retrieval and geometric measures remain descriptive because per-encoder intervals are unavailable. These rules were codified after the individual analyses as a transparent summary, not as pre-registered tests; the pre-registered analyses and outcomes are listed in Table 8 and Appendix D.15.

D.4 Temporal-Order Probe

For every knee in the evaluation pool, visits are sorted by month, and every ordered pair (a, b) with $t_a < t_b$ is retained, not only adjacent pairs. Representations are standardized with means and standard deviations computed on training scans, $\tilde{\mathbf{z}} = (\mathbf{z} - \boldsymbol{\mu}_{\mathrm{tr}})/(\boldsymbol{\sigma}_{\mathrm{tr}} + 10^{-9})$ (element-wise), and each pair contributes two antisymmetric probe rows $\boldsymbol{\phi}_{ab}^{+} = (\tilde{\mathbf{z}}_b - \tilde{\mathbf{z}}_a)/\sqrt{p}$ with label $y = 1$ and $\boldsymbol{\phi}_{ab}^{-} = -\boldsymbol{\phi}_{ab}^{+}$ with label $y = 0$. Indexing rows by r, with features $\boldsymbol{\phi}_r$ and labels $y_r \in \{0, 1\}$, the probe is an ℓ_2-regularised logistic regression without intercept fitted on the training rows $\mathcal{V}_{\mathrm{tr}}$,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{r \in \mathcal{V}_{\mathrm{tr}}} \log\left(1 + \exp(-y_r' \, \mathbf{w}^\top \boldsymbol{\phi}_r)\right) + \frac{1}{2C}\|\mathbf{w}\|_2^2, \qquad y_r' = 2y_r - 1 \in \{-1, +1\}, \tag{22}$$

with $C \in \{10^{-4}, 10^{-3}, \ldots, 10^{5}\}$ selected by five-fold cross-validation on training pairs with folds grouped by patient, maximizing mean accuracy and breaking ties toward stronger regularization. The probe is tuned separately for each representation. Temporal-order accuracy on a held-out row set $\mathcal{V}$ is

$$A_{\mathrm{ord}}(f) = \frac{1}{|\mathcal{V}|} \sum_{r \in \mathcal{V}} \mathbf{1}\big[\operatorname{sign}(\hat{\mathbf{w}}^\top \boldsymbol{\phi}_r) = y_r'\big], \tag{23}$$

and the rank AUROC of the scores $\hat{\mathbf{w}}^\top \boldsymbol{\phi}_r$ is computed on the same rows. On the development cohort, 62% of patients are held out, giving 700 training pairs and 1,162 validation pairs. The

analysis script is frozen by a SHA-256 hash recorded in the O21 pre-registration and verified by every evaluation script before execution (Appendix D.15).

D.5 Nuisance Residualization and Specificity Controls

For the temporal axis, the nuisance descriptors are the 32 battery statistics of Table 7, computed inside the common-support mask. Ten of them (the zero fraction, the six bounding-box limits, and the three support centroids) are constant within a knee after common-support masking and vanish in pair differences; the remaining 22 are standardized on training scans, differenced within pairs, and scaled like the representation rows. The residualization of Equation 10 is a multi-output ridge regression without intercept with $\alpha \in \{10^{-2}, 10^{-1}, \ldots, 10^{3}\}$ chosen by five-fold patient-grouped cross-validation on held-out R^2, fitted on training pairs and applied to training and validation pairs; the order probe is refitted on residuals with the regularization strength of the raw probe. The fraction of representation variance removed is $1 - \|R_c[\Delta\tilde{\mathbf{Z}}_{\mathrm{tr}}]\|_F^2 / \|\Delta\tilde{\mathbf{Z}}_{\mathrm{tr}}\|_F^2$, where $\Delta\tilde{\mathbf{Z}}_{\mathrm{tr}}$ stacks the training-pair differences row-wise and R_c is applied row by row. Two controls calibrate this operation. The *permutation control* breaks the correspondence between battery and representation rows, separately in the training and validation partitions, and repeats the residualization with the same α. The *random-subspace control* removes a subspace of the same rank (22) whose explained variance is closest to the variance removed by the real residualization, chosen among 400 random subsets of principal directions. The pre-registered specificity contrast is $\Delta_{\mathrm{spec}} = A^{\perp}_{\mathrm{ord}}(\text{battery}) - A^{\perp}_{\mathrm{ord}}(\text{matched random subspace})$, classified as DIFFERENT if Holm $p < 0.05$, EQUIVALENT if its interval lies within ± 0.02, and INCONCLUSIVE otherwise; results are reported in Table 11, Appendix E.2.

For the phenotype axis, the nuisance descriptors are the acquisition covariates of Appendix B.3. Both representation and battery features are replaced by residuals $Z - X_{\mathrm{acq}}\hat{\mathbf{\Gamma}}_{\mathrm{acq}}$ with $\hat{\mathbf{\Gamma}}_{\mathrm{acq}}$ estimated by least squares on the training half only (Equation 27), and probed as described in Appendix D.9.

D.6 Statistical Inference

Unless otherwise stated, intervals and tests use a patient-clustered paired bootstrap with $B = 2000$ resamples. Probes are fitted once; in each resample the held-out patients are drawn with replacement, all rows of a drawn patient enter together, and every condition (trained, initialization, residualized, reference) is scored on the same drawn rows, so that contrasts are paired. For a contrast with draws $\delta^{(\beta)}$, $\beta = 1, \ldots, B$, we report the bootstrap mean $\bar{\delta} = B^{-1}\sum_{\beta=1}^{B}\delta^{(\beta)}$, which can differ in the third decimal from the difference of the point estimates, the percentile interval $[q_{0.025}, q_{0.975}]$ of the draws, and the two-sided bootstrap p-value of Equation 6 with $G_k^{(\beta)}$ replaced by $\delta^{(\beta)}$. Holm's step-down procedure (Holm, 1979) is applied within declared families: for a family of m_{f} raw p-values ordered as $p_{(1)} \le \cdots \le p_{(m_{\mathrm{f}})}$, the adjusted values are $\tilde{p}_{(i)} = \min\{1, \max_{j \le i}(m_{\mathrm{f}} - j + 1)\, p_{(j)}\}$. The families are the three branches $\{\mathbf{h}, \mathbf{s}, \mathbf{d}\}$ for the temporal gain, the twelve labels for the phenotype gain, and the contrasts of each pre-registered gate. Spearman correlations across encoders are computed from unrounded estimates and reported with their asymptotic p-values; where the number of encoders is small, exact permutation p-values are given instead.

D.7 Low-Level Image Battery

Table 7 defines the 32 statistics. For a volume v with support $\Omega = \{v \neq 0\}$ (the set of non-zero voxels), $n = v[\Omega]$ the vector of in-support intensities, interquartile range $\mathrm{IQR} = \max(q_{75} - q_{25}, 10^{-6})$, central-difference gradients $g_i = \partial_i v$, gradient magnitude $|g|$, and Laplacian $L = \sum_i \partial_i^2 v$, the battery captures intensity distribution, sharpness, noise, and field of view, which are the channels through which acquisition differences between visits become visible. The battery serves three roles: as nuisance descriptors for temporal residualization, as a chronology reference (its own order accuracy), and as the low-level reference $\mathbf{b}$ for the phenotype axis.

D.8 Identity Retrieval

Knee identity is measured by leave-one-out retrieval over the 2,036 development scans using L2-normalized raw embeddings and cosine similarity, with a scan relevant to a query if it belongs to the

Table 7: The 32-statistic low-level image battery. Statistics marked † are constant within a knee after common-support masking and vanish in pair differences.

Index	Statistic	Definition
1	zero fraction†	fraction of grid voxels equal to 0
2	noise-floor fraction	fraction of voxels with $0 < v \le q_{10}$
3	tissue fraction	fraction of voxels with $v > q_{50}$
4–6	q_{10}, q_{50}, q_{90}	percentiles of n
7	IQR	$q_{75} - q_{25}$
8, 9	mean, standard deviation	of n
10, 11	gradient magnitude	$\mathrm{mean}_\Omega\,\lvert g\rvert$ and $\mathrm{mean}_\Omega\,\lvert g\rvert/\mathrm{IQR}$
12, 13	Laplacian variance	$\mathrm{var}_\Omega\,L$ and $\mathrm{var}_\Omega\,L/\mathrm{IQR}^2$
14	SNR proxy	$\bar{n}/(\mathrm{sd}(n) + 10^{-6})$
15–20	edge energy per axis	$\mathrm{mean}_\Omega\,\lvert g_i\rvert$ and $\mathrm{mean}_\Omega\,\lvert g_i\rvert/\mathrm{IQR}$, $i = 0, 1, 2$
21–26	bounding box†	first and last support index along each axis
27–32	centroids	support centroid† and intensity-weighted centroid along each axis

same knee (patient and side). With $\mathcal{Q}$ the set of query scans that have at least one relevant scan, $\mathcal{R}_q$ the relevant set of query q, $\mathrm{top1}(q)$ the most similar non-query scan, and $\mathrm{Prec}_q(j)$ the precision of the ranked list of q truncated at the rank of relevant scan j,

$$\mathrm{R@1} = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathbf{1}\big[\,\mathrm{top1}(q) \in \mathcal{R}_q\big], \qquad \mathrm{mAP} = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \frac{1}{|\mathcal{R}_q|}\sum_{j\in\mathcal{R}_q} \mathrm{Prec}_q(j). \tag{24}$$

The identity entry of the profile is $\Delta\mathrm{R@1} = \mathrm{R@1}(f_\theta) - \mathrm{R@1}(f_{\theta_0})$. Retrieval is computed without resampling, so identity enters the profile descriptively. The same script reports a visit-class accuracy A_{time} from a softmax readout, per-branch embedding variance, the root-mean-square cross-correlation between standardized $\mathbf{s}$ and $\mathbf{d}$, and five decomposition gates relative to a reference encoder, which are point-estimate comparisons used only in Appendix F.

D.9 Phenotype Probe and Supervised Reference

For each label, feature set, and branch, features are standardized on the training half and scaled by $1/\sqrt{p}$. The regularization strength of a no-intercept ℓ_2 logistic regression is selected from the grid in Appendix D.4 by five-fold patient-grouped cross-validation, and AUROC is evaluated on the held-out test half (40% of patients). The reported phenotype gain is the bootstrap mean of Equation 8: with $\mathrm{AUC}_\ell^{(\beta)}(f)$ the AUROC of label ℓ on bootstrap resample β,

$$G_{\mathrm{phen}}(f_\theta) = \frac{1}{12}\sum_{\ell=1}^{12}\frac{1}{B}\sum_{\beta=1}^{B}\Big[\mathrm{AUC}_\ell^{(\beta)}(f_\theta) - \mathrm{AUC}_\ell^{(\beta)}(f_{\theta_0})\Big]. \tag{25}$$

This quantity is computed from raw AUROCs, with per-label bootstrap p-values Holm-corrected over the twelve labels.

For acquisition-controlled comparisons, we use

$$X_{\mathrm{acq}} = [\mathbf{1}, \mathrm{onehot}_{-1}(\text{technologist group}), \mathrm{onehot}_{-1}(\text{scanner upgrade})], \tag{26}$$

where onehot_{-1} omits one reference level from each categorical variable so that the design matrix has full column rank. For each feature set, with Z the matrix of scan-level features (one row per scan), the least-squares coefficients are estimated using the training half only,

$$\hat{\mathbf{\Gamma}}_{\mathrm{acq}} = \arg\min_{\mathbf{\Gamma}} \|Z_{\mathrm{tr}} - X_{\mathrm{acq,tr}}\mathbf{\Gamma}\|_F^2, \qquad R_c[Z] = Z - X_{\mathrm{acq}}\hat{\mathbf{\Gamma}}_{\mathrm{acq}}, \tag{27}$$

which is the unpenalized, scan-level counterpart of the pair-level residualization in Equation 10, and the fitted transformation is then applied to the held-out test half. The same procedure is applied independently to each learned representation, its matched initialization, and the low-level battery.

The acquisition-controlled reference margin is the bootstrap mean of $U_k^{\perp}$ in Equation 3 on this axis, with $\mathrm{AUC}_{\ell}^{\perp(\beta)}$ the AUROC after residualization by Equation 27:

$$U_{\mathrm{phen}}^{\perp}(f_\theta) = \frac{1}{12}\sum_{\ell=1}^{12}\frac{1}{B}\sum_{\beta=1}^{B}\left[\mathrm{AUC}_{\ell}^{\perp(\beta)}(f_\theta) - \mathrm{AUC}_{\ell}^{\perp(\beta)}(\mathbf{b})\right]. \tag{28}$$

All ten encoders, their matched initializations, and the battery are evaluated on the same $B = 2000$ patient-clustered bootstrap resamples of the test half. Confidence intervals are the 2.5th–97.5th percentiles of the resampled label means, and per-label differences are Holm-corrected over the twelve labels. Resampling patients rather than labels preserves the dependence among labels from the same knee.

As a separate same-backbone supervised reference, which is neither acquisition-residualized nor an upper bound on the task, the same backbone is trained end-to-end with a twelve-way linear head using class-weighted binary cross-entropy and encoder-level variance and covariance regularization (weights 25 and 0.2; without them the encoder collapsed). Training uses AdamW with learning rate 3×10^{-4}, a one-cycle schedule, 60 epochs, and the same augmentation. The training pool excludes all patients in the probe test half and in a 20% inner validation split used for epoch selection. The reported value is the raw test-half AUROC at the epoch with the highest inner-validation AUROC (epoch 43; inner 0.603, test 0.548).

D.10 Geometric Descriptors and Similarity to Initialization

All cross-encoder geometric comparisons use the same estimator and the same 1,024 scans from the training cohort. Scans are embedded in batches of 16 under two augmentation draws, with the random generators re-seeded identically for every encoder. Let $Z \in \mathbb{R}^{N\times p}$ denote the column-centered embeddings of the $N = 1{,}024$ scans from the first view and $\sigma_1 \geq \sigma_2 \geq \dots$ its singular values. The participation ratio is

$$\mathrm{PR}(Z) = \frac{\left(\sum_j \sigma_j^2\right)^2}{\sum_j \sigma_j^4}, \tag{29}$$

which equals 1 when all variance lies in one direction and p when it is spread equally over all p directions. For the two augmented views of a batch of n scans, let $\hat{Z}_1, \hat{Z}_2 \in \mathbb{R}^{n\times p}$ be the column-standardized embeddings and $C = \hat{Z}_1^{\top}\hat{Z}_2/n \in \mathbb{R}^{p\times p}$ their cross-correlation matrix. The effective rank is the participation ratio of the singular values of C,

$$r_{\mathrm{eff}} = \frac{\|C\|_F^4}{\|CC^{\top}\|_F^2 + 10^{-6}}, \tag{30}$$

averaged across batches. Mean pairwise cosine is the average of $\bar{\mathbf{z}}_i^{\top}\bar{\mathbf{z}}_j$ over all pairs $i \neq j$, with $\bar{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$. Although participation ratio can also be estimated from other sample sets in the analysis pipeline, all reported cross-encoder comparisons use the estimator defined here; values obtained from different estimators are not compared.

Similarity to initialization is evaluated on 1,024 clean training-cohort scans. Let $X, Y \in \mathbb{R}^{N\times p}$ denote the column-centered, Frobenius-normalized embeddings from the matched initialization and the trained encoder. Linear CKA is

$$\mathrm{CKA}(X, Y) = \frac{\|X^{\top}Y\|_F^2}{\|X^{\top}X\|_F\,\|Y^{\top}Y\|_F} \tag{31}$$

(Kornblith et al., 2019), Procrustes similarity is the nuclear norm $\|X^{\top}Y\|_*$, and the top-k subspace overlap is

$$O_k = \frac{\|V_k(X)^{\top}V_k(Y)\|_F^2}{k} \in [0, 1], \tag{32}$$

where $V_k(\cdot) \in \mathbb{R}^{p\times k}$ contains the leading k right singular vectors; $O_k = 1$ when the two k-dimensional subspaces coincide and $O_k = 0$ when they are orthogonal.

D.11 Controlled-Change Probe

Anatomical perturbation. A displacement field is drawn with independent $\mathcal{N}(0,1)$ entries on a 5^3 grid of control points for each of its three components, upsampled trilinearly to the volume, multiplied by a Gaussian envelope $\exp\big(-\|\mathbf{u}-\mathbf{o}\|_2^2/(2\cdot 0.35^2)\big)$ in normalized coordinates $\mathbf{u}\in[-1,1]^3$ with centre $\mathbf{o}\sim\mathcal{U}[-0.6,0.6]^3$, rescaled so that its voxel-wise root-mean-square magnitude equals ε voxels, and applied by trilinear resampling. Three levels $\varepsilon\in\{1.5,3,5\}$ are used.

Nuisance perturbation. The augmentation pipeline of Appendix C.4 is scaled by a factor $\eta\in[0,2]$: magnitude parameters are multiplied by η, a range $[\mathrm{lo},\mathrm{hi}]$ centred on one is contracted toward one as $[1-(1-\mathrm{lo})\eta,\ 1+(\mathrm{hi}-1)\eta]$, the resolution factor r becomes $1+(r-1)\eta$, and application probabilities are unchanged.

Matching. For each level, η is found by bisection so that the median in-support RMS intensity change $\sqrt{\frac{1}{|\Omega|}\sum_{v\in\Omega}\big(\tau(\mathbf{x})_v-\mathbf{x}_v\big)^2}$, with Ω the common-support voxel set, of τ_N matches that of τ_A on a calibration set of 128 scans; a level is retained only if the relative mismatch is at most 10%. The achieved levels were RMS 0.0376, 0.0524, and 0.0652 with $\eta=0.060$, 0.136, and 0.187, matched within 0.3%. Perturbations are generated from seeds derived from the perturbation type, level, and batch, so every encoder sees byte-identical perturbed volumes.

Selectivity. With $D_f(\mathbf{x},\tau)=1-\cos(f(\mathbf{x}),f(\tau(\mathbf{x})))$ computed on raw (unstandardized) embeddings over all $N=2{,}036$ development scans, selectivity at the highest usable level is the ratio in Equation 9, with its 95% interval from the patient-clustered bootstrap applied to the ratio. The dissociation $S(\mathbf{d})-S(\mathbf{s})$ is computed on shared resamples and is identically zero for single-latent encoders. For the untrained network, both mean distances are below 10^{-4}, so its selectivity is reported but flagged as unresolved.

D.12 Sign-Free Change Axis

For knees with four visits, disjoint interval pairs $a<b<c<d$ provide two within-knee changes $\boldsymbol{\delta}_{ab}=\mathbf{z}_b-\mathbf{z}_a$ and $\boldsymbol{\delta}_{cd}=\mathbf{z}_d-\mathbf{z}_c$ (94 rows). Each row is scored by the alignment of the two changes, corrected by the alignment with 25 negatives,

$$\text{score}=\big|\cos(\boldsymbol{\delta}_{ab},\boldsymbol{\delta}_{cd})\big|-\frac{1}{25}\sum_{j=1}^{25}\big|\cos(\boldsymbol{\delta}_{ab},\boldsymbol{\delta}_j^{\text{neg}})\big|, \tag{33}$$

where the negatives $\boldsymbol{\delta}_j^{\text{neg}}$ are same-gap changes from different patients, and the change-axis gain is the mean score difference between the trained encoder and its initialization on identical rows. A post hoc contralateral control replaces $\boldsymbol{\delta}_{cd}$ with a same-gap change of the patient's other knee (24 rows from 18 patients). Pre-specified gates required the gain to exceed zero (G1), IET's dynamic latent to exceed every baseline at the same budget (G2, Holm-corrected), $\mathbf{d}$ to exceed $\mathbf{s}$ within IET (G3), and within-knee alignment to exceed contralateral alignment (G4).

D.13 Cross-View Predictivity Q_4

On 1,024 training-cohort scans split in half by patient, two augmented views are embedded and each is reduced to at most 64 principal components fitted on the first half. Canonical correlation analysis with 16 components is fitted on the first half, and the held-out absolute correlations $\rho_j^{\text{cca}}=|\operatorname{corr}(u_j,v_j)|$ of the j-th pair of canonical variates (u_j,v_j) are computed on the second half in the fitted order. The statistic is the mean over the leading k pairs, $Q_k=\frac{1}{k}\sum_{j=1}^{k}\rho_j^{\text{cca}}$, and Q_4 is used throughout. The untrained reference is computed for the Static SSL architecture ($Q_4=0.033$).

D.14 Subspace and Additivity Analyses

Leading directions. The trunk embedding is rotated into its principal directions estimated on the training half, and the phenotype probe is fitted on the leading $k\in\{1,2,4,\dots,128\}$ directions or on the directions beyond the leading $k\in\{1,4,16,64\}$; these probes are not acquisition-residualized and select the regularization strength once per encoder. **Trunk axes.** For the leading three principal directions, we report the explained variance, the held-out R^2 of a ridge regression predicting the

Table 8: Pre-registered analyses relevant to this paper, in chronological order, with outcomes as read.

Pre-registration	Question and gate	Outcome
O21 (frozen script hash)	Is temporal decodability specific to measured acquisition-like statistics, and does training add to the matched initialization (1,380 patients)?	Training below initialization on all branches (trunk -0.048 $[-0.075, -0.022]$); specificity DIFFERENT on 5 of 6 AUROC contrasts
O25 decorrelation	Does a decorrelation penalty increase the temporal split between $\mathbf{s}$ and $\mathbf{d}$?	Fail: correlation falls, split inverts
O26 adversaries	Do adversarial heads create a temporal split without degrading $\mathbf{s}$?	Split $+0.043$ $[+0.003, +0.081]$ obtained by degrading $\mathbf{s}$; closed
P0 budget (frozen hash)	Does ten times the budget recover IET's temporal information and clinical utility?	No: $\mathbf{d}$ -0.036 $\rightarrow$ -0.119; no positive clinical increment (small negative dJSM increment)
BENCH-1 stage 1	Do standard objectives preserve or destroy temporal structure relative to matched initialization?	No objective shows a detectable positive temporal gain (Table 1)
Change axis G2	Is the sign-free change axis specific to IET?	Fail: IET 10 minus Static SSL 10 -0.008 $[-0.076, +0.066]$
Q_k held-out ordering	Does Q_4 fixed on six encoders order four unseen encoders?	Pass: exact ordering, $p = 0.042$
Q_k rank intervention	Does manipulating a rank bottleneck move Q_4 and phenotype gain together but not PR?	Pre-registered reading: fail; final interpretation: unresolved, because the verdict changes with the PR estimator (Appendix G.4)
Q_k budget intervention	Across 10/25/50/100 epochs, $\rho(Q_4, G_{\text{phen}}) > 0$ and $> \lvert\rho(\text{PR}, G_{\text{phen}})\rvert$?	Pass ($+0.80$ vs 0.20); $n = 4$, collinear with epochs
Q_k seed perturbation	Does within-objective seed variation in Q_4 track phenotype gain?	Uninformative (1/5 objectives adequate)
Multi-seed robustness (amended: validation patients fixed)	Do seed-0 temporal verdicts reproduce at seeds 1 and 2?	No interval above zero at any seed; trunk reproduces for 4/5 objectives; Static SSL 10 loss seed-sensitive (Appendix E.10)

direction from the battery (maximized over the ridge penalty, and therefore optimistic), and the phenotype AUROC of the first direction alone. **Battery plus trunk.** After acquisition residualization of both blocks, the battery is concatenated with the full trunk or with its leading 8 or 16 directions, and the joint model is compared with the battery alone per label on shared bootstrap resamples with Holm correction over twelve labels. A dilution control repeats the procedure with the trunk of the untrained network, which should not add information.

D.15 Pre-Registration and Frozen Analyses

A subset of analyses was specified before its data were read: the hypothesis, the estimator, the gate with its possible outcomes, and, where an analysis script existed, its SHA-256 hash, which every evaluation script verifies before execution. Table 8 lists the pre-registrations that bear on the claims of this paper and their outcomes, including those that failed or were uninformative. Analyses not listed there, including the evidence rules of Appendix D.3, the dominance analysis, and the correlations of Section 5.3, are post hoc and are reported as such.

Table 9: Complete Longitudinal Information Profile for all ten encoders on the trunk representation $\mathbf{h}$, extending Table 1 with the columns held out of the main text. A_{ord} is temporal-order accuracy at initialization and after training; G_{temp} and $G^{\perp}_{\text{temp}}$ are the raw and nuisance-controlled temporal gains. R@1 is same-knee retrieval before and after training. G_{phen} is the mean MOAKS AUROC gain (Holm-significant labels in parentheses), $U^{\perp}_{\text{phen}}$ its acquisition-controlled margin over the 32-statistic reference, S controlled-change selectivity, and PR the participation ratio. Brackets are 95% patient-clustered bootstrap intervals; N.E. marks NOT ESTABLISHED. Verdicts follow Appendix D.3; branch-level and multi-seed temporal results are in Tables 10 and 16.

Encoder	$A_{\text{ord}}(\theta_0)$	$A_{\text{ord}}(\theta)$	G_{temp} [95% CI]	Verdict	$G^{\perp}_{\text{temp}}$	R@1 θ_0/θ	PR	G_{phen} (Holm)	$U^{\perp}_{\text{phen}}$ [95% CI]	S [95% CI]
Static SSL 10	0.630	0.579	-0.051 $[-0.090, -0.012]$	DEG.	-0.042	0.987/0.301	1.49	$+0.072$ (2/12)	-0.035 $[-0.064, -0.005]$	0.23 [0.22, 0.25]
Static SSL 100	0.630	0.603	-0.027 $[-0.067, +0.013]$	N.E.	-0.040	0.987/0.066	1.32	$+0.118$ (6/12)	$+0.016$ $[-0.012, +0.043]$	2.33 [2.11, 2.56]
Patient-CL 10	0.630	0.616	-0.014 $[-0.053, +0.023]$	N.E.	-0.022	0.987/0.146	1.29	$+0.068$ (3/12)	-0.042 $[-0.070, -0.014]$	0.31 [0.29, 0.34]
Patient-CL 100	0.630	0.620	-0.009 $[-0.052, +0.032]$	N.E.	-0.033	0.987/0.102	1.35	$+0.102$ (4/12)	-0.015 $[-0.043, +0.014]$	2.26 [2.08, 2.46]
Temporal SSL 10	0.630	0.650	$+0.020$ $[-0.010, +0.052]$	N.E.	-0.009	0.987/0.761	5.13	$+0.048$ (1/12)	-0.055 $[-0.084, -0.024]$	0.06 [0.06, 0.07]
Temporal SSL 100	0.630	0.620	-0.010 $[-0.040, +0.019]$	N.E.	-0.014	0.987/0.768	4.72	$+0.032$ (0/12)	-0.078 $[-0.109, -0.046]$	0.12 [0.11, 0.13]
JEPA 10	0.630	0.617	-0.013 $[-0.030, +0.004]$	N.E.	-0.005	0.987/0.966	2.04	-0.009 (0/12)	-0.113 $[-0.152, -0.073]$	0.08 [0.04, 0.21]
JEPA 100	0.630	0.620	-0.009 $[-0.028, +0.010]$	N.E.	-0.003	0.987/0.885	2.57	-0.022 (0/12)	-0.121 $[-0.159, -0.082]$	0.11 [0.07, 0.19]
IET 10	0.630	0.620	-0.009 $[-0.050, +0.031]$	N.E.	-0.019	0.987/0.954	9.67	-0.018 (0/12)	-0.123 $[-0.161, -0.088]$	1.17 [1.11, 1.24]
IET 100	0.630	0.520	-0.109 $[-0.152, -0.066]$	DEG.	-0.079	0.987/0.792	33.76	$+0.010$ (0/12)	-0.096 $[-0.126, -0.065]$	0.67 [0.65, 0.69]

Table 10: Temporal gain over the matched initialization on each branch: paired-bootstrap mean, 95% interval, and Holm-adjusted p over the three branches. Initialization accuracy is 0.630 ($\mathbf{h}$), 0.618 ($\mathbf{s}$), and 0.634 (IET $\mathbf{d}$). For single-latent encoders $d \equiv s$.

Encoder	Trunk $\mathbf{h}$	Latent $\mathbf{s}$	Dynamic latent $\mathbf{d}$
Static SSL 10	-0.051 $[-0.090, -0.012]$, 0.036	-0.045 $[-0.090, -0.001]$, 0.090	$\equiv s$
Static SSL 100	-0.027 $[-0.067, +0.013]$, 0.306	-0.035 $[-0.079, +0.006]$, 0.306	$\equiv s$
Patient-CL 10	-0.014 $[-0.053, +0.023]$, 0.498	-0.041 $[-0.084, +0.003]$, 0.231	$\equiv s$
Patient-CL 100	-0.009 $[-0.052, +0.032]$, 0.709	-0.097 $[-0.144, -0.049]$, <0.001	$\equiv s$
Temporal SSL 10	$+0.020$ $[-0.010, +0.052]$, 0.621	$+0.001$ $[-0.026, +0.028]$, 1.000	$\equiv s$
Temporal SSL 100	-0.010 $[-0.040, +0.019]$, 1.000	$+0.005$ $[-0.026, +0.036]$, 1.000	$\equiv s$
JEPA 10	-0.013 $[-0.030, +0.004]$, 0.179	-0.033 $[-0.054, -0.010]$, 0.018	$\equiv s$
JEPA 100	-0.009 $[-0.028, +0.010]$, 1.000	-0.010 $[-0.037, +0.020]$, 1.000	$\equiv s$
IET 10	-0.009 $[-0.050, +0.031]$, 0.682	-0.034 $[-0.074, +0.009]$, 0.254	-0.036 $[-0.071, 0.000]$, 0.156
IET 100	-0.109 $[-0.152, -0.066]$, <0.001	-0.104 $[-0.152, -0.055]$, <0.001	-0.119 $[-0.166, -0.075]$, <0.001

E Additional Benchmark Results

E.1 Full Longitudinal Information Profile

Table 9 gives the complete profile behind the summary in Table 1. It adds the temporal-order accuracy of each encoder at initialization and after training, the nuisance-controlled temporal gain $G^{\perp}_{\text{temp}}$, and same-knee identity retrieval R@1 before and after training. Two patterns are worth noting alongside the main text. Absolute accuracy tracks the gain poorly: Temporal SSL reaches the highest post-training accuracy (0.650) yet its gain over initialization is not distinguishable from zero, whereas encoders with lower accuracy show the same flat gains. Removing the measured low-level statistics does not uncover a hidden temporal gain either, since every $G^{\perp}_{\text{temp}}$ remains at or below zero.

E.2 Temporal Evaluation Across All Branches and Residualization Controls

Table 10 extends the temporal column of Table 1 to the projected latents. For single-latent objectives, $\mathbf{s}$ and $\mathbf{d}$ are the same tensor; for Static SSL and JEPA, this tensor is a fixed random projection of the trained trunk (Appendix C.2). Across all 22 distinct branch-level contrasts, no interval lies above zero. The significant losses are Static SSL 10 on the trunk, Patient-CL 100 and JEPA 10 on the projected latent, and all three IET 100 branches. The initialization accuracy of IET's dynamic latent (0.634) differs from that of the other latents (0.618) because its head is a separately initialized network; comparing each encoder with its own initialization is what makes these contrasts valid. IET's dynamic latent $\mathbf{d}$, which is explicitly trained to encode temporal change, loses accuracy at both budgets: its gain is -0.036 $[-0.071, 0.000]$ at 10 epochs and -0.119 $[-0.166, -0.075]$ at 100 epochs.

Table 11: Residualization controls on the trunk. Columns 2–4: change in order accuracy caused by removing the battery directions, a permuted battery, and a rank- and variance-matched random subspace (paired bootstrap mean and 95% interval for the battery). Last column: pre-registered specificity contrast (battery minus matched random subspace) with verdict.

Representation	Battery removed	Permuted	Random subspace	Specificity (verdict)
Random init	$-0.025\,[-0.052, +0.002]$	-0.036	-0.009	–
Static SSL 10	$-0.017\,[-0.054, +0.020]$	-0.021	-0.014	$-0.002\,[-0.048, +0.041]$ (inconcl.)
Static SSL 100	$-0.039\,[-0.074, -0.003]$	-0.029	-0.040	$+0.002\,[-0.035, +0.037]$ (inconcl.)
Patient-CL 10	$-0.033\,[-0.073, +0.006]$	-0.031	-0.016	$-0.017\,[-0.059, +0.026]$ (inconcl.)
Patient-CL 100	$-0.049\,[-0.080, -0.016]$	-0.049	-0.060	$+0.011\,[-0.023, +0.045]$ (inconcl.)
Temporal SSL 10	$-0.054\,[-0.090, -0.017]$	-0.044	-0.039	$-0.014\,[-0.056, +0.026]$ (inconcl.)
Temporal SSL 100	$-0.029\,[-0.066, +0.008]$	-0.037	-0.005	$-0.024\,[-0.061, +0.014]$ (inconcl.)
JEPA 10	$-0.017\,[-0.046, +0.012]$	-0.029	-0.108	$+0.091\,[+0.054, +0.129]$ (different)
JEPA 100	$-0.018\,[-0.047, +0.011]$	-0.027	-0.007	$-0.011\,[-0.049, +0.027]$ (inconcl.)
IET 10	$-0.035\,[-0.063, -0.005]$	-0.025	-0.017	$-0.018\,[-0.049, +0.013]$ (inconcl.)
IET 100	$+0.006\,[-0.014, +0.026]$	$+0.001$	$+0.008$	$-0.002\,[-0.023, +0.021]$ (inconcl.)

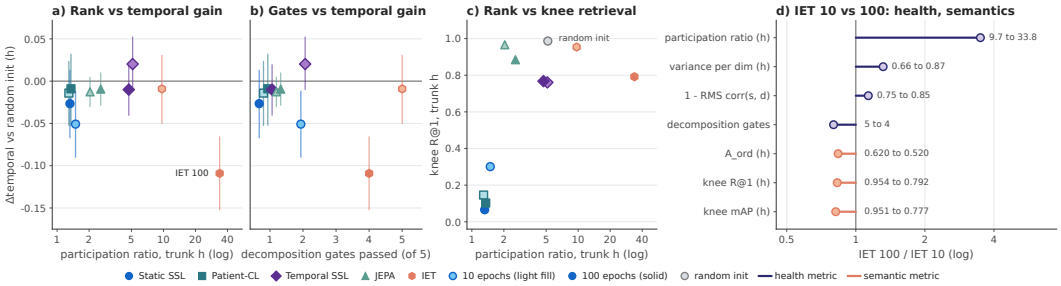

Figure 10: **Representation health does not predict information content (trunk h, one participation-ratio estimator for all encoders).** (a) Participation ratio against the temporal gain with 95% intervals. (b) Participation ratio against same-knee retrieval. (c) IET from 10 to 100 epochs: ratio of each health metric and each semantic metric, with raw values. (d) Decomposition gates passed against the temporal gain. The healthiest encoder loses the most temporal information, and rank tracks identity retention rather than learned information.

Table 11 reports the residualization controls from Appendix D.5 for the trunk embeddings. Battery residualization lowers order accuracy for the untrained network (-0.025) and for nine of the ten encoders. However, similar reductions occur with the permutation control, which removes no battery information, and with the rank- and variance-matched random subspace for most encoders. The pre-registered specificity contrast is therefore INCONCLUSIVE for nine encoders and DIFFERENT only for JEPA 10, where random-subspace removal causes a larger reduction. Battery residualization does not reveal any hidden temporal gain (Table 1), but these controls do not show that the observed reductions are specific to the measured statistics. The earlier pre-registered O21 analysis found stronger evidence of specificity on the full 1,380-patient cohort for an earlier IET configuration (Table 8). For IET 100, residualization changes order accuracy by only $+0.006$, indicating little sensitivity to removal of the battery directions.

E.3 Representation Geometry and Information Content

We further examine whether common measures of representation geometry reflect the information acquired during training. Across the ten encoders, participation ratio is associated with same-knee retrieval (Spearman $\rho = +0.67$, $p = 0.033$), but not with temporal gain ($\rho = +0.16$, $p = 0.65$), and is negatively associated with phenotype gain ($\rho = -0.66$, $p = 0.038$). IET provides the clearest example: from 10 to 100 epochs, its participation ratio increases from 9.67 to 33.76, while its temporal gain decreases from -0.009 to -0.109. These results show that healthier geometry can coexist with weaker acquired information and may reflect some information axes, such as identity retention, better than others.

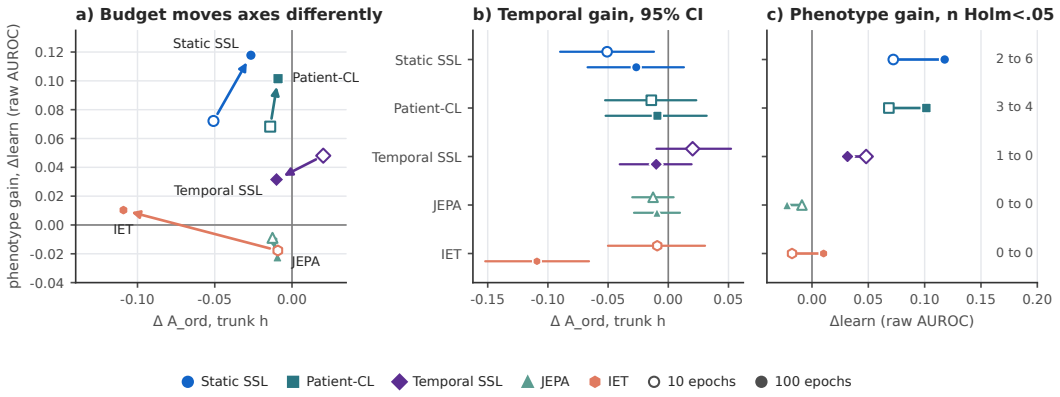

Figure 11: Budget reshapes information profiles. (a) Each objective from 10 epochs (hollow) to 100 epochs (filled) in the plane of temporal gain and phenotype gain. (b) Temporal gain with 95% intervals at both budgets. (c) Phenotype gain at both budgets with the number of Holm-significant labels. Increasing the budget improves one axis while degrading another, and the direction depends on the objective.

E.4 Optimization Budget and Information Profiles

The full objective-level comparison shows that the effect of training budget depends on the information axis. For IET, increasing the budget from 10 to 100 epochs reduces temporal gain from -0.009 to -0.109, same-knee retrieval from 0.954 to 0.792, and controlled-change selectivity from 1.17 to 0.67, while the phenotype-gain point estimate moves from -0.018 to $+0.010$. Static SSL and Patient-CL show larger phenotype gains ($+0.072 \rightarrow +0.118$ and $+0.068 \rightarrow +0.102$) and become strongly synthetic-anatomy-selective ($S : 0.23 \rightarrow 2.33$ and $0.31 \rightarrow 2.26$), while their identity retrieval decreases ($0.301 \rightarrow 0.066$ and $0.146 \rightarrow 0.102$). Temporal SSL shows a smaller phenotype gain at 100 epochs ($+0.048 \rightarrow +0.032$) with similar retrieval ($0.761 \rightarrow 0.768$), whereas JEPA decreases in both retrieval ($0.966 \rightarrow 0.885$) and phenotype gain ($-0.009 \rightarrow -0.022$). In point estimates, the five objectives therefore produce four distinct patterns across the temporal, identity, phenotype, and controlled-change axes; only Static SSL and Patient-CL share one.

E.5 Phenotype Subspace and Reference Analyses

We further examine how phenotype information is distributed within the learned representation. For Static SSL at 10 epochs, the leading eight principal directions recover the full-trunk performance (AUROC 0.604), whereas directions beyond the first sixteen are near chance (0.512). The first principal component alone explains 86.3% of the representation variance but reaches only 0.526 AUROC. Phenotype information is therefore concentrated in a small subspace without simply following the dominant variance direction.

Combining the learned representation with the low-level battery provides no validated additive benefit. The best gains are $+0.009$ at 10 epochs and $+0.023$ at 100 epochs, with no label surviving Holm correction. A separately trained same-backbone supervised model reaches 0.548 raw AUROC on the same held-out patients. Because this result is not acquisition-residualized, we report it only as a reference rather than an upper bound. Appendix E.9 gives the full subspace and fusion tables, and Appendix E.7 the geometric descriptors of every encoder.

E.6 Identity Retention

Table 12 reports same-knee retrieval together with the geometric descriptors. The untrained network retrieves the correct knee almost perfectly ($R@1 = 0.987$ on the trunk), so every objective reduces identity retrieval. The reduction is mild for JEPA and IET 10 and severe for Static SSL and Patient-CL, whose trunks are nearly collapsed. Low identity retrieval is not a defect for a longitudinal representation in itself; what matters for our argument is that it co-varies with rank (Section 5.3) and in the opposite direction from phenotype acquisition.

Table 12: Same-knee retrieval and geometric descriptors on the trunk $\mathbf{h}$ and the latent $\mathbf{s}$. PR, r_{eff}, and mean cosine use the single estimator of Appendix D.10.

	R@1		mAP		PR		r_{eff}		cos	
Encoder	**h**	**s**	**h**	**s**	**h**	**s**	**h**	**s**	**h**	**s**
Random init	0.987	0.979	0.992	0.986	5.18	4.68	2.90	2.87	0.999	0.998
Static SSL 10	0.301	0.507	0.243	0.448	1.49	5.95	1.13	2.68	0.983	0.306
Static SSL 100	0.066	0.123	0.084	0.137	1.32	5.34	1.19	2.15	0.732	0.225
Patient-CL 10	0.146	0.052	0.114	0.052	1.29	1.12	1.09	1.06	0.974	0.136
Patient-CL 100	0.102	0.028	0.113	0.050	1.35	1.56	1.29	1.30	0.794	0.128
Temporal SSL 10	0.761	0.640	0.773	0.655	5.13	10.43	3.34	5.42	0.996	0.453
Temporal SSL 100	0.768	0.722	0.766	0.719	4.72	11.67	3.21	6.43	0.993	0.544
JEPA 10	0.966	0.969	0.977	0.979	2.04	4.18	2.59	3.03	1.000	0.462
JEPA 100	0.885	0.902	0.905	0.922	2.57	4.37	2.57	3.04	1.000	0.453
IET 10	0.954	0.922	0.951	0.922	9.67	9.82	5.81	5.26	0.270	0.239
IET 100	0.792	0.745	0.777	0.735	33.76	34.50	9.63	8.90	0.137	0.180

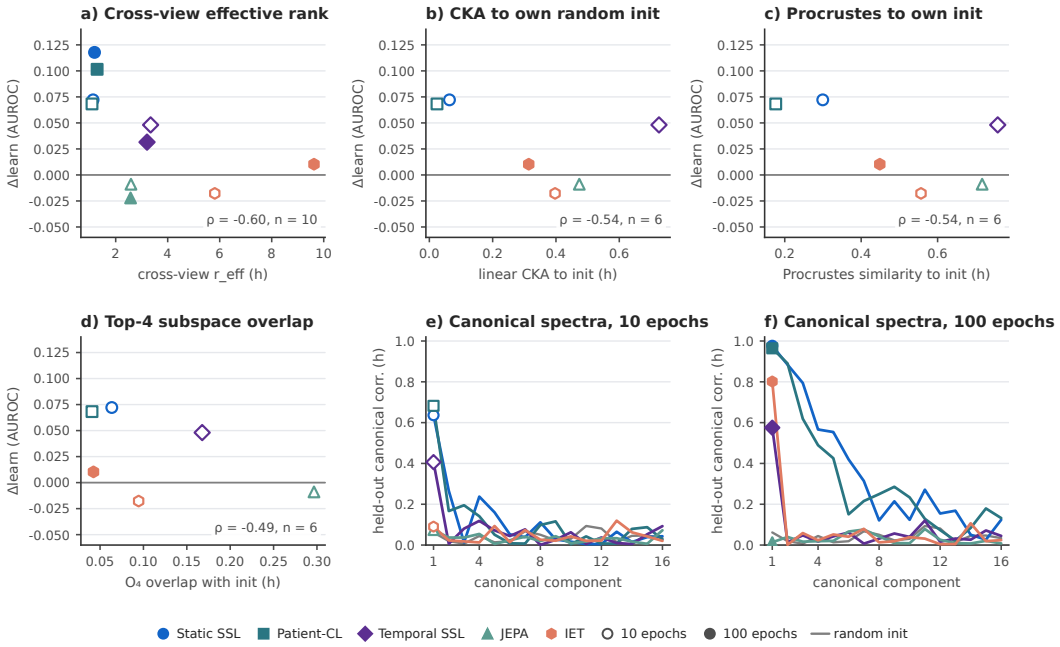

Figure 12: Label-free descriptors against the phenotype gain. (a) Cross-view effective rank (ten encoders). (b–d) Linear CKA, Procrustes similarity, and top-4 subspace overlap with the matched initialization (six encoders). (e, f) Held-out canonical correlations between two augmented views at 10 and 100 epochs, with the untrained network in grey. Spearman ρ is shown in each scatter panel.

E.7 Geometric Descriptors and Similarity to Initialization

Figure 12 tests the geometric and similarity descriptors directly against the phenotype gain. Cross-view effective rank is available for all ten encoders (Spearman $\rho = -0.60$). Linear CKA, Procrustes similarity, and top-4 subspace overlap with the matched initialization were computed for the six encoders of the first benchmark stage and correlate at -0.54, -0.54, and -0.49. None of these relationships is positive, and none is significant at this sample size. The canonical-correlation spectra of the lower panels show that phenotype-acquiring encoders develop several strong shared directions between augmented views at 100 epochs, the observation that motivated the Q_k analysis of Appendix G.4.

Table 13: Acquisition-residualized test-half AUROC per MOAKS label (trunk $\mathbf{h}$). BML: bone marrow lesion; Cart: cartilage; Ost: osteophyte; P/F/T: patella/femur/tibia; Men-M/L: medial/lateral meniscus; Syn: effusion-synovitis.

Features	BML-P	BML-F	BML-T	Cart-P	Cart-F	Cart-T	Ost-P	Ost-F	Ost-T	Men-M	Men-L	Syn	Mean
Battery	0.623	0.626	0.646	0.583	0.663	0.656	0.652	0.661	0.631	0.596	0.500	0.679	0.626
Random init	0.494	0.501	0.507	0.565	0.519	0.513	0.524	0.539	0.517	0.518	0.542	0.507	0.521
Static SSL 10	0.562	0.604	0.618	0.608	0.571	0.630	0.585	0.599	0.567	0.596	0.512	0.642	0.591
Patient-CL 10	0.581	0.589	0.596	0.586	0.583	0.629	0.578	0.522	0.577	0.596	0.518	0.653	0.584
Temporal SSL 10	0.546	0.596	0.598	0.578	0.589	0.561	0.585	0.576	0.526	0.527	0.551	0.621	0.571
JEPA 10	0.458	0.544	0.532	0.543	0.510	0.492	0.487	0.507	0.507	0.510	0.536	0.532	0.513
IET 10	0.493	0.511	0.500	0.552	0.525	0.420	0.499	0.511	0.508	0.495	0.491	0.530	0.503
Static SSL 100	0.619	0.675	0.699	0.643	0.633	0.633	0.624	0.629	0.604	0.638	0.573	0.734	0.642
Patient-CL 100	0.600	0.613	0.665	0.609	0.645	0.608	0.605	0.614	0.576	0.647	0.501	0.657	0.612
Temporal SSL 100	0.514	0.560	0.568	0.564	0.555	0.552	0.602	0.560	0.547	0.502	0.516	0.537	0.548
JEPA 100	0.461	0.508	0.492	0.538	0.542	0.489	0.475	0.489	0.509	0.512	0.535	0.514	0.505
IET 100	0.535	0.544	0.566	0.588	0.527	0.514	0.524	0.486	0.486	0.534	0.515	0.546	0.530

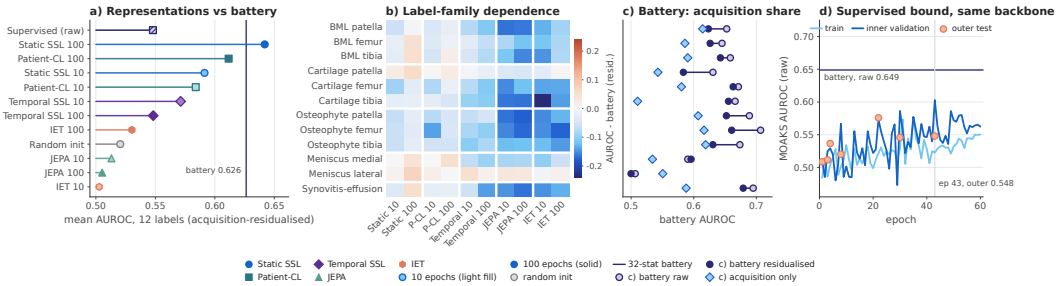

Figure 13: Phenotype information relative to a low-level reference. (a) Acquisition-residualized mean AUROC across twelve MOAKS labels relative to the 32-statistic battery. (b) Per-label differences from the battery, grouped by phenotype family. (c) Battery performance before and after acquisition residualization, with acquisition metadata alone. (d) Same-backbone supervised reference, shown separately because it is not acquisition-residualized. Learned representations and the battery perform better on different phenotype families.

E.8 Per-Label Phenotype Results

Table 13 gives the acquisition-residualized AUROC of every encoder on every label. All ten encoders, their matched initializations, and the battery were evaluated in a single run, so every reference margin in the paper is taken against one battery.

E.9 Subspace Structure of Phenotype Information

Figure 14 and Table 14 summarise where phenotype information resides in the trunk. For Static SSL and Patient-CL at 10 epochs, the AUROC of the leading k principal directions rises to its maximum at $k = 8$ and $k = 16$ respectively and declines slightly as later directions are added, while the directions beyond the leading 16 carry almost none. The first principal direction explains most of the variance of these nearly collapsed trunks (86.3% and 95.3%) and is largely predictable from the battery (R^2 0.92 and 0.90, an optimistic estimate; Appendix D.14), yet on its own it reaches only 0.526 and 0.529. IET 10, whose variance is spread over many directions, shows no concentration at any k. Joining acquisition-residualized trunk directions to the battery yields at most $+0.009$ at 10 epochs and $+0.023$ at 100 epochs with no label significant after Holm correction, while joining the full-width untrained trunk costs -0.048, which illustrates why the leading-direction joins are the fair comparison.

E.10 Multi-Seed Replication

All results in Table 1 use seed 0. To assess seed sensitivity at 10 epochs, we repeated each objective with seeds 1 and 2 using the same configuration and data. This gives fifteen 10-epoch encoders, which were evaluated using a reading rule specified before the additional runs: for each branch, the

Table 14: Phenotype information by trunk subspace. Top: mean raw AUROC of the leading k principal directions and of the directions beyond the leading 16, first-axis variance share, and first-axis AUROC. Bottom: joint (battery plus trunk) minus battery, acquisition-residualized, with the number of Holm-significant labels.

Encoder	$k = 1$	$k = 4$	$k = 8$	$k = 16$	Full	Beyond 16	PC1 variance	PC1 AUROC
Random init	0.493	0.533	0.529	0.526	0.512	0.508	0.290	0.493
Static SSL 10	0.526	0.592	0.604	0.600	0.604	0.512	0.863	0.526
Patient-CL 10	0.529	0.558	0.597	0.608	0.577	0.513	0.953	0.529
IET 10	0.498	0.520	0.524	0.504	0.520	0.506	0.186	0.498

Encoder	10 epochs: leading 8/16/full	100 epochs: leading 8/16/full
Random init	−0.001 (0) / −0.007 (1) / −0.048 (3)	−0.004 (0) / −0.009 (1) / −0.050 (2)
Static SSL	+0.003 (0) / −0.002 (1) / −0.009 (0)	+0.003 (1) / +0.023 (0) / +0.010 (0)
Patient-CL	−0.003 (1) / +0.009 (0) / −0.001 (0)	– / – / −0.001 (0)
IET 10	−0.017 (2) / −0.022 (1) / −0.066 (3)	–

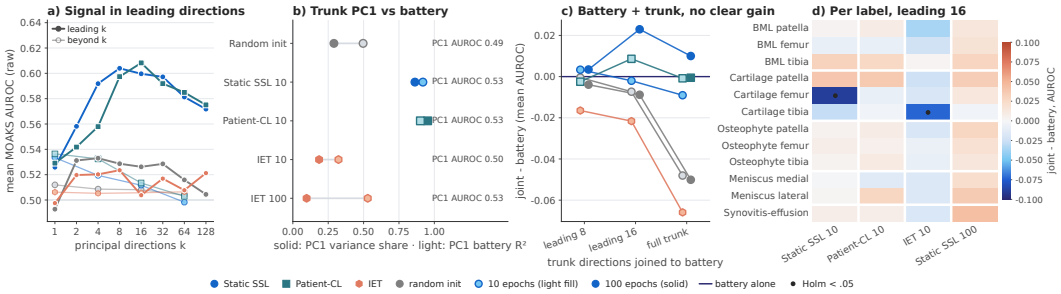

Figure 14: Subspace structure of phenotype information. (a) Mean AUROC of the leading k trunk directions and of the directions beyond k. (b) First principal direction: variance share, battery R^2, and AUROC. (c) Battery plus trunk minus battery by trunk width at 10 and 100 epochs, with the untrained trunk as a dilution control. (d) Per-label joint-minus-battery difference for the leading 16 directions; dots mark Holm $p < 0.05$.

temporal-gain interval is classed as below, containing, or above zero, and a seed-0 class *reproduces* only if seeds 1 and 2 fall in the same class; otherwise it is *seed-sensitive* and is reported, not discarded. Two design points matter. First, the frozen temporal-probe script draws its held-out patients with the encoder's configured seed, which would score each seed on different patients; it was therefore run through a wrapper that verifies the script's hash and fixes only that partition at seed 0, and on a seed-0 encoder the wrapper reproduces the original result exactly on all three branches. Second, the seed-0 encoders were trained with an earlier revision of the training code, so a difference between seed 0 and seeds 1–2 may reflect code as well as seed, whereas a difference between seeds 1 and 2 cannot.

Temporal gain. The central claim of Section 5.2 reproduces without exception: across fifteen encoders and 33 branch-level contrasts, no interval lies above zero (Table 16). The matched initialization orders visits at every seed (trunk A_{ord} 0.630, 0.601, and 0.619 for seeds 0, 1, and 2). On the trunk the verdict reproduces for four of five objectives, all of whose intervals contain zero at every seed. The one seed-0 trunk loss, Static SSL at -0.051, is seed-sensitive: seeds 1 and 2 give $+0.011$ and $+0.010$. Temporal SSL's trunk point estimate is positive at every seed ($+0.020$, $+0.033$, $+0.025$) while every interval contains zero. On the projected latents, classes move between containing and below zero for Static SSL, Patient-CL, JEPA, and IET, never above; 13 of these 15 point estimates are negative. The significance of latent-level losses is therefore seed-dependent, but no seed shows a gain.

Identity, phenotype, and geometry. Table 17 reports the remaining axes. The ordering of identity retention reproduces: JEPA and IET retain knee identity best at every seed (trunk $R@1$ between 0.949 and 0.986), Temporal SSL is intermediate (0.750–0.805), and Static SSL and Patient-CL are

Table 15: Robustness summary across seeds 0/1/2 at 10 epochs (trunk $\mathbf{h}$). Temporal: class of the temporal-gain interval per seed (below zero, contains zero, above zero). Phenotype: sign of G_{phen} per seed. Identity: rank of same-knee $R@1$ among the five objectives per seed.

Objective	Temporal (0/1/2)	Temporal verdict	Phenotype sign (0/1/2)	Identity rank (0/1/2)
Static SSL	below/contains/contains	seed-sensitive	$+/+/+$	$4/4/4$
Patient-CL	contains/contains/contains	reproduces	$+/+/+$	$5/5/5$
Temporal SSL	contains/contains/contains	reproduces	$+/+/+$	$3/3/3$
JEPA	contains/contains/contains	reproduces	$-/-/-$	$1/2/1$
IET	contains/contains/contains	reproduces	$-/-/-$	$2/1/2$
All	no interval above zero at any seed	4/5 reproduce	signs reproduce 5/5	order stable

Table 16: Three-seed replication of the temporal gain at 10 epochs, with validation patients fixed across seeds: paired-bootstrap mean and 95% interval per seed, and the pre-specified reproducibility verdict. For single-latent encoders $d \equiv s$.

Objective	Branch	Seed 0	Seed 1	Seed 2	Verdict
Static SSL	$\mathbf{h}$	$-0.051\ [-0.090, -0.012]$	$+0.011\ [-0.028, +0.048]$	$+0.010\ [-0.026, +0.044]$	seed-sensitive
Static SSL	$\mathbf{s}$	$-0.045\ [-0.090, -0.001]$	$-0.027\ [-0.065, +0.012]$	$+0.007\ [-0.028, +0.042]$	seed-sensitive
Patient-CL	$\mathbf{h}$	$-0.014\ [-0.053, +0.023]$	$+0.019\ [-0.012, +0.048]$	$+0.008\ [-0.029, +0.044]$	reproduces
Patient-CL	$\mathbf{s}$	$-0.040\ [-0.084, +0.003]$	$-0.009\ [-0.042, +0.023]$	$-0.068\ [-0.108, -0.030]$	seed-sensitive
Temporal SSL	$\mathbf{h}$	$+0.020\ [-0.010, +0.052]$	$+0.033\ [-0.003, +0.069]$	$+0.025\ [-0.009, +0.058]$	reproduces
Temporal SSL	$\mathbf{s}$	$+0.001\ [-0.026, +0.028]$	$+0.009\ [-0.023, +0.040]$	$+0.011\ [-0.017, +0.037]$	reproduces
JEPA	$\mathbf{h}$	$-0.013\ [-0.030, +0.004]$	$+0.019\ [-0.001, +0.038]$	$-0.006\ [-0.027, +0.014]$	reproduces
JEPA	$\mathbf{s}$	$-0.033\ [-0.054, -0.010]$	$+0.003\ [-0.016, +0.022]$	$-0.023\ [-0.050, +0.003]$	seed-sensitive
IET	$\mathbf{h}$	$-0.009\ [-0.050, +0.031]$	$+0.006\ [-0.033, +0.048]$	$+0.003\ [-0.034, +0.040]$	reproduces
IET	$\mathbf{s}$	$-0.034\ [-0.074, +0.009]$	$-0.061\ [-0.099, -0.023]$	$-0.009\ [-0.047, +0.029]$	seed-sensitive
IET	$\mathbf{d}$	$-0.036\ [-0.071, 0.000]$	$-0.043\ [-0.082, -0.005]$	$-0.007\ [-0.040, +0.027]$	seed-sensitive

lowest. The seed-0 Static SSL and Patient-CL encoders are more collapsed than their seed-1/2 counterparts (trunk PR 1.49 against 3.78/3.27, and 1.29 against 1.93/1.62) and retrieve correspondingly worse (0.301 against 0.588/0.633, and 0.146 against 0.480/0.309), in the direction of the positive association between rank and identity retention reported in Section 5.3. The sign of the phenotype gain reproduces for all five objectives: positive for Static SSL, Patient-CL, and Temporal SSL, and negative for JEPA and IET. IET has the highest participation ratio at every seed. Magnitudes vary substantially between seeds (for example $+0.015$ to $+0.067$ for Patient-CL), so single-seed differences between objectives should be read only where they are large. The 100-epoch encoders were not replicated and remain single-seed.

Table 15 condenses the replication into the three properties the main text relies on.

F IET Stress-Test Ablations

F.1 Design and Motivation

IET is a longitudinal self-supervised objective that we use as a stress test rather than propose as a method. Its representation is split into a subject-stable factor $\mathbf{s}$, meant to stay constant across a knee's visits, and a dynamic factor $\mathbf{d}$, meant to move predictably with elapsed time so that $\mathbf{d}_{t+\Delta} \approx T(\mathbf{d}_t, \Delta)$. IET was built before LRA, and LRA grew out of the need to check whether this design actually acquired the longitudinal information it was meant to encode. Because the objective encodes strong prior beliefs about how a longitudinal representation should behave, it is a demanding test case for an evaluation that asks what training adds rather than what a representation can decode.

The objective combines four components, each chosen to isolate one requirement and to make its failure visible; the full loss terms are given in Appendix C.2. First, the stable and dynamic factors come from two separate projection heads rather than a single latent, so that the invariance and equivariance terms act on different subspaces and cannot be satisfied by simply collapsing one into the other; a single-latent variant is retained in the ablations for comparison. Second, a cross-visit invariance term pulls together the stable factors of different visits of the same knee. This is what

Table 17: Three-seed replication at 10 epochs (trunk $\mathbf{h}$) of the non-temporal axes, all fifteen encoders re-measured together: same-knee retrieval, phenotype gain over the matched initialization, participation ratio, and Q_4, for seeds 0/1/2.

Objective	R@1 (0/1/2)	G_{phen} (0/1/2)	PR (0/1/2)	Q_4 (0/1/2)
Static SSL	$0.301/0.588/0.633$	$+0.076/+0.038/+0.057$	$1.49/3.78/3.27$	$0.288/0.160/0.135$
Patient-CL	$0.146/0.480/0.309$	$+0.067/+0.015/+0.030$	$1.29/1.93/1.62$	$0.296/0.041/0.176$
Temporal SSL	$0.761/0.750/0.805$	$+0.045/+0.052/+0.029$	$5.13/4.65/5.33$	$0.153/0.112/0.070$
JEPA	$0.966/0.949/0.986$	$-0.010/-0.020/-0.009$	$2.04/2.42/2.68$	$0.050/0.026/0.034$
IET	$0.954/0.958/0.949$	$-0.018/-0.008/-0.004$	$9.67/9.63/9.15$	$0.036/0.041/0.054$

separates IET from view-only invariance and is intended to concentrate subject identity in $\mathbf{s}$. Third, a time-equivariant transition $T(\mathbf{d}, \Delta) = \mathbf{d} + \mathbf{W}_{\text{out}}\mathbf{u}$, conditioned on the elapsed time through FiLM parameters and zero-initialised so that T begins as the identity, asks $\mathbf{d}$ to move consistently with the time gap rather than merely to be classifiable by order; we prefer equivariance to order classification alone because, as the observations below show, visit order is already decodable at initialization and so cannot by itself certify that anything was learned. An order-prediction head supplies a direct temporal-direction gradient alongside the transition. Fourth, variance and covariance penalties applied to the encoder outputs prevent dimensional collapse. We place these penalties on the encoder rather than on an exponential moving average of the covariance statistics because an early configuration that used the moving average passed only a small fraction of the anti-collapse gradient and collapsed, whereas the encoder-level penalties removed the collapse.

Three early observations shaped both the preprocessing of Appendix B.2 and the matched-initialization control. First, before registration and harmonisation, an untrained encoder ordered visits of the same knee with accuracy between 0.755 and 0.785 depending on the branch, so the temporal probe could be passed without any learning. Second, the order signal was traced successively to knee positioning, removed by registration to a common pose; to zero-voxel padding, removed by common-support masking; and to intensity and texture statistics, which neither harmonisation, degradation augmentation at any strength, nor sharpness and noise matching could remove. Third, the collapse described above was diagnosed and corrected by moving the variance and covariance penalties onto the encoder. The resulting representation passed every health check yet failed its semantic gates, and this gap is what led us to measure gains against the matched initialization instead of trusting the health checks.

F.2 Ablations

Table 18 lists the ablation series, each a single change relative to a control and each compared with its matched initialization. Four results bear directly on the claims of the main text. *The same pattern appears in the larger pre-registered O21 analysis*: on the full 1,380-patient diagnostic cohort, the trunk, stable, and dynamic latents all fell below their matched initialization by -0.048, -0.083, and -0.072 (all Holm $p \leq 0.001$). *Removing every temporal objective changes nothing*: arms without the equivariance and order losses lose as much dynamic-latent order accuracy as the full objective. *High rank is not information preservation*: an arm trained only with anti-collapse terms reached the highest dynamic-latent participation ratio among the full-length ablation arms (19.0; the separate three-epoch variance-weight check below reached 22.0) while losing -0.098 on $\mathbf{d}$. *The anti-collapse penalty is the proximate cause of the loss, and mis-tuned rather than a trade-off*: reducing the encoder variance weight tenfold improved order accuracy on every branch and simultaneously raised the dynamic-latent rank from 11.2 to 22.0. A zero-gradient control that runs the full pipeline without updating any parameter reproduces the initialization accuracy exactly (0.6299, interval $[0, 0]$ for the difference), and running only BatchNorm statistics through training moves the latents by -0.031 ($\mathbf{s}$) and -0.022 ($\mathbf{d}$), which bounds the contribution of normalization drift to any gain. The final configuration, with encoder variance weights $25/2/1$, is the IET 10 encoder of the benchmark: no branch is significantly below initialization, and none is above.

Two pre-registered attempts to create the intended stable/dynamic specialisation are shown in Figure 15. A decorrelation penalty between $\mathbf{s}$ and $\mathbf{d}$ reduced their root-mean-square cross-correlation monotonically ($0.249 \rightarrow 0.215 \rightarrow 0.178$) while the intended temporal split $A_{\text{ord}}(\mathbf{d}) - A_{\text{ord}}(\mathbf{s})$

Table 18: IET ablation series. Δ denotes order-accuracy gain over the matched initialization unless a comparison arm is named; branch in parentheses.

Arm (change)	Result	Reading
O21, pre-registered, 1,380 patients	Δ **h** -0.048 $[-0.075, -0.022]$; **s** -0.083 $[-0.113, -0.052]$; **d** -0.072 $[-0.102, -0.040]$; battery chronology 0.674 $[0.653, 0.697]$	training suppresses decodable order
Weak temporal view	vs control: **h** -0.060 (Holm 0.012), **s** -0.054 (0.028)	equivariance loses its target
Equivariance gradient fixed	vs control: **d** $+0.015$ (Holm 1.000)	loss fixed, outcome unchanged
No temporal losses (with and without dynamic augmentation)	$\Delta(\mathbf{d})$ -0.095 and -0.096 vs control -0.111; all Holm $p > 0.37$	temporal losses irrelevant
Anti-collapse terms only	$\Delta(\mathbf{d})$ -0.098 (Holm $p < 0.001$); PR **d** 19.0 vs init 5.5	high rank, lost order
Encoder variance weight $\div 10$ (3 epochs)	vs weight 250: **h** $+0.050$, **s** $+0.091$, **d** $+0.060$ (Holm ≤ 0.006); PR **d** 22.0 vs 11.2	penalty mis-tuned
Zero-gradient control	$A_{\text{ord}}(\mathbf{h}) = 0.6299$, difference interval $[0, 0]$; BatchNorm drift **s** -0.031, **d** -0.022	matched init reproduces
Final IET (weights $25/2/1$)	Δ **h** -0.009, **s** -0.034, **d** -0.036 $[-0.071, 0.000]$	neutral, not helpful
Decorrelation penalty (O25)	RMS corr $0.249/0.215/0.178$; split $+0.015/+0.013/-0.027$; strong: $\Delta(\mathbf{d})$ -0.091	decorrelation $\neq$ specialisation
Adversarial heads (O26)	split $+0.043$ $[+0.003, +0.081]$; $\Delta(\mathbf{s})$ -0.074 (Holm < 0.001); $\Delta(\mathbf{d})$ -0.047	split by degrading **s**

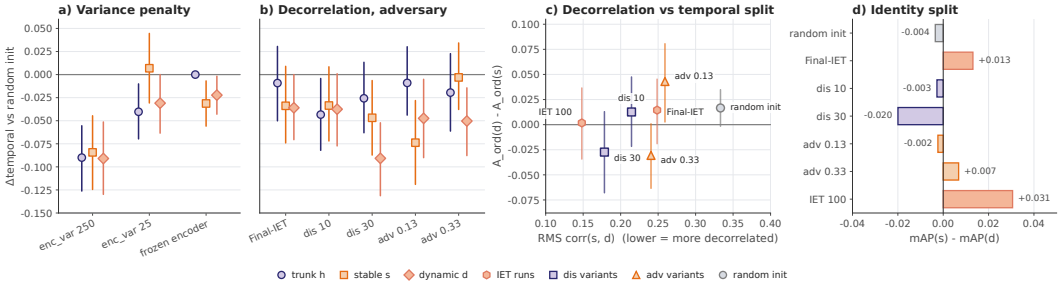

Figure 15: Mechanisms within the IET series. (a) Temporal gain per branch under different encoder variance penalties. (b) Temporal gain per branch for the final configuration and the decorrelation and adversarial variants. (c) Cross-correlation between **s** and **d** against the temporal split with 95% intervals. (d) Identity split $\text{mAP}(\mathbf{s}) - \text{mAP}(\mathbf{d})$. Lower correlation does not produce a better split.

shrank and then inverted ($+0.015 \to +0.013 \to -0.027$), and the stronger setting lost -0.091 on **d** (Holm $p < 0.001$). Adversarial heads produced a significant split ($+0.043$ $[+0.003, +0.081]$) but did so by degrading the stable latent (-0.074, Holm $p < 0.001$) rather than improving the dynamic one (-0.047). Decorrelation is not specialisation, and a split can be manufactured by destroying information on one side.

F.3 The Budget Test

A natural objection to every negative result above is under-training. It was tested by a pre-registered run of the final configuration for 100 epochs, identical in every resolved setting except the epoch count and fully annealed like the 10-epoch run, and read with the frozen analysis script. The answer was unambiguous. Temporal gains fell on every branch (Table 10). Identity retrieval on the latents fell from $0.922/0.910$ to $0.745/0.725$, while every health indicator improved: participation ratio on the diagnostic gallery rose from $10.9/13.2$ to $29.0/34.6$ for **s**/**d** (a different estimator from Table 12),

Table 19: Controlled-change selectivity S (Equation 9) with 95% patient-clustered intervals at the highest matched level. For single-latent encoders $d \equiv s$. The untrained network is unresolved (both mean distances below 10^{-4}).

Encoder	Trunk $\mathbf{h}$	Latent $\mathbf{s}$	Dynamic latent $\mathbf{d}$
Random init	$0.070\ [0.067, 0.073]$	$0.071\ [0.068, 0.074]$	–
Static SSL 10	$0.233\ [0.219, 0.250]$	$0.154\ [0.145, 0.163]$	$\equiv s$
Static SSL 100	$2.329\ [2.115, 2.556]$	$1.780\ [1.656, 1.911]$	$\equiv s$
Patient-CL 10	$0.311\ [0.287, 0.338]$	$0.526\ [0.462, 0.595]$	$\equiv s$
Patient-CL 100	$2.258\ [2.076, 2.456]$	$2.296\ [2.015, 2.603]$	$\equiv s$
Temporal SSL 10	$0.063\ [0.059, 0.066]$	$0.062\ [0.057, 0.066]$	$\equiv s$
Temporal SSL 100	$0.121\ [0.112, 0.129]$	$0.158\ [0.149, 0.168]$	$\equiv s$
JEPA 10	$0.083\ [0.041, 0.206]$	$0.231\ [0.215, 0.247]$	$\equiv s$
JEPA 100	$0.115\ [0.069, 0.193]$	$0.159\ [0.145, 0.172]$	$\equiv s$
IET 10	$1.174\ [1.113, 1.235]$	$1.099\ [1.043, 1.153]$	$0.983\ [0.936, 1.031]$
IET 100	$0.671\ [0.645, 0.695]$	$0.652\ [0.629, 0.676]$	$0.662\ [0.639, 0.683]$

the $\mathbf{s}$–$\mathbf{d}$ cross-correlation fell from 0.249 to 0.148, the equivariance gain on training batches rose from 0.335 to 0.404, and no dimension was dead. The decomposition gates passed 5 of 5 at 10 epochs and 4 of 5 at 100, while the underlying temporal split was indistinguishable from that of the untrained network at both budgets. No clinical endpoint improved (Table 22, Appendix H.2). This run is the IET 100 encoder of the benchmark.

G Additional Mechanistic Analyses

G.1 Controlled-Change Selectivity

The nuisance perturbations were calibrated to match the in-support RMS intensity change of the synthetic local deformation at three levels (0.0376, 0.0524, and 0.0652), with matching error within 0.3%. This allows S to compare responses to anatomical and nuisance perturbations of similar image-space magnitude rather than reflecting differences in perturbation strength.

Table 19 reports selectivity for every branch, and Figure 16 shows the corresponding response magnitudes. At 10 epochs, IET is the only model with trunk selectivity above one ($S = 1.17$ $[1.11, 1.24]$), whereas its dynamic latent is not selective ($S(\mathbf{d}) = 0.98$ $[0.94, 1.03]$) and is less selective than its stable latent. This does not support the intended stable–dynamic separation. At 100 epochs, IET trunk selectivity falls to 0.67, while Static SSL and Patient-CL increase from 0.23 to 2.33 and from 0.31 to 2.26, respectively. Temporal SSL and JEPA remain nuisance-dominated at both budgets across all branches.

Distances increase monotonically with perturbation level for all trained encoders (Figure 16c–e), indicating that the selectivity patterns are not driven by a single perturbation level. The wider JEPA trunk intervals arise because its representation changes very little under either perturbation.

These results also limit how S should be interpreted. IET is synthetic-anatomy-selective at 10 epochs despite showing no established phenotype acquisition, so sensitivity to the controlled deformation is not by itself evidence of anatomical or disease semantics. The relationship becomes more consistent at 100 epochs, when the models with the strongest phenotype information also show the strongest selectivity, but it does not hold uniformly across objectives or information axes.

G.2 Sign-Free Change Axis

Table 20 reports the change-axis gain of every encoder. On the projected latents, the gains of Static SSL, Patient-CL, and IET's dynamic latent are statistically indistinguishable at 10 epochs, and the pre-registered test that IET exceeds every baseline failed: IET minus Static SSL is -0.008 $[-0.076, +0.066]$. At 100 epochs Static SSL exceeds IET (-0.107 $[-0.187, -0.029]$, Holm $p = 0.018$). Because the projected latent of Static SSL is a fixed random projection of its trunk, the comparison that isolates a trained head is Patient-CL, which matches IET ($+0.297$ against $+0.295$). On the trunk the picture differs: IET retains a clearly positive gain ($+0.203$) while the baselines are

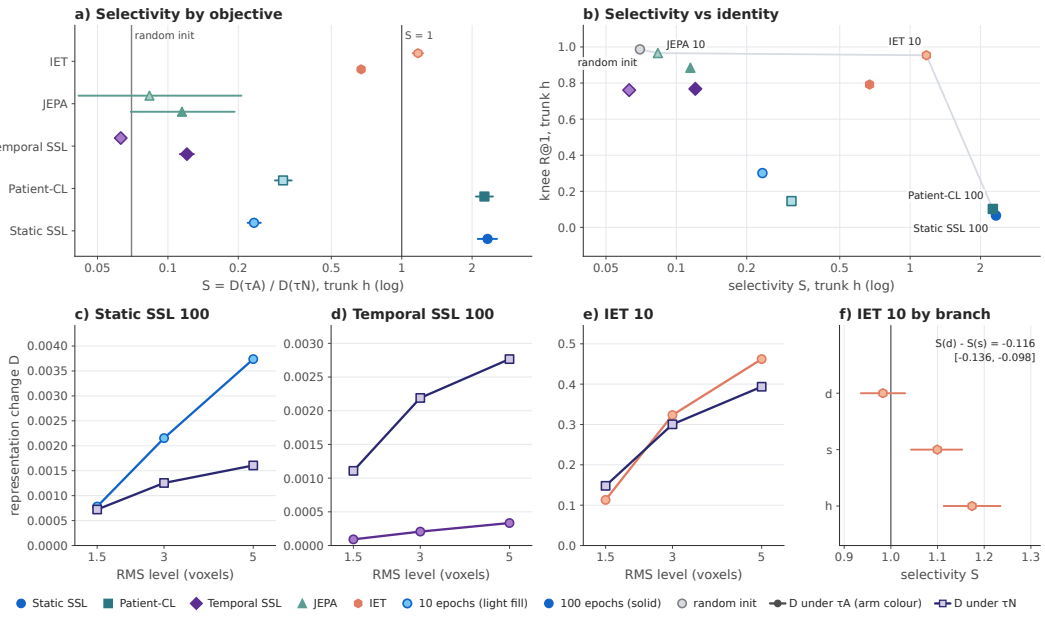

Figure 16: Controlled-change selectivity on the trunk. (a) Selectivity with 95% intervals at both budgets, with reference lines at $S = 1$ and at the untrained network. (b) Selectivity against same-knee retrieval, with the non-dominated set connected. (c–e) Representation distance under anatomical and matched nuisance perturbation across three levels for three encoders. (f) IET 10 selectivity by branch and the dynamic-minus-stable dissociation.

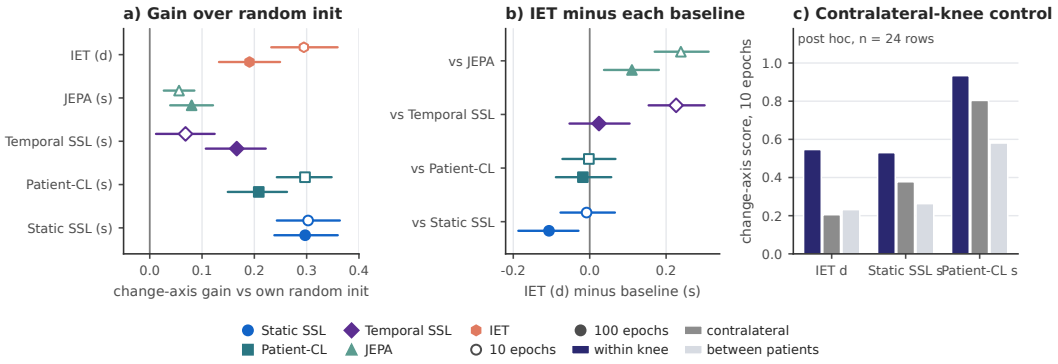

Figure 17: Sign-free change axis on the projected latents (**s** for baselines, **d** for IET). (a) Gain over the matched initialization at 10 and 100 epochs. (b) IET minus each baseline at the same budget. (c) Post hoc contralateral control: alignment within a knee, between a patient's two knees, and between patients.

between -0.034 and $+0.110$. The change axis is therefore not specific to the IET objective, but its expression depends on where in the network it is read.

The post hoc contralateral control of Figure 17c separates two kinds of structure. For IET's dynamic latent, alignment between changes of a patient's two knees (0.205) is no higher than between patients (0.232), so the axis is knee-specific. For Static SSL and Patient-CL, contralateral alignment exceeds between-patient alignment (0.378 against 0.263 and 0.804 against 0.581), so their axis is shared across a patient's knees. With only 24 contralateral rows from 18 patients, this analysis is descriptive, and the number of eligible patients cannot be increased on this dataset. Appendix G.3 summarises what this axis does and does not establish.

Table 20: Sign-free change-axis gain over the matched initialization (94 rows, 95% intervals). Bottom: pre-registered contrasts of IET 10 ($\mathbf{d}$) against each 10-epoch baseline's stronger branch, Holm-corrected.

Encoder	Trunk $\mathbf{h}$	Latent $\mathbf{s}$	Dynamic latent $\mathbf{d}$
Static SSL 10	$+0.033\,[-0.024, +0.090]$	$+0.303\,[+0.243, +0.363]$	$\equiv s$
Static SSL 100	$+0.082\,[+0.022, +0.144]$	$+0.297\,[+0.238, +0.359]$	$\equiv s$
Patient-CL 10	$+0.063\,[+0.013, +0.113]$	$+0.297\,[+0.243, +0.348]$	$\equiv s$
Patient-CL 100	$+0.110\,[+0.055, +0.166]$	$+0.208\,[+0.149, +0.262]$	$\equiv s$
Temporal SSL 10	$-0.023\,[-0.068, +0.027]$	$+0.068\,[+0.012, +0.124]$	$\equiv s$
Temporal SSL 100	$+0.008\,[-0.049, +0.066]$	$+0.166\,[+0.107, +0.222]$	$\equiv s$
JEPA 10	$-0.013\,[-0.034, +0.008]$	$+0.056\,[+0.027, +0.085]$	$\equiv s$
JEPA 100	$-0.034\,[-0.073, +0.008]$	$+0.080\,[+0.039, +0.121]$	$\equiv s$
IET 10	$+0.203\,[+0.135, +0.267]$	$+0.247\,[+0.182, +0.310]$	$+0.295\,[+0.232, +0.359]$
IET 100	$+0.164\,[+0.105, +0.221]$	$+0.211\,[+0.155, +0.263]$	$+0.191\,[+0.132, +0.249]$
IET 10 $\mathbf{d}$ *minus stronger branch of each 10-epoch baseline:*			
Static SSL 10	$-0.008\,[-0.076, +0.066]$		
Patient-CL 10	$-0.002\,[-0.071, +0.067]$		
Temporal SSL 10	$+0.226$		
JEPA 10	$+0.239$		

G.3 Generic Within-Knee Change Axis

We also examined whether repeated scans produce a consistent direction of representation change without assigning a temporal sign to that direction. Such a change axis is detectable after training, but it is not specific to an explicitly temporal objective. At 10 epochs, Patient-CL shows a gain over initialization of $+0.297$, nearly identical to the $+0.295$ gain of IET's dynamic latent (Figure 17). This suggests that the axis reflects a more general effect of learning from repeated scans rather than an IET-specific temporal representation. Per-encoder results are given in Appendix G.2.

G.4 Cross-View Predictivity Q_k

After rank-, similarity-, and overlap-based descriptors failed to track phenotype gain, we examined Q_4 (Appendix D.13) through four pre-registered tests. Table 21 and Figure 18 summarise the results. We report the full sequence because no single test is sufficient to validate Q_4 as a surrogate for phenotype acquisition.

Held-out ordering. Among the six encoders in the first stage, Q_4 correlated with phenotype gain (Spearman $\rho = 0.886$), although the association did not survive correction as the seventh label-free quantity examined (Holm-adjusted $p = 0.133$). Before phenotype gains were computed for the four 100-epoch baselines, their Q_4 values were used to predict the complete ordering of those gains. The predicted ordering was correct (one of 24 possible orderings, $p = 0.042$).

Rank intervention. We next introduced a rank-k bottleneck intended to vary Q_4 without changing participation ratio. Q_4 correctly ordered the phenotype gains of all four intervention arms, but the participation-ratio result depended on its estimator. The pre-registered analysis gave $\rho(\mathrm{PR}, G_{\mathrm{phen}}) = -1.00$ and therefore failed the uniqueness criterion, whereas the estimator used elsewhere in this paper gives $\rho = -0.80$, which would satisfy it. Because the pre-registration did not specify the estimator, we treat this intervention as unresolved.

Budget intervention. Static SSL was retrained for 10, 25, 50, and 100 epochs. Across these intervention arms, Q_4 increased from 0.239 to 0.827 and tracked phenotype gain with $\rho = +0.80$, compared with $|\rho| = 0.20$ for participation ratio, satisfying the pre-registered gate. However, the association is based on only four points (exact permutation $p = 0.333$), and Q_4 is perfectly ordered by training duration. The intervention therefore distinguishes Q_4 from participation ratio but does not distinguish it from training duration. These intervention checkpoints were trained separately from the benchmark checkpoints; device and reproducibility details are given in Appendix C.3.

Seed perturbation. We then retrained each 10-epoch objective with seeds 1 and 2 to vary Q_4 while holding the objective and training budget fixed. The pre-registered adequacy criterion required a

Table 21: Pre-registered tests of Q_4. PR is the single estimator of Appendix D.10 unless marked as training-batch. G_{phen} is the raw phenotype gain over the matched initialization. The rank-intervention and budget-intervention encoders were retrained for these tests on a second GPU and are distinct from the benchmark checkpoints; their Static SSL 10- and 100-epoch gains ($+0.061$, $+0.114$) therefore differ from those of Table 1 ($+0.072$, $+0.118$), which the held-out-ordering row uses.

Test	Values	Outcome
Held-out ordering (four 100-epoch baselines)	Q_4: Static SSL 0.805, Patient-CL 0.741, Temporal SSL 0.164, JEPA 0.022; G_{phen}: $+0.118$, $+0.102$, $+0.032$, -0.022	exact ordering, $p = 0.042$
Rank intervention (Static SSL 10 with rank 32/8/full/128)	Q_4 0.058/0.165/0.239/0.330; G_{phen} $+0.020/+0.046/+0.061/+0.065$; PR 3.48/1.84/1.95/1.22 (training-batch 3.01/1.75/1.64/1.21)	$\rho(Q_4) = +1.00$; $\rho(\text{PR}) = -0.80$ (-1.00 training-batch): unresolved
Budget intervention (Static SSL, 10/25/50/100 epochs)	Q_4 0.239/0.413/0.694/0.827; G_{phen} $+0.061/+0.089/+0.082/+0.114$; PR 1.95/1.17/1.07/1.40	pass: $+0.80$ vs 0.20; $p = 0.333$; collinear with epochs
Seed perturbation (seeds 1/2, five objectives)	between-seed $\lvert\Delta Q_4\rvert$: Static SSL 0.025, Patient-CL 0.135, Temporal SSL 0.042, JEPA 0.008, IET 0.013	uninformative (1/5 adequate)

between-seed difference greater than 0.05 for at least three of the five objectives, but only Patient-CL met this criterion (0.135). The intervention was therefore uninformative. Moreover, had the adequacy criterion been met, the observed association would not have satisfied the uniqueness condition ($\rho = +0.48$ for Q_4 versus $|\rho| = 0.89$ for participation ratio). Seed 0 was excluded from this inference because it came from a different code revision.

Taken together, these tests provide suggestive but incomplete evidence for Q_4. It tracks phenotype acquisition across encoders and under the budget intervention, and in that intervention it behaves differently from participation ratio. However, the rank intervention is estimator-dependent, the seed intervention is underpowered by its pre-registered adequacy criterion, and the budget intervention does not separate Q_4 from training duration. We therefore retain Q_4 as a mechanistic candidate rather than a validated surrogate for phenotype acquisition.

H Clinical Endpoint Analyses

H.1 Protocol

We assess clinical utility by asking how much a representation improves a clinical model beyond baseline severity, rather than by its absolute predictive performance, since baseline severity already strongly predicts every endpoint. The clinical model for dJSM has an AUROC of 0.775 $[0.719, 0.826]$, and for eight-year knee replacement 0.851 $[0.795, 0.904]$. All three analyses use five-fold outer cross-validation grouped by patient with out-of-fold scores, an inner three-fold patient-grouped search for the regularization strength maximising AUROC, and 2,000 patient-clustered paired bootstrap resamples. Representation blocks are reduced to $k = 8$ principal components fitted within training folds (16 as sensitivity), standardized, and scaled by the square root of their width so that no block dominates by dimensionality.

dJSM increment. The clinical design contains one-hot baseline Kellgren–Lawrence grade, baseline medial narrowing grade, and gap indicators; the tested model appends the principal components of the representation change $\Delta\mathbf{z} = \mathbf{z}_{t_2} - \mathbf{z}_{t_1}$. Two matched-capacity controls append the same change permuted within gap strata and Gaussian noise of the same width. Utility requires the increment interval above zero and both control intervals containing zero.

Eight-year knee replacement. A discrete-time hazard model is fitted to the person-period data using yearly intervals j, with hazard $h_j(\mathbf{x}) = \sigma(\gamma_j + \mathbf{w}^\top\mathbf{x})$, where $\sigma(\cdot)$ denotes the logistic function, γ_j is the interval-specific intercept for year j, $\mathbf{x}$ is the covariate vector, and $\mathbf{w}$ is the corresponding coefficient vector. The predicted eight-year risk is $1 - \prod_{j=1}^{8}(1 - h_j(\mathbf{x}))$. Performance is evaluated

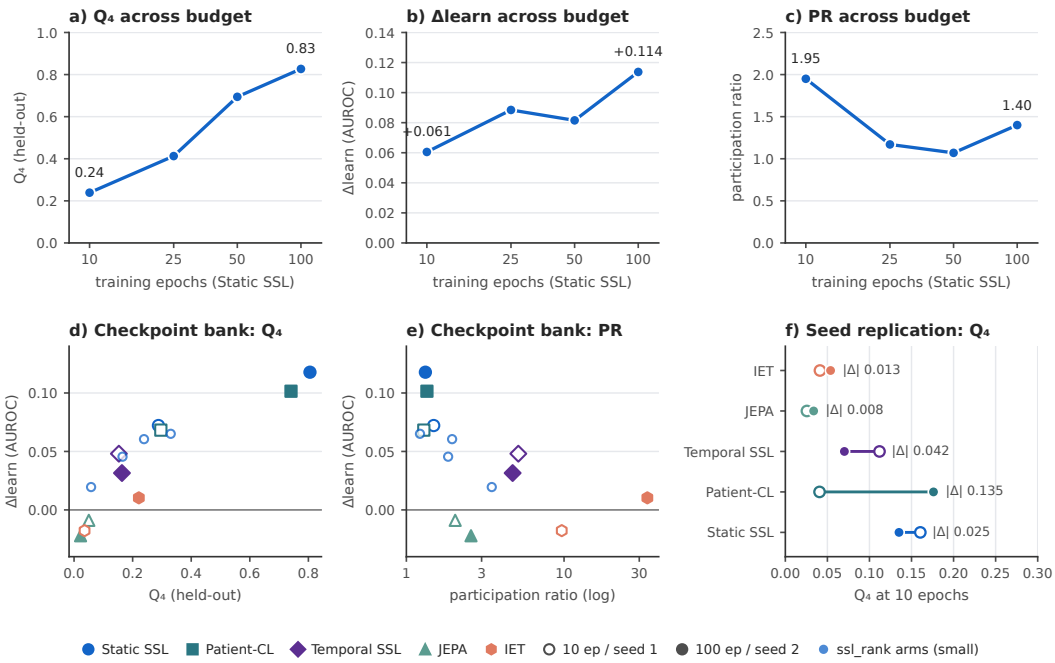

Figure 18: Cross-view predictivity Q_4. (a–c) Budget intervention on Static SSL: Q_4, phenotype gain, and participation ratio at 10, 25, 50, and 100 epochs. (d, e) Fourteen checkpoints: Q_4 and participation ratio against the phenotype gain (hollow: 10 epochs; filled: 100 epochs; small: rank-bottleneck arms). (f) Seed replication: Q_4 for seeds 1 and 2 of each 10-epoch objective.

using the cumulative/dynamic AUROC at eight years, with cases defined as individuals undergoing knee replacement by eight years and controls as those remaining event-free at eight years. For each representation branch, the index-scan representation is appended to the covariate vector and evaluated under the same protocol, with Holm correction applied across branches together with permutation and noise controls.

dCART utility (exploratory). On baseline-to-24-month pairs, the utility contrast is the AUROC of the representation change minus the better of the matched initialization and the battery. The random initialization in this script was not seeded, so its comparator is not the exact matched initialization, which is one reason this endpoint remains exploratory.

Stage 2 was run only for encoders whose temporal gain interval on the projected latent did not lie entirely below zero, which excluded Static SSL 10 and IET 100 from the benchmark stage. IET 100's clinical endpoints were evaluated separately in the pre-registered budget test (Appendix F.3), and IET 10's in earlier pre-registered analyses.

H.2 Results

Table 22 and Figure 19 report the increments. Among the evaluated encoder–endpoint combinations, none shows a statistically supported positive increment; Static SSL 10 was not eligible for stage 2 and IET 10 has no admissible knee-replacement result. Three intervals exclude zero, and all three are losses: IET 10 on dJSM (-0.023 $[-0.047, -0.004]$), IET 100 on dJSM (-0.010 $[-0.021, -0.0005]$), and Static SSL 100 on knee replacement (-0.017 $[-0.036, -0.00002]$, bootstrap $p = 0.05$, Holm $p = 0.150$), the last with an upper bound at the boundary of zero. With 154 progression events and 66 replacement events within eight years, increments smaller than about 0.05 cannot be excluded, so these results establish the absence of a large clinical increment rather than the absence of any.

Table 22: Clinical increments with 95% intervals. dCART: exploratory utility over the better of initialization and battery. dJSM: AUROC of clinical plus representation change minus clinical. TKR: eight-year cumulative/dynamic AUROC of clinical plus trunk minus clinical. †: result files exist although the stage-2 log records a failure for this encoder. n.a.: not run or no admissible result.

Encoder	dCART utility	dJSM increment	TKR increment
Static SSL 10	n.a.	n.a.	n.a.
Static SSL 100	$-0.040\,[-0.120, +0.043]$	$-0.004\,[-0.021, +0.014]$	$-0.017\,[-0.036, -0.00002]$
Patient-CL 10	$-0.020\,[-0.099, +0.060]$	$-0.008\,[-0.024, +0.008]$	$-0.016\,[-0.033, +0.001]$
Patient-CL 100	$-0.020\,[-0.102, +0.061]$	$-0.006\,[-0.034, +0.018]$	$+0.002\,[-0.010, +0.014]$
Temporal SSL 10	$-0.019\,[-0.092, +0.059]$	$+0.005\,[-0.012, +0.024]$	$-0.003\,[-0.023, +0.018]$
Temporal SSL 100†	$-0.006\,[-0.085, +0.073]$	$+0.011\,[-0.009, +0.032]$	$+0.005\,[-0.013, +0.023]$
JEPA 10	$-0.035\,[-0.119, +0.049]$	$-0.006\,[-0.019, +0.008]$	$-0.004\,[-0.027, +0.015]$
JEPA 100	$+0.020\,[-0.056, +0.097]$	$+0.003\,[-0.013, +0.018]$	$-0.024\,[-0.061, +0.002]$
IET 10	$-0.005\,[-0.083, +0.068]$	$-0.023\,[-0.047, -0.004]$	n.a.
IET 100	$+0.002\,[-0.080, +0.083]$	$-0.010\,[-0.021, -0.0005]$	$-0.003\,[-0.022, +0.015]$

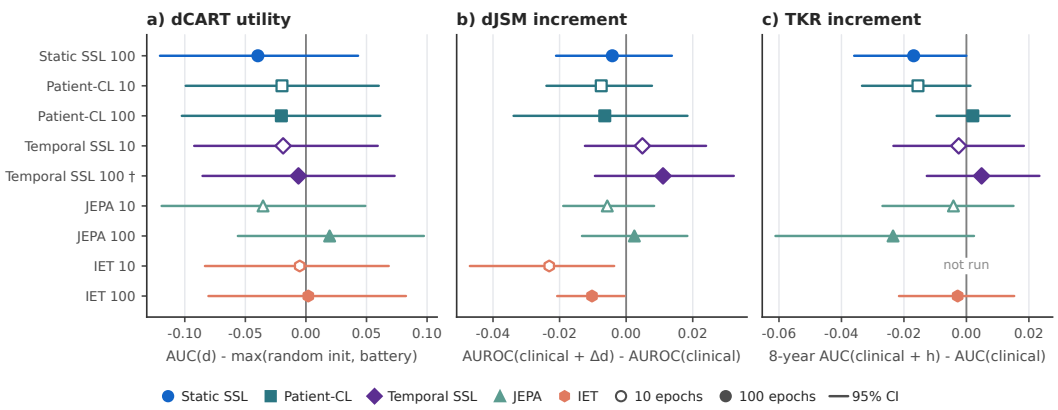

Figure 19: **Clinical increments with 95% intervals for** (a) exploratory dCART utility, (b) dJSM increment over a clinical model, and (c) eight-year knee-replacement increment over a clinical model. Hollow markers: 10 epochs; filled: 100 epochs.

I Discussion and Limitations

I.1 Scope of the Claims

LRA makes a narrow claim with broad consequences. It does not claim that longitudinal self-supervision cannot work, that the five objectives are the best possible instances of their families, or that the benchmark encoders are useless. It claims that the evidence routinely offered for longitudinal learning, namely absolute decodability, representation health, and objective satisfaction, is insufficient, and it shows on a realistic benchmark that this insufficiency changes conclusions. Every encoder that a conventional evaluation would praise for its temporal probe, rank, or equivariance fails to demonstrate a temporal gain over its own initialization, and the encoders that do acquire information acquire it on axes, phenotype, and selectivity for a controlled synthetic anatomical deformation that the conventional diagnostics rank backward.

Why can an untrained network predict visit order? Our results point to two possible sources, although we cannot fully disentangle them: the inductive bias of convolutional patch embedding and attention over a fixed grid, and low-level image drift between visits that survives registration, masking, and harmonization. The battery chronology reference, which reaches the same order accuracy as the untrained network, suggests that the second source is sufficient on its own. Either way, the implication for evaluation is the same, because both sources are present before training and neither is what a longitudinal objective is meant to supply.

The framework is not specific to knee MRI. Its requirements are a reproducible initialization, a patient- or subject-disjoint evaluation partition, information-specific probes, and, for the nuisance-controlled level, measured nuisance descriptors. Settings with repeated measurements of the same units, including other imaging modalities, longitudinal electronic health records, remote sensing time series, and video, meet these requirements.

I.2 Limitations

One study, one MRI sequence, one backbone. All results concern one MRI sequence from one study, one 3D ViT-S backbone, and linear probes. Nonlinear probes might extract more from every representation, including the untrained one, and the matched-initialization contrast would then need to be re-estimated with the same probe.

Seeds and code revision. The 10-epoch encoders were replicated with three seeds (Appendix E.10): the absence of temporal gains, the ordering of identity retention, and the sign of the phenotype gain reproduce, but the significance of individual losses is seed-dependent, and seed 0 was trained with an earlier code revision. The 100-epoch encoders, including IET 100, are single-seed.

Development evidence only. Every number is computed on the development half of the diagnostic cohort; the holdout half is sealed. The procedure that divided the two halves was not recorded as a script.

Labels and covariates. The provenance of the MOAKS export and its binarisation could not be established. The technologist group is an inferred proxy for site. Radiographic and MRI endpoints are known to be discordant, and the clinical analyses can only exclude increments larger than about 0.05.

Nuisance control. The battery captures low-level intensity, sharpness, and field-of-view statistics but not every acquisition factor. On the development cohort its residualization is not demonstrably more specific than removing a matched random subspace (Table 11). The nuisance-controlled temporal gains are therefore best read as a check that no hidden gain appears, not as an attribution of the loss to measured nuisance.

Estimators. Identity retrieval has no per-encoder interval. The participation ratio used across encoders is computed on augmented embeddings of training-cohort scans, and other estimators give different absolute values; we use a single estimator throughout and never compare values across estimators. The trunk-axis battery R^2 is optimistic because its ridge penalty is chosen on the test half.

Benchmark design. The baseline objectives do not include an encoder-level anti-collapse penalty, which helps explain the low rank of their trunk representations. For two of the five objectives, the projection head receives no training signal, so our main comparisons use the shared trunk embedding. The evidence rules in Appendix D.3 were specified after the analyses rather than pre-registered. Finally, the rank-intervention result for Q_4 depends on the participation-ratio estimator.

I.3 Recommendations

Our findings suggest five practical guidelines for evaluating longitudinal representations. (1) Compare each probe result with the matched initialization using the same evaluation protocol, and report a paired interval for the difference. (2) Evaluate several information axes and report them as a profile; do not rank encoders unless one dominates or the weighting is stated. (3) Where nuisance descriptors are measured, report the nuisance-controlled gain together with permutation and matched-subspace controls. (4) Compare against a low-level reference that involves no learning, after the same nuisance control. (5) Treat representation-health diagnostics as descriptors to be validated against gains, not as evidence of quality, and fix their estimator before comparing encoders.

I.4 Reproducibility Notes

The temporal-probe script is frozen by a SHA-256 hash that every evaluation script verifies before execution. All reported contrasts use paired bootstrap resamples shared across conditions, and deltas are bootstrap means. During the consolidation of results for this paper we found and re-

solved several inconsistencies in our internal records, and the resolutions determine the numbers reported here. Participation ratios from four estimators had been reported in different analyses; all cross-encoder values now use one estimator. Two runs of the 100-epoch phenotype probe with and without acquisition metadata had been mixed; the phenotype gain is taken from the probe without acquisition restriction, and every residualized AUROC and reference margin from a single re-run of all ten encoders with acquisition metadata, in which every encoder is compared with one battery. That re-run also replaced a Wilcoxon test over correlated labels with the patient-clustered bootstrap of Appendix D.9. The rank-intervention verdict for Q_4 had been computed with a training-batch participation ratio and is reported here as unresolved. The same-knee relevance criterion of the retrieval probe is stated explicitly, and for the two objectives with untrained projection heads, the main analysis uses the shared trunk embedding.