WHEN ANTI-COLLAPSE OBJECTIVES HAVE COLLAPSED OPTIMA

Anonymous authors
Paper under double-blind review

ABSTRACT

Every JEPA carries collapse-prevention machinery, and a recent line of work replaces the older heuristics with a single distribution-matching objective on the latent (Balestriero & LeCun, 2025; Kuang et al., 2026). We ask when such an objective is safe. An objective that reads the latent only through per-feature scalar statistics has a collapsed global optimum when its entrywise function is convex and non-increasing (softplus, hinge); for ℓ_1 the minimisers are single-entry, rank-one latents. A softplus non-negativity penalty — the form a rectified target takes when written as a penalty — placed at the encoder of tabular T-JEPA (Thimonier et al., 2025) degrades $28/28$ runs under its standard linear probe, 24 to the majority-class share, locked within five epochs. The collapse is geometric rather than informational: the re-examined latents sit on one common vector with a residual of $\sim 2\times 10^{-7}$ of its scale that still carries the label — a standardised probe recovers most of it, yet stored in bf16 or fp16 every input rounds to the same vector. Under a fair probe the placement effect survives but shrinks: 16pp for this penalty (from 26pp), 4–5pp for everyday ℓ_1 and hinge penalties (from 22pp), and no encoder-side deficit for the published Rectified-LpJEPA loss. A variance-floor repair restores information but not scale, and along the collapsed direction its restoring gradient vanishes with the residual while the published sliced-Wasserstein loss keeps it. A one-forward-pass check that scores objectives at a collapsed and a healthy latent flags every penalty that collapses deterministically, but it is a warning sign, not a certificate: in pre-registered runs the reference LEJEPA objective, which it scores as safe, collapses T-JEPA to the majority share when placed at the predictor $(6/9)$.

1 INTRODUCTION

Joint Embedding Predictive Architectures (JEPAs) (Assran et al., 2023; 2025; Thimonier et al., 2025; He et al., 2026) predict the latent representation of a masked view from a context view. Nothing in that objective forbids the trivial solution mapping every input to one latent, so every practical JEPA carries collapse-prevention machinery — EMA targets, stop-gradients, and in tabular variants regularisation tokens acknowledged to be heuristic (Thimonier et al., 2025; Mo & Tong, 2024). A recent line of work replaces these heuristics with a single *distribution-matching* objective on the latent: SIGREG (Balestriero & LeCun, 2025) matches an isotropic Gaussian via sketched projections, RDMREG, the objective of Rectified LpJEPA (Kuang et al., 2026), a sparse rectified target with a sliced two-sample loss.

This paper asks the prerequisite question the literature leaves implicit — *what makes such an objective safe?* — and answers it with a theorem, a controlled failure, a control, and a check. An anti-collapse objective is only useful if the collapsed representation is a *bad* solution under it, and that property can be violated silently: most easily when a target distribution is approximated by a *penalty* rather than realised by construction. We study one such approximation — a softplus pressure toward non-negative latents, the penalty-form stand-in for a rectified target that a reference implementation imposes with an architectural ReLU — and show its global optimum *is* the collapsed latent. We call this surrogate RECTPEN throughout, reserving RDMREG for the published objective it approximates.

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.

Writing $z_{\text{enc}} = E(x_{\mathcal{M}_c})$ for the context-encoder output and $z_{\text{pred}} = P(z_{\text{enc}})$ for the predictor output, our findings are:

1. **A deterministic, placement-dependent failure (§3–§5).** The one-sided penalty at z_{enc} degrades T-JEPA under its standard probe in $28/28$ runs (majority-exact in 24; 7 distinct (dataset, seed) units); in a fresh paired grid it collapses $15/15$ runs at z_{enc} against $0/15$ at z_{pred} (exact McNemar 6.1×10^{-5}), is untouched by four decades of loss weight ($33/33$), and survives the seven interventions we tried. Under a fair, validated probe the gap is 16pp.
2. **The collapse is geometric, not informational — until precision is lowered (§4).** Collapsed latents are one common vector plus a residual of $\sim2\times10^{-7}$ of its scale; that residual is learned (it survives a genuine float64 forward) and carries most of the label, but a bf16 or fp16 copy maps every input to a single vector.
3. **Controls with the published objectives (§6).** The Rectified LpJEPA loss does not collapse at z_{enc} ($0/9$ at 100 and 300 epochs; $0/9$ without its ReLU, which carries quality). The reference LeJEPA objective is safe at z_{enc} ($0/9$) but collapses at z_{pred} ($6/9$).
4. **A dichotomy and a warning sign (§6).** Entrywise objectives can have a collapsed global optimum (Proposition 1); distribution-matching and variance-anchored ones keep a batch-constant latent away from theirs (Proposition 2). Scoring an objective at a collapsed and a healthy synthetic latent flags every entrywise penalty that collapses deterministically — including, pre-registered, everyday ℓ_1 and hinge ($18/18$ at z_{enc}, $0/18$ at z_{pred}) — but misses stochastic and soft failures, and the reference LeJEPA objective's majority-exact one at z_{pred}.
5. **A repair that restores information but not scale (§6).** The variance floor the theory prescribes still leaves $7/18$ runs majority-exact under the standard probe, while a fair probe reads 0.68; its restoring gradient vanishes as the residual shrinks, whereas the published sliced-W_2 loss keeps a non-vanishing one.

2 Background and experimental setup

T-JEPA. T-JEPA (Thimonier et al., 2025) adapts JEPA to tabular data: a context encoder E embeds a masked view, a predictor P maps its latent to predictions of an EMA target encoder's latent for the complementary mask, and the loss is $\|P(z_{\text{enc}}) - \text{sg}(h)\|^2$. Downstream quality is a linear probe on frozen encoder features. We use the authors' tuned configuration (depth 16, width 64, batch 256, 300 epochs, probe every 20) on three OpenML datasets spanning class structure: helena (100 classes, majority share 0.0614), jannis (4, 0.4601), adult (2, 0.7607).

Objectives. SIGReg, the anti-collapse objective of LeJEPA (Balestriero & LeCun, 2025), compares projections of the batch latent with $N(0,1)$ along random directions. Our main-grid version is a *variant*: a standardised Cramér–von Mises statistic over 32 fixed directions, whereas the LeJEPA reference uses an unstandardised Epps–Pulley statistic over 256 directions resampled every step; we run that reference statistic as a control in the same harness (§6) and write SIGReg-CvM for ours where it matters. We apply both directly to the encoder output; LeJEPA itself computes it on a *projected* embedding and ablates the projector as a hyperparameter, a distinction that matters below.

The object of study in §3–§5 is what we call the **one-sided penalty**: a softplus pressure toward non-negative latents, $\mathcal{L}_{1\text{s}}(z) = \lambda_{\text{neg}}\,\mathbb{E}[\text{softplus}(-z)] + \lambda_{\ell_2}(\mathbb{E}[z^2]-1)^2$, plus a hard-threshold activation-fraction term ($\mathbb{1}[|z| > 10^{-2}]$) that contributes no gradient (kept because the implementation we started from has it; we show below it is inert). This is a *penalty-form stand-in* for a rectified target, of the kind one writes when adding non-negativity to an existing objective; it is **not** the published RDMReg, which rectifies with a ReLU and matches distributions with a sliced two-sample loss. We run that published *loss* as a control (§6); note that we apply it to $\text{ReLU}(z_{\text{enc}})$ inside the loss while the predictor and probe see z_{enc}, which reproduces the loss, not the published architecture, and we separate the ReLU from the loss by ablation. VICReg and Barlow Twins in our grid are single-view adaptations (variance/covariance, and within-view correlation), not the two-view originals. Every auxiliary enters as $\beta\,\mathcal{L}_{\text{reg}}(z)$ with $\beta = 0.1$ unless stated (in the code RectPen is the flag `rdmreg`, the published loss `rdmreg_faithful`). The rank-adaptive (AR) schedule multiplies β by $(1 - r_{\text{eff}}(z)/d)^{\gamma}$, where r_{eff} is the entropy-based effective rank (Roy & Vetterli, 2007; Garrido et al., 2023).

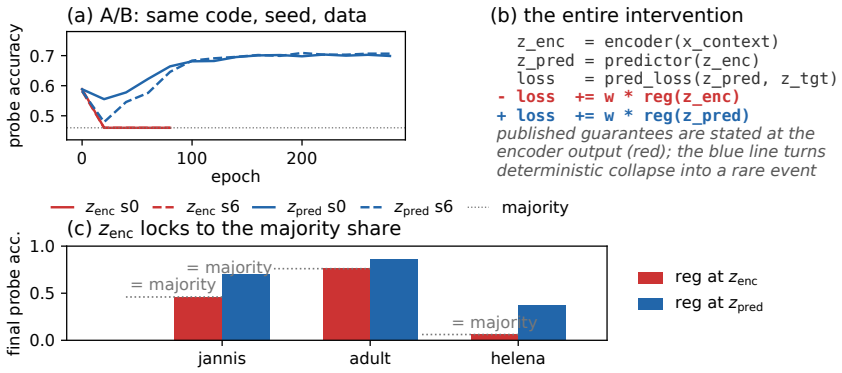

Figure 1: **A one-line placement choice decides collapse in tabular T-JEPA.** (a) Placement A/B on jannis (AR-RectPen, $\beta = 0.1$): same committed source (only `--reg_placement` differs), same seeds, sole GPU tenancy; encoder placement (red) locks at the majority share 0.4601, predictor placement (blue) recovers through the dip to 0.70+, including on a seed that collapses under encoder placement. (b) The intervention: which tensor the regulariser touches. (c) The collapse is majority-exact on all three datasets (median $|\text{probe}-\text{share}| = 8\times10^{-5}$, max 2×10^{-3}); predictor placement is risk reduction, not immunity (§7), and the phenomenon does not extend to vision or MTS JEPAs.

The single experimental variable. The placement $z \in \{z_{\text{enc}}, z_{\text{pred}}\}$. Both appear in the wild: LeJEPA states its guarantees on the encoder output while its implementation applies the loss to a *projector* output — the guaranteed and the regularised tensor differ (Balestriero & LeCun, 2025; Panchavati et al., 2026), others likewise regularise behind a head, and predictor-output placement arises naturally when code reuses the variable holding the predictor's result — which is how an earlier evaluation of ours ran for 112 runs before an audit "fixed" it to the encoder and every stable configuration began to collapse. This paper is the controlled post-mortem of that event (Appendix A).

Rigour notes. Tabular runs use seeded workers, deterministic cuDNN, a fixed sketch RNG for SIGReg, and pinned dependencies. Every A/B holds data, seed and GPU tenancy fixed and executes the *same committed source* under two values of one flag, `--reg_placement`; each arm writes a provenance sidecar whose placement-excluded config hash matches its partner's, so the arms are attestably identical except for the regularised tensor. The 163-run census ships with the code, and `src/verify_paper_numbers.py` recomputes every headline in this paper from its released artefact, failing if a claim is absent where the paper makes it. Collapse is scored two ways throughout — *strict* (within 5×10^{-3} of the majority share) and *soft* ($>$ 3pp below a no-regulariser baseline *at the same epoch budget*) — and we report both for every arm, including where soft is the less flattering number.

3 The Phenomenon: Placement Decides Collapse

Figure 1 shows the central experiment. With RectPen at z_{enc}, jannis training enters the universal early dip and never leaves: the probe locks at 0.4601. Moving the same loss to z_{pred} — one line — yields full recovery (0.699–0.706), matching our earlier 112-run predictor-placement evaluation. Seed 6, which collapses under z_{enc}, recovers under z_{pred}: *collapse basins are created by placement, not by seeds*.

Figure 8 shows the three-dataset endpoint. Every encoder-placed RectPen run terminates at the majority-class share: 0.4601 (jannis), 0.7606 against a 0.7607 share (adult), 0.0614 (helena — a hundred-class dataset reduced to a 6% constant predictor). Across every variation we ran — fixed and AR schedules, $\gamma \in \{0.25, 0.5, 2, 4\}$, $\beta \in \{0.1, 0.01\}$, five seeds, three datasets, plus the release/onset/EMA/rescue probes of §5 — $28/28$ **encoder-placed RectPen runs end degraded under the standard probe** (7 distinct (dataset, seed) units; 15 runs probe the jannis seed-0 cell) — 9.1–9.4pp below the no-regulariser baseline on adult, 24.2–35.5pp on jannis, 30.7pp on helena (its one run) — and 24 **of the** 28 **are majority-exact** under the strict criterion. The four that miss

the strict band miss it *downwards* (release-E=5 and three rescue probes, 0.3489–0.4376 against 0.4601) — worse than a constant predictor. Census membership is decided by a mechanical rule, not by the outcome (Appendix A); on tightness, 23 of these runs land within 5×10^{-4} of the share, and the predictor-era census under the same rule has 1 **strong collapse in** 110 **runs** ("strong" = strict). **The census is corroborating, not load-bearing:** every headline claim below rests on fresh, flag-attested runs. Predictor placement is still not immune ($1/9$ strict in the faithful control's predictor arm, $3/12$ in the width scan; §7). Appendix A crosswalks the historical denominators.[1]

SIGREG at z_{enc} fails stochastically instead. Across the 15 distinct encoder-era jannis (schedule, seed) cells of the A/B census (fixed and AR schedules, eight seeds), 7 collapsed (below 0.674, i.e., $>$ 3pp under the 0.704 baseline) and the rest recovered to 0.69–0.72: rate 0.47, Wilson 95% $[0.25, 0.70]$. The outcome flips between near-identical configurations (seed 2: fixed 0.458, AR 0.715; seed 0: the reverse).

A systematic grid. A fresh systematic grid (Table 5, Figure 9; not pre-registered — the harness predates the collapse criteria) runs RECTPEN, SIGREG and single-view VICReg (Bardes et al., 2022) and Barlow Twins (Zbontar et al., 2021) under both placements. Predictor placement produced no collapse for *any* regulariser or dataset ($0/42$, either definition; out-of-grid exceptions in §7), while encoder placement damages the whole family: RECTPEN $15/15$ to the exact majority share, Barlow $6/6$ soft ($2/6$ exact), VICReg and SIGREG stochastic ($3/6$, $3/15$ soft, both concentrated on jannis). SIGREG's grid rate ($3/5$ on jannis) is consistent with the census ($7/15$): a jannis-concentrated knife-edge, not a deterministic collapser. Under Holm correction only RECTPEN's paired test stays significant (2.4×10^{-4}; Barlow 0.094, VICReg/SIGREG 0.5) — the family claim is descriptive, the RECTPEN claim statistical. The severity ordering (RECTPEN, Barlow $\gg$ VICReg, SIGREG) tracks how directly each loss constrains geometry — and, as §6 shows, the objective's own optimum.

Not an artefact of one width or dataset — and the predictor arm is not clean either. A 24-run scan (100 epochs, strict criterion on *both* arms; `data/scope_scan.csv`) widens to d=128 and 256 and adds higgs (2 classes, majority 0.5286) and aloi (1000 classes, 0.001, where majority-exact means chance). Encoder placement collapses $11/12$; the twelfth (jannis, d=256) ends 11.6pp *below* the share. But predictor placement is **not** clean: $3/12$ strict, all jannis at $d \geq 128$ (0.4590, 0.4589, 0.4604 against 0.4601), and the higgs predictor arms (0.546–0.604) clear the share by only 2–8pp (no matched-budget baseline, so no soft claim). Predictor placement reduces risk, less so at larger width.

4 ANATOMY: A UNIVERSAL DIP BECOMES A BIFURCATION

The dip is not created by the regulariser. Across the 110 *predictor-era* trajectories — the right population for this question, since encoder-placed runs dip and never return — 60–76% of runs per dataset drop at least 2pp early (mean depth 4.9–6.2pp), *including unregularised baselines*, and recover in $22/22$ (helena), $37/37$ (jannis), $17/18$ (adult) of dipped cases; the one unrecovered dip is the pre-fix adult run of §3. It is a generic phase of JEPA training in which the probe signal weakens before the predictor–encoder loop establishes itself.

The encoder-placed regulariser decides the dip's exit (Figure 10). Three observations pin it down:

Schedules are bystanders. If a feedback loop through the AR schedule (weight scaled by $(1 - r_{\text{eff}}/d)^{\gamma}$) drove the collapse, scheduled runs should fail more often. They do not: $2/4$ versus $2/4$ on matched seeds, with the failing seeds *differing*. The schedule perturbs which basin a marginal run falls into; it does not create the basin. Nor does magnitude: reducing β by $100\times$ leaves the collapse intact and equally fast, and §5 sweeps four decades ($33/33$). At the dip the prediction gradient is weak, and a persistent directional bias suffices at any magnitude we tested.

The directional term drives the determinism. Decomposing RECTPEN at z_{enc} isolates the cause (jannis, three seeds each): with the softplus term removed (`noneg`; given the inert sparsity term this

[1] Untelemetried runs count as predictor-era, so five healthy post-fix runs may sit in that stratum — a misplacement that can only add healthy runs to that stratum, never remove a collapse from it.

leaves the ℓ_2 penalty alone) all three runs are healthy (0.698, 0.704, $0.704 \approx$ baseline); with *only* that term (`negonly`) all three collapse (0.450, 0.454, 0.449), though 6–11×10^{-3} *below* the share rather than onto it ($0/3$ strict, labelled so in Table 4). It is necessary and sufficient here; the symmetric remainder is benign (Figure 11). §6 explains why: at matched scale the ℓ_2 term vanishes, so the effective objective *is* the softplus term, whose global optimum is the collapsed latent. This also explains the RECTPEN/SIGREG dichotomy of §3: SIGREG matches an isotropic Gaussian and exerts no such pressure, leaving the seed-dependent knife-edge we observe.

The collapse is geometric, not informational. Re-probing the six kept checkpoints attributable to collapsed RECTPEN runs alone (`data/collapsed_probe_analysis.csv`; attribution in Appendix I) shows what the majority share hides. The latent is the batch-constant state of Proposition 1(i) to within numerical resolution: effective rank 1.00–1.11, every sample at cosine >0.99 to the batch mean, relative spread $\leq1.7\times10^{-3}$. Yet it is not constant: the residual is $\sim2\times10^{-7}$ of the latent's scale and learned, not rounding noise (a genuine float64 forward reproduces it, correlation 0.98–0.99; Appendix I). It carries the label: a raw-feature probe returns the majority share (0.460 in $5/6$), but after per-feature standardisation a linear probe reaches 0.634–0.701 and an MLP 0.690–0.705, against 0.723 for ten healthy checkpoints. **At half precision the information is gone:** stored in bf16 or fp16 before centring, all 4096 test inputs of each collapsed checkpoint round to one vector and the probe returns the majority share; centring first keeps 0.62–0.67 (healthy: 0.727 in every dtype). The failure is a collapse of *scale* that defeats the standard probe and becomes information loss at common storage precision; the one SIGREG-CvM collapse we re-probed is a different, low-rank mode (effective rank 4.0; balanced accuracy near chance even for an MLP).

5 MECHANISM: A GRADIENT-STARVED LOCK IN THE ONLINE NETWORK

Why is the collapsed branch absorbing? We ran a hypothesis tournament, eliminating candidates with targeted causal experiments:

The tournament is reported in full in Appendix E: H1 code regression, H2 pathological seeds, H3 GPU co-tenancy, H4 schedule feedback, H5 EMA hysteresis and H6 optimizer state are each refuted by a direct test, leaving H7 — the encoder enters a gradient-starved, weight-degenerate region of its own landscape — as the only survivor, and the one consistent with the forensics below.

Direct gradient measurement. Isolated from the regulariser and measured once per epoch, the prediction-loss gradient reaching the encoder starves by *eight orders of magnitude* through the danger window at z_{enc} (to $\sim4\times10^{-12}$) while the same seed at z_{pred} dips to 5×10^{-5} and recovers: the bifurcation at the level of the gradient itself (Figure 5, Appendix L).

It also explains why every rescue in §5 fails: weight noise (up to $\sigma = 0.05$) dislodges the network but yields a *wandering death* (0.34–0.40) — a degenerate region, not a single bad point.

Magnitude is not the causal variable. Sweeping β over four decades at fixed encoder placement moves the regulariser-to-prediction gradient ratio ρ from 0.20 to 1.7×10^{4}, yet all 33 runs collapse to the majority share (jannis $15/15$, adult $9/9$, helena $9/9$; Figure 12); at $\beta = 10^{-4}$ the regulariser gradient is a third of the prediction gradient. ρ escalates once the prediction gradient starves rather than causing the starvation; a threshold exists somewhere in $(0, 10^{-4})$ that we did not locate (full account in Appendix L).

No post-hoc rescue within the danger zone. Seven interventions were tried and none recovers the representation (campaign in Table 3, details Appendix A). Two entries matter: the release grid (Figure 2a) bounds the irreversibility threshold below five epochs, so *reactive* intervention is structurally too late; and the onset grid (Figure 2b) eliminates the natural prescription of a warm-up delay — encoder-placed RECTPEN collapses representations of any maturity, with a time constant that grows with maturity (220 epochs from a well-established state) but the same endpoint.

6 THE PRECONDITION: DOES THE COLLAPSED STATE MINIMISE YOUR REGULARISER?

§3–§5 anatomise one objective; here we show which property is responsible — and that it is checkable before training.

A control with the published loss. The published RDMREG (Kuang et al., 2026) realises RECTPEN's intent differently: non-negativity is *architectural* ($z = \mathrm{ReLU}(f_{\theta_2}(f_{\theta_1}(x)))$) and the objective is a sliced two-sample 2-Wasserstein distance to a Rectified Generalized Gaussian along K random directions. We implemented it and ran the *identical* A/B: three datasets $\times$ two placements $\times$ three seeds, same committed source, one flag.

The result is unambiguous (Table 6): $0/9$ strict collapses at z_{enc} (an upper bound of 34% on the rate, Appendix H) — and, for symmetry, $1/9$ at z_{pred} (jannis, 0.4575) — where RECTPEN collapses every encoder-placed run under the same datasets and code. Softly it is $1/9$ at each placement, symmetric and hence not a placement effect, against $9/9$ versus $0/9$ for the penalty it replaces. These runs use 100 epochs and the onset grid shows erosion can take 220, so we replicated all nine z_{enc} cells at the census's full 300-epoch budget: $0/9$ strict, $0/9$ soft, every trajectory *rising* from epoch 100 to 300 ($+0.1$ to $+5.2$pp; `data/faithful300_ctx.csv`). Two further pre-registered controls separate what protects it (`scripts/review_factorial_ab.sh`, 100 epochs). Without its ReLU the same loss still never reaches the majority share ($0/9$ strict) but loses 0–2pp on adult, 1–8pp on jannis and 10–15pp on helena against its paired ReLU runs: the distributional loss prevents the endpoint, the ReLU carries quality. And the reference LEJEPA objective (Epps–Pulley, 256 resampled directions) is safe at z_{enc} ($0/9$ strict, $5/9$ soft) but *collapses at* z_{pred}: $6/9$ majority-exact (adult and helena $3/3$ each) and $9/9$ soft — the opposite of the placement ordering RECTPEN shows. Its statistic scales with the token count, which differs between the tensors, and β was not re-tuned, so we report an observed, objective-specific failure. **The failure is not a property of distribution matching, nor of one placement, but of the objective–placement pair.**

The property that separates them, stated as a theorem. RECTPEN sees the latent only through *per-feature scalar statistics* — batch-means of entrywise functions (non-negativity, activation fraction, second moment). Every objective of that form is structurally blind to rank, and the two directions of the dichotomy can be proved in a few lines (proofs, and a numerical confirmation of every constant they predict, in Appendix G); Table 1 scores every objective we ran on exactly this dichotomy.

Proposition 1 (Entrywise objectives have collapsed optima)**.** *Let $\mathcal{L}(z) = \frac{1}{Bd}\sum_{ij} f(z_{ij})$ over latents with $\frac{1}{Bd}\sum_{ij} z_{ij}^2 = 1$. (i) If f is convex and non-increasing, $\mathcal{L}$ is globally minimised by $z \equiv 1$, uniquely if f is strictly convex; in particular the effective* RECTPEN *objective $f(t) = \mathrm{softplus}(-t)$ has the totally collapsed latent as its* unique *global minimiser, value $\mathrm{softplus}(-1) = 0.3133$. For the hinge $f(t) = \max(0, -t)$ the collapsed latent is* a *global minimiser but not the only one: every non-negative latent attains 0. (ii) For $f(t) = |t|$, $\mathcal{L}(z) \geq 1/\sqrt{Bd}$, with equality exactly at latents with a single nonzero entry — rank-one. (For general zero-minimised f of at most linear growth the single-entry latent still bounds the infimum by $C/\sqrt{Bd}$, but at fixed (B, d) full-rank latents can tie or beat rank-one ones; Appendix G gives a counterexample.)*

Proposition 2 (Distribution-matching and variance-anchored objectives penalise it)**.** *Let z be batch-constant ($z_{i\cdot} = c$). (i) Along every direction θ the empirical law of $z\theta$ is a point mass. Against a continuous target ν_θ the idealised Cramér–von Mises distance is then at least $1/12$ and the 2-Wasserstein distance at least $\mathrm{Var}(\nu_\theta)$, uniformly over the point's location; the finite-batch statistics we train with have their own positive floors (Appendix G). (ii) For any objective with a per-feature variance hinge (VICReg) or an on-diagonal cross-correlation term (Barlow Twins), the batch-constant latent attains that term's* maximum. *The converse holds for* off-*diagonal decorrelation: a batch-constant latent has all cross-correlations zero and therefore* minimises *such a term, so a pure redundancy-reduction objective is not protected by this proposition.*

This is the dividing line the data draw. The everyday penalties satisfying the hypotheses — softplus, hinge, ℓ_1 — admit a collapsed optimum, and the diagnostic scores each as preferring it (Table 1). Two adjacent penalties reach the same endpoint by separate arguments: activation-fraction is non-differentiable, and "KL sparsity" is minimised at its target ρ, not zero — yet a batch-constant

Objective	$\mathcal{L}$(collapsed)/$\mathcal{L}$(healthy)	verdict	outcome at z_{enc}*
RECTPEN	0.60	**prefers collapse**	$24/28$ **majority-exact**
hinge $\mathbb{E}[\text{relu}(-z)]$	0	**prefers collapse**	$9/9$ **majority-exact**
ℓ_1 $\mathbb{E}\lvert z\rvert$	0.78	**prefers collapse**	$9/9$ **collapsed**
RDMREG (published)	3.0	penalises collapse	$0/9$
SIGREG-CvM	13.5	penalises collapse	$7/15$ stochastic
SIGREG (reference)	119	penalises collapse	$0/9$; $6/9$ at z_{pred}
VICReg	35	penalises collapse	$0/6$ strict, $3/6$ soft
Barlow Twins	253	penalises collapse	$2/6$ strict

Table 1: **One forward pass separates the cases.** Each objective scored at a collapse-basin and an architecture-matched healthy latent, both at unit second moment; a ratio below 1 means collapse is cheaper. The objectives preferring collapse are exactly those that collapse T-JEPA deterministically at z_{enc}; a "penalises" verdict is not a safety certificate. *The reference SIGREG's predictor-output outcome is given after the semicolon (Appendix I). Raw values, the indifferent `noneg` ablation and the robustness sweep: Table 4.

latent at $\sigma^{-1}(\rho)$ hits it exactly. Two caveats. The hypotheses are needed ($f(t) = -|t|$ satisfies neither and has full-rank minimisers tying the collapsed one), and a pure second-moment term is *indifferent* rather than dangerous, being identically zero under the scale constraint — why deleting RECTPEN's softplus term leaves a non-collapsing objective. Distribution-matching and variance-anchored objectives keep the batch-constant latent away from their optimum (with two disclosed exceptions: pure off-diagonal decorrelation, and the released SIGREG's per-direction standardisation — Appendix G); their behaviour at rank-one latents is measured, not proved (Appendix M). The propositions order objective *values*, not training paths or stability: we claim only that reaching the collapsed state costs these auxiliary terms nothing.

The check the theorem licenses needs no training: build a collapse-basin latent ($z = |u|\, v^\top$, $u \sim \mathcal{N}(0,1)^B$, $v \sim |\mathcal{N}(0,1)|^d$) and a healthy one, renormalise both to unit second moment, and evaluate the differentiable part of the candidate objective at each. **The healthy reference must match the architecture.** A non-negativity penalty scored against a *sign-symmetric* reference is partly rewarded for the reference's negative entries. Our encoder has no output nonlinearity, so we take a sign-symmetric main reference but report both — against a full-rank *non-negative* reference of equal effective rank, RECTPEN moves from 0.50–0.59 to 0.99–1.17 (three basins; Appendix M), and the hinge to an exact tie. The published RDMREG rectifies *architecturally*, so its healthy latent is already non-negative and non-negativity pressure has nothing left to buy.

Table 4 and Figure 7 report this for every objective we ran. The ordering is not an artefact of one construction: under the architecture-matched reference, a sweep over batch/width, four basins (Appendix M tabulates three) and three seeds reproduces every verdict except two regimes the theory predicts (ℓ_1 on the constant basin; VICReg at $d \geq B$); under the non-negative reference the sign-sensitive penalties lose their preference — the architecture-dependence above, reported per reference in the released sweep. A "penalises" verdict is no guarantee either way — SIGREG penalises collapse by $13\times$ and still fails $7/15$ jannis cells, and our repaired penalties pass the check and still lock $7/18$. On our panel (three flagged; six passers: four "penalises" rows, two repaired penalties): positive predictive value $3/3$, negative predictive value $2/6$ strict ($1/6$ soft), and the rank-one margin is not monotone (Barlow: $253\times$, still $6/6$ soft). Counting "prefers collapse" as a positive and any collapse as the outcome, that is sensitivity $3/7$ and specificity $2/2$ (strict), and $3/8$ once the reference SIGREG is added as a seventh passer — which the check scores as penalising collapse $119\times$ and which nonetheless collapses $6/9$ majority-exact at z_{pred} (adult and helena $3/3$): a false negative on a majority-exact failure. On this small panel the warning is never wrong when it fires and often silent when it should.

The class, and the repair, tested. RECTPEN could be dismissed as a strawman of our own construction, so we tested two everyday members of the class, pre-registering the prediction (`scripts/entrywise_class_ab.sh`): ℓ_1 sparsity $\mathbb{E}|z|$ and hinge non-negativity $\mathbb{E}[\text{relu}(-z)]$ — the most literal penalty form of an architectural ReLU — through the identical A/B (2 penalties $\times$ 3 datasets $\times$ 3 seeds $\times$ 2 placements $=$ 36 runs). **Both collapse deterministically at** z_{enc}

and neither collapses at z_{pred}: $18/18$ against $0/18$, under *both* criteria. The soft criterion needs a same-budget comparator, so we ran the missing arm — nine no-regulariser runs at 100 epochs (`data/noreg100_baseline.csv`; budget gap 2.4–5.3pp vs. the 300-epoch grid). Against it the encoder arms are $18/18$ soft, the predictor arms $0/18$; the mismatched baseline would have read $10/18$, the gap rather than an effect.

We then asked whether the propositions are *prescriptive*. Proposition 2(ii) says a per-feature variance floor is maximised by the batch-constant latent, so adding one penalises the batch-constant state — for the hinge, whose minimisers include it, a non-negative unit-variance latent now scores strictly lower — without changing what the penalty asks pointwise; it does not penalise rank-one latents whose features vary. The training-free check agrees, flipping both objectives to the verdict "penalises collapse" against the architecture-matched reference (against the non-negative one the repaired ℓ_1 still marginally prefers rank-one, 0.99), and we pre-registered that prediction (`scripts/repair_ab.sh`) with two criteria — majority-exact, and a 3pp soft margin against each penalty's own z_{pred} arm — fixed before the runs finished. **The repair is not enough**, but not in the way the standard probe says: it cuts the majority-exact rate from $18/18$ to $7/18$ and leaves $18/18$ soft-collapsed under that probe, yet a fair, validated probe on its jannis checkpoints reads 0.68, above the unrepaired encoder arms (0.64) — it restores information but not scale. The latent gradient suggests why (Appendix I): along $z = \mathbf{1}\mu^\top + sH$ a variance floor penalises the near-constant state but its restoring derivative $\partial\mathcal{L}/\partial s$ shrinks with s ($\approx -10^{-5}$ at $s = 10^{-7}$; exactly 0 for a clamped floor), whereas the published sliced-W_2 loss keeps -1.9 at every $s \leq 10^{-2}$ — a candidate mechanism at the latent level, not a proof about training. The floor also penalises the low-rank basins beside the batch-constant one by only 1.3–1.6×, against 3.0× for the published RDMReg (descriptive, not a threshold; Appendix M). **Moving an objective's optimum does not move the optimiser**; the sign is the warning we propose, not a certificate. Full per-dataset numbers in Appendix N.

How to use it. Score a candidate objective on a synthetic collapsed batch and an architecture-matched healthy batch. If the collapsed batch scores lower, the objective's optimum is (near) collapsed, and a JEPA with a weak prediction signal can reach it. The right response to a failure is a distribution-level objective, not a patched entrywise one — and a pass still needs the placement you will use tested, as the reference SIGReg at z_{pred} shows.

7 Discussion

Why is the predictor output safer — and still not immune? Under predictor placement the gradient reaches the encoder only through P, which satisfies the constraint largely with its own parameters, turning a hard constraint on the representation into filtered pressure — near-complete protection in the 110-run predictor-era census (1 strict collapse, the pre-fix run of §3). **It is still not immunity, and not for every objective:** the reference SIGReg collapses there ($6/9$), the faithful control's predictor arm $1/9$, the width scan $3/12$. The predictor *filters* the pressure without severing it: when the dip's prediction gradient is unusually weak, residual pressure can still commit the encoder.

Implications for collapse theory. The encoder collapses under RectPen because for *this* objective the collapsed state is a good solution (0.48 vs. a healthy latent's 0.81); placement decides whether that state is reached, and once it is we measured no escape gradient (Figure 4). A guarantee is only as good as the objective actually optimised — one substitution, penalty for architecture, makes the collapsed state optimal — and even a well-formed objective can be satisfied while information is lost (SIGReg penalises collapse 13× and still fails $7/15$ jannis cells).

Implications for practice. (i) Run the precondition check before training: one forward pass *flags* every deterministic failure we observed; passing guarantees nothing (the repaired penalties pass and still lock $7/18$). (ii) If it fails, use an objective distribution-level *by construction*: a bolted-on floor leaves $7/18$ majority-exact; tuning helps least ($33/33$ across four decades of β). (iii) Prefer z_{pred} for an unchecked entrywise objective ($15/15 \to 0/15$ in the paired grid), but test the placement you use. (iv) Add a standardised probe — a raw probe mistakes scale collapse for information loss —

and centre features before storing them at half precision; a probe every 20 epochs cannot catch a lock that forms within five; report placements, and re-implement the *mechanism*, not the intent.

A boundary — and the same asymmetry, graded. A ViT I-JEPA on CIFAR-10/100 and an MTS-JEPA (soft-codebook bottleneck predictor) on MSL, run with the identical regularisers under both placements (3 seeds each; the vision arms use a leakage-free probe (Alain & Bengio, 2016)), produce *no strict collapse in any arm*, and raising β by $1000\times$ at fixed encoder placement still gives $0/12$ strict. **Every deterministic failure we observed involved an objective whose optimum is collapsed, but such an objective did not suffice**; with this few architectures, an association, not a necessary condition. The class is not inert there either: a pre-registered arm ran the *same* ℓ_1 and hinge penalties in the ViT (`scripts/vision_entrywise_ab.sh`): $0/12$ strict, as predicted — yet at z_{enc}, ℓ_1 costs 6.0–20.5pp against the budget-matched no-reg arm ($3/3$ soft) and hinge 1.8–4.5pp ($1/3$), while at z_{pred} both cost nothing ($0/6$ soft, within ± 0.5pp). The asymmetry therefore reproduces outside tabular T-JEPA; its majority-exact form does not. We state the implied reachability hypothesis as a prediction: the collapsed optimum exists in any architecture (Proposition 1 is architecture-free), and whether descent reaches it depends on the prediction gradient during the early dip — starved by eight orders of magnitude in T-JEPA (§5), never in the ViT. Weakening the vision prediction signal (smaller data, harder masks) should then yield the majority-exact endpoint; this is untested. Both probes are weak (MTS logs no downstream metric; the vision harness mean-pools before the z_{enc} regulariser; Appendix F).

8 Related work

VICReg (Bardes et al., 2022) and Barlow Twins (Zbontar et al., 2021) pioneered feature-statistics losses, SIGReg (Balestriero & LeCun, 2025) and RDMReg (Kuang et al., 2026) being their JEPA-era descendants. We study a penalty-form surrogate of the latter and run the published loss as a control: our failure is not evidence against that method. Klindt et al. (2026) prove identifiability at LeJEPA's optimum; we show the path can commit to a near-constant basin which, for an entrywise objective, *is* the optimum. "Do not put the loss on the representation" is established SSL practice (Chen et al., 2020; Grill et al., 2020; Chen & He, 2021; Bordes et al., 2023) — and even the flagship heuristic-free method retains it: LeJEPA regularises a *projected* embedding, and an EEG replication reports 0.476 against its 0.603 baseline without the projector (Panchavati et al., 2026). Our contribution is the objective-side property deciding whether violating the practice costs quality or collapses the representation's geometry, plus a training-free test. Adjacent 2025–26 work triangulates the line: SparseJEPA (Hartman & Varshney, 2025) sparsifies *weights*, outside our class; VISReg (Wu et al., 2026) pairs a variance term with a sliced-Wasserstein sketch, the very substitution our failed repair says is needed, and independently reports SIGReg's vanishing gradients under collapse; Weak-SIGReg (Akbar, 2026) uses SIGReg to recover a collapsed ViT. The ℓ_1 result echoes dead latents in sparse autoencoders (Gao et al., 2025) and gradient starvation (Pezeshki et al., 2021). Extended discussion in Appendix J.

9 Limitations

The deterministic result is anchored on T-JEPA; vision and MTS are three-seed boundary probes with disclosed asymmetries; several mechanism probes are $n{=}1$; the propositions characterise optima, not dynamics. Full list in Appendix K.

10 Conclusion

An anti-collapse regulariser whose own optimum is the collapsed state is dangerous, though not the only dangerous kind; entrywise objectives have such an optimum, distribution-level ones do not. Everyday ℓ_1 and hinge penalties collapse T-JEPA $18/18$ at the encoder against $0/18$ at the predictor, the published objective does not ($0/9$ strict at z_{enc}, at both budgets), and the repair our own theory prescribes still leaves $7/18$ majority-exact — the one-forward-pass check is a warning sign, not a certificate. The collapse is geometric — the label survives at a scale the standard probe cannot use and half precision cannot store — and a fair probe still finds a smaller placement cost. A guarantee must name the tensor and the collapse; a re-implementation, the mechanism.

AI USE STATEMENT

In this work, we used generative AI tools (Anthropic's Claude, a large language model assistant) for tasks with required disclosure: helping develop the theoretical framing, formulating and refining the mathematical claims of Propositions 1 and 2 and assisting in the writing of their proofs, proposing and refining hypotheses, designing and giving feedback on experimental methodology, implementing methods (experiment drivers, analysis scripts, and the verification gate), and interpreting results. We have *not* used generative AI tools to generate synthetic data sets; no experimental data was generated, imputed, or altered by an AI tool, and the remaining required-disclosure tasks (translation, dataset cleaning and reformatting, qualitative thematic analysis) are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting and editing the paper for readability, identifying and verifying related work and bibliographic metadata against public records, creating figures, and editing code. We have reviewed all AI-assisted work: the proofs were verified line by line and are additionally confirmed numerically by a released generator (`src/numerical_confirmation.py`), and every headline number is independently recomputable from the released artefacts via the 47-check gate described below. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyses training failure modes of self-supervised models on standard public benchmarks (OpenML tabular datasets, CIFAR-10/100, MSL); it involves no human subjects, no personally identifiable data, and no new data collection. The findings are defensive in character — they identify when an anti-collapse objective silently collapses a representation's geometry and provide a warning sign for it — and we see no foreseeable harmful application. Negative results about a published objective are reported with a faithful re-implementation of that objective run as a control, and our conclusions explicitly do not impugn the published method (§7). All artefacts we release are anonymised for double-blind review.

REPRODUCIBILITY STATEMENT

An audit gate, `src/verify_paper_numbers.py`, recomputes forty-seven pinned claims from raw released columns — the faithful control, the class experiment (strict and matched-budget soft, both placements), the repair, the census encoder population, tally and tightness, the predictor-era census, the width scan, the placement grid, the vision boundary, the collapsed-checkpoint re-probe, the pre-registered review controls, the fair re-probe, the precision boundary, the scale counterfactual, the latent gradient profile, artefact-integrity checks, the diagnostic verdict set and ratio, and the pinned majority shares — extracting each number at its exact claim site in the sources, cross-checking the compiled PDF against the verified prose, asserting the census era labels against the telemetry column they are defined by, scanning every build for retracted literals, and exiting non-zero on any mismatch, a missing artefact, a missing anchor, or a silently changed check count. `scripts/prepare_release.sh` runs it as a packaging gate that fails closed. The gate's falsification battery is itself committed (`scripts/falsify_gate.sh`): eleven attacks — CSV inversion, shifted census accuracies, a tampered majority share, a tampered claim site, a deleted artefact, a forged era label, a pasted contradictory literal, a zeroed vision CSV, a flipped faithful-300 row, a baseline collision, and wholesale dip-file falsification — are re-runnable by any reviewer, and all are caught. The gate defends against inconsistency between paper and artefacts, not against wholesale fabrication of the raw data, which no textual check can. Numbers outside the gate's forty-seven claims are recomputable from the shipped CSVs by the named analysis scripts, each of which refuses to overwrite a released artefact when its inputs are absent. The six interventional experiments carry pre-registered hypotheses in script headers committed before their data — attested by git ancestry — and two predictions failed — the repair, and the reference SIGREG at the predictor output — and are reported as failures. The training-free diagnostic reproduces bit-identically from the anonymised bundle alone, on CPU, in seconds. Raw training logs (5.7 GB) are withheld only for size and anonymity and are released at de-anonymisation; Appendix K states which claims depend on them.

REFERENCES

Habibullah Akbar. Weak-SIGReg: Covariance regularization for stable deep learning. In *ICLR 2026 Workshop on Geometry-grounded Representation Learning and Generative Modeling (GRaM)*, 2026. arXiv:2603.05924.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. ICLR 2017 Workshop Track.

Anonymous. Heuristic-free tabular JEPA: Distribution-matching regularisers as alternatives to regularisation tokens, 2026. Companion manuscript, under double-blind review; cited for provenance only, no result in this paper rests on it.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2301.08243.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2105.04906.

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*, 2023. arXiv:2206.13378.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. arXiv:2002.05709.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. arXiv:2011.10566.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. arXiv:1711.02257.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.04093.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2210.02885.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. arXiv:2006.07733.

Max Hartman and Lav Varshney. Sparsejepa: Sparse representation learning of joint embedding predictive architectures. *arXiv preprint arXiv:2504.16140*, 2025.

Yanan He, Yunshi Wen, Xin Wang, and Tengfei Ma. SC-JEPA: Stabilizing latent predictive learning for time-series anomaly prediction. In *SIAM International Conference on Data Mining (SDM)*, 2026. arXiv:2602.04643; v1 title: MTS-JEPA.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *ICLR*, 2022. arXiv:2110.09348.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? *arXiv preprint arXiv:2605.26379*, 2026.

Yilun Kuang, Yash Dagade, Tim G. J. Rudner, Randall Balestriero, and Yann LeCun. Rectified LpJEPA: Joint-embedding predictive architectures with sparse and maximum-entropy representations. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.01456.

Lukas Kuhn, Giuseppe Serra, Randall Balestriero, and Florian Buettner. LeVLJEPA: End-to-end vision-language pretraining without negatives. *arXiv preprint arXiv:2607.00784*, 2026.

Shentong Mo and Shengbang Tong. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2410.19560.

Saarang Panchavati, Uddhav Panchavati, Hiroki Nariai, Corey Arnold, and William Speier. Laya: A LeJEPA approach to EEG via latent prediction over reconstruction. *arXiv preprint arXiv:2603.16281*, 2026. v2.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *NeurIPS*, 2021. arXiv:2011.09468.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Hugo Thimonier, José Lucas De Melo Costa, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. T-JEPA: Augmentation-free self-supervised learning for tabular data. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05016.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *ICML*, 2021. arXiv:2102.06810.

Haiyu Wu, Randall Balestriero, and Morgan Levine. VISReg: Variance-invariance-sketching regularization for JEPA training. *arXiv preprint arXiv:2606.02572*, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020. arXiv:2001.06782.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.03230.

A Full run census and provenance

One protocol note: every tabular arm inherits upstream T-JEPA's full-data (transductive) covariate preprocessing; re-preprocessing inductively shifts endpoints by at most 1.7pp (`data/inductive_vs_transductive.csv`), about half the 3pp soft margin, shared by all arms, and unable to mask or mimic a majority-share collapse. The census (`data/collapse_census.csv`; 163 completed runs) classifies every run by dataset, placement, regulariser, schedule, seed, and collapse outcome (strong = the strict criterion, within 5×10^{-3} of the majority share — the extraction script's 1pp label coincides with it on every one of the 163 rows; soft: more than 3pp below the dataset's baseline mean). Headline counts: encoder-placed RectPen $24/28$ strong collapses (population 28 under the mechanical telemetry rule of §3; membership is never decided by the outcome) (jannis 7, adult 8 — including all γ and β variants —

helena 1, plus release/onset/EMA/rescue probes); encoder-placed SIGREG on jannis $7/15$ distinct A/B-census cells collapsed (rate 0.47, Wilson 95% $[0.25, 0.70]$; including the two telemetried seed-fix extension runs the full encoder-era tally is $9/17$); predictor-era census 110 runs with 1 strong collapse — the untelemetried adult run whose hand-promotion to the encoder era an earlier draft made on completion-mtime grounds and this version withdraws (its start timestamp precedes the fix's first attestation). *The membership rule, in full.* Census membership is decided by a *fully mechanical rule*, never the outcome: a run is encoder-era iff its setting carries a placement prefix or its log carries the rank telemetry that entered the code in the same commit as the placement fix. Telemetry is uneditable after the fact and re-derivable from the shipped census; it separates outcomes perfectly within the RECTPEN family (36 runs, zero crossover) and is visibly not outcome-shaped, also moving four neutral helena SIGREG runs into the encoder era. On tightness, 23 land within 5×10^{-4} of the share (of 24; median 8×10^{-5}); the exception is the onset probe (0.4621). `src/collapse_stats.py` recomputes population, tally and tightness from the released census. Under the same rule the predictor-era census is 1 **strong collapse in** 110 **runs** ("strong" = the strict criterion), that one being an untelemetried adult run whose start timestamp precedes the placement fix and which is also the only dipped predictor-era run that never recovers (§4). *Corrections to this population, in print.* Two. (i) An earlier version reported $22/22$: the denominator was the count of collapses, so the figure was circular; the honest denominator is the whole encoder-era RECTPEN population, giving $24/28$. (ii) The hand-promotion described above is withdrawn, which moves one strict collapse out of the encoder era and into the predictor era (the encoder tally becomes 24 of 28 and the predictor era gains its single collapse, 1 of 110) — a correction against our own interest, and the reason the rule now admits no hand-coded exceptions. Dip statistics on the same 110: $\geq$ 2pp dips in 60–76% of runs per dataset, mean depth 4.9–6.2pp, recovery $22/22$, $37/37$, $17/18$ (helena/jannis/adult); the one unrecovered dip is that same run.

Denominator crosswalk. Four predictor-placement run counts appear in this paper; each is a different pre-specified population, and none was chosen post hoc: (i) **112** — the runs of the original, pre-audit predictor-placement evaluation (§2 of the main paper; back-referenced once in §3), a historical count only. (ii) **110** — the predictor-era census stratum under the mechanical telemetry rule ($31 + 49 + 30$ across helena/jannis/adult); *this is the only predictor-era denominator the main text uses*, and its 1 strict collapse is the withdrawn straggler. (iii) **90** — the pre-declared canonical stratum (seeds 0–4 of the core settings; `data/run_manifest.csv`). (iv) **98** — the pooled predictor-era manifest: the canonical 90 plus the 8 diagnostic seed-extension runs. An earlier draft quoted rates on (iii) and (iv) ($1/90$, $3/8$, $4/98$); those strata predate the telemetry rule, the $3/8$ misclassification is retracted in §3, and the main paper now quotes only (ii) plus the flag-attested experiments ($1/9$ faithful, $3/12$ scope). The raw counts for all four populations ship in the repository.

B FORENSICS METHODOLOGY

For each checkpoint (epochs 0, 19, 300; `src/ckpt_forensics.py`) we compute: per-matrix stable rank $\|W\|_F^2/\|W\|_2^2$ over context-encoder weights; the distribution (median, p99, max) of Adam's $\hat{v}$ over encoder parameters; and the cosine between flattened context and target encoders.

Avoiding circularity: how checkpoints are labelled. The collapsed/healthy label used in Figure 4 is *not* derived from any weight statistic. Checkpoints store no configuration, so we map each checkpoint to its originating run by start timestamp: both the checkpoint directory and the run's `wandb` offline directory are stamped at launch, and they agree to within 0–1 second for $25/25$ matched jannis runs (`src/map_ckpt_to_run.py`, `data/ckpt_run_map.csv`); three stale directories with no log within 5 seconds are dropped. The label is then the run's final linear-probe accuracy (collapsed if within 1pp of the majority share or more than 3pp below baseline). The forensics figure therefore plots a weight/optimizer statistic (y) against a downstream-accuracy grouping (x) measured independently. Grouped this way over 60 epoch-300 checkpoints: collapsed median $\hat{v} \in [10^{-28}, 10^{-17}]$ versus healthy $\sim 3 \times 10^{-8}$; minimum stable rank at epoch 19, collapsed 2.6–4.9 versus healthy 9.7–11.6. Latent effective-rank trajectories of collapsing runs (logged each epoch): $15.2 \to 2.5$, $15.5 \to 1.4$, $8.2 \to 1.1$ (of $d = 64$).

C PREDICTOR-PLACEMENT SUMMARY (SELF-CONTAINED)

So that the safety of predictor placement does not rest on an external reference, we summarise the 110 predictor-era T-JEPA runs directly (the seedfix grid; AR and fixed schedules, three datasets, seeds 0–7, the γ and the cls-token variants). Across all 110 runs there is a single majority-share endpoint — an adult RECTPEN-only run without rank telemetry whose recorded start timestamp precedes the placement fix's first attestation, i.e., a genuine predictor-era strict collapse under the pre-fix code ($1/110$; Appendix A). Predictor placement is near-complete, not complete, protection.

dataset	n	acc. range	mean $\pm$ sd	strong collapses
helena (100 cls, maj. 0.0614)	31	0.341–0.383	0.369 ± 0.009	0
jannis (4 cls, maj. 0.4601)	49	0.671–0.717	0.705 ± 0.009	0
adult (2 cls, maj. 0.7607)	30	0.758–0.864	0.852 ± 0.018	$1^{\dagger}$

†the predictor era's single strict collapse; an earlier draft's promotion of this run to the encoder era is withdrawn (Appendix A).

The per-setting jannis means under predictor placement cluster tightly around the no-reg baseline (0.704): RECTPEN +cls 0.705 (n=20), RECTPEN-only 0.708, SIGREG-only 0.705, SIGREG +cls 0.703, AR-MIX 0.707 — the regularisers neither collapse nor strongly help at this placement, which is exactly the benign behaviour their design intends and the sharp complement to the $24/28$ encoder-placement collapse.

D REPRODUCTION RECIPES

Deterministic collapse (one command). T-JEPA tuned config on jannis with `--reg_type rdmreg --reg_weight 0.1 --n_cls_tokens 1 --np_seed 0`, regulariser at the encoder output: locks at 0.4601 by epoch 20. Any seed, any $\gamma \in [0.25, 4]$, any $\beta \in [10^{-3}, 10^{-1}]$ reproduces it.

Placement rescue (one flag). The placement is a committed flag, `--reg_placement {ctx,pred}` (default `pred`); `scripts/reproduce_placement_ab.sh` runs both placements of the same config at repository HEAD (`ctx` $\rightarrow$ majority-class collapse, `pred` $\rightarrow$ recovery to 0.70+), with placement written into each log name and the working-tree state recorded in each sidecar; validated at repository HEAD on three seeds (`data/placement_ab_head_validation.csv`: `ctx` $= 0.4601$, the majority share, on all three; `pred` 0.706–0.716). The validation file carries, per arm, the repository and submodule commit hashes, a code-scoped clean-tree flag (`code_dirty` = false for every arm — the tracked code paths were committed at run time), and a placement-excluded config hash that matches within each seed pair, so the two arms are attestably identical except for `--reg_placement`. Each arm also writes a provenance sidecar (`*.prov.json`: repository and submodule commit hashes, dirty-state flags, the full argument vector, a config hash computed with the placement token excluded — identical across arms by construction — and the post-run log SHA-256), so the claim that the two arms differ only in placement is attestable from the artefacts alone. Pre-specified statistics (Wilson CIs, exact McNemar on the paired grid, sign-flip permutation and bootstrap deltas) are regenerated by `src/collapse_stats.py` into `data/collapse_stats.json`. The repository also includes the epoch-windowed regulariser flags (`--reg_release_epoch`, `--reg_onset_epoch`) and rescue flags (`--rescue_perturb_std`, `--rescue_reset_opt`) used in §5, and the census scripts that generated every figure from raw logs.

Vision generalisation (one script). The vision ViT I-JEPA harness is self-contained in `vision_jepa/train_vjepa.py` and reuses the identical regulariser code. `scripts/vision_placement_ab.sh` runs the CIFAR-10 placement A/B ($\{$none, RECTPEN, SIGREG $\} \times \{z_{\text{enc}}, z_{\text{pred}}\} \times 3$ seeds); `scripts/vision_scaleup.sh` runs the larger ViT (d=256, depth 8) on CIFAR-100; `scripts/vision_strong.sh` runs the strengthened-recipe grid (augmentation, warmup-cosine, EMA ramp, leakage-free probe) and `--recipe strong_noaug` its no-augmentation decomposition. Every run is evaluated with a leakage-free

linear probe (classifier fit on train-set features, accuracy measured once on the held-out test set; `linear_probe_proper`). Each run logs per-epoch encoder r_{eff} and probe accuracy as JSON lines; `src/aggregate_vision.py` produces `data/vision_placement_results.csv` (Table 2). No placement collapses at either scale; encoder-placed RECTPEN raises the encoder's effective rank but matches the baseline probe.

dataset	arm	probe	r_{eff}	collapse
CIFAR-10, ViT d=128 *(chance* 0.10*)*				
CIFAR-10	no-reg (z_{enc})	0.465	30.3	0/3
CIFAR-10	RECTPEN z_{enc}	0.440	38.5	0/3
CIFAR-10	RECTPEN z_{pred}	0.462	28.2	0/3
CIFAR-10	SIGREG z_{enc}	0.464	28.5	0/3
CIFAR-10	SIGREG z_{pred}	0.461	29.9	0/3
CIFAR-100, ViT d=256 *(chance* 0.01*)*				
CIFAR-100	no-reg (z_{enc})	0.251	38.4	0/3
CIFAR-100	RECTPEN z_{enc}	0.246	47.2	0/3
CIFAR-100	RECTPEN z_{pred}	0.237	32.6	0/3

Table 2: Vision placement A/B (base recipe), three seeds per cell, all evaluated with a leakage-free probe. *No placement collapses* at either scale: every arm — including encoder-placed RECTPEN— lands well above chance (4–5× on CIFAR-10, ∼25× on CIFAR-100). Encoder-placed RECTPEN raises the encoder effective rank (38.5 vs. 28.2 at z_{pred} on CIFAR-10; 47.2 vs. 32.6 on CIFAR-100) yet matches the baseline probe, so here it reshapes the spectrum without harming information. A strengthened recipe (augmentation, warmup-cosine, EMA ramp) and its no-augmentation variant reproduce this null (0/3 throughout; `data/vision_placement_results.csv`). Note: an under-trained *test-fit* probe spuriously scores encoder-placed RECTPEN at chance — the reason we fit the probe on training features throughout.

Compute. A jannis/helena cell is ∼5h and an adult cell ∼1.5–3.5h on one RTX 3090; the full study (census of 163 runs, of which ∼40 were new causal probes) consumed roughly 250 GPU-hours; the subsequent interventional programme (faithful control, β sweeps, entrywise class, repair, scope scan, matched baseline) adds ∼200 more, and the 300-epoch faithful replication ∼45, for ∼500 in total on a 4×3090 node.

E HYPOTHESIS TOURNAMENT (FULL)

Why is the collapsed branch absorbing? We eliminated candidate explanations with targeted tests; H7 is the only survivor.

H1: code regression. The collapse first appeared after a code audit. But the committed code produces it deterministically as a function of the `--reg_placement` value alone, at fixed seed: flipping that one flag — with no other code change — both causes and removes it. A version regression cannot explain an effect toggled by a single runtime flag. Refuted.

H2: pathological seeds. Seeds that collapse under z_{enc} recover fully under z_{pred}. Refuted.

H3: GPU co-tenancy. The A/B arms ran with sole GPU tenancy and identical data; the outcome tracks the flag, not the device. Refuted.

H4: schedule feedback. Fixed and rank-adaptive schedules collapse at the same rate (2/4 vs 2/4 seeds, different seeds). Refuted.

H5: EMA hysteresis. Slowing the target to m=0.9999 and releasing the regulariser does not recover a collapsed run. Refuted.

H6: optimizer state. Clearing Adam state and releasing the regulariser does not recover a collapsed run. Refuted.

H7: gradient-starved weight degeneracy. The survivor — a localisation, not a proof. Collapsed encoders occupy a region where the prediction loss supplies essentially no gradient (Adam second moments 10^{9}–$10^{20}\times$ smaller than in healthy runs) and the weight matrices are degenerate (minimum stable rank $\sim$10 $\rightarrow$ 2.6–4.9 by epoch 19). There is no gradient to follow out, regardless of the target, the optimizer, or the regulariser's continued presence.

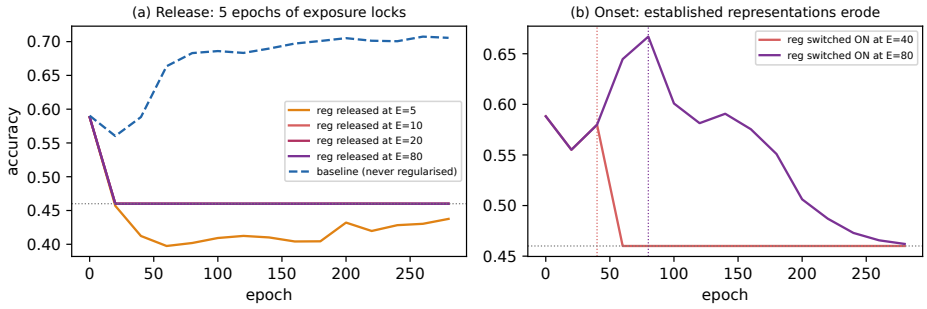

Figure 2: (a) Release grid: RECTPEN active only for the first E epochs, then off. Even E=5 locks permanently; 295 regulariser-free epochs do not recover. (b) Onset grid: RECTPEN switched on at epoch E *after* healthy training. At E=40 an established 0.580 representation dies within 20 epochs; at E=80 a 0.667 representation erodes to the majority share over 220 epochs.

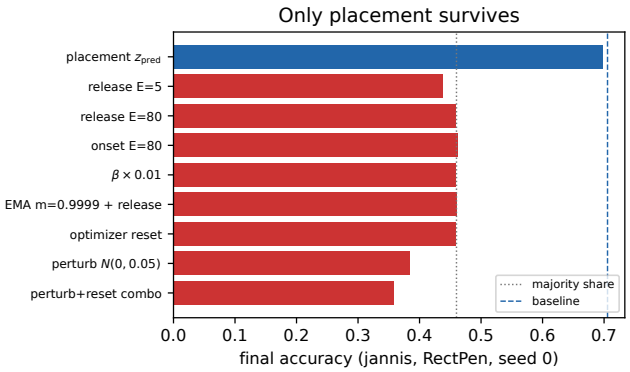

Figure 3: Final accuracy under every intervention attempted (jannis, RECTPEN, seed 0). Only placement survives.

F BOUNDARY PROBES IN FULL: VISION AND TIME SERIES

The main text summarises these probes in §7; we give the full account here. The deterministic RECTPEN result is supported by 28 runs over wide hyperparameter variation; the mechanism probes (release, onset, EMA, optimizer, noise) are anchored to it, and the two we replicated at a second seed (release E=10, onset E=80) reproduced the collapsed endpoint (0.446/0.460 vs. seed-0's 0.460/0.462). **We asked whether the same placement effect appears in other JEPA architectures, and it does not.** We built a ViT I-JEPA on CIFAR-10 (d=128, depth 4, EMA target) and a $\sim$4$\times$ larger one on CIFAR-100 (d=256, depth 8), with the identical regularisers under both placements (3 seeds each; `vision_jepa/`), evaluating every run with a leakage-free linear probe (Alain & Bengio, 2016) (classifier fit on *train-set* features, accuracy measured once on the held-out test

set). *No placement collapses*: on CIFAR-10 all arms — no-reg, RECTPEN and SIGREG at z_{enc} and z_{pred} — reach probe 0.44–0.465 (chance 0.10); on CIFAR-100 all reach 0.24–0.25 (chance 0.01). Encoder-placed RECTPEN does visibly reshape the spectrum — its encoder effective rank runs higher than the predictor arm's (38–48 vs. 28–36) — but this costs *no* downstream accuracy, so in this setting the regulariser inflates rank harmlessly rather than destroying information. (A cautionary methodological note: an under-trained, test-fit linear probe reports these same non-collapsed RECTPEN-z_{enc} representations at chance; the effect is an artefact of the probe, not of the representation, which is why we fit the probe on training features throughout.) **The entrywise class in vision, pre-registered.** The vision arms above use RECTPEN and SIGREG. To test the *class* rather than one objective we ran the same ℓ_1 and hinge penalties — the tabular definitions verbatim — through the ViT harness under both placements, 3 seeds each, at the placement A/B's 60-epoch budget, with criteria fixed in the committed driver before the runs (`scripts/vision_entrywise_ab.sh`; per-run values in `data/vision_entrywise.csv`). Strict collapse: $0/12$, the pre-registered prediction. Against the budget-matched no-reg arm (0.4648): ℓ_1 at z_{enc} reaches 0.405, 0.302, 0.260 (-6.0, -16.2, -20.5pp; $3/3$ soft) and at z_{pred} 0.462, 0.467, 0.468 ($0/3$ soft); hinge at z_{enc} 0.447, 0.442, 0.420 (-1.8 to -4.5pp; $1/3$ soft) and at z_{pred} 0.463, 0.470, 0.465 ($0/3$). Two things follow. The placement asymmetry is not a tabular artefact — it appears in vision at up to 20pp for ℓ_1 — while the majority-exact endpoint is: no vision cell comes within 16pp of chance. And severity does not track the diagnostic margin: hinge scores a *stronger* preference for collapse than ℓ_1 (0 versus 0.78) yet damages the ViT less, the same non-monotonicity we report for Barlow in Appendix M. We likewise ran an implementation of MTS-JEPA (He et al., 2026) — renamed SC-JEPA in its published revision; we keep the name of the version we implemented — (multi-resolution predictor with a soft-codebook bottleneck, d=128) on the MSL benchmark, {RECTPEN, SIGREG} $\times$ $\{z_{\text{enc}}, z_{\text{pred}}\} \times 3$ seeds: again *neither* placement collapses ($r_{\text{eff}} \approx$ 7–12 of 128, identical across placements to within 0.05), and removing the bottleneck (and raising the weight $10\times$) does not induce it (`data/mts_bottleneck_p1_results.csv`), so the bottleneck is not the protective factor. **The vision boundary is not a weak-regulariser artefact — and we establish this by intervention.** A natural objection to a negative result is that the regulariser was simply too weak. We therefore raised it: holding encoder placement fixed on CIFAR-10, we swept $\beta \in \{0.1, 1, 10, 100\}$ (three seeds each), a $1000\times$ range. *No run collapses under the strict criterion* ($0/12$; probe never within 5pp of the 0.10 chance level). We apply our own soft criterion here too rather than reporting only the favourable one, and we anchor it to a budget-matched comparator: the sweep runs 40 epochs while the placement A/B's no-reg baseline (0.4648) runs 60, so we measure against the sweep's own β=0.1 arm (0.4622, same budget). Against it, β=10 and β=100 *are* soft collapses ($3/3$ each, means 0.417 and 0.402 — 4.6 and 6.0pp down), while $\beta \in \{0.1, 1\}$ are not. So vision degrades — gracefully, monotonically, and only once the regulariser outweighs the prediction signal by two orders of magnitude — where T-JEPA is destroyed outright at $\beta = 10^{-4}$. That is the contrast we claim: not that vision is untouched, but that the same objective at the same placement produces a 6pp quality cost in one regime and a majority-exact, unrecovered collapse of the latent's scale in the other. This is also where our result departs from projector-head folklore, which predicts exactly the graceful cost we observe in vision (Bordes et al., 2023). The precondition of §6 is consistent with both halves: RECTPEN ranks collapse above health everywhere, so its optimum is collapsed in vision too, but a strong prediction signal keeps the ViT away from it. An objective with a collapsed optimum accompanied every deterministic failure we saw and did not suffice for one; that is an observed association, not a necessary condition. One asymmetry to disclose, as we do for MTS below: the vision harness mean-pools tokens before the regulariser at z_{enc}, so the two vision arms do not regularise identical supports either — the vision result is a boundary probe, and we treat it as one. Two caveats make MTS the weakest of our three domains, and we state them rather than lean on it: (i) our MTS harness zero-fills the non-context positions before the regulariser sees the context tensor, whereas at z_{pred} the regulariser sees the predictor's normalised outputs, so the two arms do not regularise the same support — a weaker intervention than the tabular one-flag A/B; and (ii) the harness logs no downstream anomaly-detection metric, so "no collapse" rests on the encoder's effective rank rather than on a task probe. The deterministic placement collapse is therefore *specific to the tabular* T-JEPA *regime we study* (d=64, a weak prediction signal at the epoch-20–40 dip) and does not universalise to these vision or time-series JEPAs. A working account, which our data support but do not prove, is that collapse requires the distribution-matching gradient to dominate a weak prediction gradient; where the prediction task is strong (augmented vision, multi-resolution MTS) the encoder never enters the dip and the regulariser never gains purchase. A companion manuscript (Anonymous, 2026) reports the same null for MTS-JEPA regulariser variants; being un-

der double-blind review it is cited for provenance only, and the boundary claim here rests on our own runs alone. Whether *any* large-scale vision/video JEPA — where LEJEPA reports broad success (with the loss on a projected embedding) — can be driven into the tabular regime is left open; our two-placement comparison ($r_{\text{eff}}^{z_{\text{enc}}}$ vs. $r_{\text{eff}}^{z_{\text{pred}}}$, plus a leakage-free probe) is a concrete diagnostic for future work.

G PROOFS: WHICH OBJECTIVES HAVE COLLAPSED OPTIMA

Throughout, $z \in \mathbb{R}^{B\times d}$ is a batch latent, $N = Bd$, and the scale is fixed by $\frac{1}{N}\sum_{ij} z_{ij}^2 = 1$ (the constraint RECTPEN's own ℓ_2 term imposes, and the normalisation under which the diagnostic of §6 is computed).

Proof of Proposition 1(i). Let f be convex and non-increasing and write $\bar{z} = \frac{1}{N}\sum_{ij} z_{ij}$. By Jensen, $\frac{1}{N}\sum_{ij} f(z_{ij}) \geq f(\bar{z})$. By Cauchy–Schwarz, $\bar{z} \leq \left(\frac{1}{N}\sum_{ij} z_{ij}^2\right)^{1/2} = 1$, and since f is non-increasing, $f(\bar{z}) \geq f(1)$. Hence $\mathcal{L}(z) \geq f(1)$ for every admissible z, with equality attained by $z \equiv 1$. If f is strictly convex, Jensen is an equality only when all z_{ij} coincide, and the scale constraint then forces $z_{ij} = \pm 1$; monotonicity selects $+1$. For $f(t) = \mathrm{softplus}(-t)$, which is strictly convex and decreasing, the totally collapsed latent $z \equiv 1$ is therefore the unique global minimiser, with value $\log(1 + e^{-1}) = 0.3133$. □

Proof of Proposition 1(ii). For $f(t) = |t|$: $\sum_{ij} z_{ij}^2 = N$ gives $\|z\|_1 \geq \|z\|_2 = \sqrt{N}$, so $\mathcal{L}(z) = \|z\|_1/N \geq 1/\sqrt{N}$, with equality iff at most one entry is nonzero; such a latent, $\sqrt{N}\, e_i e_j^\top$, is rank-one. A full-rank Gaussian latent sits at $\sqrt{2/\pi} = 0.798$, far above. □

What does not generalise. For any $f \geq 0$ with $f(0) = 0$ and $f(t) \leq C|t|$ the single-entry latent gives $\mathcal{L} \leq C/\sqrt{N}$, which bounds the infimum — but it does *not* order full-rank against rank-one latents at fixed (B, d), and an earlier version of this proposition that said so was false. Counterexample: $B = d = 2$, $a = 2/\sqrt{3}$, $f(t) = \min\{|t|, \big||t| - a\big|\}$. The latent $a\left(\begin{smallmatrix}1 & 1\\ 1 & 0\end{smallmatrix}\right)$ is rank-two, satisfies the scale constraint and has $\mathcal{L} = 0$; a zero-loss latent needs exactly three entries equal to $\pm a$, and every such 2×2 matrix has rank two, so the rank-one minimum is positive (0.120 by grid search; `src/numerical_confirmation.py`). The ordering holds for ℓ_1 exactly and not for the class.

Remark (which basin). The two parts predict *different* collapsed states: (i) the constant latent for softplus/hinge-type penalties, (ii) a sparse rank-one latent for ℓ_1. The robustness sweep reflects exactly this within the matched (sign-symmetric) reference: ℓ_1 prefers the rank-one basin in every non-constant construction and is within 3% of a tie on the constant one, while softplus and hinge prefer every collapsed construction; per-reference counts are in `data/collapsed_state_loss_robustness.json`. Per-unit mean-activation ("KL sparsity") penalties fall outside the proposition's hypotheses (they are minimised at the target ρ, not at zero, and act on a per-unit statistic after a nonlinearity), but reach the same endpoint by a one-line separate argument: a batch-constant latent at $\sigma^{-1}(\rho)$ attains the target exactly, so their optimum is collapsed as well.

Proof of Proposition 2(i). Let $z_{i\cdot} = c$ for all i and let θ be any direction. The empirical distribution of $z\theta$ is the point mass $\delta_{c^\top\theta}$, while the target ν_θ has no atom of mass one, so any divergence separating distinct laws is positive. One implementation caveat we state rather than gloss: SIGREG as released standardises each projection before the Cramér–von Mises comparison, making it invariant to per-direction affine maps — so its zero set is a family of laws, not $\{\nu\}$, and in particular it is blind to uniform shrinkage of the latent. The conclusion at the batch-constant latent still holds (a point mass standardises to a point mass under the ϵ-clamped variance), but the proposition's separation hypothesis describes the idealised objective, not the standardised implementation. For the idealised Cramér–von Mises distance the value at location a is $\int (\mathbb{1}[t \geq a] - F_\nu(t))^2\, dF_\nu(t) = \big(p^3 + (1-p)^3\big)/3$ with $p = F_\nu(a)$, minimised at $p = \frac{1}{2}$ to $1/12$: a bound uniform in the location. For 2-Wasserstein, $W_2^2(\delta_a, \nu_\theta) = \mathrm{Var}(\nu_\theta) + (a - \mathbb{E}\nu_\theta)^2 \geq \mathrm{Var}(\nu_\theta)$, again uniform. *The implemented statistics are not these integrals.* Our standardised Cramér–von Mises variant averages $(\Phi(x_{(i)}) - i/(B+1))^2$ over the sorted batch; at an exactly constant projection every standardised value is 0 and the statistic is

$\frac{1}{B}\sum_i(\frac{1}{2}-\frac{i}{B+1})^2=\frac{B-1}{12(B+1)}$ — 0.0827 at $B=256$, *below* $1/12$, still bounded away from zero, and increasing to $1/12$ with B. The sliced-W_2 estimator of RDMREG compares against sorted target samples, so its value at a point mass is the target sample's empirical variance, positive with probability one. Averaging over directions keeps each floor. □

Proof of Proposition 2(ii). For a batch-constant latent every per-feature standard deviation is 0, so VICReg's variance term $\frac{1}{d}\sum_j \max(0, 1-\mathrm{std}_j)$ equals its maximum 1; the empirical cross-correlation matrix is undefined at exactly zero variance and, under the usual ϵ-regularisation, its diagonal is 0, so Barlow Twins' on-diagonal term $\sum_j(1-C_{jj})^2$ is maximal over the range attainable by our single-view adaptation. □

The converse, and a limit of the dividing line. Part (ii) covers variance hinges and *on*-diagonal terms only. *Off*-diagonal decorrelation runs the other way: a batch-constant latent has every cross-correlation exactly zero, so it *minimises* any redundancy-reduction term. Measured at B=256, d=64: the off-diagonal cross-correlation is 0.000000 at the batch-constant latent against 0.2583 at a healthy isotropic one. A *pure* decorrelation objective — covariance-level, not entrywise, with no variance floor — therefore has the totally collapsed latent among its global minimisers and sits on the dangerous side of the dividing line despite reading the covariance. VICReg and Barlow Twins escape this only because they pair the decorrelation term with a variance floor or an on-diagonal term; strip that pairing and the protection goes with it. We state this because it bounds the proposition: "reads the covariance" is not by itself sufficient for safety, and the dividing line is between objectives that constrain the *spread* of each feature and those that do not, rather than between entrywise and non-entrywise ones.

Numerical confirmation of both propositions. Because the propositions carry the paper's central claim, we verify each numerically as well as proving it. Every constant below is generated and asserted by `src/numerical_confirmation.py` (CPU, seconds; exit 0 = all confirmed); all latents are at unit second moment.

Proposition 1(i), the bound and its attainment. The predicted optimal value under the hard constraint is $f(1)=\mathrm{softplus}(-1)=0.3132617$, and the effective RECTPEN objective at $z\equiv 1$ evaluates to exactly that. Two honesty notes. A *projected* search cannot beat a bound proved on its own constraint set, so we claim no adversarial evidence from one — the closed-form evaluation is the confirmation. And the objective actually trained anchors scale *softly* ($0.1(\mathbb{E}[z^2]-1)^2$), under which the global minimiser is still a constant latent but at $c^*=1.213$ with value 0.2825: the hard-constraint statement is an idealisation, and what matters for the paper — the minimiser is collapsed — survives the relaxation.

Proposition 1(ii). At $(B,d)=(256,64)$ the single-entry latent evaluates to exactly $1/\sqrt{Bd}=0.00781$, the proved minimum; a single-active-row rank-one latent measures 0.0994, 0.0512, 0.0249, 0.0126 at $B=64,256,1024,4096$ (the $C/\sqrt{B}$ witness rate); a full-rank isotropic latent sits at 0.797, matching $\sqrt{2/\pi}=0.798$. The counterexample above is asserted too.

Proposition 2(i). The idealised bound is $1/12=0.0833$. The implemented variant at an exactly constant projection gives $(B-1)/(12(B+1))=0.0827, 0.0830, 0.0832$ at $B=256,512,1024$ (float64); in float32 the ϵ-clamped standardisation turns rounding noise into spread and the measured values are 0.0866–0.0879. Both are bounded away from zero; neither is the idealised constant, and an earlier version conflated them.

Proposition 2(ii), the maxima. At a batch-constant latent VICReg measures 0.9801 against its ceiling of 1 (isotropic: ≈0.24) and Barlow Twins measures exactly 1.0000, its maximum (isotropic: 0.0012).

What the propositions do and do not say. They characterise *optima of the auxiliary term over free latents* — not the dynamics of a parameterised encoder trained jointly with a predictor, an EMA target and Adam, and not local stability or basins of attraction, which we do not analyse. What Proposition 1 establishes is that for these objectives the auxiliary term never prefers a healthy latent to the collapsed one; that is consistent with the observed lock-in once the prediction gradient starves (§5), but it is not a proof that training reaches the global optimum, and the four-decimal repetition is

of *probe accuracy*, which any latent that makes the linear probe predict the majority class produces. Proposition 2 says a distribution-level objective never *wants* the collapsed state; it does not say the optimiser cannot arrive there for other reasons — which is exactly SIGREG's stochastic failure, a dynamics effect outside the scope of any statement about optima, and the reason the diagnostic is a warning sign and not a certificate.

H STATISTICAL SCOPE AND OUTCOME DEFINITIONS

Outcome definitions, per experiment. "Strict" is the same everywhere: final linear-probe accuracy within 5×10^{-3} of the dataset's majority share (the census's historical 1pp label coincides with it on all 163 rows). "Soft" is *not* one quantity, and we do not pool it across experiments. Its comparator is: the 300-epoch no-reg baseline for the census, the placement grid and the 300-epoch faithful replication; the matched 100-epoch no-reg baseline for the class experiment, the 100-epoch faithful A/B and the review controls; each penalty's own z_{pred} arm for the repair; the in-sweep β=0.1 arm for the vision weight sweep; and the 60-epoch no-reg arm for the vision class arm. The margin is 3pp throughout. Strict and soft are not nested — a run can fall far below the majority share and still fail the strict band (the four release/rescue probes of §3) — so neither is a severity scale. Every 300-epoch run's final probe is at epoch 280 (cadence 20).

What the tests condition on. The exact McNemar tests and bootstrap intervals of the placement grid treat the 42 (dataset, regulariser, seed) pairs as the sampling units; they are conditional on these three datasets and this recipe and are not inferences about a population of datasets. At the dataset level the RECTPEN effect is consistent in all three datasets (paired $z_{\text{pred}} - z_{\text{enc}}$ gain $+9.9$pp adult, $+30.5$pp helena, $+24.8$pp jannis; every one of the 15 pairs positive, minimum $+9.4$pp), but with $n = 3$ a sign test cannot exceed $p = 0.25$, and we make no dataset-population claim. Likewise the $33/33$ β sweep spans three datasets and three seeds per setting, and the 28-run census spans seven (dataset, seed) units: they measure robustness to configuration, not independent replications.

What a zero count bounds. Zero strict collapses is not a safety result. Under an independent-runs model the two-sided 95% Clopper–Pearson upper bound is 33.6% for $0/9$ (the faithful control at either budget), 26.5% for $0/12$ (the vision arms) and 18.5% for $0/18$. We report absence of the deterministic majority-exact endpoint, which at these rates would be unmissable given $24/28$ for RECTPEN, not an equivalence test.

I FAIR RE-EVALUATION, PRECISION AND GRADIENTS

The standard T-JEPA probe is a linear classifier on raw encoder features trained for a fixed budget. As shown in §4, a collapsed latent can defeat it while still carrying the label. This appendix re-evaluates every paired experiment under one fair protocol, locates where the label information lives numerically, and measures the gradient each regulariser supplies near collapse. All of it runs on existing checkpoints and ships with the code.

One protocol for both placements. For every jannis run with a surviving checkpoint (the placement grid, the ℓ_1/hinge class experiment, the faithful control and the repair; `src/fair_reprobe_placement.py`) we extract the EMA target-encoder features — the representation T-JEPA probes — and fit an L2 logistic regression on the run's own seeded train split, choosing $C \in \{10^{-2}, 1, 10^{2}\}$ on its validation split, once on raw features and once after per-feature z-scoring with training statistics. *Raw* reproduces the paper's collapsed numbers (0.460 for encoder-placed RECTPEN); *fair* is the z-scored readout.

Attribution. T-JEPA names a checkpoint directory by configuration and launch second, and the checkpoint files carry no run identity, so runs launched in parallel can share a directory and overwrite each other's epochs. `src/attribute_ckpts.py` lists every jannis run launched within 5 s of each directory (`data/ckpt_attribution.csv`); we use a checkpoint only if every such run belongs to the arm it is reported under, and count a shared directory once. Of the 56 matched runs this leaves 40 distinct checkpoints — none in the SIGREG-CvM encoder arm — and per-seed pairing only where both arms are attributable to single runs. The same rule restricts the re-probe

of §4 to six collapsed RECTPEN checkpoints; of the three collapsed checkpoints in the precision check below, one shares its directory with concurrently launched SIGREG runs, and all three are batch-constant to numerical resolution whichever run wrote them.

experiment	objective	n_{enc}	n_{pred}	raw z_{enc}	raw z_{pred}	fair z_{enc}	fair z_{pred}
class	hinge	3	3	0.460	0.685	0.643	0.693
class	ℓ_1	1	2	0.460	0.678	0.640	0.680
faithful	RDMREG (published loss)	1	2	0.690	0.573	0.695	0.668
grid	Barlow (1-view)	3	3	0.657	0.715	0.692	0.719
grid	RECTPEN	4	4	0.460	0.722	0.562	0.723
grid	SIGREG-CvM	0	2	—	0.718	—	0.716
grid	VICReg (1-view)	3	3	0.666	0.715	0.697	0.718
repair	hinge + floor	3	0	0.587	—	0.678	—
repair	ℓ_1 + floor	3	0	0.460	—	0.682	—

The placement effect survives the fair protocol but shrinks. For RECTPEN the gap is 16.1pp (0.562 vs. 0.723) against 26.2pp under the raw probe, positive in each of its three seed pairs; for the everyday penalties it is 5.0pp for the hinge (all three seed pairs positive) and 4.0pp for ℓ_1 (arm means only), against 22–23pp raw; for single-view Barlow and VICReg 2–3pp. For the published RDMREG loss the one attributable encoder-arm checkpoint (0.695) reads above the predictor arm (0.668, which contains the control's one predictor-side collapse): no encoder-side deficit, on thin evidence. So scale collapse accounts for most of the raw gap of the everyday penalties and for about 40% of RECTPEN's. The repaired penalties, which the raw probe scores as collapsed, read 0.678 and 0.682 — above the unrepaired encoder arms (0.64) and near the predictor arms: the floor restores information without restoring scale.

Where the information lives. `src/fp64_forward_check.py` runs three collapsed checkpoints and a healthy control with weights and inputs in float64 (TF32 off), and repeats the readout after storing the features at lower precision before or after centring.

checkpoint	residual	fp32/fp64 corr.	fp64 probe	bf16 store→centre	bf16 centre→store	distinct (bf16)
collapsed (fix)	$2.4e-07$	0.990	0.658	0.460	0.660	1/4096
collapsed (fix)	$2.1e-07$	0.988	0.654	0.460	0.657	1/4096
collapsed (fix)	$1.7e-07$	0.981	0.659	0.460	0.667	1/4096
healthy (fix)	$6.5e-01$	1.000	0.732	0.727	0.727	4096/4096

The collapsed residual is $\sim 2\times 10^{-7}$ of the latent's scale in both dtypes, correlates 0.98–0.99 between fp32 and fp64 and is batch-size invariant, so it is computed by the network rather than produced by rounding. It is also only one or two fp32 ulps above the common vector, which is why half precision destroys it: stored in bf16 (or fp16) before centring, all 4096 inputs of each collapsed checkpoint become one vector and the probe returns the majority share, while centring first preserves the readout. The healthy control is unchanged under every dtype and order.

A counterfactual isolates scale as the cause (`src/scale_offset_counterfactual.py`, `data/scale_offset_counterfactual.csv`). Taking the latent H of a healthy checkpoint and the common vector μ of a collapsed one, both uniquely attributable, we probe $Z' = \mathbf{1}\mu^\top + sH$ for s from 1 to 10^{-8}. The label information in Z' is by construction that of the healthy latent, yet the raw probe returns the majority share (0.460) for every $s \le 10^{-4}$ and bf16 storage maps all 2048 probed inputs to one vector at the same scales, while the z-scored probe reads 0.709–0.711 at every s. The measured collapsed residual ($\sim 2\times 10^{-7}$) lies deep in that regime: scale alone suffices to produce the raw-probe and half-precision observations.

The restoring gradient near collapse. `src/regularizer_gradient_profile.py` evaluates each regulariser along $z(s) = \mathbf{1}\mu^\top + sH$ (H centred, unit per-feature variance) and records $\partial\mathcal{L}/\partial s$, whose negative sign means the term pushes the residual back up.

regulariser	$\partial\mathcal{L}/\partial s$ at s=1	at s=10^{-3}	at s=10^{-7}
repair: ℓ_1 + floor (ϵ=10^{-4}, additive)	0.96	-0.1	-1.0×10^{-5}
repair: hinge + floor (ϵ=10^{-4}, additive)	0.68	-0.1	-1.0×10^{-5}
variance hinge, additive ϵ=10^{-4} (VICReg form)	0	-0.1	-1.0×10^{-5}
variance hinge, clamped ϵ=10^{-4} (our adaptation)	-0.41	0	0
variance hinge, additive ϵ=10^{-8}	0	-1	-1.0×10^{-3}
SIGREG (reference, EP)	-80	-0.98	-9.8×10^{-5}
RDMREG sliced W_2 (published loss)	-0.21	-1.9	-1.9

Every variance-floor term penalises the near-constant state, yet its restoring derivative scales like $s/\sqrt{s^2+\epsilon}$ and is $\approx10^{-5}$ at $s = 10^{-7}$ — exactly zero for a clamped floor — and the reference SIGREG's derivative also shrinks with s. The sliced-W_2 loss of the published RDMREG compares sorted projections with sorted target samples, so its restoring derivative does not vanish: it is -0.2 near the loss's own optimum (s=1) and settles at -1.9 for every $s \leq 10^{-2}$. This is consistent with the repair's failure and with RDMREG being the one clean survivor, but it is measured on free latents only; the encoder Jacobian, the prediction loss and the optimiser are not included, and we do not claim it as the cause.

The review controls, in full. `scripts/review_factorial_ab.sh` fixed three predictions before its 36 runs (100 epochs; criteria as in Appendix H):

objective	placement	adult	jannis	helena	strict / soft
SIGREG (reference, EP)	ctx	0.839	0.613	0.274	0/9 / 5/9
SIGREG (reference, EP)	pred	0.761	0.419	0.059	6/9 / 9/9
SIGREG-CvM (ours)	ctx	0.848	0.545	0.203	0/9 / 6/9
RDMREG loss, no ReLU	ctx	0.844	0.585	0.215	0/9 / 6/9
majority share		0.761	0.460	0.061	
no-reg, 100 ep		0.828	0.651	0.339	

P1 (reference SIGREG at z_{enc}, no strict collapse) and P3 (RDMREG's loss without its ReLU, no strict collapse) held. P2 (reference SIGREG at z_{pred}, no strict collapse) failed: $6/9$ runs are majority-exact. The training-free check scores this objective as penalising collapse ($119\times$; `data/diagnostic_sigreg_ep.csv`), so it is a false negative of the check. Two differences between the placements may matter and were not controlled: the reference statistic is multiplied by the number of flattened tokens, which differs between z_{enc} and z_{pred}, and β was not re-tuned per placement.

J EXTENDED RELATED WORK

JEPA collapse machinery and its principled replacements. I-JEPA (Assran et al., 2023) and V-JEPA 2 (Assran et al., 2025) rely on EMA targets and asymmetric stop-gradients; Mo & Tong (2024) document that these alone are insufficient and connect JEPA to contrastive objectives; T-JEPA (Thimonier et al., 2025) adds learnable regularisation tokens. VICReg (Bardes et al., 2022) and Barlow Twins (Zbontar et al., 2021) pioneered collapse prevention through feature-statistics losses; SIGREG (Balestriero & LeCun, 2025) and RDMREG (Kuang et al., 2026) are their JEPA-era descendants with distributional targets and theoretical framing. Our results add a cautionary entry — the principled replacements can be the collapse vector — with an important qualification: we study a penalty-form surrogate of the rectified objective and, as a control, the published RDMREG itself, and the distinction carries the result (§6). We state this prominently because a reader could otherwise take our failure as evidence against that method, which it is not. Our released code includes the faithful RDMREG implementation and single-view adaptations of VICReg and Barlow Twins under both placements.

Concurrent theory and scale. Klindt et al. (2026) *prove* that LEJEPA's alignment-plus-Gaussian objective linearly identifies the world's latent variables at the optimum, with Gaussianity uniquely enabling the guarantee. This is complementary rather than contradictory: identifiability characterises the *optimum*, whereas we show the *optimisation path* can commit — within five epochs,

and without recovery under the interventions we tried — to a near-constant basin, and that for a badly formed objective that basin *is* the optimum. A guarantee about the destination says nothing about which basin SGD reaches, and says nothing at all if the destination is wrong. Meanwhile Kuhn et al. (2026) deploy per-modality distributional regularisation in large-scale non-contrastive vision-language pretraining — decisions of exactly the kind we study are now being made at scale, typically reported only in code.

Relation to projector-head folklore, and to gradient balancing. SSL has long applied its losses behind a projection head (Chen et al., 2020), a predictor (Grill et al., 2020; Chen & He, 2021), or an expander (Bardes et al., 2022), taking *pre-head* features downstream; Bordes et al. (2023) study this systematically, so "do not put the loss directly on the representation" is established practice. Our contribution is not that practice but three things it does not contain: (i) in JEPAs, violating it with the wrong objective does not merely cost quality — it produces *deterministic, majority-class-exact, irreversible* collapse with sub-five-epoch lock-in; (ii) the flagship heuristic-free line states its guarantees on the encoder output while its implementations apply the loss to a *projector* output (Balestriero & LeCun, 2025; Panchavati et al., 2026) — the guaranteed and the regularised tensor differ, the folklore head is retained but undocumented as load-bearing, and our results measure what happens without it; (iii) we identify the objective-side property that decides which happens, and give a training-free test for it. Our ρ is the magnitude-only special case of the per-loss gradient norms used by GradNorm (Chen et al., 2018) and the gradient-surgery line (Yu et al., 2020), reported as a correlate precisely because it does not separate outcomes alone. On dynamics, Tian et al. (2021) analyse how the predictor and stop-gradient shape collapse in non-contrastive SSL — which our predictor-as-adapter reading informally re-derives — and Jing et al. (2022) characterise dimensional collapse, the spectral failure mode our vision result shows can be absent while the representation is healthy.

Rank diagnostics and evaluation hygiene. The entropy-based effective rank (Roy & Vetterli, 2007) was repurposed by RankMe (Garrido et al., 2023) as an SSL quality metric. We use it as a collapse telltale and find a caveat: healthy runs show transient deep rank dips, so instantaneous rank is not a sufficient statistic for failure. A companion manuscript (Anonymous, 2026) (under double-blind review; cited for provenance, not evidence) documents the seed-determinism and metric-implementation pitfalls our rigour notes inherit; the discovery reported here emerged from that audit — an argument that aggressive internal replication is instrumentation, not overhead.

K FULL LIMITATIONS

(i) The *deterministic* result is anchored on T-JEPA (tabular, d=64); vision (two ViT scales) and MTS-JEPA enter as three-seed boundary probes, and large-scale vision/video JEPAs remain untested (see Scope). (ii) Several mechanism probes are $n = 1$ against a deterministic phenomenon; the headline claims (placement A/B, determinism, knife-edge) carry $n \geq 8$. (iii) "No rescue" is bounded by the seven interventions tried; targeted surgery might yet succeed and would further localise the mechanism. (iv) Both non-tabular probes are weaker than the tabular A/B: the MTS arms regularise different token supports and the harness logs no downstream metric, and the vision harness mean-pools before the z_{enc} regulariser (§7). (v) The faithful-RDMReg control is 27 runs (9 at z_{enc} at 100 epochs, replicated $9/9$ at the full 300-epoch budget (final probe at epoch 280, cadence 20), 9 at z_{pred}); it establishes the absence of the deterministic majority-exact collapse, which is unmissable at this n given the $24/28$ contrast, but it is not a power-matched equivalence test. (vi) The precondition of §6 is validated only on the objectives of Table 4 (sensitivity $3/7$, specificity $2/2$ there); it is a warning sign, not a safety test, and not a mathematical necessary condition for safe training. (vii) This study originates in an audit of prior work; the 112 predictor-placement runs predate the recognition of placement as a variable. The 163-run census, per-run manifests, and one-command reproduction recipes ship with the (anonymised) code; raw training logs, which contain absolute paths, are retained in the full repository and released after de-anonymisation.

L DIRECT GRADIENT MEASUREMENT, IN FULL

The ρ intervention, in full. A natural hypothesis is that collapse is driven by the regulariser *overpowering* the prediction signal, i.e., by the ratio of the two gradients reaching the encoder, $\rho = \|\nabla_E \mathcal{L}_{\text{reg}}\| / \|\nabla_E \mathcal{L}_{\text{pred}}\|$. ρ is linear in β, so sweeping β at *fixed* encoder placement moves ρ over orders of magnitude with everything else held constant ($\beta \in \{10^{-4}, \ldots, 1\}$ on jannis, $\{10^{-4}, 10^{-2}, 1\}$ on adult and helena, three seeds each). The hypothesis fails everywhere: *all* 33 *runs collapse to the majority share* (jannis $15/15$, adult $9/9$, helena $9/9$) while dip-median ρ spans 0.20 to 1.7×10^4 (Figure 12). At $\beta = 10^{-4}$ the regulariser gradient is a third of the prediction gradient — weaker than in any healthy predictor-placed run we measured — and the representation still locks, so magnitude does not govern the outcome: ρ is a by-product that escalates once the prediction gradient starves. The $\beta{=}0$ control is healthy ($0/18$), so a threshold exists somewhere in $(0, 10^{-4})$ that we did not locate and would not recommend operating near. What governs the outcome is where the objective's minimiser lies (§6): shrinking β slows the descent without changing where the gradient points.

Because Adam's second moment is an exponential moving average and could in principle be stale, we also measure the *instantaneous* prediction-loss gradient reaching the encoder, $\|\nabla_{\text{enc}} \mathcal{L}_{\text{pred}}\|$, once per epoch and isolated from the regulariser (Figure 5). Two runs with identical seed and starting gradient (3.9×10^{-3}) diverge through the danger window: the encoder-placement run starves to $\sim4\times10^{-12}$ — an eight-order-of-magnitude collapse — and stays there, while the predictor-placement run dips to 5×10^{-5} then recovers to $\sim$1–9. This is the bifurcation at the level of the gradient itself: in the collapsed branch the prediction loss provides nothing for the encoder to follow, confirming the starvation directly rather than through the optimizer's EMA.

M PER-BASIN MARGINS: WHY THE PRESCRIBED REPAIR IS NOT ENOUGH

The main text reports the repair's outcome; this appendix gives the measurement that explains it (`data/collapse_basin_margins.csv`, produced by `src/collapsed_state_loss.py`). Instead of one collapse basin we evaluate each objective at three — the batch-constant latent, a rank-one all-positive latent, and a rank-two all-positive latent — and report $\mathcal{L}(\text{basin})/\mathcal{L}(\text{healthy})$, so that a value below 1 means the objective *prefers* that basin.

	vs. signed healthy			vs. non-negative healthy			
objective	const	r1	r2	const	r1	r2	outcome at z_{enc}
hinge $\mathbb{E}[\text{relu}(-z)]$	0.00	0.00	0.00	1.0	1.0	1.0	9/9 strict
ℓ_1 $\mathbb{E}\|z\|$	1.01	0.80	0.94	1.01	0.80	0.94	9/9
RECTPEN	0.50	0.59	0.53	0.99	1.17	1.05	24/28
hinge + variance floor	2.37	1.30	1.31	2.48	1.36	1.37	4/9, all 9 soft
ℓ_1 + variance floor	2.20	1.45	1.59	1.50	0.99	1.08	3/9, all 9 soft
RDMREG (published)	5.84	3.02	2.54	7.48	3.87	3.30	0/9
SIGREG	383	13.6	9.29	369	13.1	8.96	7/15 stochastic
VICReg	3.95	34.0	14.4	5.07	43.7	18.6	0/6 strict
Barlow Twins	805	252	179	810	253	180	2/6 strict

Two readings. First, the repair does what Proposition 2(ii) says it does: the batch-constant latent goes from the global optimum (0.00 for hinge) to $2.4\times$ worse than health. Second, that is not where the runs end up. The low-rank basins beside it move only to 1.3–$1.6\times$, while every objective that survives training in our study penalises rank-one by at least $2.5\times$ and usually by one or two orders of magnitude more. A repair that raises the exact optimum a theorem names, without raising its neighbourhood, buys a reduction in the majority-exact rate ($18/18 \to 7/18$) and no quality at all ($18/18$ still more than 3pp below their matched z_{pred} arms).

We report this as the reading the numbers support rather than a demonstrated mechanism. One caveat is visible in the table: `rdmreg_noneg` (see `data/collapse_basin_margins.csv`, where every ratio is 1.0 by the indifference floor) does *not* collapse in $3/3$ runs — its effective objective evaluates to ≈0 at both latents, so no margin is defined for it. Establishing a reliable

intervention	tests	outcome	verdict
remove reg after E epochs	$E \in \{5, 10, 20, 80\}$	locked; E=5 wanders at 0.40	locked within $E \leq 5$
delay onset to epoch E	$E \in \{40, 80\}$	established reps erode to majority	no safe window
weight reduction	$\beta \in \{0.01, 0.001\}$	identical-speed collapse	direction, not magnitude
slow EMA + release	$m \in \{0.999, 0.9999\}$	locked	H5 refuted
optimizer reset + release	Adam state cleared	locked	H6 refuted
weight noise + release	$\sigma \in \{0.01, 0.05\}$	wandering death 0.34–0.40	region, not point
noise + reset combo	σ=0.02 + reset	wandering death	—
placement $\mathbf{z}_{\text{pred}}$	2 datasets, incl. collapsing seeds	**healthy** (0.704, 0.859)	only survivor

Table 3: The rescue campaign (jannis, RECTPEN, seed 0): every post-hoc intervention attempted, and its outcome. Only placement survives.

margin threshold would need a larger family of objectives than the ones we ran, and we do not claim one.

N THE CLASS EXPERIMENT AND THE REPAIR, IN FULL

The class, tested on penalties people actually use. RECTPEN could be dismissed as a strawman of our own construction, so we tested two everyday members of the class, pre-registering the prediction (`scripts/entrywise_class_ab.sh`): the ℓ_1 sparsity penalty $\mathbb{E}|z|$ and the hinge non-negativity penalty $\mathbb{E}[\text{relu}(-z)]$ — the most literal penalty form of an architectural ReLU — each with the same ℓ_2 scale anchor, at $\beta = 0.1$, through the identical placement A/B on all three datasets (3 seeds; 36 runs). **Both collapse deterministically at z_{enc} and neither collapses at z_{pred}:** hinge $9/9$ (jannis 0.458, adult 0.7606, helena 0.061 — the majority shares) against $0/9$ (0.65, 0.84, 0.34); ℓ_1 $9/9$ against $0/9$. (An earlier draft read the theorem's two basins in these endpoints; as stated in the main text, the per-dataset deviations overlap and reverse order on jannis, so that attribution is withdrawn — a linear probe cannot distinguish the basins in any case.) The failure is therefore not a property of one contrived loss: it is what a per-feature scalar penalty does to a JEPA encoder whose prediction gradient is weak at some point in training.

The repair the theorem prescribes — and why it is not enough. If the propositions are prescriptive they should say how to *fix* a failing objective, and Proposition 2(ii) appears to: a per-feature variance floor is *maximised* by the batch-constant latent, so adding one to an entrywise penalty destroys the collapsed optimum while changing nothing about what the penalty asks for pointwise. We used it prospectively: before launching any job the training-free check scored the repaired objectives and flipped both verdicts (hinge 0.000 vs. 0.399 becomes 0.553 vs. 0.418; ℓ_1 0.623 vs. 0.798 becomes 1.176 vs. 0.817), and we pre-registered those numbers with the prediction that the repaired penalties would not collapse (`scripts/repair_ab.sh`). Run at z_{enc} on three datasets $\times$ three seeds against the identical unrepaired arms that collapse $9/9$: **moving the optimum is not sufficient to move the optimiser**: the variance floor penalises the batch-constant state (its term is maximal there; for the hinge this makes a non-negative unit-variance latent strictly better than any batch-constant one), but not rank-one latents, whose per-feature variance can be positive — the training-free check flips both objectives to *penalises collapse* — yet $7/18$ runs still lock at the majority share and *all* $18/18$ fall more than 3pp below their matched z_{pred} arm. This is the sharpest evidence we have for the scope limit we state throughout: the propositions characterise *optima*, and an objective whose optimum is healthy can still be optimised into a collapsed state. The practical reading is not "bolt a variance term onto an entrywise penalty" but "use an objective that is distribution-level by construction". repaired hinge: majority-exact in $4/9$, soft in $9/9$ — jannis 0.504 (unrepaired 0.458, z_{pred} 0.654), adult 0.760 (unrepaired 0.761, z_{pred} 0.839), helena 0.073 (unrepaired 0.061, z_{pred} 0.342); repaired ℓ_1: majority-exact in $3/9$, soft in $9/9$ — jannis 0.439 (unrepaired 0.459, z_{pred} 0.661), adult 0.761 (unrepaired 0.761, z_{pred} 0.840), helena 0.050 (unrepaired 0.060, z_{pred} 0.334).

O FULL RESULT TABLES AND FIGURES

Objective	$\mathcal{L}$(collapsed)	$\mathcal{L}$(healthy)	ratio	verdict	outcome at z_{enc} *
RECTPEN	0.482	0.806	0.60	**prefers collapse**	$24/28$ **majority-exact**
RECTPEN (negonly)	0.482	0.806	0.60	**prefers collapse**	$3/3$ **collapsed**, $0/3$ exact
hinge $\mathbb{E}[\text{relu}(-z)]$	0	0.399	0	**prefers collapse**	$9/9$ **majority-exact**
ℓ_1 $\mathbb{E}\lvert z\rvert$	0.623	0.798	0.78	**prefers collapse**‡	$9/9$ **collapsed**
RECTPEN (noneg)	0	0	—	indifferent	$0/3$
RDMREG (published)	0.744	0.246	3.0	penalises collapse	$0/9$
SIGREG-CvM	3.1×10^{-3}	2.3×10^{-4}	13.5	penalises collapse	$7/15$ stochastic
SIGREG (reference, EP)	122.4	1.03	119	penalises collapse	$0/9$; $6/9$ at z_{pred}
VICReg	8.64	0.247	35	penalises collapse†	$0/6$ strict, $3/6$ soft
Barlow Twins	0.313	1.2×10^{-3}	253	penalises collapse	$2/6$ strict

Table 4: **A training-free precondition predicts deterministic collapse.** Each objective evaluated at a collapse-basin latent and a healthy latent, both renormalised to unit second moment so no objective can separate them by scale alone, and scored on the *differentiable* part of the loss only (B=256, d=64, 8 draws; `src/collapsed_state_loss.py`). The four objectives that prefer the collapsed state are exactly the four that collapse T-JEPA deterministically at z_{enc}. At matched scale RECTPEN's ℓ_2 term vanishes identically, so its effective objective *reduces* to the softplus term — which is why the top two rows coincide and why the term ablation flips the outcome. †VICReg's three exceptions in the robustness sweep are all $d \geq B$ with an exactly constant latent, where its variance hinge saturates; we report the regime rather than average it away. *For the reference SIGREG (`data/diagnostic_sigreg_ep.csv`) the outcome after the semicolon is at z_{pred}, where it collapses despite the verdict (Appendix I). "Penalises" does not mean safe: VICReg, Barlow and SIGREG-CvM all fail softly or stochastically. ‡ℓ_1 prefers the sparse rank-one basin ($90/90$ constructions) and is within 3% of a tie on the constant one, as Proposition 1(ii) predicts. The hinge and ℓ_1 rows are the class experiment of §6: penalties in everyday use, run through the identical A/B (`data/entrywise_class_ab.csv`).

regulariser	placement	n	strict collapse	soft collapse
BARLOW	ctx	6	2/6	6/6
BARLOW	pred	6	0/6	0/6
RECTPEN	ctx	15	15/15	15/15
RECTPEN	pred	15	0/15	0/15
SIGREG	ctx	15	0/15	3/15
SIGREG	pred	15	0/15	0/15
VICREG	ctx	6	0/6	3/6
VICREG	pred	6	0/6	0/6

Table 5: Systematic placement grid (fresh runs, disjoint from the census; the harness predates the collapse criteria, so unlike the five interventional experiments it is *not* pre-registered). Four distribution-matching regularisers $\times$ two placements; RECTPEN/SIGREG at 5 seeds over {jannis, adult, helena}, VICReg/Barlow at 3 seeds over {jannis, adult}. *strict* collapse = within $5\text{e}{-3}$ of the majority share; *soft* $=>$ 3pp below the no-reg baseline. Predictor placement shows zero collapses for every regulariser and dataset ($0/42$); encoder placement degrades or collapses across the entire family. The no-reg control (18 runs across the three datasets) collapses 0 times under either definition, so the failure is introduced by the encoder-placed regulariser, not by JEPA training itself. The grid is fully paired by (dataset, regulariser, seed): exact McNemar over the 42 pairs gives $p = 1.5\times10^{-5}$ (strict) and 1.5×10^{-8} (soft) pooled; RECTPEN alone $p = 6.1\times10^{-5}$ (Holm-adjusted 2.4×10^{-4}), paired accuracy delta $+21.7$pp, bootstrap 95% $[+17, +26]$ (`src/collapse_stats.py`).

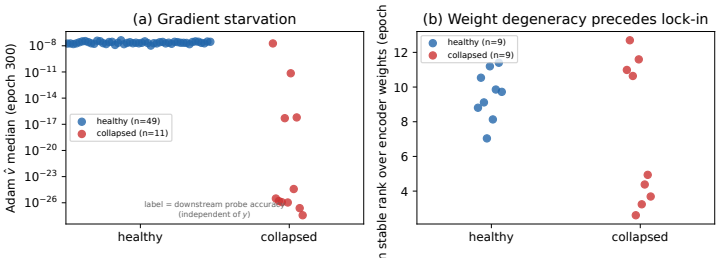

Figure 4: Checkpoint forensics, with each checkpoint labelled *collapsed* or *healthy* by its run's downstream probe accuracy — a signal independent of the plotted y-axis statistics (mapping in Appendix B). (a) Median Adam second moment $\hat{v}$ over encoder parameters at epoch 300: healthy runs sit at $\sim 3\times10^{-8}$; collapsed runs at 10^{-17}–10^{-28} — the collapsed encoder receives essentially zero gradient. (b) Minimum stable rank over encoder weight matrices at epoch 19: collapsed runs (red) have already degenerated before lock-in completes.

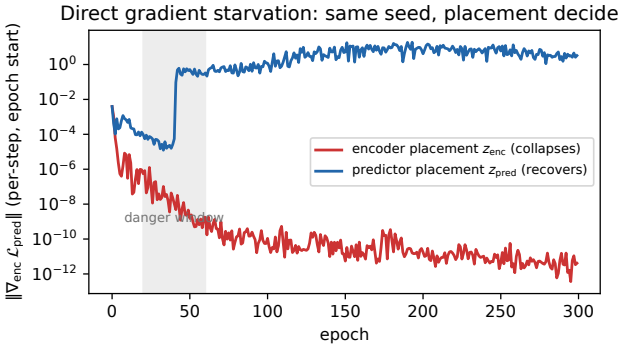

Figure 5: Instantaneous prediction-loss gradient norm on the encoder (logged once per epoch, isolated from the regulariser). Same seed and identical starting value; placement decides whether the gradient starves to 10^{-12} (encoder placement, collapse) or recovers past the danger window (predictor placement).

Dataset	Majority	z_{enc} mean (n)	z_{pred} mean (n)	strict collapses
adult	0.7607	0.8557 (3)	0.8421 (3)	0/3 / 0/3
helena	0.0614	0.3346 (3)	0.3356 (3)	0/3 / 0/3
jannis	0.4601	0.6339 (3)	0.5934 (3)	0/3 / 1/3
pooled	—	0.6080 (9)	0.5904 (9)	**0/9** / 1/9

Table 6: **The published RDMREG does not collapse at the placement its own line of work uses.** Same harness, same flag, same seeds as Figure 1, at a 100-epoch budget, replicated in full at the census's 300-epoch budget (0/9 strict and 0/9 soft at z_{enc}; every trajectory rises from the epoch-100 to the final probe (Figure 6) — the probe cadence of 20 places it at epoch 280 — +0.1 to +5.2pp — `data/faithful300_ctx.csv`); final linear probe, strict criterion (within 5e−3 of the majority share), reported as $z_{\text{enc}}/z_{\text{pred}}$. Encoder placement is healthy on all three datasets — helena is 5.4× the majority share, adult +9.5pp over it. The one strict endpoint is a *predictor*-placed jannis seed (0.4575), the same rare z_{pred} case documented in §7; encoder placement produced none. Contrast: encoder-placed RECTPEN collapses 28/28, 24 of them majority-exact. Raw values: `data/faithful_rdmreg_ab.csv`.

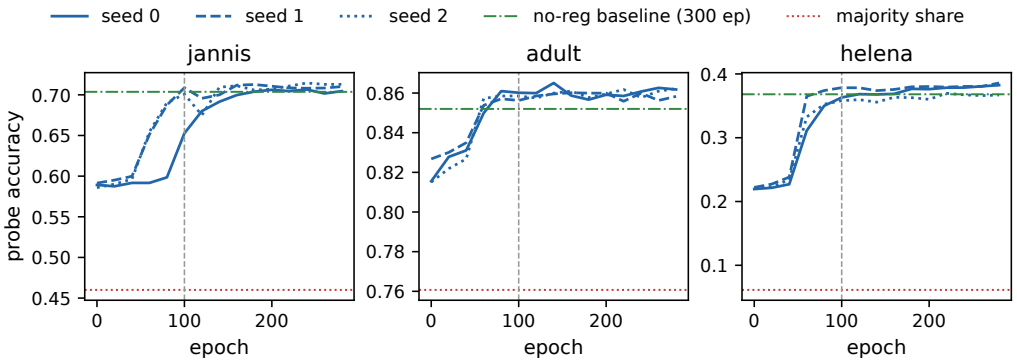

Figure 6: **The budget-parity replication, in full.** All nine encoder-placed (z_{enc}) runs of the published RDMReg at the census's 300-epoch budget (blue; one panel per dataset, three seeds each), against the 300-epoch no-reg baseline (green dash-dot) and the majority share (red dotted). The grey dashed line marks epoch 100, the original control's budget: every trajectory *rises* from there to the final probe (+0.1 to +5.2pp), the opposite of the late-onset erosion the onset grid licenses one to fear. Raw values: `data/faithful300_ctx.csv`.

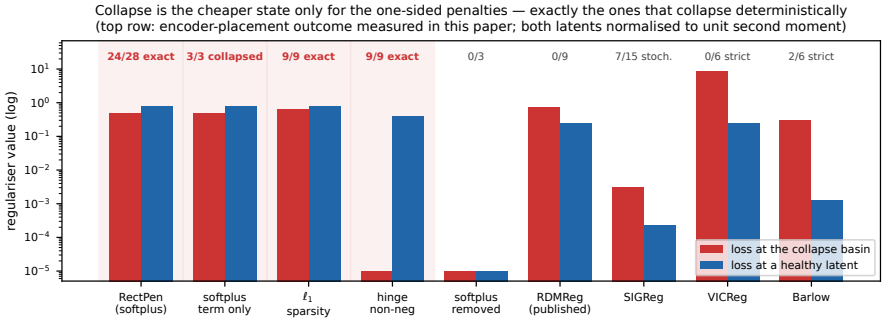

Figure 7: **Does the collapsed state minimise the regulariser?** Each objective evaluated at a collapse-basin latent (dark) and a healthy latent (light), log scale, with the measured T-JEPA encoder-placement outcome above each pair. The four objectives for which collapse is the *cheaper* state (shaded) are exactly the four that collapse T-JEPA deterministically; the published RDMReg scores collapse 3.0× worse than health and does not collapse.

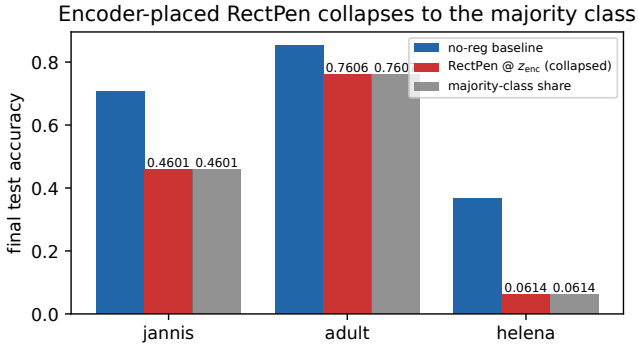

Figure 8: Encoder-placed RectPen collapses to the majority class on all three datasets. The collapsed accuracy reproduces the majority share (median deviation 8×10^{-5}, max 2×10^{-3}) — the latent is constant up to a tiny residual, and the raw-feature linear probe defaults to the majority class; a standardised probe still recovers most of the label information (§4).

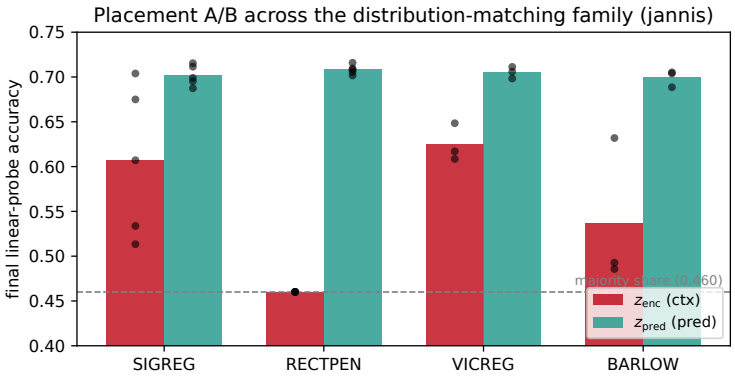

Figure 9: Placement A/B across the distribution-matching family on jannis. Encoder placement (z_{enc}, red) drives every regulariser below the no-reg baseline — to the majority share for RECTPEN and Barlow — while predictor placement (z_{pred}, green) is healthy for all four. Points are per-seed final probe accuracies.

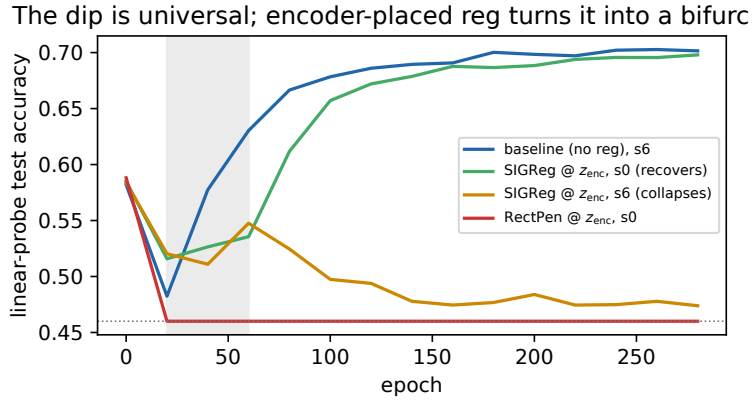

Figure 10: The accuracy dip at epochs 20–40 is universal — the unregularised baseline (blue) traverses and recovers from it. Encoder-placed regularisation converts the dip into a bifurcation: RECTPEN (red) always exits on the collapsed branch; SIGREG exits on either branch depending on seed (green vs. orange).

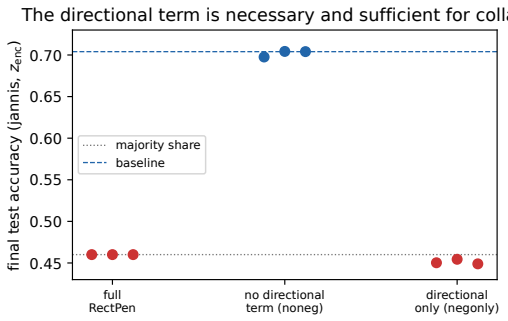

Figure 11: RECTPEN term ablation at encoder placement (jannis, three seeds each). Removing the directional non-negativity term (`noneg`) eliminates the collapse; keeping only it (`negonly`) still collapses every run (0.449–0.454), though 6–11$\times 10^{-3}$ below the majority share, i.e. 0/3 by the strict criterion. In this ablation the directional term is necessary and sufficient.

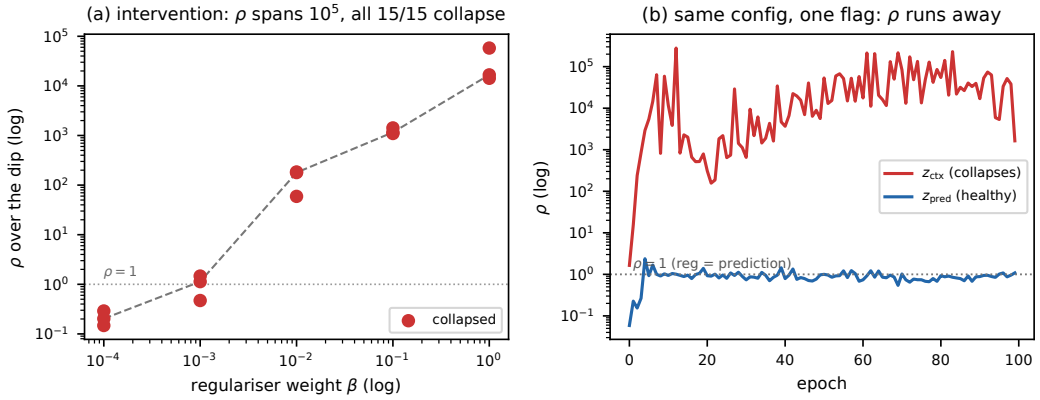

Figure 12: **Gradient *magnitude* does not decide collapse; direction does.** (a) Intervention at fixed encoder placement: sweeping the regulariser weight $\beta \in [10^{-4}, 1]$ moves the encoder-reaching gradient ratio $\rho = \|\nabla_E \mathcal{L}_{\text{reg}}\| / \|\nabla_E \mathcal{L}_{\text{pred}}\|$ across five orders of magnitude (dashed line: per-β median), yet every one of the 15 runs still locks at the majority share — including at $\rho \approx 0.2$, where the regulariser's gradient is weaker than the prediction gradient. No loss-weight setting avoids the collapse. (b) The same quantity within a matched pair differing only in placement: at z_{enc} it escalates past 10^5 as the prediction gradient starves, at z_{pred} it holds near 1. ρ therefore describes the collapsed branch but does not cause or predict it.