POWERJEPA: FROM TAIL-AWARE ALIGNMENT TO THE DOWNSTREAM UTILITY OF GLOBAL SPECTRAL STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning general-purpose embeddings requires preserving semantic correspondences while retaining variation useful across downstream tasks. Despite LeJEPA's success in visual representation learning, its direct extension to general-purpose text embeddings yields limited downstream performance in our experiments. Empirical analysis identifies insufficient alignment of difficult positive pairs as a key limitation. Motivated by this finding, we introduce PowerJEPA, combining tail-aware power-mean alignment, selective hard-negative discrimination, and Sketched Isotropic Gaussian Regularization (SIGReg) to shape local and global embedding geometry. Experiments across text (MS MARCO and NLI) and image (ImageNet) modalities demonstrate its effectiveness. We further examine how the global representational capacity shaped by SIGReg, measured by effective dimension, relates to downstream performance. Spectral interventions on frozen text representations show that retrieval is more sensitive to truncation than clustering and reranking, while larger effective dimension does not consistently improve performance. Theory further shows that spectra with identical effective dimension can have different regularized effective degrees of freedom. We therefore combine effective dimension with spectral curvature into a capacity–shape description for assessing downstream utility. A calibrated predictor using both descriptors achieves a Spearman correlation of 0.97 with mean MTEB scores. These results connect geometric representation learning with calibrated diagnostics of downstream transfer.

1 INTRODUCTION

Learning general-purpose embeddings requires preserving semantic correspondences while supporting tasks with different demands on representation geometry (Gupta et al., 2024; Lee et al., 2025). Contrastive learning approaches this goal through positive attraction and negative repulsion, commonly understood through alignment and uniformity (Wang & Isola, 2020; Koromilas et al., 2024). However, in-batch negatives can include semantically related examples, introducing conflicting supervision (Lee et al., 2025; Kim et al., 2025). More fundamentally, neither strong average alignment nor a broadly dispersed embedding distribution alone specifies which variations remain useful for downstream transfer (Koromilas et al., 2024; Gupta et al., 2024; Zhang et al., 2024). This motivates studying both how representation geometry is learned and how its usefulness can be diagnosed (Koromilas et al., 2024; Park et al., 2023; Kalibhat et al., 2024).

LeJEPA provides a promising starting point by combining predictive alignment with Sketched Isotropic Gaussian Regularization (SIGReg), avoiding negative-based competition as the mechanism for preventing collapse (Balestriero & LeCun, 2025). Its success in visual self-supervised learning motivates extending this principle to general-purpose text embeddings. However, our MS MARCO experiments reveal that the MSE-based adaptation leaves a tail of poorly aligned positive pairs, highlighting a challenge in aligning heterogeneous query–passage pairs. Figure 1(a) projects a selected subset of these difficult pairs into a shared PCA space. The separation between each query and its matched positive passage makes the residual alignment error visible. We further examine the benefits of improving alignment for these difficult positive pairs in our experiments.

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.

To address this challenge, we propose POWERJEPA, replacing mean-squared alignment with a tail-aware power mean. For a batch of B positive pairs, let d_i denote the distance between their normalized embeddings. We define $\mathcal{L}_p = \left(\frac{1}{B}\sum_{i=1}^{B} d_i^p\right)^{1/p}$. Its distance-gradient weights are proportional to d_i^{p-1}, so choosing $p > 2$ emphasizes difficult positives more strongly than mean-squared alignment, without discrete positive mining. Moment-based bounds connect this objective to upper-tail distances, supporting its interpretation as a tail-aware alignment surrogate. Figure 1(b) shows the same pairs under power-mean alignment, with improved alignment and fewer cases in which another sampled passage outranks the matched positive.

We retain SIGReg to discourage distributional collapse and promote global spectral diversity. When hard negatives are available, we further incorporate a relative-margin objective to resolve ambiguous local orderings between matched positives and competing negatives. Together, power-mean alignment, optional hard-negative margins, and SIGReg form a joint-embedding objective that addresses three complementary aspects of representation geometry: positive-tail alignment, local discrimination, and global spectral diversity.

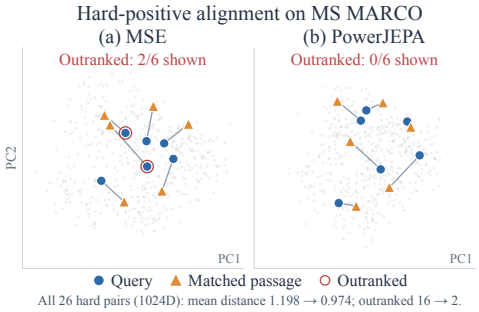


Figure 1: **Hard-positive alignment on MS MARCO.** Six pairs sampled at equally spaced ranks within the 26 largest MSE positive distances among 512 training pairs, shown with a shared PCA and scale. Lines connect queries to matched passages; gray points are sampled passages. Red rings mark queries whose matched passage is outranked (ties included) in the original 1024D space.

Beyond preventing collapse, we examine how the capacity of the global embedding space shaped by SIGReg relates to downstream transfer, using effective dimension as a proxy (Garrido et al., 2023). On MS MARCO-trained embeddings, our representative MTEB study shows that retrieval is more sensitive to spectral truncation than clustering and reranking, while larger effective dimension is not consistently beneficial. Theory further shows that spectra with identical effective dimension and total variance can have different regularized effective degrees of freedom. This motivates examining whether spectral shape provides information about downstream utility beyond effective dimension. We characterize this shape using *spectral curvature*, defined as the contrast between head and tail decay slopes. On 12 additional text encoders excluded from descriptor selection and predictor fitting, adding curvature improves the calibrated predictor's Spearman correlation with mean scores over shared MTEB tasks from 0.69 to 0.97. Effective dimension and spectral curvature thus provide complementary predictive information, forming a *capacity–shape description* of the global embedding space. Our contributions are threefold:

- We find that directly adapting LeJEPA to supervised text representation learning leaves difficult positive pairs poorly aligned. Motivated by this limitation, we introduce POWERJEPA, combining tail-aware power-mean alignment, SIGReg, and optional hard-negative margins. POWERJEPA substantially outperforms the LeJEPA baseline on text benchmarks and achieves competitive results on image benchmarks.
- We investigate how the capacity of SIGReg-shaped text embedding spaces relates to downstream transfer. Through spectral interventions, we reveal task-dependent responses and the limitations of effective dimension as a standalone diagnostic.
- We develop a *capacity–shape description* for text representations by combining effective dimension and spectral curvature. Motivated by regularized-capacity analysis, this description improves calibrated prediction of aggregate downstream performance on additional text encoders over effective dimension alone.

2 RELATED WORK

Representation learning. Joint-embedding methods pair alignment or prediction with regularization to prevent collapse. LeJEPA combines predictive alignment with SIGReg (Balestriero & LeCun,

2025); VISReg and QQWorld further develop distributional regularization for visual representations and world models (Wu et al., 2026a; Yu et al., 2026). We extend this paradigm to general-purpose text embeddings, where methods such as NV-Embed and Qwen3 Embedding instead emphasize LLM encoders and task-aware contrastive training (Lee et al., 2025; Zhang et al., 2025).

Representation geometry and quality assessment. Prior measures of representation quality include alignment–uniformity and effective rank (Wang & Isola, 2020; Garrido et al., 2023). Recent studies examine task-dependent dimensional redundancy, uninformative embedding subspaces, and topological structure (Tsukagoshi & Sasano, 2025; Wu et al., 2026b; Shestov et al., 2026). We examine whether spectral curvature predicts downstream performance beyond effective dimension.

3 PRELIMINARY

LeJEPA. LeJEPA combines predictive alignment with Gaussian distributional regularization (Balestriero & LeCun, 2025). For B samples with V views, its objective is

$$\mathcal{L}_{\text{LeJEPA}} = \frac{1-\lambda}{BV}\sum_{i=1}^{B}\sum_{v=1}^{V}\|z_{i,v}-\mu_i\|_2^2 + \frac{\lambda}{V}\sum_{v=1}^{V}\text{SIGReg}(Z_v), \tag{1}$$

where

$$\text{SIGReg}(Z_v) = \frac{B}{M}\sum_{m=1}^{M}\int_{\mathbb{R}}\left|\widehat{\phi}_{v,m}(t) - e^{-t^2/2}\right|^2 e^{-t^2/2}\,\mathrm{d}t. \tag{2}$$

Here, $z_{i,v} \in \mathbb{R}^d$ denotes a view embedding, $Z_v = (z_{i,v})_{i=1}^B$, and $\mu_i = V_g^{-1}\sum_{v=1}^{V_g} z_{i,v}$ averages the global views. The empirical characteristic function $\widehat{\phi}_{v,m}(t) = B^{-1}\sum_{i=1}^{B}\exp(\mathrm{i}t\,a_m^\top z_{i,v})$ uses directions a_m sampled uniformly from the unit sphere. SIGReg matches these projected distributions to $\mathcal{N}(0,1)$ to discourage collapse, while $\lambda \in (0,1)$ balances alignment and regularization.

Spectral Diagnostics of Representation Quality. For an embedding matrix Z with nonzero singular values $\sigma_1, \ldots, \sigma_r$, RankMe (Garrido et al., 2023) measures dimensional diversity through effective rank:

$$r_{\text{eff}}(Z) = \exp\Big(-\sum_{j=1}^{r} q_j \log q_j\Big), q_j = \frac{\sigma_j}{\sum_{\ell=1}^{r}\sigma_\ell}. \tag{3}$$

4 POWERJEPA: TAIL-AWARE REPRESENTATION LEARNING

PowerJEPA addresses three complementary aspects of representation geometry: positive-tail alignment, local discrimination, and global distributional diversity. It combines power-mean positive alignment with SIGReg and, when hard-negative supervision is used, a selective boundary objective.

Representation spaces and objective. Let $h(x) = f_\theta(x) \in \mathbb{R}^d$ denote the unnormalized text embedding, including pooling and any exported embedding head. Define $\mathcal{N}(v) = v/\max\{\|v\|_2, \epsilon\}$ and $z(x) = \mathcal{N}(h(x))$. Here, note that text embeddings are normalized, while image embeddings are not. Since our main experiments focus on text, we present the formulation using normalized embeddings for now. Positive alignment and local discrimination operate on z, while SIGReg acts on unnormalized features $r(x) = \psi(h(x))$, where ψ is the identity or a regularization head. Figure 2 summarizes the framework. Given B positive pairs (q_i, p_i), the text objective is

$$\mathcal{L} = a\widehat{\mathcal{L}}_p + \gamma\mathcal{L}_{\text{bdry}} + \lambda\mathcal{L}_{\text{SIG}}, \tag{4}$$

where $a > 0$ and $\gamma, \lambda \geq 0$; $\gamma = 0$ in configurations without hard-negative supervision.

Tail-aware positive alignment. For positive-pair distances $d_i = \|z(q_i) - z(p_i)\|_2$, we use

$$\widehat{\mathcal{L}}_p = \left(\frac{1}{B}\sum_{i=1}^{B} d_i^p\right)^{1/p}, \qquad p > 1. \tag{5}$$

This is a power mean across pairwise Euclidean distances, not an ℓ_p norm across feature coordinates. For positive distances and $\widehat{\mathcal{L}}_p > 0$,

$$\frac{\partial\widehat{\mathcal{L}}_p}{\partial d_i} = \frac{1}{B}\widehat{\mathcal{L}}_p^{1-p} d_i^{p-1}, \qquad \frac{\partial\widehat{\mathcal{L}}_p/\partial d_i}{\partial\widehat{\mathcal{L}}_p/\partial d_j} = \left(\frac{d_i}{d_j}\right)^{p-1}. \tag{6}$$

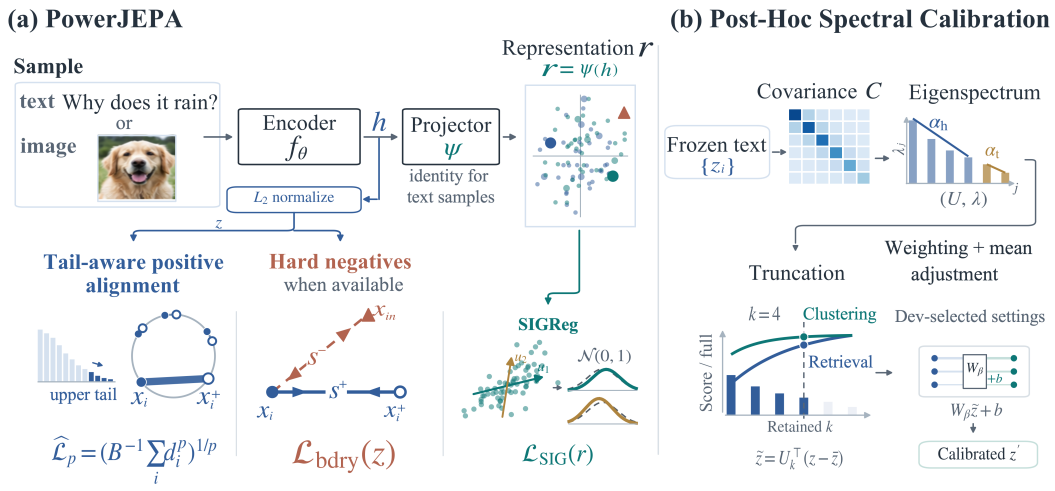


Figure 2: PowerJEPA and spectral diagnostics. (a) Power-mean alignment emphasizes difficult positives, optional hard-negative supervision corrects local orderings, and SIGReg regularizes the global feature distribution. Note that L2 regularization is applied only to text data. (b) Effective dimension and spectral curvature describe frozen text embeddings for downstream analysis.

Thus $p > 2$ increases the relative emphasis on large residuals compared with squared-distance alignment, without discrete positive mining. At $p = 2$, the objective is root-mean-square distance rather than MSE; their relative distance weights coincide, but their overall gradient scales differ. The implementation adds a small constant before taking the root for numerical stability.

Let $D \geq 0$ be a positive-pair residual and $M_p = (\mathbb{E}D^p)^{1/p} < \infty$. Denote the average of the largest α probability mass by $\mathrm{CVaR}_\alpha^{\mathrm{upper}}(D)$, with fractional weighting at atoms when necessary.

Proposition 4.1 (L_p duality and positive-tail control)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $1 < p < \infty$, $q = p/(p-1)$, and $D \in L_p(\mathbb{P})$ with $D \geq 0$ almost surely. Set $M_p = \|D\|_{L_p}$. Then*

$$M_p = \sup_{\substack{w \geq 0 \\ \|w\|_{L_q} \leq 1}} \mathbb{E}[wD]. \tag{7}$$

If $M_p > 0$, a maximizer is $w_p^* = (D/M_p)^{p-1}$. Moreover, for every $t > 0$ and $\alpha \in (0, 1)$,

$$\mathbb{P}(D \geq t) \leq \min\left\{1, \left(\frac{M_p}{t}\right)^p\right\}, \qquad \mathcal{T}_\alpha(D) \leq \alpha^{-1/p} M_p. \tag{8}$$

The factor $\alpha^{-1/p}$ is sharp uniformly over probability spaces and nonnegative L_p residuals. For the uniform empirical distribution over a batch, $M_p = \widehat{\mathcal{L}}_p$.

These standard moment bounds follow from Markov's and Hölder's inequalities (Appendix A) and motivate power-mean alignment as a tail-aware surrogate. For a fixed batch, it also approaches the maximum distance as $p \to \infty$. The empirical bounds do not themselves guarantee population improvement; increasing p can also amplify erroneous positive supervision.

Selective local discrimination. For each query, let $\mathcal{H}_i$ contain selected high-similarity documents from a candidate pool, excluding its paired positive. Define $s_i^+ = z(q_i)^\top z(p_i)$ and $s_{in}^- = z(q_i)^\top z(n)$. Our soft-margin configuration uses

$$\mathcal{L}_{\text{bdry}} = \frac{\sum_i \sum_{n \in \mathcal{H}_i} \mathrm{softplus}\left(\frac{s_{in}^- - s_i^+ + m}{T}\right)}{\sum_i |\mathcal{H}_i|}, \tag{9}$$

where $m \geq 0$ and $T > 0$ control the relative margin and smoothness. Selection uses detached scores; selected similarities are recomputed for differentiation. We set the term to zero when no

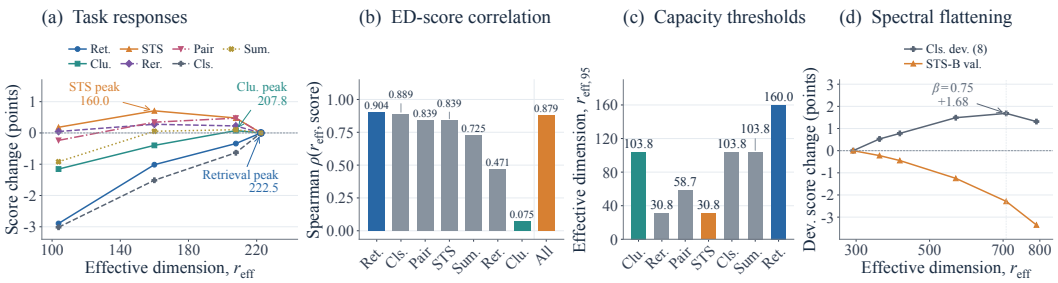


Figure 3: Task-dependent spectral utility. (a) Performance changes versus effective dimension under PCA truncation of the frozen PowerJEPA encoder across all MTEB(eng, v2) tasks. (b) ED–score Spearman correlations across historical training results on all tasks with `subset_size=1000`. (c) Lowest tested effective dimension retaining 95% of full-representation performance. (d) Full-rank spectral flattening on the same encoder, evaluated on classification development tasks and STS-Benchmark validation; effective dimension uses a separate diagnostic protocol.

negatives are selected. The loss emphasizes ambiguous orderings and decays for well-separated pairs, without assigning a repulsive term to every unpaired example. Selection does not guarantee the absence of false negatives. Earlier reference configurations use a distance-based hinge margin; the boundary variant is specified for each experimental cohort.

Global distribution regularization. We retain SIGReg (Balestriero & LeCun, 2025), which matches random one-dimensional projections of unnormalized features to a standard Gaussian. Pooling query and positive features gives $\{r_j\}_{j=1}^{N}$ with $N = 2B$. For random unit directions u_ℓ, define $\widehat{\phi}_\ell(t) = N^{-1}\sum_j \exp(\mathrm{i}tu_\ell^\top r_j)$ and minimize

$$\mathcal{L}_{\text{SIG}} = \frac{N}{L}\sum_{\ell=1}^{L}\sum_{k=1}^{K} w_k \left|\widehat{\phi}_\ell(t_k) - e^{-t_k^2/2}\right|^2, \tag{10}$$

where the quadrature weights w_k include the Gaussian window. SIGReg discourages collapse and directional concentration in its input distribution; it does not directly optimize the effective dimension of the normalized inference

5 Global Dimensionality and Downstream Utility

Beyond collapse prevention, we examine how global representation dimensionality relates to downstream utility. Spectral interventions on an MS MARCO-trained PowerJEPA encoder motivate our analysis of regularized capacity and task-signal allocation.

Effective dimension. Let C be the centered covariance of normalized inference embeddings, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ and positive trace. Following Equation 3, we define $D_{\text{eff}}(C) = r_{\text{eff}}(C)$, applying effective rank to covariance eigenvalues rather than feature-matrix singular values. This measures the effective number of occupied directions.

Empirical motivation: task-dependent dimensional requirements. We freeze the PowerJEPA encoder and retain 32–1024 leading covariance directions estimated from 512 MS MARCO training triples, evaluating all MTEB(eng, v2) tasks. At rank 128, reranking matches full-representation performance, whereas retrieval remains 2.89 points below it (Figure 3(a)). The lowest tested effective dimension retaining 95% of full-representation performance is also higher for retrieval than clustering: 160.0 versus 103.8 (Figure 3(c)). These observations motivate examining why spectral directions have different utility across tasks. Appendix B.4 provides the full protocol.

Effective dimension and regularized capacity. Spectral diversity differs from the capacity accessible to a regularized learner. At regularization scale $\eta > 0$, the effective degrees of freedom are

$$d_\eta(C) = \operatorname{tr}\left[C(C+\eta I)^{-1}\right] = \sum_{j=1}^{d}\frac{\lambda_j}{\lambda_j+\eta}. \tag{11}$$

Table 1: Comparison of MS MARCO-trained encoders on our MTEB(eng, v2) evaluation suite. "+ hard negatives" indicates explicit hard-negative supervision; both InfoNCE variants use in-batch negatives. TERM uses $\tau = 10$. Scores are multiplied by 100, and Avg. weights all tasks equally. Best available scores in each column are bold. † denotes exploratory post-hoc spectral calibration of the preceding model with the encoder frozen.

Method	Cls.	Clust.	Pair	Rerank	Retr.	STS	Sum.	Avg.
InfoNCE	64.00	44.56	74.64	43.97	42.13	73.83	22.29	55.96
InfoNCE + hard negatives	64.94	44.57	**76.16**	**43.98**	**43.57**	**74.28**	23.93	56.72
LeJEPA (w projector)	53.90	46.84	51.71	40.73	27.04	64.98	15.58	46.98
LeJEPA (w/o projector)	54.03	**47.28**	52.70	40.47	28.83	65.02	13.12	47.52
LeJEPA (w/o projector) + hard negatives	61.69	45.48	69.10	43.55	38.20	70.19	19.76	53.47
VISReg + hard negatives	61.81	44.01	70.68	43.23	38.17	71.25	23.46	53.63
TERM + SIGReg + hard negatives	63.92	45.60	73.80	43.50	42.19	72.50	**25.59**	55.86
PowerJEPA (ours)	61.91	46.36	67.80	42.63	40.47	70.63	25.17	54.31
PowerJEPA (ours) + hard negatives	64.92	45.26	73.84	43.68	42.24	72.59	24.38	55.98
+ post-hoc spectral calibration†	**67.38**	45.10	74.37	43.96	42.65	73.74	24.51	**56.81**

This quantity depends on eigenvalue magnitudes relative to η. For example, $D_{\text{eff}}(cC) = D_{\text{eff}}(C)$ for every $c > 0$, whereas $d_\eta(cC)$ varies with c at fixed η. Effective dimension alone therefore does not determine regularized capacity.

Task-dependent spectral utility. Even for a fixed spectrum, the benefit of retaining a direction depends on its task signal. The following sequence model formalizes this dependence.

Proposition 5.1 (Spectral risk and direction-retention criterion)**.** *Fix $d \geq 1$, $C \in \mathbb{S}^d_{++}$ with eigenvalues $\lambda_j > 0$, $a \in \mathbb{R}^d$, $n \geq 1$, $\sigma \geq 0$, and $\eta > 0$. In an orthonormal eigenbasis of C, let*

$$Y = a + \tfrac{\sigma}{\sqrt{n}}\xi, \qquad \mathbb{E}\xi = 0, \qquad \mathbb{E}[\xi\xi^\top] = I_d.$$

Set $\nu = \sigma^2/n$, $w_j = \lambda_j/(\lambda_j + \eta)$, and $\widehat{a} = \operatorname{diag}(w)Y$. Then

$$\mathcal{R}_\eta(C, a) := \mathbb{E}\|\widehat{a} - a\|_2^2 = \textstyle\sum_{j=1}^d \left(\frac{\eta}{\lambda_j+\eta}\right)^2 a_j^2 + \nu \sum_{j=1}^d \left(\frac{\lambda_j}{\lambda_j+\eta}\right)^2. \tag{12}$$

Let $\mathcal{R}_\eta^{(-j)}(C, a)$ denote the risk obtained by replacing w_j with zero while leaving all other weights unchanged. Then

$$\mathcal{R}_\eta(C, a) < \mathcal{R}_\eta^{(-j)}(C, a) \quad \Longleftrightarrow \quad a_j^2 > \nu \tfrac{\lambda_j}{\lambda_j + 2\eta}. \tag{13}$$

Retaining a direction reduces risk precisely when its task signal exceeds a spectrum-dependent noise threshold. Different signal allocations can therefore produce different risks under the same covariance spectrum, providing a mechanism for task-dependent responses to spectral truncation. Effective dimension summarizes spectral diversity but does not by itself determine downstream utility. The proof appears in Appendix A.

6 Experiments

We evaluate PowerJEPA primarily on text representation learning using MS MARCO and AllNLI, examining downstream transfer (Section 6.1) and ablations (Section 6.2). We then extend the evaluation to ImageNet-1K retrieval (Section 6.3). Finally, we study task-dependent dimensional requirements and spectral calibration (Section 6.4).

6.1 Text Representation Learning: Downstream Transfer

Training and Evaluation. We train separate text encoders on MS MARCO (Nguyen et al., 2016) for retrieval-oriented learning and AllNLI, comprising SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018), for sentence-level transfer, especially STS. On AllNLI, entailment

Table 2: Comparison of AllNLI-trained encoders on the nine STS tasks in our MTEB(eng, v2) evaluation suite. We report Spearman correlation with cosine similarity, multiplied by 100; Avg. is the unweighted mean across tasks. STS-B denotes STSBenchmark. InfoNCE uses in-batch negatives in addition to explicit hard negatives; TERM uses $\tau = 10$. Best scores in each column are bold. All methods use explicit hard-negative supervision.

Method	STS12	STS13	STS14	STS15	STS17	STS22.v2	STS-B	SICK-R	BIOSSES	Avg.
InfoNCE	70.45	80.32	74.43	**83.52**	**85.90**	47.42	**82.43**	**77.18**	75.47	75.24
LeJEPA (w/ projector)	59.86	65.86	59.76	74.04	77.51	41.74	68.21	58.15	72.13	64.14
LeJEPA (w/o projector)	62.11	72.28	62.59	73.73	77.86	56.77	70.66	60.30	75.70	68.00
TERM + SIGReg	68.17	75.12	67.67	78.43	84.56	49.43	74.91	62.03	**77.40**	70.86
PowerJEPA (ours)	**72.69**	**82.68**	**75.07**	82.39	82.62	**57.15**	78.60	72.72	75.47	**75.49**

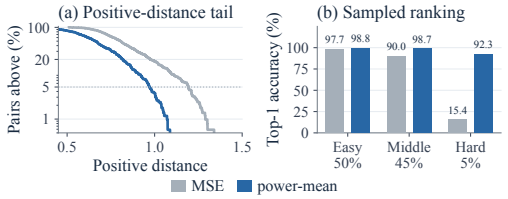


Figure 4: Difficult-positive alignment on 512 MS MARCO development pairs. (a) Positive-distance tail. (b) Top-1 retrieval of the 26 pairs most difficult under MSE.

Table 3: Component analysis of PowerJEPA. Q95, margin violations, and effective dimension (ED) use 8192 fixed MS MARCO training triples with L2-normalized embeddings; violations use an L2 margin of 0.3.

Configuration	Mean	Q95	Violation (%)	ED
Without power alignment	54.95	1.034	28.7	441.0
Without boundary term	54.28	0.997	41.4	361.3
Without SIGReg	53.89	0.561	36.2	125.4
PowerJEPA	55.98	0.939	27.6	310.9

pairs serve as positives and contradiction examples as hard negatives when enabled. All encoders use Qwen3-0.6B (Yang et al., 2025) with LoRA (Hu et al., 2021) and produce 1024-dimensional embeddings. PowerJEPA is trained for 10 epochs using AdamW (Loshchilov et al., 2017), a global batch size of 2048, learning rates of 5×10^{-5} on MS MARCO and 10^{-4} on AllNLI, and the power exponent of $p = 18$ on MS MARCO and $p = 5$ on AllNLI.

We evaluate on a fixed 39-task suite from MTEB(eng, v2) (Muennighoff et al., 2023), covering classification (Cls.), clustering (Clust.), pair classification (Pair), reranking (Rerank), retrieval (Retr.), semantic textual similarity (STS), and summarization (Sum.). We multiply each task's primary metric by 100 and report task-family averages and an equally weighted mean across all tasks. Appendix B provides training configurations and task-level scores.

Baselines. We compare with LeJEPA (Balestriero & LeCun, 2025), which combines MSE alignment with SIGReg and is evaluated with and without a predictor; VISReg (Wu et al., 2026a), which replaces SIGReg with regularization of embedding center, scale, and projected distribution; and Tilted Empirical Risk Minimization (TERM) (Li et al., 2020) + SIGReg, where TERM uses exponential tilting to assign greater importance to high-loss examples. We also include InfoNCE (Oord et al., 2018), a strong baseline for text representation learning that uses in-batch negatives as strong baselines.

Downstream Task Performance. The LeJEPA results show that removing the projector improves overall text representation performance. PowerJEPA performs substantially better than LeJEPA, supporting the effectiveness of our method. TERM also emphasizes the high-loss tail of positive pairs, but performs worse than PowerJEPA. The NLI results show a similar pattern: removing the projector improves LeJEPA, while PowerJEPA achieves the strongest overall STS performance.

6.2 Ablation Studies

Through experiments on MS MARCO, we examine whether PowerJEPA improves difficult-positive alignment, and how its objective terms contribute to transfer.

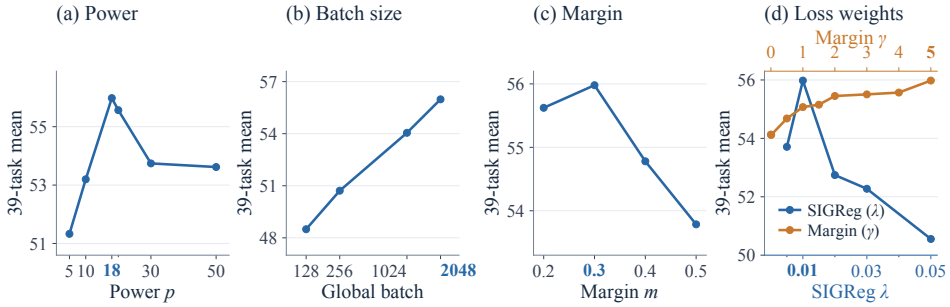


Figure 5: Hyperparameter comparisons on the test task: (a) power p, (b) batch size, (c) boundary margin m, and (d) loss weights. Scores are unweighted task averages multiplied by 100. In (d), the blue bottom axis shows SIGReg weight λ, and the orange top axis shows boundary weight γ.

Difficult Positives. Power alignment places greater relative weight on positive pairs with large distances. We compare the main PowerJEPA encoder with the MSE-based encoder on the same 512 development pairs, using their exported, L2-normalized embeddings. Figure 4 shows a shorter positive-distance tail and substantially better paired-positive ranking on the pairs most difficult under MSE. All distances and rankings are measured in the original 1024-dimensional space.

Complementary Components. We remove each objective term from PowerJEPA to compare its effect on transfer with its effect on representation geometry. Table 3 shows that removing power alignment worsens Q95 and reduces the task mean. Removing the boundary term increases margin violations and lowers the mean. Removing SIGReg produces the smallest Q95, but effective dimension falls sharply and transfer deteriorates. The full objective thus balances positive-pair alignment, negative-relative ordering, and distributional diversity.

Hyperparameter Comparisons. Figure 5 summarizes transfer across different hyperparameter settings on the test task. It shows that the highest mean occurs at $p = 18$ in panel (a), while larger global batches yield higher means, up to 2048, in panel (b). Panel (c) shows the strongest displayed result at margin $m = 0.3$. Panel (d) compares SIGReg and boundary loss weights, respectively.

6.3 Extension to Image Representation Learning

We extend PowerJEPA to supervised image retrieval on ImageNet-1K and compare it with LeJEPA in Table 4. PowerJEPA consistently outperforms Lepanto across all four metrics, with gains of +1.29 CMC@1, +0.13 CMC@5, +1.25 backbone mAP, and +1.76 projector mAP. Appendix C describes the implementation and evaluation protocol, and reports additional unsupervised experiments with online linear-probe evaluation.

Table 4: Supervised ImageNet-1K retrieval (%).

Method	CMC@1	CMC@5	mAP	Proj. mAP
LeJEPA	46.27	65.72	26.98	27.71
PowerJEPA (ours)	**47.56**	**65.85**	**28.22**	**29.47**

6.4 Task-Dependent Dimensional Requirements and Spectral Calibration

The preceding experiments show a trade-off in the global representation space expanded by SIGReg: a higher effective dimension does not necessarily improve transfer, while an overly restricted space limits representational diversity. We therefore study how many representation directions different tasks need, then examine whether adapting the spectrum improves downstream performance.

Task-dependent dimension–performance associations. Complementing the frozen-encoder interventions in Section 5, Figure 3(b) measures ED–score Spearman correlations across historical training checkpoints on all tasks with `subset_size=1000`. The association is stronger for retrieval than clustering, consistent with their differing responses to truncation. These correlations describe task-dependent associations across checkpoints, without isolating the effect of ED from other training-induced changes.

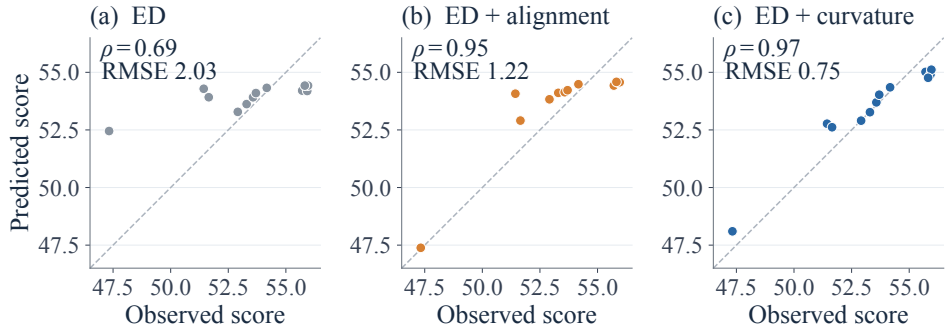


Figure 6: Predicted versus actual task means on 12 held-out configurations using ED, ED plus alignment, and ED plus curvature. dashed diagonals indicate perfect prediction.

Post-Hoc Spectral Calibration. On PowerJEPA, spectral flattening with $\beta = 0.75$ raises effective dimension from 291.98 to 708.43, improving mean classification development performance by 1.68 points but lowering STSBenchmark validation performance by 2.29 points. At $\beta = 1$, dimension reaches 790.75 while both scores decline, demonstrating task-dependent benefits (Figure 3(d)). We therefore explore post-hoc calibration through truncation, spectral reweighting, and mean adjustment. Spectra are fitted on MS MARCO and optionally unlabeled target training text, with task-specific settings selected on development data. Without encoder updates, calibration raises PowerJEPA's test task average with hard negatives from 55.98 to 56.81, slightly exceeding the corresponding uncalibrated InfoNCE baseline (56.72; Table 1).

Predicting Transfer with Effective Dimension and Spectral Curvature. We test whether spectral curvature complements effective dimension (ED) in predicting transfer. Of 39 PowerJEPA and InfoNCE configurations, 27 are used for metric selection and quadratic ridge regression to predict mean test-task scores, and 12 are held out for evaluation. Metrics use 1024-dimensional embeddings from the same 512 MS MARCO development query–passage pairs. Curvature contrasts log-spectrum decay slopes over ranks 8–128 and 129–512. All baselines share the fitting procedure and data split. Adding curvature to ED raises Spearman correlation from 0.692 to 0.972 and reduces RMSE from 2.034 to 0.753 points (Table 5; Figure 6). Compared with ED plus alignment, ED plus curvature also lowers prediction error for classification, pair classification, reranking, retrieval, and STS.

Table 5: Task-mean prediction on 12 held-out configurations.

Metrics	Spearman	RMSE
ED	0.692	2.034
Curvature	0.573	1.129
Alignment + uniformity	0.888	2.108
ED + alignment	0.951	1.216
ED + uniformity	0.776	1.939
ED + single slope	0.902	1.570
ED + curvature	**0.972**	**0.753**
ED + alignment + curvature	0.972	0.787

7 Conclusion

We study representation learning through positive-tail alignment, local boundary correction, and global distribution regularization. Power-mean alignment makes difficult-positive emphasis explicit, while SIGReg and selected negative supervision address complementary geometric constraints. Frozen-encoder interventions reveal that better training-domain pairwise geometry need not imply better transfer and that downstream tasks respond differently to spectral changes. Effective dimension and head-to-tail spectral curvature provide a compact capacity–shape description with useful, but conditional, predictive information across multiple text task families. The resulting perspective connects learning objectives to measurable geometry without reducing representation quality to a single monotonic statistic.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript and assist with code implementation. We reviewed all AI-assisted work, and two authors verified and tested the AI-generated code for correctness. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

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

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 632–642, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. In *International Conference on Learning Representations*, volume 2024, pp. 49001–49024, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Neha Kalibhat, Kanika Narang, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Measuring self-supervised representation quality for downstream classification using discriminative features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13031–13039, 2024.

Myunsoo Kim, Seong-Woong Shim, and Byung-Jun Lee. Falcon: False-negative aware learning of contrastive negatives in vision-language pretraining. *arXiv e-prints*, pp. arXiv–2505, 2025.

Panagiotis Koromilas, Giorgos Bouritsas, Theodoros Giannakopoulos, Mihalis Nicolaou, and Yannis Panagakis. Bridging mini-batch and asymptotic analysis in contrastive learning: From infonce to kernel-based losses. *arXiv preprint arXiv:2405.18045*, 2024.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pp. 79310–79333, 2025.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5):5, 2017.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human-generated machine reading comprehension dataset. 2016.

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

Young-Jin Park, Hao Wang, Shervin Ardeshir, and Navid Azizan. Quantifying representation reliability in self-supervised learning models. *arXiv preprint arXiv:2306.00206*, 2023.

Aleksei Shestov, Anton Klenitskiy, Daria Denisova, Amurkhan Dzagkoev, Daniil Petrovich, Andrey Savchenko, and Maksim Makarenko. Topological metric for unsupervised embedding quality evaluation. In *European Conference on Information Retrieval*, pp. 596–605. Springer, 2026.

Hayato Tsukagoshi and Ryohei Sasano. Redundancy, isotropy, and intrinsic dimensionality of prompt-based text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25915–25930, 2025.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pp. 1112–1122, 2018.

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

Songhao Wu, Zhongxin Chen, Yuxuan Liu, Heng Cui, Cong Li, and Rui Yan. Your unembedding matrix is secretly a feature lens for text embeddings. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5454–5463, 2026b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhoushun Yu, Xiaoyu Hu, and Xiangyu Xu. Qqworld: Quantile-quantile matching for world model regularization. *arXiv preprint arXiv:2607.28415*, 2026.

Jihai Zhang, Xiang Lan, Xiaoye Qu, Yu Cheng, Mengling Feng, and Bryan Hooi. Learning the unlearned: Mitigating feature suppression in contrastive learning. In *European Conference on Computer Vision*, pp. 35–52. Springer, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A Proofs and Additional Analysis

A.1 Power-Mean Alignment

Gradient allocation. The following proposition characterizes how the power-mean objective distributes its distance-gradient weights.

Proposition A.1 (Gradient allocation)**.** *For $p > 1$ and $\widehat{\mathcal{L}}_p > 0$,*

$$\frac{\partial \widehat{\mathcal{L}}_p}{\partial d_i} = \frac{1}{B}\left(\frac{d_i}{\widehat{\mathcal{L}}_p}\right)^{p-1}. \tag{14}$$

For $d_i, d_j > 0$, the ratio of their distance derivatives is $(d_i/d_j)^{p-1}$.

Proof. Let $S = B^{-1}\sum_i d_i^p$, so that $\widehat{\mathcal{L}}_p = S^{1/p}$. For $S > 0$, differentiation gives

$$\frac{\partial \widehat{\mathcal{L}}_p}{\partial d_i} = \frac{1}{p} S^{1/p-1} \frac{p}{B} d_i^{p-1} \tag{15}$$

$$= \frac{1}{B} \widehat{\mathcal{L}}_p^{1-p} d_i^{p-1}. \tag{16}$$

Taking the ratio for two positive distances proves the claim. □

For the MSE objective $B^{-1}\sum_i d_i^2$, the corresponding ratio is d_i/d_j. Thus, when $d_i > d_j$, choosing $p > 2$ increases the relative distance-gradient weight assigned to the larger residual. The resulting parameter gradients additionally depend on each pair's Jacobian:

$$\nabla_\theta \widehat{\mathcal{L}}_p = \sum_i \frac{\partial \widehat{\mathcal{L}}_p}{\partial d_i} \nabla_\theta d_i, \tag{17}$$

where the distance derivatives exist.

Positive-tail control.

Proof of Proposition 4.1. Markov's inequality applied to D^p yields

$$\Pr(D \geq t) = \Pr(D^p \geq t^p) \leq \frac{\mathbb{E}D^p}{t^p} = \left(\frac{M_p}{t}\right)^p. \tag{18}$$

Because quantiles are nondecreasing,

$$Q_{1-\alpha}(D) \leq \frac{1}{\alpha} \int_{1-\alpha}^1 Q_u(D)\, \mathrm{d}u. \tag{19}$$

Hölder's inequality gives

$$\mathrm{CVaR}_\alpha^{\mathrm{upper}}(D) \leq \alpha^{-1/p} \left(\int_{1-\alpha}^1 Q_u(D)^p\, \mathrm{d}u\right)^{1/p} \tag{20}$$

$$\leq \alpha^{-1/p} \left(\int_0^1 Q_u(D)^p\, \mathrm{d}u\right)^{1/p} \tag{21}$$

$$= \alpha^{-1/p} M_p, \tag{22}$$

where the last equality uses the quantile representation of the pth moment.

For the empirical distribution assigning probability $1/B$ to each d_i, its pth moment is $B^{-1}\sum_i d_i^p = \widehat{\mathcal{L}}_p^p$. □

Relation to the maximum residual. Writing $d_{\max} = \max_i d_i$, we have

$$\frac{d_{\max}^p}{B} \leq \frac{1}{B}\sum_i d_i^p \leq d_{\max}^p. \tag{23}$$

Therefore,

$$B^{-1/p} d_{\max} \leq \widehat{\mathcal{L}}_p \leq d_{\max}. \tag{24}$$

This proves both $d_{\max} \leq B^{1/p}\widehat{\mathcal{L}}_p$ and convergence to $d_{\max}$ as $p \to \infty$. Increasing p does not automatically tighten every moment bound, because M_p also changes with p.

Numerical evaluation. For $d_{\max} > 0$, an equivalent scaled expression is

$$\widehat{\mathcal{L}}_p = d_{\max} \left[\frac{1}{B}\sum_i \left(\frac{d_i}{d_{\max}}\right)^p\right]^{1/p}. \tag{25}$$

All powered ratios lie in $[0, 1]$, avoiding overflow from large powers. If all distances vanish, the objective is zero. At $p = 2$, this objective is the root mean square distance: its relative distance-gradient weights match MSE, although their overall scale differs.

A.2 Equal Dimension and Different Regularized Capacity

Proof. Fix $\eta > 0$ and first consider three positive eigenvalues $x = (x_1, x_2, x_3)$. Define

$$T(x) = \sum_{j=1}^{3} x_j, \qquad S(x) = \sum_{j=1}^{3} x_j \log x_j. \tag{26}$$

The entropy of the normalized spectrum is

$$H(x) = -\sum_{j=1}^{3} \frac{x_j}{T(x)} \log \frac{x_j}{T(x)} = \log T(x) - \frac{S(x)}{T(x)}. \tag{27}$$

Hence fixing both T and S fixes effective dimension $\exp(H)$.

Choose

$$x^{(0)} = (\eta, 2\eta, 4\eta), \qquad v = (1, -2, 1). \tag{28}$$

At $x^{(0)}$, the gradients

$$\nabla T = (1, 1, 1), \qquad \nabla S = (1 + \log x_1, 1 + \log x_2, 1 + \log x_3) \tag{29}$$

are linearly independent. Moreover,

$$\nabla T(x^{(0)})^\top v = 0, \tag{30}$$

$$\nabla S(x^{(0)})^\top v = \log \eta - 2\log(2\eta) + \log(4\eta) = 0. \tag{31}$$

The implicit function theorem therefore gives a smooth local curve $x(s)$ in the positive orthant satisfying

$$x(0) = x^{(0)}, \qquad x'(0) = v, \tag{32}$$

$$T(x(s)) = T(x^{(0)}), \qquad S(x(s)) = S(x^{(0)}). \tag{33}$$

Now define

$$F_\eta(x) = \sum_{j=1}^{3} \frac{x_j}{x_j + \eta}. \tag{34}$$

Along this curve,

$$\frac{\mathrm{d}}{\mathrm{d}s} F_\eta(x(s))\Big|_{s=0} = \sum_{j=1}^{3} \frac{\eta v_j}{(x_j^{(0)} + \eta)^2} \tag{35}$$

$$= \frac{1}{\eta}\left(\frac{1}{4} - \frac{2}{9} + \frac{1}{25}\right) = \frac{61}{900\eta} > 0. \tag{36}$$

Thus, for sufficiently small positive s, the diagonal matrices with spectra $x^{(0)}$ and $x(s)$ have identical trace and effective dimension but different d_η.

For $d > 3$, append $d - 3$ eigenvalues equal to η to both spectra. The added contributions to T, S, and d_η are identical. Equation 27, with sums over all d coordinates, preserves equality of effective dimension, while the difference in d_η remains nonzero. Both matrices are positive definite and have rank d. □

A.3 Task-Dependent Spectral Risk

Proof of Proposition 5.1. For each coordinate,

$$\widehat{a}_j - a_j = -\frac{\eta}{\lambda_j + \eta} a_j + \frac{\lambda_j}{\lambda_j + \eta} \frac{\sigma}{\sqrt{n}} \xi_j. \tag{37}$$

Since $\mathbb{E}\xi_j = 0$ and $\mathbb{E}\xi_j^2 = 1$, the cross term vanishes after taking expectations, giving

$$\mathbb{E}(\widehat{a}_j - a_j)^2 = \left(\frac{\eta}{\lambda_j + \eta}\right)^2 a_j^2 + \frac{\sigma^2}{n}\left(\frac{\lambda_j}{\lambda_j + \eta}\right)^2. \tag{38}$$

Summing over coordinates proves Equation 12. □

A.4 Batch Statistics in SIGReg

Let $X_1, \dots, X_N$ be i.i.d. projected samples with characteristic function $\phi(t)$, and define

$$\widehat{\phi}(t) = \frac{1}{N} \sum_{j=1}^{N} \exp(\mathrm{i}tX_j). \tag{39}$$

Independence and $|\exp(\mathrm{i}tX_j)| = 1$ imply

$$\mathbb{E}|\widehat{\phi}(t) - \phi(t)|^2 = \frac{1 - |\phi(t)|^2}{N}. \tag{40}$$

For a fixed target characteristic function $\phi_0(t)$, unbiasedness further gives

$$\mathbb{E}|\widehat{\phi}(t) - \phi_0(t)|^2 = |\phi(t) - \phi_0(t)|^2 + \frac{1 - |\phi(t)|^2}{N}. \tag{41}$$

Consequently,

$$\mathbb{E}\Big[N|\widehat{\phi}(t) - \phi_0(t)|^2\Big] = N|\phi(t) - \phi_0(t)|^2 + 1 - |\phi(t)|^2. \tag{42}$$

Although the unscaled estimation variance decreases with N, the SIGReg objective includes an N multiplier. At distributional match, the expected scaled discrepancy in Equation 42 is independent of N. Moreover, pooled query–positive views are correlated, requiring additional covariance terms, and changing batch size at fixed epochs changes the number of optimizer updates. These identities motivate examining batch statistics but do not establish a monotonic relationship between batch size, effective dimension, and downstream performance.

B Additional Experimental Details

B.1 Training Recipes and Evaluation Scope

All use Qwen3-0.6B and seed 42. The exported encoder includes any inference-time residual embedding head. This head is distinct from the query-side predictor used only by the MSE-based training objective.

MS MARCO configurations. The $p = 18$ recipe combines normalized power alignment, union SIGReg with an additional query regularizer, L2 triplet loss with weight 5 and margin 0.3, self-hard-negative symmetric supervision with weight 0.5, and a global hard-document softplus term with weight 0.1, $k = 8$, cosine margin 0.05, and temperature 0.05. Its residual output MLP coefficient decays from 1.0 to 0.2.

NLI configurations. The main $p = 5$ recipe uses learning rate 10^{-4}, SIGReg weight 0.01, additional query regularization weight 0.01, L2 triplet weight 3, margin 0.3, and a residual output head. Premise–entailment/contradiction triplets are grouped by anchor. One positive and, when enabled, one negative are sampled per instance.

Evaluation scope. We use the intersection of 37 evaluated tasks, excluding ImdbClassification and Touche2020Retrieval.v3 from the 39-task evaluation pool. Task-family means are recomputed from the included task-level results. Family sizes are 6, 8, 3, 2, 8, 9, and 1 in the order used in Table 1.

B.2 Frozen Spectral Interventions

The covariance reference contains 1024 normalized embeddings from the query and positive sides of 512 MS MARCO training triples. Let μ denote their mean, C their centered covariance, and U_k the leading k covariance eigenvectors. The implemented transformation is

$$T_k(v) = \mathcal{N}\big(\mu + c_k U_k U_k^\top (v - \mu)\big), \qquad c_k = \sqrt{\frac{\mathrm{tr}(C)}{\sum_{j=1}^{k} \lambda_j}}, \tag{43}$$

Table 6: Complete spectral-intervention results on the 13-task capped evaluation. PCA-1024 is identity. ED and Q95 are measured on the spectral fitting sample.

Transform	Directions	ED	Q95	Subset mean
Bottom	512	296.2	1.756	61.36
PCA	32	30.6	0.746	57.05
PCA	64	58.4	0.794	60.41
PCA	128	103.4	0.850	62.23
PCA	256	160.1	0.891	62.72
PCA	512	209.5	0.943	63.28
PCA	1024	225.5	0.953	63.62
Random	64	53.5	0.978	59.59

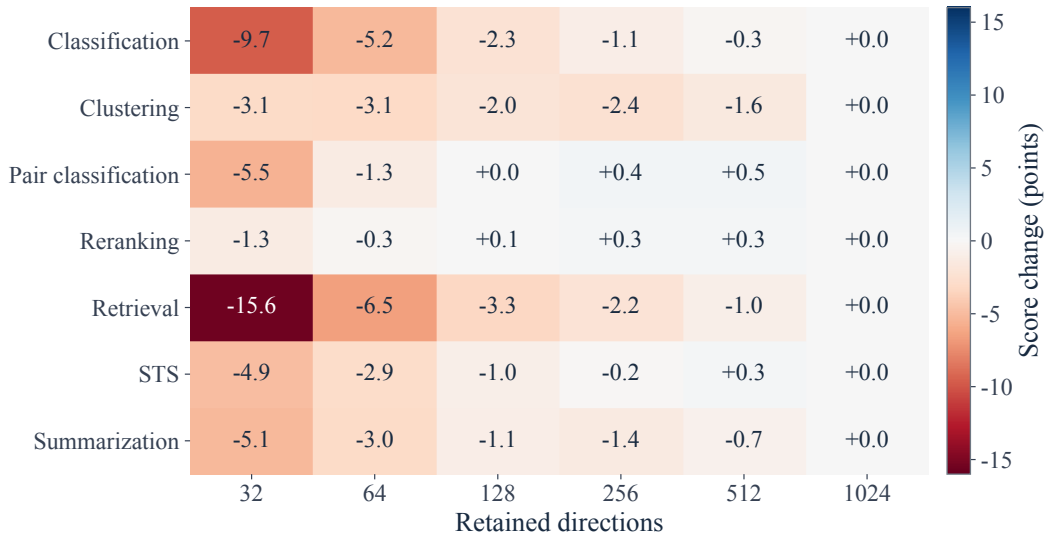


Figure 7: Responses of all seven task families to PCA truncation, relative to identity. Evaluation uses the same 13-task capped protocol. Zero denotes the identity score, and each family uses its respective evaluation metric.

where $\mathcal{N}$ denotes L2 normalization.

This transformation preserves the reference centered trace before final normalization and reconstructs embeddings in the original 1024-dimensional space. The full-rank transform returns identity. The retained direction count does not equal the output width, and the output lies in the span of the reference mean and retained eigenvectors. Directly exporting centered PCA coordinates would constitute a different transformation.

The mechanism study uses 13 tasks with `subset_size=1000` and seed 42. Retrieval corpora and queries are capped, so its retrieval results do not represent full-corpus evaluation. Effective dimension and pair geometry are measured on the same reference set used for spectral fitting. We evaluate one random-64 subspace; the different transformations do not constitute independent training replicates.

Bottom-spectrum geometry is particularly sensitive to finite-sample covariance estimation. A separate fit/measurement split produces less severe boundary degradation on held-out geometry samples. The bottom-512 result therefore does not imply that all low-variance directions lack semantic information. Reranking and pair classification also exhibit non-monotonic responses at intermediate truncation levels.

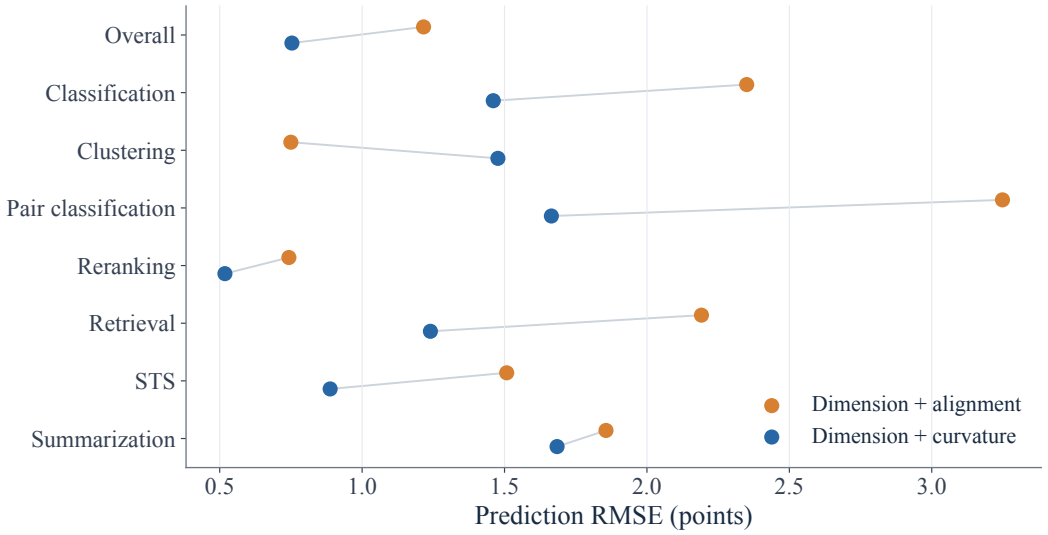


Figure 8: Task-family prediction errors on the same 12 additional configurations and independent development probe. ED plus curvature reduces error for most families, but increases clustering RMSE relative to ED plus alignment.

B.3 Post-Hoc Calibration

The post-hoc policies are exploratory follow-ups informed by the preceding experiments. Source policies fit spectra to source-training embeddings. Domain policies additionally permit unlabeled target training text and use development data to select task-family-specific settings. The reported policies do not select each test task's best transformation using test scores.

We include one common source transform as a control; it does not improve the branch without boundary supervision. Identity scores differ slightly from the original export evaluations because the calibration study uses its own inference implementation. All calibration gains are therefore computed relative to the identity result within that study. The contrastive references do not receive an equivalent calibration procedure.

B.4 Full-Task Spectral Interventions on the Main $p = 18$ Encoder

We retain the eight settings of the original intervention study: PCA with 32, 64, 128, 256, 512, and 1024 directions, random-64, and bottom-512. The encoder is the main $p = 18$ PowerJEPA checkpoint with its saved embedding head. The full-rank PCA setting is implemented as identity and evaluated once.

The source sample contains the same 512 MS MARCO training triples selected with seed 20260914 for the original $p = 20$ experiment. We re-encode this sample using the $p = 18$ encoder. Normalized query and positive embeddings supply 1024 vectors for estimating the mean μ and centered covariance $C = U\Lambda U^\top$. No downstream text or labels are used for this fit.

Let B have orthonormal columns spanning the retained directions. PCA uses the leading eigenvectors; bottom-512 uses the trailing 512 eigenvectors. Random-64 uses the reduced QR basis of a 1024×64 standard Gaussian matrix generated with seed 20260914. For an initially normalized embedding z, the transform is

$$T_B(z) = \mathrm{norm}_2\left[\mu + s_B B B^\top (z - \mu)\right], \qquad s_B = \sqrt{\frac{\mathrm{tr}(C)}{\mathrm{tr}(B^\top C B)}}.$$

The scaling matches centered variance on the fitting sample before final L2 normalization. The output width remains 1024. Effective dimension and positive-distance Q95 are measured on the transformed source fitting sample.

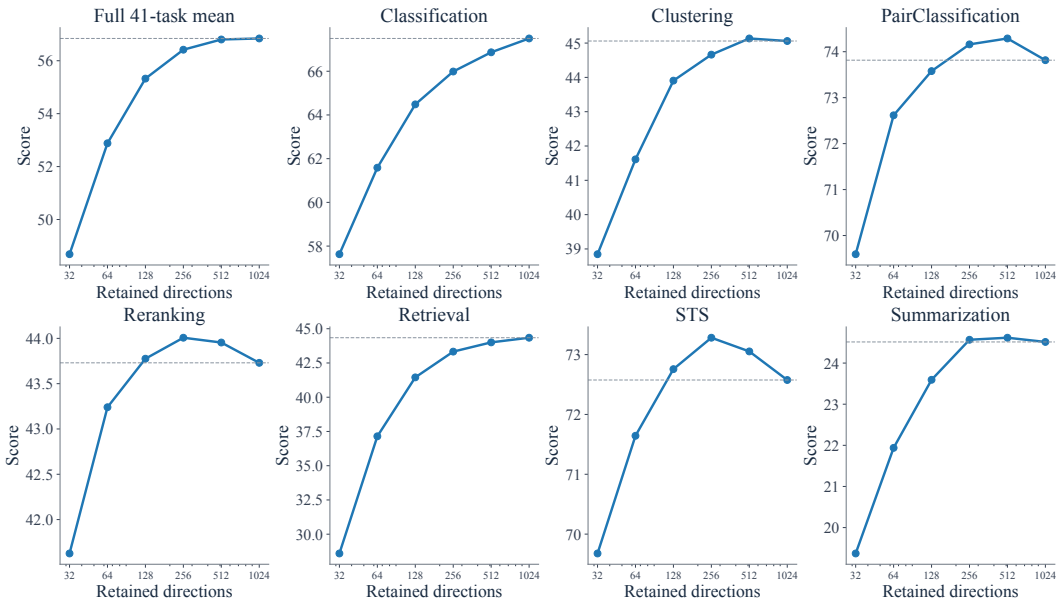


Figure 9: Leading-direction interventions on the frozen main $p = 18$ encoder across all seven task families and the full 41-task mean. Dashed lines indicate each panel's full-representation score. All curves use the same six retained ranks and complete official evaluation data.

All eight settings are evaluated on all 41 official tasks in MTEB(eng, v2), using the official evaluation splits and subsets, evaluator seed 42, maximum sequence length 512, and encoding batch size 64. Retrieval corpora, queries, and relevance judgments are not truncated. Each setting acts on all embeddings used by its evaluator, including classification training and test embeddings and retrieval query and document embeddings.

Figure 3 in the main text retains the original two-panel structure. Its left panel reports within-family changes from identity for retrieval, clustering, and reranking; its right panel reports effective dimension against the full 41-task mean for all eight settings. Figure 9 below provides the complete leading-direction curves for all seven task families and the overall mean. Separate common-39 and common-37 aggregates are exported for comparisons requiring those fixed task sets.

Only raw embeddings with matching model-file fingerprints, input text, and encoding role are reused from the existing $p = 18$ cache; scores are recomputed for every task and intervention. All settings were fixed before examining this rerun's test results. Completion requires 328 successful task–setting evaluations. These transformations change direction selection and other aspects of geometry jointly, so their effects do not identify an independent causal contribution of effective dimension.

Increasing retained rank from 32 to 1024 raises the full-task mean from 48.69 to 56.84. Retrieval changes by 15.75 score points, clustering by 6.21, and reranking by 2.10, each in its own metric. At rank 64, the mean is 52.88 and source-sample Q95 is 0.760, compared with 56.84 and 0.925 for the full representation. PCA-64 also exceeds random-64, whose mean is 51.24. These observations separate task-dependent responses from a universal preference for lower distance or greater effective dimension.

B.5 Post-Hoc Calibration Protocol

The candidate transformations include truncation, spectral reweighting, and mean adjustment. Source policies fit spectra on MS MARCO; domain policies additionally use unlabeled target training text. Task-specific settings use development supervision. All reported gains compare with identity under the same inference implementation. Source calibration increases the 37-task mean from 55.95 to 56.53, domain calibration to 56.79, and refined development selection to 56.80.

This table retains the calibration study's existing inference protocol and 37-task aggregation. Its absolute means should not be compared directly with the 41-task intervention study or the main table's 39-task aggregation. InfoNCE does not receive equivalent calibration, so these results do not establish superiority over calibrated InfoNCE.

B.6 Transfer Prediction: Protocol and Additional Analysis

B.6.1 Protocol

Effective dimension and spectral curvature are measured on fixed upstream embedding probes. Curvature contrasts log-spectrum decay slopes fitted over ranks 8–128 and 129–512. Descriptor selection and predictor calibration use 27 discovery encoders. The 12 additional configurations are excluded from both stages and cover regularization scope, schedules, batch and hard-negative settings, and self-symmetric training. Three use $p = 18$, including the main encoder; the prediction study spans multiple training configurations. Their descriptors are measured using 512 independent development pairs. All descriptor baselines use the same degree-two ridge learner and calibration set to predict the shared 37-task mean.

B.6.2 Additional Comparisons and Limitations

Curvature alone has Spearman correlation 0.573 and RMSE 1.129, below the joint descriptor's correlation of 0.972 and RMSE of 0.753. ED plus alignment reaches 0.951 and 1.216; ED plus a single slope reaches 0.902 and 1.570. These comparisons support complementary information in spectral curvature beyond dimension and average alignment.

Relative to ED plus alignment, ED plus curvature reduces prediction error for classification, pair classification, reranking, retrieval, and STS, but increases clustering RMSE from 0.750 to 1.477. The results depend on the nonlinear learner and do not establish reliable selection among intermediate checkpoints or generalization across backbones. Although descriptor measurement requires no downstream labels, fitting the performance predictor uses calibration encoders' downstream scores.

C Image Representation Learning Details

C.1 Implementation Details

Model architecture. We use a ViT-B/16 backbone (Dosovitskiy, 2020) implemented via `timm`, configured with `global_pool=token` and stochastic depth rate 0.1. The backbone outputs a 768-dimensional CLS token. A 3-layer MLP projector maps the CLS vector to 256 dimensions through hidden dimensions $768 \to 2048 \to 2048 \to 256$, with GELU activations and Synchronized BatchNorm between layers. The alignment loss operates on the unnormalized 256-d projector output, while retrieval evaluation uses both the 768-d backbone CLS (primary metric) and 256-d projector features.

Training setup. Both methods are trained on ImageNet-1K (Deng et al., 2009) training images (1,281,167 images, 1000 classes). Each input image is resized to 224×224. For each query image, a positive is uniformly sampled from the same class (a different image). We train for 100 epochs on 32 GPUs with per-GPU batch size 16 (global batch 512 pairs), using AdamW (lr $= 5 \times 10^{-4}$, weight decay 0.05), cosine schedule with 5-epoch linear warmup and minimum-LR divisor 1000, gradient clipping at 1.0, and BF16 mixed precision.

Both methods share the same SIGReg regularizer:

$$\mathcal{L} \;=\; 0.98\,\mathcal{L}_{\text{align}} + 0.02\,\mathcal{L}_{\text{SIGReg}}. \tag{44}$$

C.2 Evaluation Protocol

Retrieval evaluation uses the full ImageNet-1K validation set (50,000 images). For each query, all other images of the same class form the relevant set (49 positives per query out of 50 images per class). Similarity is cosine distance on ℓ_2-normalized features. The query itself is excluded from the gallery. We report metrics:

- **CMC@1**: fraction of queries whose top-1 retrieval is a true positive;
- **CMC@5**: fraction with at least one true positive in the top 5;
- **mAP**: mean average precision across all queries.

Evaluation is performed in a distributed manner: each rank processes a shard of queries against the full gallery, and partial sums are reduced before computing the final metrics.

C.3 Unsupervised Representation Learning

Setup. We additionally evaluate PowerJEPA in the unsupervised setting of SIGReg, using ViT-B/16 on ImageNet-1K with the multi-crop protocol (4 global views at 224×224 and 6 local views at 96×96). The global batch size is 512 (32 GPUs $\times$ 16 images), with base learning rate 9×10^{-4}, cosine schedule over 400 epochs (5-epoch linear warmup, minimum-LR divisor 1000), and SIGReg weight $\lambda = 0.02$. Representation quality is measured by *online linear-probe* top-1 accuracy on the ImageNet-1K validation set.

Table 7: Unsupervised ImageNet-1K online linear-probe top-1 accuracy (%)..

Method	Top-1 Acc.
LeJEPA (SIGReg)	74.76
PowerJEPA	75.00

Results. As shown in Table 7, PowerJEPA achieves 75.00% top-1 accuracy, providing a modest improvement of 0.24 percentage points over LeJEPA in unsupervised visual representation learning.