

On the Geometry of Multimodal Saturation: Riemannian VICReg

Nessim Ben Abbes

Protectline

nessim.benabbes@protectline.fr

Duc Han Le

Télécom Paris

duc-han.le@telecom-paris.fr

Sabri Mtibaa

Protectline

sabri.mtibaa@protectline.fr

Van-Tam Nguyen

Télécom Paris

van-tam.nguyen@telecom-paris.fr

Abstract

In self-supervised learning, a third modality should improve, or at least preserve, performance. Across nine image–text–tabular datasets, we show that it instead harms performance: the trimodal model underperforms its own best bimodal subset in 55.6% of paired runs under VICReg. The same failure occurs in 51.1% of paired runs under SimSiam. We call this failure multimodal saturation. We propose that the failure lies in the alignment geometry. Riemannian VICReg (R-VICReg) generalizes classical VICReg: it aligns views by squared geodesic distance on learnable negative-curvature product factors and recovers VICReg exactly as curvature vanishes. Over the same 45 paired runs, R-VICReg raises the probability that the third modality helps from 44.4% to 64.4%, with gains concentrated where VICReg saturates.

Keywords: multimodal self-supervised learning, hyperbolic geometry, product manifolds, readout capacity, packing bounds, VICReg, representational geometry

1 Introduction

Multimodal self-supervised learning (SSL) inherits a monotonicity expectation from the scaling literature: an additional modality should improve, or at worst preserve, representation quality (Baltrušaitis et al., 2019; Girdhar et al., 2023). The expectation is reasonable: a trimodal encoder observes strictly more than any bimodal subset. Yet in masked-view non-contrastive multimodal SSL, after identical pretraining and finetuning, the trimodal model underperforms its own best bimodal subset in 55.6% of paired runs under VICReg (Bardes et al., 2022) and 51.1% under SimSiam (Chen and He, 2021). Figure 3 reports the per-dataset rates behind these aggregates. We call this failure multimodal saturation. Related non-monotonic effects are usually attributed to optimization: gradient imbalance, dominant modalities, fusion schedules (Wu et al., 2022; Magal et al., 2025; Chaudhuri et al., 2025; Xu et al., 2025).

The optimization account is useful but incomplete. It does not identify which component of the objective carries the failure. We propose that the missing variable is the metric inside the invariance term.

In non-contrastive SSL, the invariance term contracts embeddings of two views of the same input toward each other (Grill et al., 2020; Zbontar et al., 2021; Wang and Isola, 2020). Under heavy masking or modality dropout, partial observations can lose the evidence that distinguishes two semantic states (He et al., 2021; Ebrahimi et al., 2023; Ma et al., 2022). The metric for this contraction also measures downstream separability. The invariance metric sets alignment tightness

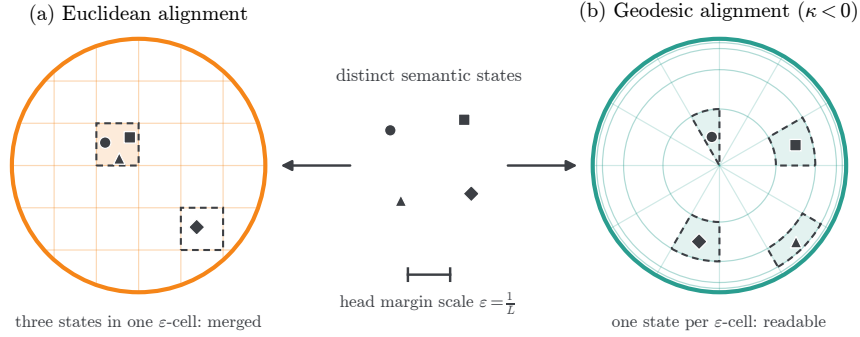


Figure 1: The alignment geometry decides whether aligned states stay readable. Masked-view alignment contracts each semantic state (one marker shape per state) inside the same radius budget, and a downstream head resolves states only at its margin scale $\varepsilon = \frac{1}{L}$. (a) Under Euclidean alignment, several states land in one ε -cell and merge: induced redundancy. (b) Under geodesic alignment at $\kappa < 0$, the same budget offers exponentially more ε -cells, so each state keeps its own. Lemmas 8–10 prove the two growth laws.

and readout capacity. Simply put, it determines how many semantic states a downstream head can distinguish at its margin.

In Euclidean space, readout capacity grows only polynomially with radius. When demand exceeds that capacity, distinct states collide at the margin scale (Figure 1). The added modality becomes redundant at the representation level while remaining informative at the input level. We name this mechanism induced redundancy: the alignment geometry manufactures redundancy that the data did not contain.

The diagnosis suggests a targeted intervention rather than a new objective. Constant negative curvature buys exponential volume growth in radius and therefore exponential readout capacity (Corollary 3; proofs in Appendix A). We hold the pipeline fixed (encoders, fusion, masking, projector dimension, regularization weights, finetuning) and replace only the metric of the invariance term.

Riemannian VICReg (R-VICReg) generalizes VICReg. It maps the projector output onto a product of low-dimensional constant negative-curvature factors with learnable curvature, aligns views by squared product-geodesic distance, and computes VICReg-style variance and covariance penalties in tangent coordinates at batch Fréchet means. The full objective has an exact VICReg limit (Proposition 4). Curvature is one continuous, learnable geometric degree of freedom, not an architectural commitment.

While R-VICReg is not a drop-in replacement for VICReg, we find that its gains concentrate where the Euclidean objective is capacity-limited. That selectivity is the falsifiable footprint we test.

Related work. Joint-embedding SSL without negatives first avoided collapse through architectural asymmetry (BYOL, SimSiam) (Grill et al., 2020; Chen and He, 2021), then through variance–covariance constraints (Barlow Twins, VICReg) (Zbontar et al., 2021; Bardes et al., 2022), and analyses of alignment–uniformity, spectra, and information content followed (Wang and Isola, 2020; Mialon et al., 2022; Balestrieri and LeCun, 2022; Schwartz-Ziv et al., 2023). Masked modeling extended the recipe from language and vision (Devlin et al., 2018; He et al., 2021; Xie et al., 2021) to multimodal encoders (Geng et al., 2022; Bachmann et al., 2022) and to the LANISTR-style image–text–tabular scaffold we adopt (Ebrahimi et al., 2023); as modalities accumulated, non-monotonic gains appeared and were attributed to optimization: missing-modality sensitivity, collapse, greedy reliance, negative co-learning (Ma et al., 2022; Chaudhuri et al., 2025; Wu et al., 2022; Magal et al., 2025). In parallel, hyperbolic representation learning grew from hierarchy embedding (Nickel and Kiela, 2017; Ganea et al., 2018) to product and mixed-curvature spaces with learnable per-factor curvature (Gu et al., 2019; Skopek et al., 2020; Bachmann et al., 2020), entered contrastive and cross-modal SSL (Ge et al., 2023; Desai et al., 2023; Franco et al., 2023; Durrant and Leontidis, 2023), and reached geometry-adaptive VICReg variants that kernelize the objective or penalize curvature (Sepanj et al., 2025; Ghoghogh et al., 2025). The two lines have not met at the invariance term: multimodal responses change balancing schedules, fusion, or regularizers and keep the alignment metric Euclidean, while the hyperbolic line targets hierarchy rather than readout capacity at the downstream margin. That capacity vocabulary comes from metric entropy

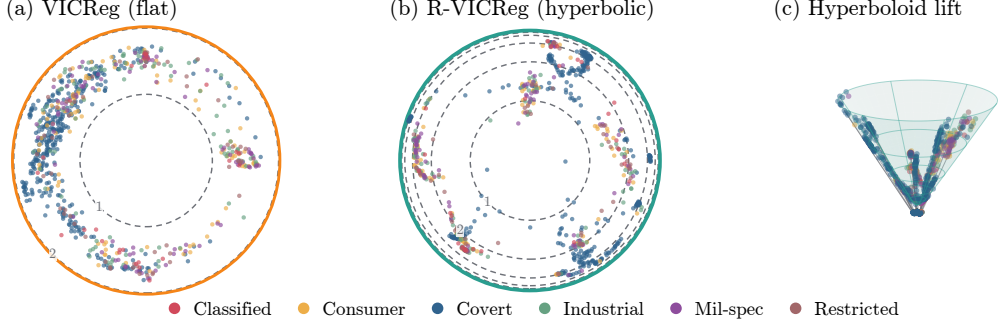


Figure 2: Three views render the learned metric of one trimodal CSGO run (seed 456). (a) In the flat VICReg projector, equal-distance rings stay evenly spaced. (b) In the most-utilized R-VICReg factor ($\kappa_{32} = -0.988$), the rings at the same geodesic distances crowd toward the boundary. (c) The same points sit on the hyperboloid above their Poincaré-ball shadow (construction in Appendix D).

and large-margin classification in metric spaces (Kolmogorov and Tikhomirov, 1961; von Luxburg and Bousquet, 2004; Gottlieb et al., 2014).

Contributions. We give (i) multimodal saturation, documented over nine datasets and two Euclidean objectives; (ii) a proposed account, induced redundancy, formalized by a readout-capacity theorem (Theorem 2) placing readable concepts between $\mathcal{N}_{\text{pack}}(B, \frac{1}{L} + 2\delta)$ and $\mathcal{N}_{\text{pack}}(B, \frac{1}{L})$, so the excess above capacity counts forced merges; (iii) R-VICReg, a generalization of VICReg with geodesic invariance on a learnable-curvature product manifold, Fréchet-centered tangent regularization, and an exact VICReg limit; and (iv) experimental results that report what the metric change does to multimodal performance.

2 Saturation as a Readout-Capacity Deficit

We read multimodal saturation as a margin-scale failure: after finetuning, distinct semantic states are no longer separated enough for a downstream head to discriminate them robustly. The invariance metric pulls views of one state together and measures how far different states remain; both roles enter the capacity count.

Throughout, $(\mathcal{M}, d)$ is the representation space. We suppose codes lie in a closed ball $B(o, R)$; normalization and weight decay motivate this radius budget. $\mathcal{N}_{\text{pack}}$ denotes packing numbers (Definition 5). A semantic state is realized as a set of codes, and the invariance term shrinks that set’s diameter.

Definition 1 (Aligned concepts and readability)

Let $K \in \mathbb{N}$ and $\delta \geq 0$. A δ -aligned K -concept family in $B(o, R)$ is a collection of nonempty sets $C_1, \dots, C_K \subseteq B(o, R)$ with $\text{diam}(C_i) \leq \delta$ for every $i \in \{1, \dots, K\}$.

The family is $(L, 1)$ -readable if there exist L -Lipschitz functions $h_1, \dots, h_K : \mathcal{M} \rightarrow \mathbb{R}$ such that, for all $i \in \{1, \dots, K\}$ and all $c \in C_i$,

$$h_i(c) \geq \max_{j \in \{1, \dots, K\} \setminus \{i\}} h_j(c) + 1.$$

Theorem 2 (Readout capacity)

Fix $o \in \mathcal{M}$, $R > 0$, $L > 0$, and $\delta \geq 0$.

(i) *Obstruction:* if a δ -aligned K -concept family in $B(o, R)$ is $(L, 1)$ -readable, then

$$K \leq \mathcal{N}_{\text{pack}}(B(o, R), \frac{1}{L}; d).$$

(ii) *Guarantee:* if $s_1, \dots, s_K \in B(o, R)$ satisfy $d(s_i, s_j) \geq \frac{1}{L} + 2\delta$ for all $i \neq j$, then every δ -aligned K -concept family with $s_i \in C_i$ for every $i \in \{1, \dots, K\}$ is $(L, 1)$ -readable.

The proof is in Appendix A.

Exceeding the packing capacity forces merges. If K concepts are realized as a δ -aligned family and $K > \mathcal{N}_{\text{pack}}(B(o, R), \frac{1}{L})$, then under any L -Lipschitz heads at most $\mathcal{N}_{\text{pack}}(B(o, R), \frac{1}{L})$ of them

attain unit margin. This follows by applying (i) to the subfamily on which the margin holds. At least $K - \mathcal{N}_{\text{pack}}(B(o, R), \frac{1}{L})$ concepts therefore fail on some view. Induced redundancy is this count: redundancy manufactured by the geometry, not present in the data. Saturation, in this lens, is not noise. It is pigeonholing.

Curvature sets the growth law. The theorem sandwiches the readable count between two packings of one ball, at scales that differ by 2δ . Geometry enters through one quantity: ball-packing growth in radius. Appendix A proves matching bounds for both geometries:

$$\mathcal{N}_{\text{pack}}(B_{\mathbb{R}^n}(0, R), \varepsilon) \leq \left(1 + \frac{2R}{\varepsilon}\right)^n, \quad \mathcal{N}_{\text{pack}}(B_{\mathbb{H}_\kappa^n}(R), \varepsilon) \geq c_\varepsilon e^{(n-1)\sqrt{-\kappa}R}. \quad (1)$$

The Euclidean bound grows polynomially in R ; the hyperbolic bound grows exponentially. Both bounds are two-sided (Lemmas 8–10). Products inherit the exponential lower bound factor-wise, so bits add across factors (Proposition 11).

Corollary 3 (Radius cost of bits)

Fix $L > 0$ and $\delta \geq 0$.

- (i) In $\mathbb{R}^n$ the readable-concept count grows polynomially in R : hosting $\log K$ bits needs a radius exponential in $(\log K)/n$.
- (ii) In $\mathbb{H}_\kappa^n$ with $\kappa < 0$ it grows exponentially in R : the required radius is linear in $\log K$.
- (iii) At fixed (n, R, L, δ) , replacing the flat metric by curvature $\kappa < 0$ multiplies the guaranteed count by a factor that grows exponentially in R .

The proof is in Appendix A.

Demand compounds while flat supply crawls. Compositional states multiply across modalities, so demanded bits add per modality. A third modality is where additive demand can cross a logarithmic flat supply. Tightening alignment narrows the sandwich, but only the metric changes the growth law. The invariance term of R-VICReg shrinks δ in the metric whose capacity grows exponentially. Its variance and covariance terms prevent degenerate collapse into a vanishing region that would destroy capacity in any geometry.

The theorem gives both an obstruction and a guarantee. It does not certify that a given dataset’s demand exceeds the Euclidean supply. The saturation diagnostic addresses that question.

The supply side leaves a visible signature: curvature buys capacity, alignment spends it, and the codes crowd into the exponential-volume shell. Figure 2 shows the signature on a trained run: 71% of codes sit beyond geodesic radius 2.

The design principle has a falsifiable footprint: help where the flat objective saturates and impose mild cost where it does not. Section 5 tests this footprint.

3 Riemannian VICReg

Masked-view scaffold. We adopt a LANISTR-style pretrain-then-finetune pipeline for image (I), text (T), and tabular (Tab) inputs (Ebrahimi et al., 2023). Each minibatch yields an unmasked view $a_1(x)$ and a masked view $a_2(x)$ via modality-specific masking; modality encoders feed single-layer projections into a fusion Transformer, and a projector g maps the fused CLS representations to codes $z_1, z_2 \in \mathbb{R}^D$. Pretraining combines weighted unimodal masked-reconstruction losses with a masked-view alignment loss $\mathcal{L}_{\text{MMM}}$; the compared objectives (SimSiam, VICReg, R-VICReg) instantiate only $\mathcal{L}_{\text{MMM}}$, and finetuning attaches a classifier to the fusion CLS embedding. Hyperparameters and the minibatch algorithm are in Appendix E.

3.1 Geodesic invariance on a product of curved factors

R-VICReg represents the code $z_i \in \mathbb{R}^D$ as concatenated tangent vectors. We partition it into F factors $z_i^{(f)} \in \mathbb{R}^d$ with $D = Fd$. Each factor carries a learnable curvature $\kappa_f < 0$, and the exponential map at the origin sends it to the κ -stereographic model: $x_i^{(f)} = \exp_0^{\kappa_f}(z_i^{(f)})$. Closed-form exp, log, Möbius addition, and geodesic distance are collected in Table 1 (Appendix B).

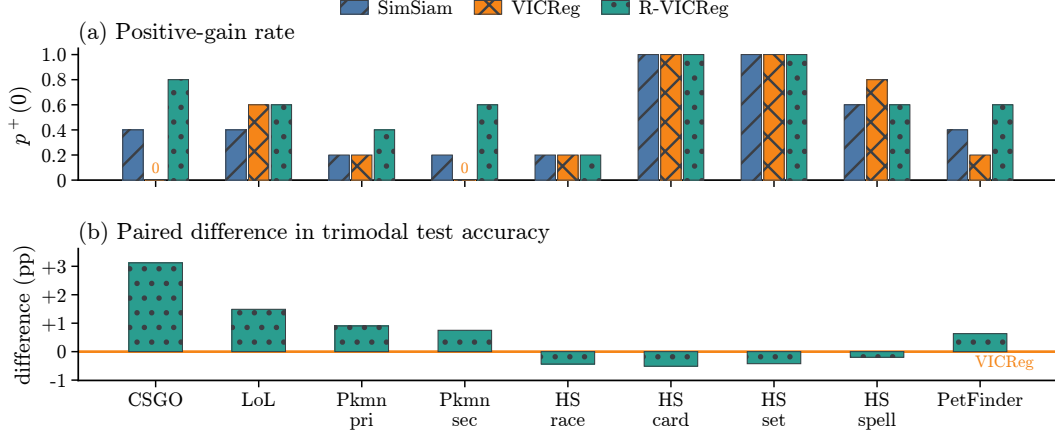


Figure 3: Panel (a) reports the per-dataset positive-gain rate $p^+(0) = \Pr(\Delta_{\text{tri}} > 0)$ over five seeds, so rates are multiples of 0.2. VICReg scores 0 of 5 on CSGO and Pokemon Secondary (marked 0). Panel (b) reports the paired difference in trimodal test accuracy between R-VICReg and VICReg: bars give the five-seed mean and the orange line the VICReg baseline.

The curved factors are not an arbitrary choice. We read each factor’s code as a location paired with its alignment uncertainty; the metric invariant under sufficient statistics is then the Fisher–Rao metric, and for isotropic Gaussian factors it is hyperbolic (Appendix C). Euclidean VICReg is the slice on which the uncertainty coordinate is frozen, and R-VICReg unfreezes that coordinate with one learnable curvature per factor.

The product manifold is constant-curvature factor-wise, not globally; distances combine by the product metric, and Proposition 11 bounds its packing capacity from below.

The invariance term replaces Euclidean mean squared error with normalized squared geodesic distance,

$$\mathcal{L}_{\text{inv}} := \mathbb{E} \left[\frac{1}{F} \sum_{f=1}^F \frac{d_{\kappa_f}(x_1^{(f)}, x_2^{(f)})^2}{4d} \right], \quad (2)$$

where the 4 matches the stereographic convention $d_0(x, y)^2 = 4 \|x - y\|_2^2$ and $1/(Fd) = 1/D$ matches VICReg’s normalization.

3.2 Variance and covariance in Fréchet-centered tangent coordinates

Variance and covariance statistics require a vector space, but a curved factor is not one. For each factor, with $q_1, \dots, q_{2B}$ the factor codes of both views in a batch of size B , the batch Fréchet mean $\mu_f := \text{FM}(\{q_j\}_{j=1}^{2B})$ is the minimizer of $\sum_j d_{\kappa_f}(\mu, q_j)^2$. We approximate it by three origin-initialized Karcher iterations (Karcher, 1977),

$$\mu \leftarrow \exp_{\mu}^{\kappa_f} \left(\frac{1}{2B} \sum_{j=1}^{2B} \log_{\mu}^{\kappa_f}(q_j) \right),$$

and detach the result. We then log-map both views to the tangent space at μ_f and concatenate across factors,

$$t_{i,b}^{(f)} := \log_{\mu_f}^{\kappa_f}(x_{i,b}^{(f)}), \quad t_{i,b} := [t_{i,b}^{(1)}; \dots; t_{i,b}^{(F)}] \in \mathbb{R}^D.$$

The standard VICReg penalties act on these tangent codes,

$$\mathcal{L}_{\text{var}} := \frac{1}{2} \sum_{i \in \{1,2\}} \mathbb{E}[\text{ReLU}(\gamma - \text{std}(t_i))], \quad \mathcal{L}_{\text{cov}} := \sum_{i \in \{1,2\}} \frac{1}{D} \sum_{a \neq b} \text{Cov}(t_i)_{ab}^2. \quad (3)$$

Here $\text{std}(t_i)$ is the per-coordinate standard deviation over the batch, the expectation in $\mathcal{L}_{\text{var}}$ averages the D coordinates, and $\text{Cov}(t_i)$ is the $D \times D$ batch covariance of the tangent codes. The variance term prevents individual tangent coordinates from collapsing; the covariance term limits redundancy between coordinates. The full objective is $\mathcal{L}_{\text{rvicreg}} = \lambda_{\text{inv}} \mathcal{L}_{\text{inv}} + \lambda_{\text{var}} \mathcal{L}_{\text{var}} + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}}$ with the exact VICReg weighting $(\lambda_{\text{inv}}, \lambda_{\text{var}}, \lambda_{\text{cov}}) = (25, 25, 1)$, $\gamma = 1$.

Proposition 4 (Exact VICReg limit)

Fix a minibatch of codes and let $\kappa_f \rightarrow 0$ for every $f \in \{1, \dots, F\}$.

- (i) $\exp_0^{\kappa_f}(z_i^{(f)}) \rightarrow z_i^{(f)}$ and $d_{\kappa_f}(x_1^{(f)}, x_2^{(f)})^2 \rightarrow 4 \left\| z_1^{(f)} - z_2^{(f)} \right\|_2^2$, so $\mathcal{L}_{\text{inv}} \rightarrow \frac{1}{D} \|z_1 - z_2\|_2^2$ in the batch average.
- (ii) Every Karcher iterate μ_f converges to the Euclidean batch mean and $\log_{\mu_f}^{\kappa_f}$ to subtraction, so the tangent codes converge to the centered codes.
- (iii) Variance and covariance statistics are translation-invariant, so $\mathcal{L}_{\text{var}}$ and $\mathcal{L}_{\text{cov}}$ converge to their VICReg values.

Hence $\mathcal{L}_{\text{rvicreg}} \rightarrow \mathcal{L}_{\text{vicreg}}$ (Lemma 12, Appendix B).

VICReg is the boundary case of the whole objective: the family is continuous in the geometry, and VICReg’s loss weights transfer unchanged.

Learnable curvature and stability. Per-factor curvature is parameterized as $\kappa_f = -\kappa_{\max} \sigma(\theta_f)$ with $\kappa_{\max} = 1.5$, so $\kappa_f \in (-\kappa_{\max}, 0)$ while $\theta_f \in \mathbb{R}$ trains by backpropagation; we initialize θ_f by inverse sigmoid to target $\kappa_0 = -1$ with small per-factor jitter. For stability in the Poincaré ball of radius $1/\sqrt{-\kappa_f}$, we clip tangent norms before the exponential map and project points exceeding the boundary. We monitor boundary utilization $u = \sqrt{-\kappa_f} \|x^{(f)}\|_2 \in [0, 1)$: values near 1 mean embeddings occupy the high-volume region where the exponential capacity is spent.

4 Experimental Design

Datasets and modality sweep. We use nine trimodal image–text–tabular classification datasets: eight from MuG-Bench (Lu et al., 2023), namely CSGO ($n=767$), LoL ($n=1,006$), Pokemon Primary and Secondary type ($n=719$ each), Hearthstone Race (8,568), CardClass (13,153), Set (13,087), and Spell (3,924), plus PetFinder ($n=11,996$ training examples, 80/5/15 split) (Kaggle and PetFinder.my, 2019). For each dataset and method, we train all seven modality subsets $\{I, T, \text{Tab}, (I, T), (I, \text{Tab}), (T, \text{Tab}), (I, T, \text{Tab})\}$ under the same pipeline. These subsets measure what the third modality adds over every bimodal alternative.

Saturation metrics. For each run (dataset $\times$ seed), let $S(\cdot)$ be downstream test accuracy. The best bimodal baseline is $B^* := \max\{S(I, T), S(I, \text{Tab}), S(T, \text{Tab})\}$, and the trimodal marginal gain is $\Delta_{\text{tri}} := S((I, T, \text{Tab})) - B^*$. A run is saturated at threshold τ if $\Delta_{\text{tri}} \leq \tau$. We sweep $\tau \in \{0, 0.005, 0.01\}$ and report the reliability $p^+(\tau) := \Pr(\Delta_{\text{tri}} > \tau)$ across paired runs. For R-VICReg against VICReg, a rescue at threshold τ is a paired run with $\Delta_{\text{tri}}^{\text{VICReg}} \leq \tau$ and $\Delta_{\text{tri}}^{\text{R-VICReg}} > \tau$. A harm is the reverse. Saturation is a run-level failure event, so we treat it as one: the primary endpoint is how often the third modality helps, not the average accuracy shift.

Protocol and controls. Each method contributes $N = 45$ paired runs (nine datasets, five seeds). Each run pretrains 50 epochs, selects the checkpoint with lowest pretraining loss, finetunes 50 epochs, selects by validation accuracy, and reports test accuracy. SimSiam serves as a second Euclidean baseline: it shares no variance–covariance regularizer with VICReg, so saturation under both objectives implicates Euclidean alignment rather than a regularizer.

5 Results

5.1 Euclidean alignment saturates; R-VICReg restores reliability

Does saturation arise from one objective, or from the geometry? At $\tau = 0$, VICReg saturates in 55.6% of runs, while SimSiam saturates in 51.1%. Two Euclidean objectives with different anti-collapse mechanisms fail at comparable rates. This result supports the geometric account, but not a regularizer-specific one. Monotonicity violation is not an edge case; it is the modal outcome.

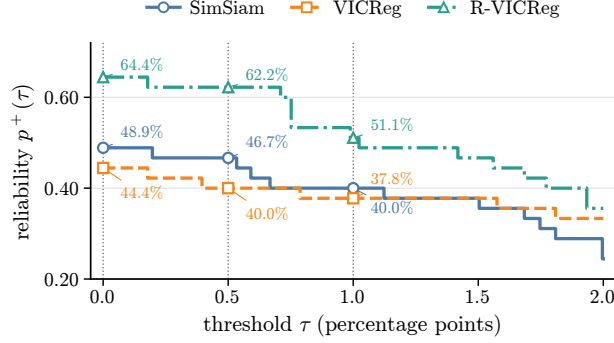


Figure 4: Reliability $p^+(\tau) = \Pr(\Delta_{\text{tri}} > \tau)$ over $N = 45$ paired runs per method, swept over $\tau \in \{0, 0.005, 0.01\}$: R-VICReg stays above both baselines at every threshold, so the ranking is no artifact of $\tau = 0$.

At $\tau = 0$, R-VICReg attains 64.4% (29/45) against 44.4% (20/45) for VICReg: a twenty-point shift in the probability that a third modality is worth adding, at fixed dimension, architecture, and loss weights. The gains cost no absolute accuracy: mean trimodal accuracy is 0.627 for R-VICReg, 0.621 for VICReg, and 0.619 for SimSiam, and Figure 3(b) reports the per-dataset paired differences. Figure 4 reports $p^+(\tau)$ across the sweep.

5.2 The gains are selective, and the harms are real

Does R-VICReg rescue saturated runs without harming unsaturated ones? Figure 5 (Appendix D) answers. At $\tau = 0$, R-VICReg rescues 13 of the 25 runs VICReg saturates and harms 4 of the 20 it does not.

Where $\Delta_{\text{tri}}^{\text{VICReg}} \leq 0$, mean paired improvement δ is +1.38 points ($n = 25$); elsewhere it is -0.76 ($n = 20$). This split is the footprint Section 2 predicts; a uniform improvement would have argued against the capacity account. The method pays off where capacity binds and costs a little where it does not.

5.3 What the learned geometry does and does not explain

Does the learned geometry stay hyperbolic in use? It does. The learned curvature stays negative on every dataset, with mean $\kappa = -0.934$ and range $[-0.992, -0.820]$, and boundary utilization averages 0.987. The codes live in the region where volume grows exponentially, the regime that Figure 2 renders.

The per-dataset pattern matches the account. Gains concentrate where VICReg saturates most, and the two tasks that never saturate show no difference. Figure 3 gives the breakdown, and Appendix D details the runs where R-VICReg is worse.

Which saturated runs get rescued? The geometry does not say. Rescued and not-rescued runs are indistinguishable in curvature magnitude (0.946 vs. 0.955) and in boundary utilization (0.987 vs. 0.983). The mechanism evidence is aggregate, not run-level.

6 Discussion

The intervention modifies one object: the metric in which masked views are aligned. The rest of the pipeline remains unchanged. The invariance metric therefore co-determines what survives into the representation. However, if a bimodal-versus-trimodal sweep shows $\Delta_{\text{tri}} > 0$ under the Euclidean objective, capacity is not binding and switching the metric will more often cost a little accuracy. R-VICReg is conditioned on its diagnostic: a checkable condition turns geometric theory into a decision rule.

Limitations. The real limitation is predictive. Curvature does not reveal rescue: rescued and not-rescued runs share the same learned geometry (Section 5.3), so the hypothesis holds in aggregate, but we cannot say in advance which saturated runs a metric change will save. The rescue is not

guaranteed, and the harm is real: 4 of the 20 unsaturated runs got worse. Attribution is incomplete as well: the geodesic objective couples the two views position-dependently, and separating capacity from that coupling is left to future work.

7 Conclusion

Across nine trimodal datasets under masked-view non-contrastive SSL, the monotonicity expectation fails in half of the paired runs. We proposed a geometric account: the invariance metric sets alignment tightness and readout capacity. Changing only that metric raised reliability from 44.4% to 64.4%, with thirteen rescues against four harms concentrated where Euclidean alignment saturates. That selectivity is what the proposal predicted.

References

- C. Atkinson and A. F. S. Mitchell. Rao’s distance measure. *Sankhyā: The Indian Journal of Statistics, Series A*, 43(3):345–365, 1981. URL <https://www.jstor.org/stable/25050476>.
- Nihat Ay, Jürgen Jost, Hồng Vân Lê, and Lorenz Schwachhöfer. *Information Geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge*. Springer, Cham, 2017. doi: 10.1007/978-3-319-56478-4.
- Gregor Bachmann, Gary Bécigneul, and Octavian Ganea. Constant curvature graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 486–496. PMLR, 2020. URL <https://proceedings.mlr.press/v119/bachmann20a.html>.
- Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision – ECCV 2022*, volume 13697 of *Lecture Notes in Computer Science*, pages 348–367. Springer, 2022. doi: 10.1007/978-3-031-19836-6_20. URL https://doi.org/10.1007/978-3-031-19836-6_20.
- Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35:26671–26685, 2022.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607. URL <https://doi.org/10.1109/TPAMI.2018.2798607>.
- Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL <https://arxiv.org/abs/2105.04906>.
- Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 7555–7577. PMLR, 2025. URL <https://proceedings.mlr.press/v267/chaudhuri25a.html>.
- Jeff Cheeger and David G. Ebin. *Comparison Theorems in Riemannian Geometry*. AMS Chelsea Publishing. American Mathematical Society, Providence, RI, 2008. doi: 10.1090/chel/365. Reprint of the 1975 edition.
- Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. URL <https://arxiv.org/abs/2011.10566>.
- N. N. Chentsov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1982. doi: 10.1090/mmono/053.

- Sueli I. R. Costa, Sandra A. Santos, and João E. Strapasson. Fisher information distance: A geometrical reading. *Discrete Applied Mathematics*, 197:59–69, 2015. doi: 10.1016/j.dam.2014.10.004.
- Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7694–7731. PMLR, 2023. URL <https://proceedings.mlr.press/v202/desai23a.html>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. URL <https://arxiv.org/abs/1810.04805>.
- Aiden Durrant and Georgios Leontidis. Hmsn: Hyperbolic self-supervised learning by clustering with ideal prototypes. *arXiv preprint arXiv:2305.10926*, 2023. URL <https://arxiv.org/abs/2305.10926>.
- Sayna Ebrahimi, Sercan O. Arik, Yihe Dong, and Tomas Pfister. Lanistr: Multimodal learning from structured and unstructured data. *arXiv preprint arXiv:2305.16556*, 2023. URL <https://arxiv.org/abs/2305.16556>.
- Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. In *International Conference on Learning Representations*, 2023. URL <https://arxiv.org/abs/2303.06242>.
- Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *arXiv preprint arXiv:1805.09112*, 2018. URL <https://arxiv.org/abs/1805.09112>.
- Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6840–6849, 2023. URL <https://arxiv.org/abs/2212.00653>.
- Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022. URL <https://arxiv.org/abs/2205.14204>.
- Benyamin Ghogh, M. Hadi Sepanj, and Paul Fieguth. Self-supervised learning by curvature alignment. *arXiv preprint arXiv:2511.17426*, 2025. URL <https://arxiv.org/abs/2511.17426>.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023.
- Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014. doi: 10.1109/TIT.2014.2325840.
- 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL <https://arxiv.org/abs/2006.07733>.
- Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=HJxeWnCcf7>.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. URL <https://arxiv.org/abs/2111.06377>.

- Kaggle and PetFinder.my. Petfinder.my adoption prediction. Kaggle competition dataset, 2019.
- Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977. doi: 10.1002/cpa.3160300502.
- A. N. Kolmogorov and V. M. Tikhomirov. ϵ -entropy and ϵ -capacity of sets in functional spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961. doi: 10.1090/trans2/017/10.
- Jiaying Lu, Yongchen Qian, Shifan Zhao, Yuanzhe Xi, and Carl Yang. Mug: A multimodal classification benchmark on game data with tabular, textual, and visual fields. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5332–5346. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.354. URL <https://aclanthology.org/2023.findings-emnlp.354/>.
- Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? *arXiv preprint arXiv:2204.05454*, 2022. URL <https://arxiv.org/abs/2204.05454>.
- Nicholas Magal, Minh Tran, Riku Arakawa, and Suzanne Nie. Negative to positive co-learning with aggressive modality dropout. *arXiv preprint arXiv:2501.00865*, 2025. URL <https://arxiv.org/abs/2501.00865>.
- Grégoire Mialon, Randall Balestriero, and Yann LeCun. Variance covariance regularization enforces pairwise independence in self-supervised representations. *arXiv preprint arXiv:2209.14905*, 2022. URL <https://arxiv.org/abs/2209.14905>.
- Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *arXiv preprint arXiv:1705.08039*, 2017. URL <https://arxiv.org/abs/1705.08039>.
- John G. Ratcliffe. *Foundations of Hyperbolic Manifolds*, volume 149 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2 edition, 2006. ISBN 9780387331973.
- M. Hadi Sepanj, Benyamin Ghogh, Saed Moradi, and Paul Fieguth. Kernel vicreg for self-supervised learning in reproducing kernel hilbert space. *arXiv preprint arXiv:2509.07289*, 2025. URL <https://arxiv.org/abs/2509.07289>.
- Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G. J. Rudner, and Yann LeCun. An information theory perspective on variance-invariance-covariance regularization. In *Advances in Neural Information Processing Systems*, volume 36, pages 33965–33998, 2023. URL <https://arxiv.org/abs/2303.00633>.
- Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=S1g6xeSKDS>.
- Lene Theil Skovgaard. A Riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, 11(4):211–223, 1984. URL <https://www.jstor.org/stable/4615960>.
- Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004. URL <https://www.jmlr.org/papers/v5/luxburg04a.html>.
- Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 2020. URL <https://proceedings.mlr.press/v119/wang20k.html>.
- Nan Wu, Stanisław Jastrzębski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. *arXiv preprint arXiv:2202.05306*, 2022. URL <https://arxiv.org/abs/2202.05306>.
- Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021. URL <https://arxiv.org/abs/2111.09886>.

Shaoxuan Xu, Menglu Cui, Chengxiang Huang, Hongfa Wang, and Di Hu. Balancebenchmark: A survey for multimodal imbalance learning, 2025. URL <https://arxiv.org/abs/2502.10816>.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 2021. URL <https://arxiv.org/abs/2103.03230>.

A Packing, Covering, and Growth-Law Proofs

Definition 5 (Packing number)

For a metric space $(\mathcal{Z}, \rho)$ and $S \subseteq \mathcal{Z}$,

$$\mathcal{N}_{\text{pack}}(S, \varepsilon; \rho) := \max \left\{ K \in \mathbb{N} : \exists z_1, \dots, z_K \in S, \min_{i \neq j} \rho(z_i, z_j) \geq \varepsilon \right\}. \quad (4)$$

Definition 6 (Lipschitz functions)

For metric spaces $(\mathcal{Z}, \rho)$ and $(\mathcal{W}, \eta)$, $h : \mathcal{Z} \rightarrow \mathcal{W}$ is L -Lipschitz if $\eta(h(z), h(z')) \leq L\rho(z, z')$ for all $z, z' \in \mathcal{Z}$.

Definition 7 (Metric ball)

For $(\mathcal{Z}, \rho)$, $z \in \mathcal{Z}$, $R > 0$: $B_{\mathcal{Z}}(z, R) := \{u \in \mathcal{Z} : \rho(u, z) \leq R\}$.

We now prove Theorem 2.

Proof. (i) Let $h_1, \dots, h_K$ witness readability, and fix $i \neq j$ in $\{1, \dots, K\}$, $c \in C_i$, and $c' \in C_j$. The function $g := h_i - h_j$ is $2L$ -Lipschitz as a difference of two L -Lipschitz functions, and the readability condition gives $g(c) \geq 1$ and $g(c') \leq -1$. Hence

$$2 \leq g(c) - g(c') \leq 2L d(c, c'),$$

so $d(c, c') \geq \frac{1}{L}$. Choosing one code $c_i \in C_i$ for each $i \in \{1, \dots, K\}$ therefore produces K points of $B(o, R)$ that are pairwise at distance at least $\frac{1}{L}$, and $K \leq \mathcal{N}_{\text{pack}}(B(o, R), \frac{1}{L}; d)$ follows from Definition 5.

(ii) For each $i \in \{1, \dots, K\}$ define

$$h_i(x) := \min\{1, \max\{0, 1 - L(d(x, s_i) - \delta)\}\}.$$

Each h_i is L -Lipschitz as a composition of 1-Lipschitz clamps with an L -Lipschitz function of distance. Fix $i \in \{1, \dots, K\}$ and $c \in C_i$. The conditions $s_i \in C_i$ and $\text{diam}(C_i) \leq \delta$ give $d(c, s_i) \leq \delta$, so $h_i(c) = 1$. For $j \in \{1, \dots, K\}$ with $j \neq i$, the triangle inequality gives

$$d(c, s_j) \geq d(s_i, s_j) - d(s_i, c) \geq \frac{1}{L} + 2\delta - \delta = \frac{1}{L} + \delta,$$

so $1 - L(d(c, s_j) - \delta) \leq 0$ and $h_j(c) = 0$. The readability condition therefore holds with margin 1. □

Lemma 8 (Euclidean ball packing is polynomial in radius)

For $\mathbb{R}^n$ with $d_E(x, y) = \|x - y\|_2$ and any $R \geq \varepsilon > 0$,

$$\left(\frac{R}{\varepsilon}\right)^n \leq \mathcal{N}_{\text{pack}}(B_{\mathbb{R}^n}(0, R), \varepsilon; d_E) \leq \left(1 + \frac{2R}{\varepsilon}\right)^n. \quad (5)$$

For the stereographically scaled metric $d_0(x, y) = 2\|x - y\|_2$ the upper bound becomes $(1 + 4R/\varepsilon)^n$.

Proof. Upper bound: if $z_1, \dots, z_K$ is an ε -packing of $B_{\mathbb{R}^n}(0, R)$, the balls $B(z_i, \varepsilon/2)$ are disjoint and contained in $B(0, R + \varepsilon/2)$. Comparing volumes gives

$$K(\varepsilon/2)^n \leq (R + \varepsilon/2)^n.$$

Lower bound: a maximal ε -packing is an ε -cover, since any point at distance at least ε from every center could be added to the packing; and any ε -cover of $B(0, R)$ needs at least

$$\frac{\text{vol}(B(0, R))}{\text{vol}(B(0, \varepsilon))} = \left(\frac{R}{\varepsilon}\right)^n$$

balls. □

Both hyperbolic lemmas rest on one imported fact: the volume of a geodesic ball of radius t in $\mathbb{H}_\kappa^n$ is (Ratcliffe, 2006; Cheeger and Ebin, 2008)

$$\text{Vol}(B(t)) = \text{Vol}(\mathbb{S}^{n-1}) \int_0^t \left(\frac{\sinh(\sqrt{-\kappa} s)}{\sqrt{-\kappa}} \right)^{n-1} ds. \quad (6)$$

Lemma 9 (Hyperbolic ball packing is exponential in radius)

Fix $n \geq 2$, $\kappa < 0$, $\varepsilon > 0$. There exist $c_\varepsilon > 0$ and R_0 (depending on n, κ, ε) such that for all $R \geq R_0$,

$$\mathcal{N}_{\text{pack}}(B_{\mathbb{H}_\kappa^n}(R), \varepsilon) \geq c_\varepsilon \exp((n-1)\sqrt{-\kappa} R). \quad (7)$$

Proof. A maximal ε -packing of $B(R)$ is also an ε -cover, since any point at distance at least ε from every center could be added to the packing. All ε -balls have volume $\text{Vol}(B(\varepsilon))$ by homogeneity, so

$$\mathcal{N}_{\text{pack}}(B_{\mathbb{H}_\kappa^n}(R), \varepsilon) \geq \frac{\text{Vol}(B(R))}{\text{Vol}(B(\varepsilon))}.$$

Since $\sinh$ is increasing and $\sinh u \geq e^u/4$ for $u \geq \ln 2$, Equation (6) gives, for $R \geq R_0 := 1 + \ln 2/\sqrt{-\kappa}$,

$$\text{Vol}(B(R)) \geq \text{Vol}(\mathbb{S}^{n-1}) \int_{R-1}^R \left(\frac{\sinh(\sqrt{-\kappa} s)}{\sqrt{-\kappa}} \right)^{n-1} ds \geq c_{n,\kappa} e^{(n-1)\sqrt{-\kappa} R},$$

and the claim follows with $c_\varepsilon := c_{n,\kappa}/\text{Vol}(B(\varepsilon))$. □

Lemma 10 (Hyperbolic ball packing is at most exponential)

Fix $n \geq 2$, $\kappa < 0$, $\varepsilon > 0$. There exists $c'_\varepsilon > 0$ (depending on n, κ, ε) such that for all $R \geq 1$,

$$\mathcal{N}_{\text{pack}}(B_{\mathbb{H}_\kappa^n}(R), \varepsilon) \leq \frac{\text{Vol}(B(R + \varepsilon/2))}{\text{Vol}(B(\varepsilon/2))} \leq c'_\varepsilon \exp((n-1)\sqrt{-\kappa} R). \quad (8)$$

Proof. Centers of an ε -packing carry pairwise-disjoint geodesic $\varepsilon/2$ -balls contained in $B(R + \varepsilon/2)$; by homogeneity all $\varepsilon/2$ -balls have equal volume, giving the first inequality. Since $\sinh u \leq e^u/2$, Equation (6) gives $\text{Vol}(B(t)) \leq C_{n,\kappa} e^{(n-1)\sqrt{-\kappa} t}$. Substituting $t = R + \varepsilon/2$ absorbs $e^{(n-1)\sqrt{-\kappa} \varepsilon/2}$ into c'_ε . □

Proposition 11 (Product-hyperbolic packing lower bound)

Let $\mathcal{Z} = \prod_{f=1}^F \mathbb{H}_{\kappa_f}^d$ with $\kappa_f < 0$ and product metric $\rho(z, z') = (\sum_f \rho_f(z_f, z'_f)^2)^{1/2}$. For any sets S_f ,

$$\mathcal{N}_{\text{pack}}\left(\prod_f S_f, \varepsilon; \rho\right) \geq \prod_f \mathcal{N}_{\text{pack}}(S_f, \varepsilon; \rho_f), \quad B_{\mathcal{Z}}(o, R) \supseteq \prod_f B_{\mathbb{H}_{\kappa_f}^d}(o_f, R/\sqrt{F}). \quad (9)$$

Consequently there exist $c_\varepsilon > 0$, $R_0 > 0$ such that for $R \geq \sqrt{F} R_0$,

$$\mathcal{N}_{\text{pack}}(B_{\mathcal{Z}}(o, R), \varepsilon) \geq (c_\varepsilon)^F \exp\left((d-1) \frac{R}{\sqrt{F}} \sum_{f=1}^F \sqrt{-\kappa_f}\right). \quad (10)$$

Proof. For the packing inequality, take per-factor ε -packings $P_f \subseteq S_f$; distinct points of $\prod_f P_f$ differ in some factor f^* by at least ε , so their product distance is at least ε . For the ball containment, if $\rho_f(z_f, o_f) \leq R/\sqrt{F}$ for all f then $\rho(z, o) \leq R$. Apply Lemma 9 in each factor at radius $R/\sqrt{F}$ and multiply. $\square$

We now prove Corollary 3.

Proof. Apply Theorem 2 at the two sandwich scales $\varepsilon_1 = \frac{1}{L}$ and $\varepsilon_2 = \frac{1}{L} + 2\delta$.

- (i) In $\mathbb{R}^n$, Lemma 8 bounds both ends polynomially in R . Readability of K concepts forces

$$K \leq (1 + 2LR)^n,$$

so $R \geq \frac{1}{2L}(K^{1/n} - 1)$. A packing of size $(R/\varepsilon_2)^n$ exists at scale ε_2 , so a radius linear in $\varepsilon_2 K^{1/n}$ suffices. In both directions the radius needed for $\log K$ bits scales as $e^{(\log K)/n}$.

- (ii) In $\mathbb{H}_\kappa^n$, Lemmas 9 and 10 bound both ends by $e^{(n-1)\sqrt{-\kappa}R}$ up to constants depending only on (n, κ, L, δ) , so

$$\log \mathcal{N}_{\text{pack}} = (n-1)\sqrt{-\kappa}R + O(1)$$

at either scale, and the required radius is linear in $\log K$.

- (iii) For the fixed-budget comparison, divide the hyperbolic guarantee at scale ε_2 by the Euclidean obstruction at scale ε_1 . The ratio is at least

$$c_{\varepsilon_2} e^{(n-1)\sqrt{-\kappa}R} (1 + 2LR)^{-n},$$

which grows exponentially in R . $\square$

B κ -Stereographic Operations

All operations are closed-form and autodiff-compatible. Table 1 collects them with their flat limits, following the presentation of the mixed-curvature literature (Skopek et al., 2020). Scalar κ -trigonometry, with the convention $u/\|u\|_2 := 0$ at $u = 0$ throughout:

$$\tan_\kappa(r) = \begin{cases} \tanh(\sqrt{-\kappa}r)/\sqrt{-\kappa} & \kappa < 0, \\ r & \kappa = 0, \end{cases} \quad \arctan_\kappa(r) = \begin{cases} \text{atanh}(\sqrt{-\kappa}r)/\sqrt{-\kappa} & \kappa < 0, \\ r & \kappa = 0. \end{cases} \quad (11)$$

The exponential map at μ is the one used by the Karcher iteration of Section 3. The next lemma proves the limits and covers the rest of the objective.

Lemma 12 (Full-objective flat limit)

For one factor, let $q_1, \dots, q_{2B}$ be the factor codes of both views and $\bar{q} := \frac{1}{2B} \sum_j q_j$. Set $\mu^{(0)} = 0$ and, for $s \geq 0$,

$$\mu^{(s+1)} = \exp_{\mu^{(s)}}^\kappa \left(\frac{1}{2B} \sum_{j=1}^{2B} \log_{\mu^{(s)}}^\kappa(q_j) \right).$$

For any fixed number of iterations $t \geq 1$, as $\kappa \rightarrow 0$,

$$\mu^{(t)} \longrightarrow \bar{q}, \quad t_{i,b}^{(f)} = \log_{\mu^{(t)}}^\kappa(x_{i,b}^{(f)}) \longrightarrow z_{i,b}^{(f)} - \bar{z}_f, \quad (12)$$

where $\bar{z}_f$ is the Euclidean mean of both views' factor codes. Consequently $\mathcal{L}_{\text{var}}$ and $\mathcal{L}_{\text{cov}}$ converge to their VICReg values and, with the invariance limit above, $\mathcal{L}_{\text{rvicreg}} \rightarrow \mathcal{L}_{\text{vicreg}}$.

Table 1: κ -stereographic operations used by R-VICReg and their flat limits. Every limit is uniform on compacta (Lemma 12); the distance limit squares to the invariance-term limit $d_\kappa(x, y)^2 \rightarrow 4 \|x - y\|_2^2$ of Proposition 4.

Operation	Closed form	Limit $\kappa \rightarrow 0$
Exponential map at 0	$\exp_0^\kappa(u) = \tan_\kappa(\ u\ _2) \frac{u}{\ u\ _2}$	u
Logarithmic map at 0	$\log_0^\kappa(x) = \arctan_\kappa(\ x\ _2) \frac{x}{\ x\ _2}$	x
Möbius addition	$x \oplus_\kappa y = \frac{(1 - 2\kappa\langle x, y \rangle - \kappa\ y\ _2^2)x + (1 + \kappa\ x\ _2^2)y}{1 - 2\kappa\langle x, y \rangle + \kappa^2\ x\ _2^2\ y\ _2^2}$	$x + y$
Geodesic distance	$d_\kappa(x, y) = 2 \arctan_\kappa(\ (-x) \oplus_\kappa y\ _2)$	$2\ x - y\ _2$
Conformal factor	$\lambda_\mu^\kappa = 2/(1 + \kappa\ \mu\ _2^2)$	2
Logarithmic map at μ	$\log_\mu^\kappa(x) = \frac{2}{\lambda_\mu^\kappa} \log_0^\kappa((- \mu) \oplus_\kappa x)$	$x - \mu$
Exponential map at μ	$\exp_\mu^\kappa(v) = \mu \oplus_\kappa \exp_0^\kappa(\lambda_\mu^\kappa v/2)$	$\mu + v$
Karcher step	$\mu \leftarrow \exp_\mu^\kappa\left(\frac{1}{2B} \sum_j \log_\mu^\kappa(q_j)\right)$	$\mu \leftarrow \frac{1}{2B} \sum_j q_j$

Proof. All codes in the fixed minibatch have bounded norms, and only finitely many iterations are taken, so every quantity below ranges over a compact set and every limit is uniform on compacta. As $\kappa \rightarrow 0$, the scalar maps $\tan_\kappa, \arctan_\kappa \rightarrow \text{id}$ uniformly on compacta, hence $\exp_0^\kappa, \log_0^\kappa \rightarrow \text{id}$; the conformal factor satisfies $\lambda_\mu^\kappa \rightarrow 2$; and the Möbius formula gives $x \oplus_\kappa y \rightarrow x + y$. Therefore $\log_\mu^\kappa(x) = \frac{2}{\lambda_\mu^\kappa} \log_0^\kappa((- \mu) \oplus_\kappa x) \rightarrow x - \mu$ and $\exp_\mu^\kappa(v) = \mu \oplus_\kappa \exp_0^\kappa(\lambda_\mu^\kappa v/2) \rightarrow \mu + v$. For the iterates: $\mu^{(1)} \rightarrow 0 + \frac{1}{2B} \sum_j (q_j - 0) = \bar{q}$, and if $\mu^{(s)} \rightarrow \bar{q}$ then $\log_{\mu^{(s)}}^\kappa(q_j) \rightarrow q_j - \bar{q}$, whose average tends to 0, so $\mu^{(s+1)} \rightarrow \bar{q} + 0 = \bar{q}$; induction gives $\mu^{(t)} \rightarrow \bar{q}$ for every fixed $t \geq 1$. For the tangent codes: $x_{i,b}^{(f)} = \exp_0^\kappa(z_{i,b}^{(f)}) \rightarrow z_{i,b}^{(f)}$ and $\bar{q} \rightarrow \bar{z}_f$, so $t_{i,b}^{(f)} \rightarrow z_{i,b}^{(f)} - \bar{z}_f$; concatenation over factors gives $t_{i,b} \rightarrow z_{i,b} - \bar{z}$. Finally, subtracting the common batch vector $\bar{z}$ from every column changes neither per-dimension standard deviations nor covariance matrices, and both are continuous in the finitely many tangent codes, so $\mathcal{L}_{\text{var}}$ and $\mathcal{L}_{\text{cov}}$ converge to the VICReg values computed on z_1, z_2 . The detachment of μ is immaterial in the limit: it is a batch-dependent constant whose only effect is the common shift. The weighted sum then converges term by term. $\square$

C The Canonical Geometry of Location–Uncertainty Codes

This appendix motivates, from information geometry, why a product of learnable negative-curvature factors is a canonical geometry for codes that carry per-factor location and uncertainty. It is motivation for the design; Theorem 2 and all empirical claims are independent of it.

The modeling axiom is stated explicitly: suppose each factor’s code is read as a statistical experiment. It is a location–scale state (μ, σ) with $\mu \in \mathbb{R}^m$ and $\sigma > 0$. It combines a fitted position with its alignment uncertainty. Suppose the geometry placed on codes must be invariant under sufficient statistics and monotone under Markov morphisms. Data processing cannot increase statistical distinguishability.

Proposition 13 (Fisher–Rao geometry of location–uncertainty codes)

Under the modeling axiom, the following classical facts apply.

- (i) *Invariant metric: the metric is the Fisher–Rao metric, uniquely up to a positive scale on each independent factor (Chentsov, 1982; Ay et al., 2017).*
- (ii) *Gaussian factors are hyperbolic: for the isotropic Gaussian family $\mathcal{N}(\mu, \sigma^2 I_m)$ on $\mathbb{R}^m$, the Fisher–Rao line element is*

$$ds^2 = \frac{\|d\mu\|_2^2 + 2m d\sigma^2}{\sigma^2},$$

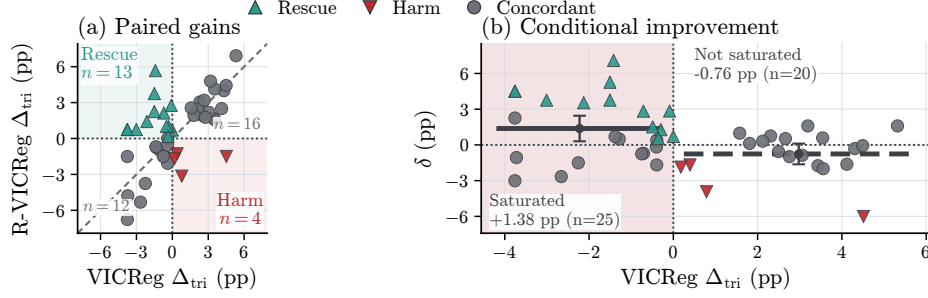


Figure 5: A rescue is a paired run that VICReg saturates ($\Delta_{\text{tri}} \leq \tau$) and R-VICReg does not; a harm is the reverse. (a) The quadrants at $\tau = 0$ hold 13 rescues, 4 harms, and 28 concordant runs. (b) The paired improvement δ is positive where VICReg saturates and negative where it does not; bars show 95% Student- t intervals.

and the substitution $y = \sqrt{2m} \sigma$ gives

$$ds^2 = 2m \frac{\|d\mu\|_2^2 + dy^2}{y^2},$$

the Poincaré half-space metric scaled by $2m$, hence $\mathbb{H}^{m+1}$ with constant curvature $-1/(2m)$ (Atkinson and Mitchell, 1981; Skovgaard, 1984; Costa et al., 2015). General regular location-scale families give scaled hyperbolic factors (Atkinson and Mitchell, 1981).

- (iii) Independent factors compose: a product of F independent location-scale factors carries the product Fisher-Rao geometry $\prod_f \mathbb{H}_{\kappa_f}^{m_f+1}$ with $\kappa_f < 0$, and the per-factor scale freedom of (i) is one curvature magnitude per factor. A learnable κ_f parameterizes that freedom, and only that freedom.

Corollary 14 (The fixed-uncertainty slice is flat)

On the slice of a factor with σ held fixed, the induced metric is

$$ds^2|_{d\sigma=0} = \frac{\|d\mu\|_2^2}{\sigma^2},$$

which is flat. Euclidean VICReg is, in this reading, the Fisher-Rao geometry of codes whose uncertainty coordinate is frozen; unfreezing that coordinate forces negative curvature.

Within this reading, tightening alignment (smaller σ) moves codes toward the ideal boundary and expands the accessible Fisher volume; the boundary-utilization statistic of Section 3 measures the occupancy of that region.

D Per-dataset Reliability and Learned Geometry

The aggregate reliability hides substantial dataset heterogeneity. Figure 3 (Section 5) reports the per-dataset positive-gain fractions and paired trimodal-accuracy differences, Figure 5 shows the paired quadrants behind the rescue counts, and Figure 6 reports the learned geometry.

Where the gains are, and where they are not. The per-dataset positive-gain advantage of R-VICReg over VICReg is +0.80 on CSGO, +0.60 on Pokemon Secondary, +0.40 on PetFinder, and +0.20 on Pokemon Primary. The two largest Hearthstone tasks, CardClass and Set, are positive in every seed for every method and show no difference: no capacity problem, no effect. On Hearthstone Spell, R-VICReg is worse (−0.20), as the capacity account predicts where flat supply is not binding.

The distribution improves in the bulk, not the extreme tail. R-VICReg improves the mean (+0.008 vs. +0.004), median (+0.010 vs. −0.003), and 10th percentile (−0.027 vs. −0.034) of Δ_{tri} , but its 5th percentile (−0.046) and minimum (−0.068) are worse than VICReg’s (−0.038): the intervention shifts the typical case, not the worst case. Percentiles use linear interpolation.

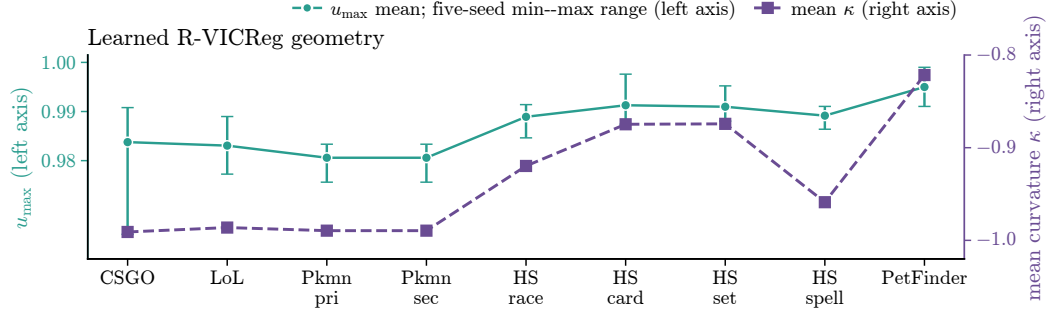


Figure 6: The learned R-VICReg geometry stays hyperbolic on every dataset: curvature remains negative and boundary utilization stays high. The left axis reports $u_{\max}$, with bars spanning the five-seed minimum–maximum range; the right axis reports the mean learned curvature κ .

Seeing the learned metric. Figure 2 makes one learned geometry visible: the trimodal CSGO run at seed 456. Panel (a) draws the flat VICReg projector with the angle from the top two principal directions and the radius equal to the true embedding norm; rings at equal distances stay evenly spaced. Panel (b) draws the most-utilized R-VICReg factor ($\kappa_{32} = -0.988$) with the angle from the top two tangent principal directions (67% of the variance) and the radius equal to the exact geodesic distance from the origin. The same rings now crowd toward the boundary circle. The crowding is the exponential volume of Section 2 at work: 71% of the points lie beyond geodesic radius 2, and at the median occupied radius the factor’s geodesic sphere carries 11.4 times the flat-metric area. Panel (c) lifts the same points onto the constant-curvature hyperboloid above their Poincaré-ball shadow.

E Algorithm, Hyperparameters, and Reproducibility

Algorithm 1: R-VICReg masked-view loss $\mathcal{L}_{\text{MMM}}$ (one minibatch)

```

construct views  $a_1(x)$ ,  $a_2(x)$ ; encode, fuse, and project to  $z_1, z_2 \in \mathbb{R}^D$ 
reshape  $z_i$  into factors  $z_i^{(f)} \in \mathbb{R}^d$ ,  $f = 1, \dots, F$ 
map  $x_i^{(f)} \leftarrow \exp_0^{\kappa_f}(z_i^{(f)})$  and compute  $\mathcal{L}_{\text{inv}}$  by Equation (2)
for  $f = 1$  to  $F$  do
     $\mu_f \leftarrow \text{FM}(\text{cat}(x_1^{(f)}, x_2^{(f)}))$  (three origin-initialized Karcher iterations, detached)
     $t_i^{(f)} \leftarrow \log_{\mu_f}^{\kappa_f}(x_i^{(f)})$  for  $i \in \{1, 2\}$ 
end
concatenate  $t_i \leftarrow [t_i^{(1)}; \dots; t_i^{(F)}]$ ; compute  $\mathcal{L}_{\text{var}}, \mathcal{L}_{\text{cov}}$  by Equation (3)
return  $\lambda_{\text{inv}}\mathcal{L}_{\text{inv}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}} + \lambda_{\text{cov}}\mathcal{L}_{\text{cov}}$ 

```

All datasets are public (MuG-Bench and PetFinder); MuG-Bench uses the provided splits and PetFinder an 80/5/15 split. Image and text encoders are initialized from pretrained checkpoints and finetuned during multimodal pretraining; VICReg and R-VICReg share a VICReg-style projector with BatchNorm, while SimSiam uses its own projector and predictor. The downstream classifier operates on the fusion CLS embedding without the SSL projector. The full sweep (nine datasets, seven modality subsets, three methods, five seeds) comprises 945 completed runs whose recorded run times sum to approximately 200 GPU-hours, executed on two NVIDIA RTX 4090 GPUs. The code will be open-sourced.

Table 2: Hyperparameters shared by all experiments.

Hyperparameter	Value
Pretrain / finetune epochs	50 / 50
Batch size	64
Optimizer / learning rate / weight decay	AdamW / 10^{-4} / 0.02
Warmup epochs / gradient clip	5 / 1.0
Text / image / tabular reconstruction weights	(1.0, 1.0, 0.01)
Alignment weight λ_{MMM}	1.0
$(\lambda_{\text{inv}}, \lambda_{\text{var}}, \lambda_{\text{cov}}), \gamma$	(25, 25, 1), 1.0
Projection dimension D	256
R-VICReg factors $F \times d$	64×4
Curvature bound and initialization	$\kappa_{\text{max}} = 1.5, \kappa_0 = -1$
Fréchet mean steps (detached)	3
Seeds per dataset and method	5