COMPOSITIONAL OBJECTIVES: LEARNING STRUCTURE IN STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Intelligence is defined in many ways. One of these definitions defines intelligence as the pursuit of learnable novelty. However, learnable novelty can be meaningless without the ability to compose the learned structures to take action and achieve goals. Learnable novelty builds on epiplexity, which is a way to measure learnable structure in data through a bounded observer. In this paper, we investigate a closed-form spectral approximation to compute epiplexity. We use a fixed-trace constraint and find that the epiplexity objective prefers a more uniform distribution of spectral mass rather than concentrating it in a small number of directions. However, a representation may spread information across many directions without organizing that information into features useful for a particular task. To address this gap, we propose a compositional objective whose observer measures the relationships between the parts and interactions of an image. We compare it with the original spectral objective given only the masked parts. In our ImageNet training runs, the spectral objective with masked parts produces an almost maximally spread representation while achieving the strongest frozen-feature classification performance on most evaluations, more than doubling the linear-probe accuracy of the whole-image baseline. Across multiple image benchmarks, changing what the observer sees matters more than adding relation and interaction tokens. At the same time, our prediction-oriented compositional objective produces substantially better held-out observer prediction but relatively weaker classification, revealing that spectral diversity, predictability, and downstream utility are distinct properties. These results suggest that the usefulness of spectral spreading depends not only on how much structure is preserved, but on which relationships the observer makes available to the objective.

1 INTRODUCTION

Intuitively, we recognize that one part of intelligence is the ability to find patterns in observations. Another part is the ability to reduce these observations into efficient, reusable representations. Much of our understanding of perception, learning and reasoning has placed these two parts at the center of cognition and machine learning. Information-theoretic approaches, like Minimum Description Length (Rissanen, 1978; Grünwald, 2007), further formalize the connection between pattern extraction and compression as extracting structural regularity in data. However, these abilities are constrained by the computational bounds of the entity or model that is observing the data. Computationally constrained theories of information formalize precisely this distinction between what is present in data and what can be used.

One such formalization is epiplexity (Finzi et al., 2026). It is defined as the description length in bits of the time-bounded MDL-minimizing program. Intuitively, epiplexity can be thought of as the amount of structure that the best model extracts from data, given constraints of time and compute bounds. This separates structural information from the residual, time-bounded entropy.

Zhang & Levin (2026) introduce the idea of learnable novelty and make the concept of epiplexity operational through a differentiable spectral approximation. Their observer combines a fixed nonlinear feature map with a fitted linear readout. This readout then predicts a trainable encoder's representation, and the encoder is optimized to increase the spectral score of the fitted readout. The observer thus participates in learning rather than merely evaluating its outcome: its construction

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.

shapes the signal through which the representation is trained. This provides a concrete setting in which to ask how the organization of the observer influences the organization of what is learned. Prior work shows that the choice of views affects which task information is retained in contrastive learning (Tian et al., 2020).

Together, these developments provide a way to measure learnable structure and use its approximation as a training objective. However, we find that a criterion for extracting structure is not necessarily a criterion for making that structure useful. Our theoretical formulation examines the distinction through a counterexample in which representations share the same covariance spectrum while differing in task utility.

The issue is not that spectral objectives cannot reward relationships, but that spectral diversity alone does not specify which relationships should be learned. We therefore turn to the observer as a place to introduce a more explicit objective on how the whole is explained. Consider a tree. A learner may capture recurring leaf shapes, bark textures, and branch forms in reusable descriptions. However, a catalog of these parts does not explain how a particular tree is put together. To trace a path from a leaf to the trunk, the learner must represent which branches connect to which. Knowing the pieces is not knowing the whole. So, to summarize, we can decompose intelligence into three parts. The first is the ability to find patterns. The second is to reduce these patterns into efficient, reusable parts. And the third, our proposition: composition.

In this paper, we make three contributions:

1. We analyze the spectral approximation to epiplexity and show that, under a fixed-trace constraint, the objective favors a more uniform distribution of spectral mass. We complement this analysis with a counterexample to show that the uniform distribution of spectral mass does not necessarily contribute to downstream utility.
2. We train four different models on the full ImageNet-1K training set and evaluate performance on CIFAR-100, DTD, EuroSAT, CLEVR, and SpatialSense. We find that input pressure is the leading reason for improved performance over any architectural change.
3. Through our experimentation, we find that three properties of a representation can diverge: how broadly it distributes variance, how well an observer can predict it, and how effectively a downstream learner can use it.

2 Background

2.1 Epiplexity

Finzi et al. (2026) define the time-bounded minimum description length of a data object D as

$$\mathrm{MDL}_T(D) = \min_{P \in \mathcal{P}_T} \left\{ \ell(P) + \mathbb{E}_D[-\log_2 p_P(D)] \right\}, \tag{1}$$

where $\mathcal{P}_T$ contains programs satisfying the time bound T, $\ell(P)$ is the length of P in bits, and $\mathbb{E}_D$ is the expectation over the data object D, treated as a random variable. The two terms balance the bits needed to describe the model against those needed to encode the data using it. Epiplexity is $S_T(D) = \ell(P_T^\star)$, the description length in bits of the program $P_T^\star$ that minimizes this time-bounded objective.

For random data objects $\mathcal{U}, \mathcal{V}$, conditional epiplexity is $S_T(\mathcal{V} \mid \mathcal{U}) = \ell(P^\star_{\mathcal{V}|\mathcal{U}})$, where

$$P^\star_{\mathcal{V}|\mathcal{U}} \in \arg\min_{P \in \mathcal{P}_T^{\mathcal{U}}} \left\{ \ell(P) + \mathbb{E}_{(\mathcal{U},\mathcal{V})}[-\log_2 p_P(\mathcal{V} \mid \mathcal{U})] \right\} \tag{2}$$

and $\mathcal{P}_T^{\mathcal{U}}$ contains admissible conditional probabilistic programs satisfying the time bound T. It is the description length in bits of the program minimizing the expected combined cost of describing the model and encoding $\mathcal{V}$ given $\mathcal{U}$. Epiplexity can be estimated using prequential and requential coding, but neither naturally provides the cheap, differentiable, per-batch signal required for direct gradient-based representation learning. This motivates a closed-form surrogate that can be optimized during training.

2.2 THE SPECTRAL OBSERVER

Zhang & Levin (2026) make the conditional epiplexity quantity tractable using a fixed observer and a closed-form readout. We describe it below. For a batch $X = (x_1, \ldots, x_N)$, $z_i = E_\theta(x_i) \in \mathbb{R}^D$, $\|z_i\|_2 = 1$, $h_i = \phi(x_i) \in \mathbb{R}^m$, where E_θ is the trainable, unit-normalized encoder and ϕ is a fixed random nonlinear feature map. Originally, the scored targets are a system's outputs; here they are the learned representations z_i. Collect these into matrices $Z_\theta \in \mathbb{R}^{N\times D}$ and $H \in \mathbb{R}^{N\times m}$. These observer features and targets are then standardized, giving $\widetilde{H}$ and $\widetilde{Z}_\theta$ (*Appendix A*).

The observer fits a linear readout by ridge regression (Hoerl & Kennard, 1970), using a quadratic penalty of the form studied in Tikhonov regularization (Tikhonov & Arsenin, 1977):

$$W_\lambda = \underset{W\in\mathbb{R}^{m\times D}}{\arg\min} \left\{ \|\widetilde{Z}_\theta - \widetilde{H}W\|_F^2 + \lambda\|W\|_F^2 \right\} = \left(\widetilde{H}^\top\widetilde{H} + \lambda I_m\right)^{-1}\widetilde{H}^\top\widetilde{Z}_\theta, \qquad \lambda > 0, \tag{3}$$

so that the observer's predictions in the preprocessed target coordinates are $\widehat{\widetilde{Z}}_\theta = \widetilde{H}W_\lambda$. The spectral approximation to epiplexity is

$$S^\phi(Z_\theta \mid X) = \frac{1}{2}\sum_j \log_2\left(1 + \eta\, s_j(W_\lambda)^2\right), \qquad \eta > 0, \tag{4}$$

where $s_j(W_\lambda)$ denotes a singular value of the fitted readout and η is the spectral resolution parameter. The encoder is trained by

$$\min_\theta \; -S^\phi(Z_\theta \mid X). \tag{5}$$

Clearly, they have proposed an elegant approach to make epiplexity usable. There is now a way to extract patterns and their compressed forms from data. However, we still need to answer what these spectral scores are rewarding in a representation.

3 WHAT THE SPECTRAL OBJECTIVE REWARDS

3.1 SPECTRAL PREFERENCE UNDER A FIXED-TRACE CONDITION

We show that the spectral objective tends to reward a flatter spectrum. The log-determinant score depends both on the total magnitude of the readout and on how that magnitude is distributed across spectral directions. To isolate the second effect, consider a fixed total readout energy: $\mathrm{tr}(W_\lambda^\top W_\lambda) = \|W_\lambda\|_F^2 = \tau$. The following proposition characterizes the resulting spectral preference. The bound follows from the concavity of the logarithm and Jensen's inequality (Boyd & Vandenberghe, 2004).

Proposition 3.1 (Fixed-trace spectral preference)**.** *Let $A \in \mathbb{R}^{D\times D}$ be positive semidefinite, with $\mathrm{tr}(A) = \tau > 0$, $\mathrm{rank}(A) \le r$, $1 \le r \le D$. For $\eta > 0$,*

$$\frac{1}{2}\log_2\det(I_D + \eta A) \le \frac{r}{2}\log_2\left(1 + \frac{\eta\tau}{r}\right). \tag{6}$$

Equality holds if and only if A has exactly r nonzero eigenvalues, each equal to τ/r. Proof is in Appendix B.1.

Applying Proposition 3.1 to $A = W_\lambda^\top W_\lambda$ shows that, *at fixed readout energy, the score favors distributing spectral mass evenly across the available directions*. Moreover, the bound is increasing in r, so at fixed energy the score also prefers using more directions. Since $W_\lambda \in \mathbb{R}^{m\times D}$ has rank at most $\min(m, D)$, when $m \ge D$ all D readout directions are feasible and the bound becomes

$$S^\phi(Z_\theta \mid X) \le \frac{D}{2}\log_2\left(1 + \frac{\eta\tau}{D}\right). \tag{7}$$

This motivates us to answer the next question: whether a uniform distribution of spectral mass helps a representation's downstream task utility.

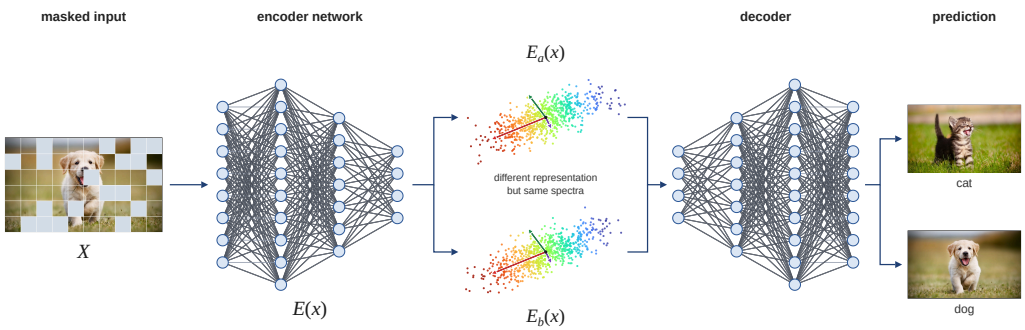


Figure 1: Identical spectra need not give the same prediction. Two representations $E_a(x)$ and $E_b(x)$ of the same masked input share a spectrum but lead the decoder to different outputs. A conceptual illustration, not trained models.

3.2 Equal spectra do not imply equal task utility

We propose a counterexample, considering two different representations of the same input, with the same covariance spectra, and show that they can perform differently on the same downstream task (Figure 1).

Proposition 3.2. *There exist two unit-norm representations and a fixed observer such that their representation covariances are identical and isotropic, their spectral scores are identical, but one representation permits perfect linear classification while the other permits only chance-level classification. Proof is in Appendix B.2.*

This counterexample ultimately shows us that the spectrum by itself does not identify which features of the input occupy the representation's directions. So, under the spectral objective S^{ϕ}, we find that there is no signal that forces the representation to be useful for downstream tasks. This motivates us to think about what the observer rewards, and in Section 4 we elucidate our core contribution of a Compositional Objective.

4 The Compositional Objective

4.1 Formalization of the Compositional Objective

In this section we introduce the notion of a compositional objective. We construct this objective by providing inputs that are decomposed into multiple parts. These parts form three kinds of tokens (Figure 2):

1. *View tokens*: These are inputs that are masked differently. Every different mask makes a different view.
2. *Relation tokens*: We posit that the observer needs to have information to understand how each view relates to each other view.
3. *Interaction tokens*: For every view token pair and its corresponding relation token, we construct a fixed projection of the multiplicative interaction $(Uu_i) \odot (Vu_j) \odot (T\,[r_{ij}; 1])$, where u_i, u_j are the two views, r_{ij} is their relation, and U, V, and T are fixed Gaussian matrices.

We study two main ablations. The first, $S^{\phi_{\text{comp}}}$, keeps the original epiplexity objective S^{ϕ} of Section 2.2 and passes its observer only the view tokens. The second combines all three kinds of tokens into the observer's feature map, ϕ_{comp}, which our compositional objective S_{comp} uses. Rather than retaining the spectral objective of S^{ϕ} and changing only its input to ϕ_{comp}, S_{comp} measures how much representation variance the observer actually explains. We try to distinguish between using the decomposed input as the only signal for composition, versus altering the observer itself to fit the input uniquely.

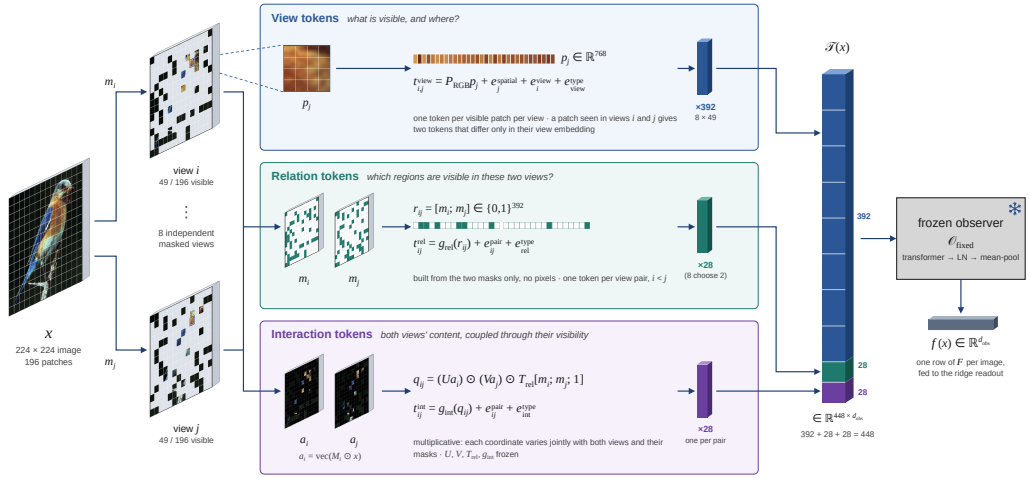


Figure 2: The compositional observer's input tokens. Eight masked views of a 196-patch image give 392 view tokens, 28 relation tokens from the visibility masks of each view pair, and 28 interaction tokens that couple both views' content with their visibility. A frozen transformer maps the 448 tokens to one feature vector $f(x)$.

Centering and the shared readout. Our objective uses its own preprocessing and readout, distinct from those of Section 2.2. Let $f_i = \phi_{\text{comp}}(x_i; \mathcal{M}_i) \in \mathbb{R}^m$ denote the compositional observer features. We stack the centered features $f_i - \bar{f}$, where $\bar{f}$ is their batch mean, into $F \in \mathbb{R}^{N \times m}$, and the centered encoder outputs into $\bar{Z}_\theta = Z_\theta - \mathbf{1}_N \mu_Z^\top \in \mathbb{R}^{N \times D}$, where μ_Z is the batch mean of the encoder outputs. Neither is standardized. We define the covariances

$$C_F = \frac{1}{N} F^\top F, \qquad C_Z = \frac{1}{N} \bar{Z}_\theta^\top \bar{Z}_\theta, \qquad C_{FZ} = \frac{1}{N} F^\top \bar{Z}_\theta. \tag{8}$$

Centering makes the target mean the reference predictor, so S_{comp} measures prediction of variation across examples rather than of average activations. The readout solves a mean-error ridge problem,

$$B = \underset{B' \in \mathbb{R}^{m \times D}}{\arg\min} \left\{ \frac{1}{N} \|\bar{Z}_\theta - F B'\|_F^2 + \gamma \|B'\|_F^2 \right\} = (C_F + \gamma I_m)^{-1} C_{FZ}, \qquad \gamma > 0. \tag{9}$$

Relative to the readout W_λ of equation 3, B uses centering without standardization, $u_Z = 1$, and $\gamma = \lambda / N$. The observer remains fixed, while the readout is recomputed from the current encoder outputs.

4.2 S_{comp}: PREDICTIVE COVARIANCE GAIN

A spectrally rich readout need not explain a substantial amount of representation variation. S_{comp} addresses this distinction by comparing the target covariance with the regularized covariance remaining after prediction (Figure 3).

Define $E = \bar{Z}_\theta - FB$ and

$$R = \frac{1}{N} E^\top E + \gamma B^\top B = C_Z - C_{FZ}^\top (C_F + \gamma I_m)^{-1} C_{FZ}, \tag{10}$$

where R is the unexplained variance. Further, the objective is

$$S_{\text{comp}}(Z_\theta \mid X) = \frac{1}{2} \log_2 \frac{\det(C_Z + \epsilon I_D)}{\det(R + \epsilon I_D)}, \qquad \epsilon > 0. \tag{11}$$

Here, ϵ provides a positive covariance floor. The matrix R includes the ridge penalty and is therefore *not simply the covariance of prediction errors*. The score measures reduction in regularized covariance volume relative to the constant-mean predictor. As $\gamma, \epsilon \to 0$, S_{comp} approaches

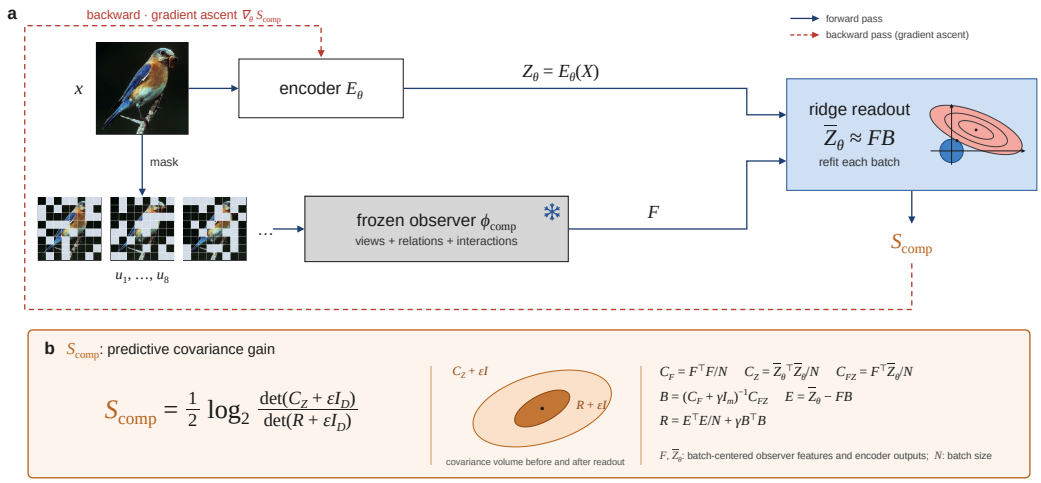


Figure 3: The compositional objective. (a) The encoder E_θ sees the whole image, the frozen observer ϕ_{comp} builds features F from masked views, and a ridge readout B, refitted each batch, predicts the centered encoder outputs; the encoder ascends S_{comp}. (b) S_{comp} compares the target covariance $C_Z + \epsilon I_D$ with the regularized remaining cost $R + \epsilon I_D$.

$\frac{1}{2} \log_2 \det C_Z / \det(C_Z - C_{FZ}^\top C_F^{-1} C_{FZ})$, the mutual information in bits between the observer features and the representation when they are jointly Gaussian, with nonsingular marginal covariances and positive-definite conditional covariance (Bach & Jordan, 2002). The finite-regularization score S_{comp} is a regularized covariance-volume measure.

For positive γ and ϵ,

$$S_{\text{comp}} \geq 0, \qquad S_{\text{comp}} = 0 \iff C_{FZ} = 0. \tag{12}$$

Proof is in Appendix C.5. Thus, the score vanishes when the centered observer features have no linear cross-covariance with the representation.

Unlike S^ϕ, which *scores the spectral content of the readout*, S_{comp} *scores the reduction in regularized target covariance the readout achieves*, so the two need not favor the same representations. The encoder minimizes

$$\mathcal{L}(\theta) = -S_{\text{comp}}(Z_\theta \mid X), \tag{13}$$

and compositionality enters only through the observer's features.

5 Experiments

We test the three contributions on full ImageNet. Section 5.2 tests the spectral preferences of Section 3 and *Appendix C*, Section 5.3 tests whether performance follows the observer's input, and Section 5.4 tests whether spread, predictability, and utility diverge.

5.1 Setup

Pretraining. We pretrain on the full ImageNet-1K (Russakovsky et al., 2015) training set (1,281,167 images) with a ViT encoder (Dosovitskiy et al., 2021) whose 192-dimensional output is unit-normalized. Each method is trained for ten epochs (33,360 updates at batch size 384) with two pretraining seeds, without class labels. All results evaluate the final checkpoint with the encoder frozen. Training details and hardware are in *Appendix D*.

Objectives and notation. S^ϕ is the original spectral objective of Section 2.2, with a whole-image observer and its reference preprocessing conventions. $S^{\phi_{\text{comp}}}$ applies the same spectral score to an observer that sees only the eight masked views of Section 4.1, without relation or interaction tokens. S_{comp} is the objective of Section 4.2, using the views, relations, and interactions of Section 4.1. In all three methods, the trainable encoder receives the whole image.

Evaluation. We report ImageNet linear-probe top-1 and top-5 accuracy and 5-nearest-neighbor accuracy, a separate low-label ImageNet evaluation with 20 labels per class, and transfer to CIFAR-100 (Krizhevsky, 2009), DTD (Cimpoi et al., 2014), and EuroSAT (Helber et al., 2019). For relational evaluation, we report SpatialSense (Yang et al., 2019) and a relational-question subset of CLEVR (Johnson et al., 2017), using readouts on encoder features and metadata.

5.2 The spectral objective spreads the representation

Proposition 3.1 predicts that the spectral score rewards a flatter spectrum. With the masked-view observer, $S^{\phi_{\text{comp}}}$ produces a nearly isotropic representation, with effective rank 175.23 of 192 and isotropy 0.869 (Table 2, Figure 4). S_{comp}, whose score is convex in its improvement eigenvalues (*Appendix C.5*), instead concentrates its spectrum, with effective rank 27.07 and isotropy 0.112. Each geometry follows the preference of its score. With the whole-image observer, the same spectral score, S^{ϕ}, reaches isotropy 0.194: the spectrum the score can attain depends on what the observer sees. The most spread representation is not the most useful on every task: it trails S_{comp} on DTD and S^{ϕ} on both relational evaluations (Table 1), as Proposition 3.2 allows.

5.3 Observer input drives performance

Changing what the spectral observer sees, from the whole image (S^{ϕ}) to masked views ($S^{\phi_{\text{comp}}}$), produces the largest gains in Table 1 (Figure 5). ImageNet top-1 accuracy rises from 3.19% to 7.02%, top-5 from 9.57% to 17.21%, low-label accuracy from 1.94% to 2.23%, CIFAR-100 from 8.32% to 9.68%, and EuroSAT from 63.50% to 68.02%; each difference is positive in both seeds. DTD is unchanged (11.22% and 11.20%). The architectural change in S_{comp}, which adds relation and interaction tokens and replaces the spectral score, does not add to these gains: S_{comp} is below $S^{\phi_{\text{comp}}}$ on every classification evaluation except DTD, where it is highest (11.33%). On SpatialSense and the CLEVR relational subset, the three methods are within 0.6 points of one another, and their ordering changes across seeds.

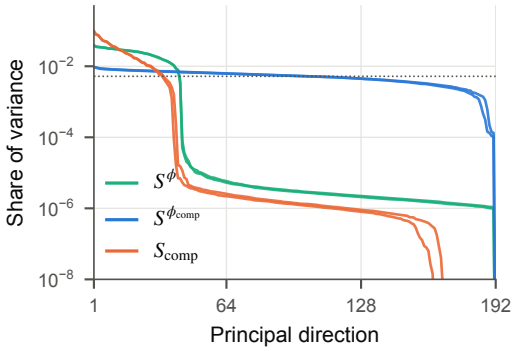


Figure 4: Variance share per principal direction of the low-label ImageNet test features, one line per seed. Dotted: isotropic share $1/192$.

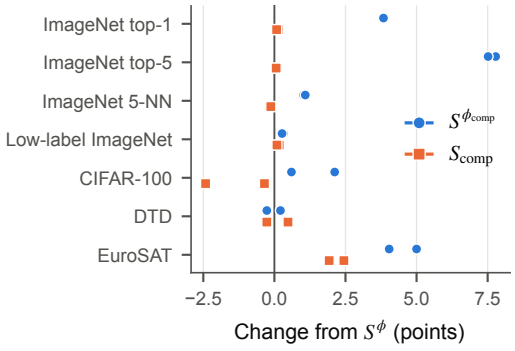


Figure 5: Accuracy change from S^{ϕ} of the same seed on each evaluation, one mark per seed.

5.4 Spread, predictability, and utility diverge

Held-out predictability gives a third ordering. Under a common frozen observer, a ridge readout reduces held-out prediction error by 83.51% for S_{comp}, compared with 18.56% for $S^{\phi_{\text{comp}}}$ (Table 2; Figure 6), yet S_{comp} is less accurate on most evaluations. $S^{\phi_{\text{comp}}}$ is the most spread and the most accurate on six of the seven classification evaluations, but it is the less predictable of the two; S_{comp} is the most predictable and the least spread; and S^{ϕ} is best on the two relational readouts. Held-out error reduction is an evaluation metric, not the training score. Spread, predictability, and utility therefore order the methods differently.

Table 1: Frozen-encoder accuracy (%) after ten epochs on ImageNet-1K, averaged over two seeds. Bold marks the best of the three methods; per-seed values are in *Appendix D*.

Evaluation	S^{ϕ}	$S^{\phi_{\text{comp}}}$	S_{comp}
ImageNet top-1	3.19	**7.02**	3.30
ImageNet top-5	9.57	**17.21**	9.64
ImageNet 5-NN	1.62	**2.67**	1.48
Low-label ImageNet	1.94	**2.23**	2.07
CIFAR-100	8.32	**9.68**	6.94
DTD	11.22	11.20	**11.33**
EuroSAT	63.50	**68.02**	65.69
SpatialSense	**52.84**	52.75	52.32
CLEVR relational subset	**44.49**	44.44	44.27

Table 2: Representation diagnostics and counted training compute, averaged over two seeds. Geometry uses the low-label ImageNet test features; held-out reduction uses the common compositional observer and is unavailable for S^{ϕ}.

Measurement	S^{ϕ}	$S^{\phi_{\text{comp}}}$	S_{comp}
Effective rank (of 192)	39.27	175.23	27.07
Isotropy	0.194	0.869	0.112
Held-out error reduction (%)	–	18.56	83.51
Counted TFLOPs per update	3.826	5.055	5.810
Counted PFLOPs per run	127.62	168.64	193.82

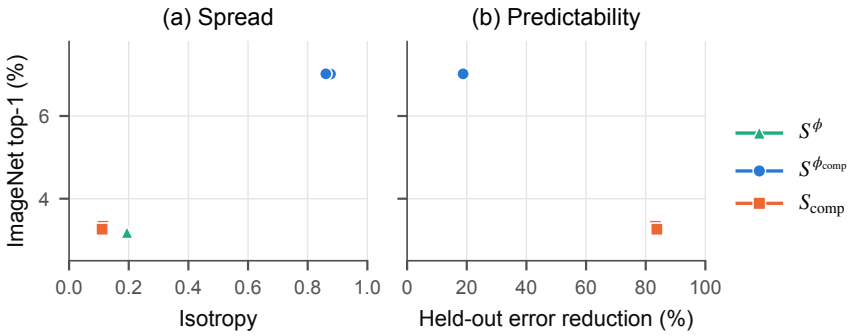


Figure 6: ImageNet top-1 accuracy against (a) isotropy and (b) held-out error reduction, one mark per method and seed.

5.5 Limitations

Two seeds permit replication and paired comparisons, but not strong significance claims. The change from S^{ϕ} to $S^{\phi_{\text{comp}}}$ also changes image preprocessing and score calibration (*Appendix D*), so its gains cannot be attributed to the observer's input alone, and the comparison between $S^{\phi_{\text{comp}}}$ and S_{comp} does not separate the relation and interaction tokens from the score. The relational readouts combine encoder features with metadata and do not isolate the encoder. Counted training compute differs across methods (Table 2).

6 Related Work

Our work builds on the minimum description length principle (Rissanen, 1978; Grünwald, 2007) and computationally bounded notions of information (Xu et al., 2020; Finzi et al., 2026), and on Zhang & Levin (2026), whose learnable-novelty estimator resolves the failure modes of curiosity and surprise-minimization objectives (Schmidhuber, 2010; Burda et al., 2019; Friston, 2010). These frameworks measure how much structure a bounded observer can extract, but not which structure a

representation should retain. Spectral and redundancy-reduction objectives (Yu et al., 2020; Zbontar et al., 2021; Bardes et al., 2022) and spectral proxies for representation quality (Garrido et al., 2023; Agrawal et al., 2022) reward spread; our counterexample shows that spread alone cannot determine utility, paralleling the finding that maximizing mutual information does not by itself yield useful representations (Tschannen et al., 2020). Dimensional collapse can also occur in contrastive representation learning (Jing et al., 2022). Masked and context-prediction methods (He et al., 2022; Assran et al., 2023) learn to predict from visible parts, and part–whole architectures build composition into the encoder (Sabour et al., 2017; Hinton, 2023; Locatello et al., 2020); our observer is instead fixed and random, in the tradition of reservoir computing (Jaeger & Haas, 2004; Maass et al., 2002). Fixed random features followed by a fitted linear model also appear in random-feature methods (Rahimi & Recht, 2007). Our main contribution is to analyze what spectral epiplexity objectives reward: we show that spectral spread alone cannot determine utility, and we find empirically that spread, predictability, and utility can diverge.

7 Conclusion

We started with a simple question: when the spectral approximation to epiplexity is used to train a representation, what is it actually rewarding? In theory, the answer is spread. At fixed readout energy, the score prefers to distribute spectral mass evenly across as many directions as it can, yet two representations with identical spectra can be useful in entirely different ways. Spread alone does not tell us what a representation is good for.

Our experiments on full ImageNet-1K bear this out. The biggest lever we found was not a new architecture but what the observer gets to see: switching the spectral observer from the whole image to masked views more than doubled linear-probe top-1 accuracy. Adding relation and interaction tokens and switching to our predictive score, S_{comp}, did not add to that gain, even though it made the representation far more predictable. Spread, predictability, and utility turned out to be three different properties, and they rank the same methods in three different orders.

To return to the tree from the introduction: recognizing the leaves is not the same as knowing how they connect, and an objective that only measures how much structure is present cannot tell the two apart. Building objectives that reward the connections themselves, and testing them across more seeds, at larger scale, and on tasks that directly probe compositional generalization, is where we want to go next.

AI use statement

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, and implement methods. We have not used generative AI tools to generate synthetic datasets, clean or reformat datasets, support qualitative and thematic data analysis, or interpret results; these uses are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, create artifacts, draft parts of a research paper, and summarize or analyze existing literature. We have reviewed all AI-assisted work. We have manually verified the proofs, and we have thoroughly vetted the code, results, and claims made in the paper. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

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

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding

predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002. URL https://jmlr.org/papers/v3/bach02a.html.

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

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://stanford.edu/~boyd/cvxbook/.

Yuri Burda, Harrison Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations*, 2019.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://www.robots.ox.ac.uk/~vgg/publications/2014/Cimpoi14/cimpoi14.pdf.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Marc Finzi, Shikai Qiu, Yiding Jiang, Pavel Izmailov, J. Zico Kolter, and Andrew Gordon Wilson. From entropy to epiplexity: Rethinking information for computationally bounded intelligence. *arXiv preprint arXiv:2601.03220*, 2026. URL https://arxiv.org/abs/2601.03220.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010.

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

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. URL https://arxiv.org/abs/1709.00029.

Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *Neural Computation*, 35(3):413–452, 2023.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL https://doi.org/10.1080/00401706.1970.10488634.

Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

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

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1612.06890.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. URL https://arxiv.org/abs/1409.0575.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, 2017.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html.

Andrey N. Tikhonov and Vasiliy Y. Arsenin. *Solutions of Ill-Posed Problems*. V. H. Winston & Sons, Washington, DC, 1977. ISBN 0470991240. URL https://books.google.com/books?id=ECrvAAAAMAAJ.

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

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. SpatialSense: An adversarially crowdsourced benchmark for spatial relation recognition. In *IEEE/CVF International Conference on Computer Vision*, 2019. URL https://gfx.cs.princeton.edu/pubs/Yang_2019_ACB/index.php.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems*, 2020.

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

Yanbo Zhang and Michael Levin. Intelligence from learnable novelty. *arXiv preprint arXiv:2607.18433*, 2026. URL https://arxiv.org/abs/2607.18433.