A UNIFYING FRAMEWORK FOR CONCEPT-BASED REPRESENTATIONAL SIMILARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

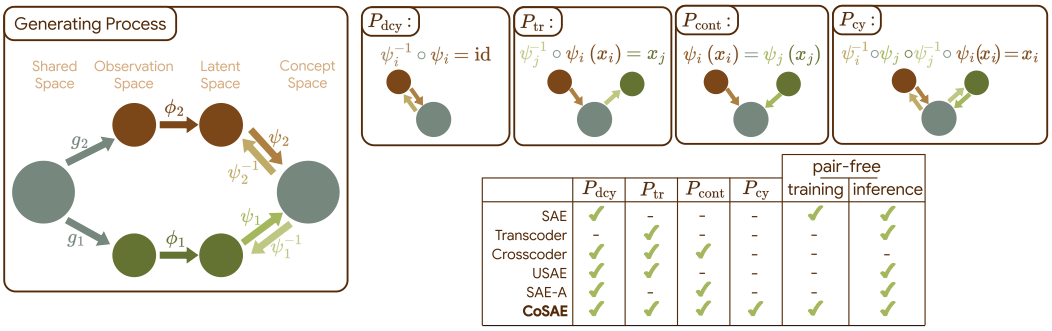
Learned representations across models and modalities often exhibit striking structural similarities, suggesting shared underlying concept decompositions. However, concept alignment remains poorly defined: existing approaches optimize different objectives under the same terminology, obscuring what is actually aligned. We propose a unifying framework that decomposes alignment along two axes: *what* is aligned (representations vs. concepts) and *at what level* (instance-wise vs. distributional). This induces four corresponding properties—instance-wise and distributional variants of translation and concept consistency—and reveals precisely which of these guarantees existing methods provide. We further introduce *InterVenchA*, an intervention-based benchmark that separately measures extraction quality, translation quality, and concept consistency—which we validate against human perception of concept alignment. Using this framework, we show that commonly assumed equivalences between alignment objectives fail in practice: optimizing one property does not reliably recover the others, and, on real embeddings, purely unsupervised objectives fail to recover meaningful instance-level alignment. As a constructive consequence, a model, the Coupled SAE (CoSAE), matches or exceeds specialized alignment methods such as crosscoders and Universal SAEs on their own target metrics, by jointly enforcing the properties identified by our framework without any architectural innovation. Moreover, as little as 1% paired data is sufficient to recover instance-level alignment when anchoring distributional objectives. Overall, concept alignment is fundamentally multi-objective, and progress requires treating it as such. Our framework and benchmark provide the means to define, measure, and compare concept alignment methods on this basis.

1 INTRODUCTION

A recurring finding in modern deep learning is that internal representations learned by different models are often far less arbitrary than their parameterizations suggest. Across architectures, random initializations, and training objectives, independently trained networks frequently exhibit substantial geometric similarity [1; 2]. This phenomenon extends across modalities: models such as CLIP [3], LiT [4], and SigLIP [5] learn representations that align modalities strongly enough to support zero-shot transfer. Together, these results suggest that learned representations share non-trivial structure across systems, including across modalities.

If learned representations do share non-trivial structure, the key question is how to characterize it. Representational similarity metrics (e.g., CCA [6], CKA [1], and variants [7; 8; 9]) provide a coarse geometric view of representational overlap. They do not, however, identify which features are shared across systems, nor whether those features admit a common interpretable decomposition. Sparse autoencoders (SAEs) [10; 11; 12] provide one such decomposition by expressing neural activations in terms of sparse, interpretable features—*concepts*—and recent work has begun extending this approach across models and modalities [13; 14; 15]. This line of work naturally raises the question of *concept alignment*, which we use informally to mean that different systems capture the same underlying structure in their representations. In practice, however, this notion splits along two distinct axes. The first asks whether two systems' latent spaces share aligned geometries, such that one's representations can be mapped to the other's through their respective *concept* dictionaries. We refer to this as *translation*. The second asks for something stronger: that the underlying concepts themselves match one-to-one across systems. We refer to this as *concept consistency*. Existing methods, however,

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.



	P_{dcy}	P_{tr}	P_{cont}	P_{cy}	pair-free training	pair-free inference
SAE	✓	-	-	-	✓	✓
Transcoder	-	✓	-	-	-	✓
Crosscoder	✓	✓	✓	-	-	-
USAE	✓	✓	-	-	-	✓
SAE-A	✓	-	✓	-	-	✓
CoSAE	✓	✓	✓	✓	✓	✓

Figure 1: Summary of the proposed framework. **Left:** Illustration of a generative process where observations stem from a shared space through modality-specific generators g_i. Feature extractors (ϕ_i) learn to invert this generative process up to some transform ψ_i. **Top right:** Different properties of ψ_i that achieve different types of alignment. **Bottom right:** Summary of previous SAE-based methods for concept extraction, along with which property they are trained to satisfy. *Pair-free* indicates whether instance-wise pairs of inputs are required for the alignment during training and inference of each SAE.

rarely make clear which of these two properties they target: different methods optimize different objectives under the same terminology (see [16] for a survey), leaving it unclear which properties are enforced, and when apparent alignment reflects genuinely shared structure.

Any alignment objective must specify both *what* is being aligned—representations (through translation) or concepts—and *at what level* alignment is required—for individual instances or in distribution. In this paper, we argue that these distinctions define a framework for concept alignment. Existing methods instantiate different parts of this space. Our framework makes these choices explicit, providing a rigorous language for comparing existing methods—from standard SAEs [10] and transcoders [17] to crosscoders [13], Universal SAEs (USAEs) [14], and SAE-Aligned (SAE-A) [15] (Figure 1).

Building on this framework, we make the following contributions:

- **A unifying framework for concept alignment.** We formalize concept alignment with four distinct properties—translation, concept consistency, instance-wise and distributional alignment—and use this decomposition to contextualize existing methods in a common design space.
- **A diagnostic benchmark for concept alignment.** We introduce *InterVenchA*, which naturally extends SAEBench [18] evaluation to independently measure extraction quality, translation quality, and concept consistency.
- **Human validation of the benchmark.** We show through a controlled human study that our alignment score agrees with human perception of concept alignment, confirming that it captures a meaningful notion of alignment rather than an arbitrary proxy.
- **An empirical audit of alignment objectives.** On both synthetic and real embeddings, we show that several commonly assumed relationships between alignment properties do not hold in practice: translation does not reliably yield concept consistency (and vice versa), and, on real embeddings, distributional matching does not guarantee instance-level alignment.
- **Correct diagnosis beats architectural novelty.** We introduce the Coupled SAE (CoSAE), which combines only well-known regularization terms yet matches or exceeds specialized alignment methods such as crosscoders [13] and Universal SAEs (USAEs) [14] on their own target metrics despite using scarce supervision, requiring as little as 1% of the data to be paired.

Taken together, these findings point to a simple picture: concept alignment is a multi-dimensional problem; common surrogate objectives only capture part of it; and the strongest practical performance arises from combining complementary regularizers with minimal anchoring supervision. Beyond the specific SAE instantiation studied here, we hope this framework provides a clearer basis for defining, evaluating, and optimizing concept alignment.

2 Related Work

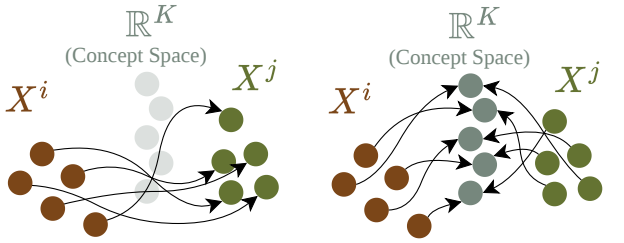
Representation Alignment. A large body of work has studied representation similarity, aiming to characterize whether two neural representations encode similar information. Early approaches rely on statistical or geometric comparisons between representation spaces, such as Canonical Correlation Analysis (CCA) [6; 8], Centered Kernel Alignment (CKA) [1], and related methods [19; 7; 9]. These methods provide a coarse-grained view of similarity by identifying shared linear subspaces or measuring invariance under orthogonal transformations. While these approaches have been instrumental in establishing that independently trained models often converge to similar representations [9; 20; 21], they remain limited in their interpretability [9; 22]. In particular, they characterize the degree to which representations are similar, but not *how* or *why*. These limitations have motivated a shift from coarse, single-score metrics toward finer-grained analyses that operate at the level of individual features or concepts [13].

Interpretability: Towards Concept-based Analysis. The interpretability field aims at understanding the internal workings of neural networks. Many tools have been developed to dissect learned representations [23; 24; 25]. The overall effort is currently shifting from feature visualization [24; 26] and attribution methods [27; 28; 29] to concept-based explanations [30; 31; 32; 12]. This shift exemplifies the trend of moving from input-based explanations to internal explanations. Concept superposition and extraction originated in sparse coding [33; 34; 35; 36; 37], and are grounded in compressed sensing and random projection theory [38; 39]. It is therefore typically framed as a dictionary learning problem [40; 41; 42; 43]: Sparse autoencoders (SAEs) have now become the default approach for concept extraction [44; 45; 46; 47; 48] following their empirical success in interpretability for both vision [49; 47; 50] and language modeling [10; 11; 51].

Concept-based Representation Alignment. Recent work moved from global geometric similarity to concept-level similarity. Instead of comparing representations as monolithic objects, these approaches aim to decompose them into interpretable features and study how these features align across models or modalities. SAEs have emerged as a central tool in this setting, enabling the extraction of structured, often interpretable feature dictionaries. Crosscoders [13] enforce a shared coding space across representations through architectural constraints, while Universal SAEs (USAEs) [14] (resp. Aligned SAEs [15]) rely on regularization through translation (resp. concept consistency) objectives to align feature spaces across models. Despite sharing a common goal, these methods differ significantly in the properties they enforce and in the assumptions they implicitly rely on. More details on how these methods fit into our framework are provided in Appendix A.

3 Framework Formulation

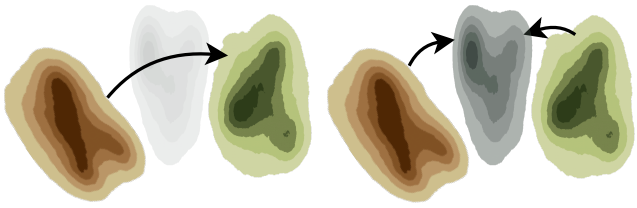
Concept Alignment. Let I be an index set whose elements correspond to representation instances. Each $i \in I$ may, for example, index the same modality embedded in different models, different layers of a single model, checkpoints of a single model at different training steps, modality-specific instantiations of an abstract idea, etc.



(a) P_{tr} : $(\psi_j^{-1} \circ \psi_i)(\boldsymbol{x}_i) = \boldsymbol{x}_j$ (b) P_{cont} : $\psi_i(\boldsymbol{x}_i) = \psi_j(\boldsymbol{x}_j)$

Figure 2: Element-wise alignment.

Let X be a set of data points equipped with a measure μ, and $\phi_i : X \to \mathbb{R}^d$ be a representation function for each $i \in I$. Denote by $X^i := \phi_i(X)$ the set of representations of X under ϕ_i, and by $\mu_i := \phi_i \# \mu$ the pushforward measure on X^i induced by ϕ_i. Let $\psi_i : X^i \to \mathbb{R}^K$ be concept extraction functions with corresponding decoders $\psi_i^{-1} : \mathbb{R}^K \to X^i$ abusively denoted with the inverse notation for readability. Under this setting, and with $\boldsymbol{x}_i = \phi_i(x)$ and $\boldsymbol{x}_j = \phi_j(x)$, we can define the following alignment properties of *translation* and *concept consistency* at the instance level (Figure 2).



(a) $P_{\text{tr}}^{\text{dist}} : (\psi_j^{-1} \circ \psi_i)\#\mu_i = \mu_j$ (b) $P_{\text{cont}}^{\text{dist}} : \psi_i\#\mu_i = \psi_j\#\mu_j$

Figure 3: Distributional alignment.

If the X^i are such that instance-wise correspondences are not available, i.e., we do not have access to the underlying X, we can translate these into distributional alignment properties as follows. Let $(\psi_j^{-1} \circ \psi_i)\#\mu_i$ define a measure on X^j, and $\psi_i\#\mu_i$ define a measure on the concept space $\mathbb{R}^K$. We define the distributional alignment properties corresponding to translation and concept consistency in Figure 3.

In addition to these alignment properties, we can also define standard autoencoding properties, effectively stating that ψ_i^{-1} is indeed the inverse of ψ_i on the support of μ_i. The *demi-cycle* property P_{dcy} is the standard reconstruction objective of autoencoders, while the full *cycle* P_{cy} is an alternative way to couple autoencoders without requiring instance-wise correspondences through reconstruction only. These properties are defined in Figure 4.



(a) $P_{\text{dcy}} : (\psi_i^{-1} \circ \psi_i)(\boldsymbol{x}_i) = \boldsymbol{x}_i$ (b) $P_{\text{cy}} : (\psi_i^{-1} \circ \psi_j \circ \psi_j^{-1} \circ \psi_i)(\boldsymbol{x}_i) = \boldsymbol{x}_i$

Figure 4: Self reconstruction: demi and full cycle.

3.1 Alignment Dualities

Proposition 1. *Suppose injectivity of the decoders, and suppose* P_{dcy}. *Then* $P_{\text{tr}} \Leftrightarrow P_{\text{cont}}$

Proof. See Appendix J or [52]. □

Note that in the framework of SAEs, the injectivity assumption is directly violated by the overcompleteness of the dictionary and the linearity of the decoder. However, the sparsity assumption goes in the other direction, as injectivity is restricted on the sparse support of the concept space. For this reason, we empirically check whether this equivalence holds (Section 4.2). It does not. A second asymmetry arises between instance-wise and distributional alignment:

Proposition 2. $P_{\text{tr}} \Rightarrow P_{\text{tr}}^{\text{dist}} \quad \wedge \quad P_{\text{cont}} \Rightarrow P_{\text{cont}}^{\text{dist}}$

Proof. See Appendix J. □

In other words, the instance-wise properties imply their distributional counterparts (Proposition 2), but the converse generally fails without strong assumptions on the distributions and function class. In Section 4.3, we show that a few anchor pairs can partially restore this bridge in practice.

Finally, while cycle consistency has been used as a surrogate in pair-free translation settings [53; 54], we show in Section 4.2 that it is not a reliable proxy for alignment in our setting. Recent works on joint concept recovery focus only on subsets of these properties, with little to no discussion of the others and their interactions [13; 14; 15]. This motivates the need for a unifying framework to understand the properties and assumptions underlying the different methods used in the literature, and to design new methods that better enforce these properties for future concept alignment studies.

Additional assumptions. In the SAE framework, this work relies on several other assumptions described here and expanded in Appendix F. First, we assume that representations follow the linear representation hypothesis, and that latent space is the projection of a compressed, sparse and higher dimensional space called the concept space. These are the assumptions underlying compressed sensing, and are standard in the literature. Then, we assume a generative form of the platonic representation hypothesis (PRH), stating that independently trained models, across different domains, learn to recover the same concept space.

A complementary controlled experiment with known shared and private factors (Appendix I) confirms this picture: instance-wise translation regularization recovers shared structure without erasing private, domain-specific factors.

3.2 Operationalizing the framework

Regularization. Each of the properties defined above can be operationalized as regularization terms in the training loss:

$$\mathcal{L}_{\text{SAE}} = \alpha_{\text{dcy}}\mathcal{L}_{\text{dcy}} + \alpha_{\text{cy}}\mathcal{L}_{\text{cy}} + \alpha_{\text{tr}}\mathcal{L}_{\text{tr}} + \alpha_{\text{tr}}^{\text{dist}}\mathcal{L}_{\text{tr}}^{\text{dist}} + \alpha_{\text{cont}}\mathcal{L}_{\text{cont}} + \alpha_{\text{cont}}^{\text{dist}}\mathcal{L}_{\text{cont}}^{\text{dist}} \tag{1}$$

Linear Models. In the linear regime, where both the encoders and decoders are linear maps, the different objectives and their combinations collapse to a few analytical solutions. **(A)** Without coupling, i.e., with only reconstruction, we recover PCA solutions for each domain independently (Lem. 1). **(B)** Introducing coupling through concept consistency (Lems. 3 and 4) or translation (Lems. 5 and 6) both lead to the same solution of the CCA for the encoders, while decoders are some variation of optimal linear readouts of the CCA projections. This perfectly illustrates the duality between P_{tr} and P_{cont} in the linear case. **(C)** Distributional alignment fails to introduce any meaningful coupling, and only enforces covariance matching (Appendix E.2.3). **(D)** Finally, although cycle consistency appears to introduce coupling, it actually fails to do so in the linear case, and only enforces PCA solutions for each domain independently (Appendix E.2.4).

SAE. As discussed in Section 2, over the recent years, the SAE framework for sparse coding has become the default concept extraction in the community. We adopt the *batchtopk* architecture throughout this work, with details in Appendix H.2.

Reconstruction. As is standard practice [44], the demi-cycle consistency property P_{dcy} can simply be operationalized as an ℓ_2 reconstruction loss (Equation (2)). The translation P_{tr} and cycle properties P_{cy} can also be operationalized as an ℓ_2 loss on the translated representations (Equations (3) and (4)).

$$\mathcal{L}_{\text{dcy}} := \mathbb{E}_{\boldsymbol{x}_i\sim\mu_i}\|(\psi_i^{-1}\circ\psi_i)(\boldsymbol{x}_i) - \boldsymbol{x}_i\|^2 \tag{2}$$

$$\mathcal{L}_{\text{tr}} := \mathbb{E}_{\boldsymbol{x}\sim\mu}\|(\psi_j^{-1}\circ\psi_i)(\boldsymbol{x}_i) - \boldsymbol{x}_j\|^2 \tag{3}$$

$$\mathcal{L}_{\text{cy}} := \mathbb{E}_{\boldsymbol{x}_i\sim\mu_i}\|(\psi_i^{-1}\circ\psi_j\circ\psi_j^{-1}\circ\psi_i)(\boldsymbol{x}_i) - \boldsymbol{x}_i\|^2 \tag{4}$$

Concept Consistency. Crosscoders [13] enforce P_{cont} not through regularization of disjoint encoders, but through summation of all encoders' outputs to create a common code: $\psi_i'(\boldsymbol{x}_i) = \sum_{j\in I}\psi_j(\boldsymbol{x}_j)$. This is a very strong constraint, since interpretability is imposed only on the aggregated code rather than on the individual encoder outputs. The duality between P_{tr} and P_{cont} is here afforded for free by the fact that the code of $\boldsymbol{x}_i$ already contains all information about $\boldsymbol{x}_j$. We show in Section 4.4 that the standalone ψ_i are not able to perform translation in general, and do not even satisfy P_{cont}. To resolve this tension, we replace the summation in ψ_i' with a regularization term for P_{cont}. We simply use an ℓ_2 loss on the difference between the codes (Equation (5)).

$$\mathcal{L}_{\text{cont}} := \mathbb{E}_{x\sim\mu}\left\|\psi_i(\boldsymbol{x}_i) - \psi_j(\boldsymbol{x}_j)\right\|^2 \tag{5}$$

Distributional alignment. When instance-wise correspondences are unavailable, enforcing $P_{\text{tr}}^{\text{dist}}$ or $P_{\text{cont}}^{\text{dist}}$ requires matching distributions rather than individual samples. A natural approach is to minimize a discrepancy between pushforward measures, for instance, using optimal transport. However, distributional distances like the Wasserstein distance are both computationally expensive and poorly behaved in high dimensions and mini-batch settings [55; 56], making them ill-suited for large-scale SAE training.

A practical alternative is provided by the Cramer-Wold theorem, which states that two distributions are equal if and only if their one-dimensional projections along all directions are equal. We take inspiration from Balestriero and LeCun [57] and use empirical characteristic function (ECF) based tests for our distributional losses, as defined in Appendix G. These losses are equivalent to sliced Maximum Mean Discrepancies (MMDs):

$$\mathcal{L}_{\text{tr}}^{\text{dist}} := \text{MMD}\left((\psi_j^{-1}\circ\psi_i)\#\mu_i, \mu_j\right) \quad \mathcal{L}_{\text{cont}}^{\text{dist}} = \text{MMD}\left(\psi_i\#\mu_i, \psi_j\#\mu_j\right) \tag{6}$$

4 Framework Validation

4.1 Experimental Protocol

4.1.1 Metrics

***InterVench* for Concept Extraction.** For SAE quality in real embeddings, we do not have access to ground truth concepts. Therefore, we rely on standard metrics, derived from SAEBench [18], as a proxy for concept extraction quality. SAEBench is a battery of tests designed to evaluate the quality of concepts extracted by SAEs in the context of language models. We adapt the relevant metrics to our setting and use them to evaluate the quality of the concepts extracted by our CoSAE under different regularization regimes.

Specifically, we focus on reconstruction and intervention-based metrics from SAEBench, which we further extend and factorize into the following classes. First, *InterVench-Latent* metrics evaluate the quality of the dictionary to characterize latent geometry. For example, probes are parameterized as sparse combinations of dictionary atoms. An overall high score on these metrics means that **the dictionary** is both **a good descriptor of latent geometry** and **an actionable tool** for downstream task uses. Second, *InterVench-Concept* metrics evaluate **the quality of the concept space as its own representation space.** We report a single aggregate *extraction score* M_{extract}. More details in Appendix B.

***InterVenchA* for Concept Alignment.** We extend *InterVench* from concept extraction to concept alignment by turning self-reconstruction metrics into transfer metrics. Concretely, we replace self-reconstruction with cross-domain translation, and reinterpret intervention-based metrics as tests of whether concepts learned in one domain transfer meaningfully to another. This yields two complementary views of alignment: *InterVench-Latent* measures translation quality independently of concept consistency (with an aggregate score M_{tr}), while *InterVench-Concept* measures concept consistency independently of translation (through the aggregate score M_{cont}). The mean of M_{tr} and M_{cont} is denoted by the *alignment score* M_{align}. More details can be found in Appendix C.

Human Validation. Because M_{align} is built from proxy metrics rather than ground truth concepts, we validate it directly against human judgment. To measure whether this benchmark tracks a human-perceptible notion of alignment, we ran a human evaluation study (Appendix D) in which annotators are shown a reference set of maximally activating images for a given atom and must pick, among two candidate sets, the one from the corresponding atom in another domain. We compare three cells: a *control* (no alignment regularization), an *aligned bad* cell (atoms with high M_{extract} but low M_{align}), and an *aligned good* cell (high M_{extract} and high M_{align}). A crossed random-effects binomial mixed model shows that participants are significantly more likely to identify the correct match in the aligned good cell than in either the aligned bad or control cells (odds ratios 1.50 and 2.75 respectively, both with 95% credible intervals excluding 1), confirming that M_{align} agrees with human judgment of alignment quality.

4.1.2 Datasets and Models

Synthetic DGPs. We use two complementary synthetic settings. The first follows precisely a generative formulation of the platonic representation hypothesis (Hypotheses 1 and 2) motivating our framework, and evaluates recovery under controlled transformations of sparse latent factors (Hypotheses 1 and 2 and Appendix H.1). The second uses a separate linear generator with fully shared, partially shared, or entirely private factors. It distinguishes within-view factor recovery from cross-view correspondence and evaluates sharedness using ground-truth matching, permutation-calibrated dependence tests, and intervention controls. Its protocol and results are reported in Appendix I.

Vision Encoders. We conduct a battery of cross-model alignment experiments on vision models. We select a ViT (`google/vit-base-patch16-224`) [58], DinoV2 (`facebook/dinov2-base`) [59], and SigLIP (`google/siglip-base-patch16-224`) [5] from Hugging Face [60]. The choice of these models was primarily to ensure a clean comparison with Thasarathan et al. [14]. These models are used to extract embeddings for the ImageNet training and validation splits. SAEs are trained on these embeddings (details in Appendix H.2). For evaluation of both concept extraction and alignment quality, we rely on metrics aggregated in *InterVenchA*.

Regularization	*InterVenchA*— Vision									
	Translation						Concept Consistency			
	R^2	Faith	SP	Unlearn	TPP	M_{tr}	SP	Unlearn	TPP	M_{cont}
(a) $\mathcal{L}_{\text{dcy}}$	0.000	0.686	0.000	0.000	0.000	0.137	0.000	0.000	0.000	0.000
(b) $\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}$	0.466	1.185	0.535	0.880	0.575	0.728	0.190	0.241	0.221	0.217
(c) $\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}}$	0.000	0.781	0.295	0.812	0.733	0.524	0.000	0.000	0.000	0.000
(d) $\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}$	0.000	0.700	0.000	0.000	0.000	0.140	0.460	0.826	0.521	0.602
(e) $\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.000	0.701	0.000	0.000	0.000	0.140	0.000	0.000	0.000	0.000
(f) $\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}} + \mathcal{L}_{\text{cont}}$	0.471	1.182	0.545	0.924	0.544	0.733	0.469	0.877	0.510	0.619
(g) $\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.000	0.714	0.000	0.000	0.000	0.143	0.000	0.000	0.000	0.000
(h) mixed (1 in 10)	0.524	1.151	0.522	0.923	0.523	0.729	0.727	0.935	0.779	0.813
(i) mixed (1 in 100)	0.422	1.093	0.627	0.885	0.632	0.732	0.715	0.931	0.737	0.795
(j) mixed (1 in 1000)	0.410	0.946	0.467	0.602	0.591	0.602	0.353	0.821	0.458	0.544

Table 1: Detailed *InterVenchA* metrics on **vision embeddings** under different regularization regimes. Similar tables for *InterVench* and for the language and multimodal settings can be found in Appendix K.

LLMs and VLMs. We also run similar experiments on tokens extracted from Large Language Models. We select Llamas [61] and Gemmas [62], and extract tokens at the 75-th percent layer as is common practice. For token alignment across different tokenizers, we adopt the same procedure as [63]. Finally, we also conduct cross-modal alignment experiments on Vision Language Models. More details in Appendix H. Only the vision encoder's settings results are reported in the main body of this paper. See Appendix K for the other settings.

4.2 Alignment Duality in Practice

We train a battery of CoSAEs with different regularizations to isolate the contribution of each and measure whether alignment dualities (Section 3.1) hold empirically. Table 1 shows the detailed results for *InterVenchA* in the vision setting. Due to our hyperparameter selection protocol (Appendix H.2), concept extraction quality remains comparably high across the board. Details are reported in Appendix K, along with details for other modalities (LLMs and VLMs).

Translation vs Concept Consistency. We first ask whether regularization for one of translation or concept consistency is sufficient to get the other for free, or whether we need to explicitly regularize both. Neither direction of the duality holds in practice: optimizing translation improves translation, and optimizing concept consistency improves concept consistency. But neither reliably recovers the other (see Table 1, lines *(b)* and *(d)*). The practical implication is clear: both properties must be explicitly regularized.

On the Effect of Cycle Consistency. The second duality we investigate is that of alignment and cycle consistency. As mentioned above, [53; 54] use cycle consistency as a surrogate for translation in NLP, enabling pair-free training. We ask whether cycle consistency is a faithful surrogate for alignment in our framework. Training an additional SAE with regularizations $\mathcal{L}_{\text{dcy}}$ and $\mathcal{L}_{\text{cy}}$ leads to a model with $M_{\text{tr}} = 0.072$ and $M_{\text{cont}} = 0.000$. This model therefore fails to recover concept alignment, suggesting that the assumptions underlying the duality are not satisfied in practice.

We additionally measure the R^2 associated to the cycle consistency property (Appendix B), denoted by R^2_{cy}. All measured models have values between $R^2_{\text{cy}} \in [0.4, 0.6]$, with the exception of *(b)* $\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}$, for which $R^2_{\text{cy}} = 0.121$. The takeaway is blunt: in our setting, cycle consistency is not a faithful surrogate for alignment. It can be high even when alignment is poor—and low despite strong alignment. Therefore, cycle consistency should not be used as a surrogate for alignment in our framework.

Instance-wise vs Distributional Alignment. Finally, we ask whether distributional objectives behave as faithful surrogates for instance-wise alignment. (*i*) In all settings, we find that, unsurprisingly, instance-wise regularization (Table 1 lines *(b)* and *(d)*) leads to both instance-wise and distributional

alignment (Appendix K). (*ii*) However, the converse is false: distributional objectives do not recover instance-wise alignment. This is exactly the failure mode expected from the unidentifiability of the transport problem, and it explains why purely distributional alignment is insufficient in practice. Results in Table 1 lines *(c)* and *(e)*, and in Appendix K. We show in the next section that this does not mean that distributional objectives are entirely useless, particularly in the case of scarce supervision.

4.3 Anchoring Alignment with Scarce Supervision

A common issue with instance-wise regularization in multimodal settings is the scarcity of high-quality matching pairs. Therefore, it is desirable to rely on distributional regularization for the bulk of the training. However, results from the previous experiments suggest that translation and concept consistency regularizations are complementary, and that distributional objectives are not faithful surrogates for instance-wise alignment. This motivates the need for a combination of regularization terms to achieve maximal alignment quality in the wild. Anchoring distributional alignment using a few matched pairs could be enough to resolve the unidentifiability of the transport problem. In this experiment, we study the behavior of such mixed regimes. Specifically, we compute $\mathcal{L}_{\text{tr}}$ and $\mathcal{L}_{\text{cont}}$ only on a small subset of held-out matching pairs—with 1 in 1000 amounting to one per batch on average. The proportion of matched pairs is denoted by p. Refer to Appendix H.3 for full details.

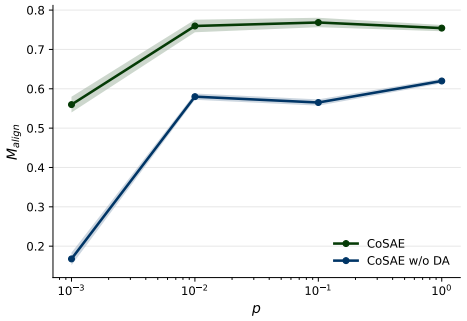


Figure 5: **(i)** High alignment score can be recovered with scarce supervision. **(ii)** CoSAEs trained with distributional alignment (in green) recover alignment scores of full supervision, contrary to those trained without (in blue).

Table 1 lines *(f)* to *(j)* show that a tiny amount of pairing is enough to resolve the ambiguity of purely distributional objectives. With as few as 1 in 100 matched pairs, we recover strong instance-wise alignment—matching or even exceeding full instance-wise supervision—while preserving concept extraction. Note that anchor points alone are sufficient to recover high alignment metrics, with the exact effect of distributional objectives being shown in Figures A3 and 5.

4.4 The Alignment Recipe

Method	Eval	M_{extract} (↑)	M_{tr} (↑)	M_{cont} (↑)
Crosscoder	shared z	0.887	0.596	<u>0.794</u>
Crosscoder	standalone z	0.860	0.319	0.000
USAE	–	**0.895**	<u>0.731</u>	0.435
Ours (CoSAE)	–	<u>0.893</u>	**0.732**	**0.795**

Table 2: Our method (Table 1 line *(i)*) matches or exceeds both crosscoders and USAE.

We conduct a battery of ablation studies to understand the contribution of each regularization term to the overall performance of the CoSAE (Table 1 line *(i)*) on both concept extraction and alignment. We train SAEs with all combinations of regularization terms, and evaluate them on our *InterVenchA* for both concept extraction and alignment quality. See Appendix K for the full table of results. We also include a comparison with the original USAE and crosscoder methods. For the crosscoder, we distinguish two evaluation settings. In the first one, we consider the unmodified crosscoder, i.e., where the codes are the sum of all encoders' standalone codes. In the second evaluation setting, the codes are no longer aggregated, and each encoder is evaluated for its capacity to produce standalone codes.

Our CoSAE matches or exceeds both the USAE and the crosscoder on their respective target metrics (M_{tr} and M_{cont}, respectively). CoSAE achieves the best performance on both, despite crosscoder enforcing *perfect* P_{cont} by design through code aggregation. Moreover, when evaluated with standalone codes, the crosscoder's performance drops sharply, including a collapse in M_{cont}. This indicates that code aggregation can mask weak standalone encoders, and that alignment observed during aggregated evaluation does not necessarily transfer to the individual encoders.

4.5 Cross-Modal Transfer with Scarce Supervision

As a final validation, we evaluate whether the alignment learned by our CoSAE translates into competitive cross-modal transfer. We consider a setting analogous to CLIP training, replacing the vision–vision alignment of Section 4.1.2 with a vision–text setting (DINOv2 and all-roberta-large), while keeping the same training procedure. We extract CLS tokens on the subset of LAION provided by [64]. In this regime, the translation from vision to text representations effectively acts as a learned projector, comparable to the MLP heads used by [64]. We therefore compare these settings.

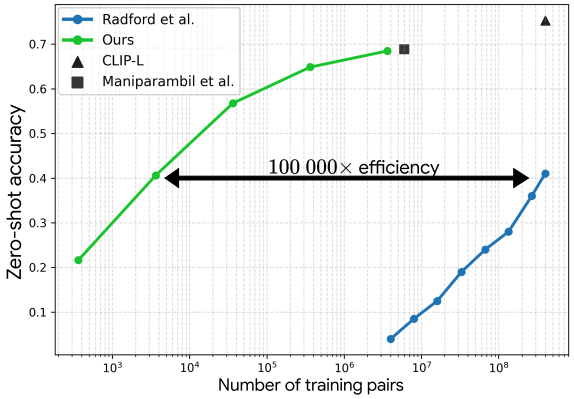


Figure 6: **Concept alignment in our CoSAE is strong enough to support competitive downstream behavior**. ImageNet zero-shot accuracy of our CoSAE trained on unimodal backbones.

We evaluate zero-shot ImageNet accuracy obtained from the aligned representations (see Figure 6). Although CoSAE is not optimized for contrastive cross-modal training, it matches the reported zero-shot accuracy of [64]. This suggests that the alignment learned by CoSAE is strong enough to support competitive downstream transfer, despite relying on sparse autoencoders rather than a dense projector.

5 Discussion

We revisit representational similarity through a concept-based lens and argue that alignment should not be treated as a single objective. Instead, it is a structured combination of properties—translation and concept consistency, each at both instance and distribution levels—and progress depends on keeping these distinctions explicit. Empirically, widely assumed equivalences break down: translation does not imply concept consistency, cycle consistency is not a reliable surrogate, and, on real embeddings, distributional objectives alone fail to recover instance-wise alignment. Alignment is therefore inherently multi-dimensional and cannot be captured by any single objective. These claims are backed by quantitative metrics: *InterVenchA*. We validate them against human judgment, confirming that M_{align} tracks a perceptually meaningful notion of alignment rather than an internal proxy.

CoSAE is the constructive consequence of this view: rather than betting on a single surrogate, it jointly enforces complementary objectives. Strong alignment emerges only in this regime, with distributional regularization becoming effective once anchored by a few paired examples. In this setting, CoSAE matches or exceeds prior SAE-based approaches while requiring only minimal supervision.

Our work also has several limitations. First, the framework focuses on shared structure and does not explicitly model idiosyncratic or domain-specific features beyond self-reconstruction. Second, our evaluation in real embeddings necessarily relies on proxy metrics rather than ground-truth concepts, so semantic alignment and functional transfer may not always coincide. Our human study mitigates this for the alignment score specifically, but extraction quality and the underlying concepts themselves remain unverified against ground truth. This is inherent to unsupervised dictionary learning. Third, the proposed objective combines multiple regularization terms, introducing hyperparameter sensitivity that deserves further study. Finally, our empirical validation, while broad, still covers only a limited set of architectures and datasets. Extending these findings to larger-scale and more heterogeneous settings remains an important direction for future work.

Overall, concept alignment is best understood not as a single objective, but as a coupled family of objectives. Meaningful alignment emerges not from optimizing one surrogate well, but from respecting this structure in both evaluation and method design. We hope this perspective clarifies the assumptions underlying existing methods and lays a foundation for more robust, interpretable alignment techniques in the future.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript for readability, including correcting typos and grammar. All authors reviewed the final text, and we take responsibility for the full content of this work, including any text produced with the aid of generative AI.

REFERENCES

[1] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pages 3519–3529, 2019.

[2] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 225–236, 2021.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763, 2021.

[4] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18123–18133, 2022.

[5] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023.

[6] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. ISSN 00063444. URL http://www.jstor.org/stable/2333955.

[7] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, 2007.

[8] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[9] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 1556–1568, 2021.

[10] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[11] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[12] Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 54805–54818, 2023. doi: 10.52202/075280-2391.

[13] Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 2024.

[14] Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267, pages 59304–59325, 2025.

[15] Grégoire Dhimoïla, Thomas Fel, Victor Boutin, and Agustin Martin Picard. Cross-modal redundancy and the geometry of vision–language embeddings. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

[16] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. doi: 10.1145/3728458.

[17] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 24375–24410, 2024.

[18] Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267, pages 29223–29264, 2025.

[19] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:249, 2008.

[20] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pages 20617–20642, 2024.

[21] A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A Efros. Back into Plato's cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026.

[22] Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

[23] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6541–6549, 2017.

[24] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11), 2017. doi: 10.23915/distill.00007.

[25] Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of the IEEE International Conference on data science and advanced analytics (DSAA)*, pages 80–89, 2018.

[26] Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom Rousseau, Rémi Cadène, Lore Goetschalckx, Laurent Gardes, and Thomas Serre. Unlocking feature visualization for deep network with MAgnitude Constrained Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 37813–37826, 2023. doi: 10.52202/075280-1646.

[27] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833, 2014. doi: 10.1007/978-3-319-10590-1_53.

[28] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, page 151. BMVA Press, 2018.

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3319–3328, 2017.

[30] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[31] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for CNN models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 11682–11690, 2021.

[32] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[33] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[34] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[35] Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, 2004. doi: 10.1016/j.conb.2004.07.007.

[36] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 19, 2006.

[37] Ilias Rentzeperis, Luca Calatroni, Laurent U Perrinet, and Dario Prandi. Beyond ℓ_1 sparse coding in V1. *PLoS Computational Biology*, 19(9):e1011459, 2023.

[38] William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984. doi: 10.1090/conm/026/737400.

[39] Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson–Lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. IEEE, 2017.

[40] Ron Rubinstein, Alfred M Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[41] Michael Elad. *Sparse and redundant representations: From theory to applications in signal and image processing*. Springer New York, 2010. doi: 10.1007/978-1-4419-7011-4.

[42] Ivana Tošić and Pascal Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2): 27–38, 2011.

[43] Bogdan Dumitrescu and Paul Irofti. *Dictionary learning algorithms and applications*. Springer, 2018.

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

[45] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

[46] Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

[47] Thomas Fel, Ekdeep Singh Lubana, Jacob S Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[48] Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

[49] Liv Gorton. The missing curve detectors of InceptionV1: Applying sparse autoencoders to InceptionV1 early vision. *arXiv preprint arXiv:2406.03662*, 2024.

[50] Thomas Fel, Binxu Wang, Michael A Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep S Lubana, Talia Konkle, Demba Ba, et al. Into the rabbit hull: From task-relevant concepts in DINO to Minkowski geometry. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

[51] Viacheslav Surkov, Chris Wendler, Mikhail Terekhov, Justin Deschenaux, Robert West, and Caglar Gulcehre. Unpacking SDXL Turbo: Interpreting text-to-image models with sparse autoencoders. In *Mechanistic Interpretability for Vision at CVPR 2025 (Non-proceedings Track)*, 2025.

[52] Benjamin Devillers, Léopold Maytié, and Rufin VanRullen. Semi-supervised multimodal representation learning through a global workspace. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):7843–7857, 2025.

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

[54] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[55] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015.

[56] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

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

[58] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[59] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

[60] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[62] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[63] Thomas Jiralerspong and Trenton Bricken. Cross-architecture model diffing with crosscoders: Unsupervised discovery of differences between LLMs. *arXiv preprint arXiv:2602.11729*, 2026.

[64] Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, and Noel E O'Connor. Harnessing frozen unimodal encoders for flexible multimodal alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29847–29857, 2025.

[65] Julian Minder, Clément Dumas, Caden Juang, Bilal Chughtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=yFdNygEryH.

[66] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[67] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

[68] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[69] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research (JMLR)*, 13(1): 723–773, 2012.

[70] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 97–105. PMLR, 2015.

[71] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1718–1727. PMLR, 2015.

[72] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[73] Luke Marks, Alasdair Paren, David Krueger, and Fazl Barez. Enhancing neural network interpretability with feature-aligned sparse autoencoders. *arXiv preprint arXiv:2411.01220*, 2024.

[74] Sai Sumedh R Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba E. Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

[75] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.

[76] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[77] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2023.

[78] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[79] T. W. Epps and Lawrence B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. doi: 10.1093/biomet/70.3.723.

[80] Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

[81] Peter H. Westfall and S. Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, New York, 1993.

[82] Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010. doi: 10.2202/1544-6115.1585.

A Extended Related Work

A.1 Unifying existing methods

Figure 1 summarizes the different methods for concept extraction, with which combination of properties they are designed to satisfy, as well as whether they require instance-wise correspondences for training and inference. We discuss these methods in more detail below. Our framework can be used to regularize for any combination of properties.

SAE. The vanilla SAE framework only considers isolated representation instances and therefore only enforces the demi-cycle consistency property P_{dcy} through the reconstruction loss $\mathcal{L}_{\text{dcy}}$. This is the most basic concept extraction method, and does not enforce any particular alignment between the concepts extracted from different representation instances.

Transcoders. [17] only enforce the translation property P_{tr} through the translation loss $\mathcal{L}_{\text{tr}}$. This method was designed to learn a sparse alternative to a dense MLP, not to compare the conceptual alignment between two representations.

Crosscoders. As discussed above, crosscoders [13] were designed for concept alignment. As a result, they enforce P_{cont} architecturally and regularize for P_{dcy} (and equivalently P_{tr}). Given their strong architectural constraint for a perfect satisfaction of P_{cont}, they trivially satisfy the alignment duality of Proposition 1. Indeed, they introduce $\psi'_i = \sum_{j \in I} \psi_j$. Minder et al. [65] finds that in selective encoders such as top-k or batchtopk, enforcing selection on ψ'_i creates a strong pressure for the ψ_i to satisfy P_{cont} in the form of an "L_0 budget". Splitting is therefore strongly disfavored. However, it is not impossible, for example, in the case where feature activation energies span several orders of magnitude and noise on high energy features outweighs the signal on low energy ones. In such a scenario, high-energy features benefit from splitting in order to perfectly reconstruct their instance-specific noises at the expense of some L_0 budget that is therefore not spent on low-energy features. On top of these hypothetical failure modes, we show in Section 4.4 that the architectural trick of introducing ψ'_i renders ψ_i incapable of acting as an isolated encoder, failing to satisfy all properties of reconstruction P_{dcy}, translation P_{tr}, and concept consistency P_{cont}.

USAE. The Universal SAE (USAE) [14] was introduced as an alternative to crosscoders in the context of vision models, with the same motivation of learning a shared dictionary. In their case, they studied cross-model alignment. Their method discarded the architectural trick of crosscoders and replaced it with $\mathcal{L}_{\text{tr}}$ to regularize for translation. However, as shown by Devillers et al. [52], this is not sufficient to guarantee the concept consistency property P_{cont} unless the decoders ψ_i^{-1} are injective. In the case of SAEs, decoders are linear maps from a K-dimensional concept space to a d-dimensional latent space, with $K > d$, and are therefore not injective on the full space. However, what matters is injectivity on the support of the shared concept space $\bigcup_{i \in I} \text{supp}(\psi_i \# \mu_i)$. In practice, due to the sparsity constraints of the encoders coupled with translation constraints, failing to align codes would require the decoders to have split features and each encoder to use an arbitrary subset of these split features. A sufficient condition to fall in such a case is for the SAE to be locked in a local minimum. As an illustrative example, take the case where a shared high-energy feature f was independently discovered by both SAEs at indices k_i and k_j respectively, as part of the reconstruction objective. Then, with high probability, ψ_i at index k_j corresponds to a low-energy feature. In such a case, the translation loss dominates the reconstruction one for this feature, such that ψ_i learns to not encode meaningful information there, and the decoder ψ_i^{-1} learns to copy f at index k_j. This locks the SAEs in a local minimum where, in order to preserve the constraints of sparsity, reconstruction, and translation, this feature needs to stay split. Note that there is no equivalent of the L_0 budget pressure of crosscoders that would prevent such a failure mode, since here, the encoders' sparsity is enforced separately for all $i \in I$. Adding a regularization term for P_{cont} is therefore necessary to guarantee alignment in concept space.

SAE-A. The Aligned SAE (SAE-A) [15] was introduced in the context of multimodal encoders explicitly tasked to learn a shared representation across modalities. As a result, they share the weight of image and text SAEs: $\psi_{\text{img}} = \psi_{\text{txt}}$ and enforce both P_{dcy} for regular concept extraction and P_{cont} for alignment. They formulate their DGP and assumptions under the distributional case of $P_{\text{cont}}^{\text{dist}}$ for generality. However, due to the existence of matching pairs between the two modalities in CLIP-like

encoders [3], they operationalize it as a regularization term on the codes of matching pairs. They further show that a very small weight on this regularization term is sufficient to get the necessary inductive bias and align codes.

A.2 THE COSAE

Given the unifying framework described above, it is now immediately possible to design new methods with arbitrary combinations of regularizations based on which properties are relevant under a specific experimental setting. One can even consider a mixture of instance-wise and distributional regularization terms, e.g., based on the availability and quality of matching pairs of data points. This framework also allows for possible curriculum learning schemes for maximal sample efficiency.

B *InterVench*

The *InterVench* is a battery of tests designed to evaluate the quality of concepts extracted by SAEs. It is based on SAEBench [18], a similar benchmark for evaluating disentanglement and interpretability of learned representations in the context of large language models. We adapt the relevant metrics to our setting and use them to evaluate the quality of the concepts extracted by our CoSAE under different regularization regimes. Specifically, we consider the following metrics, which target *latent geometry* and *concept space* quality, respectively. All of our metrics are intervention-based.

R^2 reconstruction. We measure the R^2 of the reconstruction of the latent representations from the codes, estimated as $R^2 = 1 - \frac{\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|^2}$, where $\boldsymbol{x}$ is the original latent representation, $\widehat{\boldsymbol{x}}$ is the reconstructed representation from the codes, and $\bar{\boldsymbol{x}}$ is the mean of the original representations. This metric's evaluation is in *latent space*.

Faithfulness. Quantifies how much of the information encoded in the latent representations is captured by the codes. Measured as the drop in model's loss when replacing the original latent representations with the reconstructed ones from the codes: $\frac{\widehat{\mathcal{L}}-\mathcal{L}_0}{\mathcal{L}-\mathcal{L}_0}$, where $\mathcal{L}$ is the original model's loss, $\widehat{\mathcal{L}}$ is the loss of the model composed with the SAE, and $\mathcal{L}_0$ is the same as $\widehat{\mathcal{L}}$ but the codes are patched with zeros. This metric's evaluation is in *latent space*.

In the vision setting, we consider the training loss for SigLIP, and the loss of a linear probe on ImageNet for ViT (its original loss), as well as for DinoV2 (as we do not have access to its original loss).

Sparse Probing. Measures the extent to which codes capture downstream task-relevant information in a sparse manner. We split this metric into two, one acting in *latent space* and the other in *concept space*. For the former, we parameterize a standard probe $\boldsymbol{x}^T\boldsymbol{\theta}$ as a sparse linear combination of dictionary atoms $\boldsymbol{\theta} \coloneqq \sum_{i=0}^{K-1} \alpha_i \boldsymbol{d}_i \in \mathbb{R}^d$—$\mathrm{supp}((\alpha_i)) \ll d$—instead of as a dense combination of canonical basis vectors $\boldsymbol{\theta} \coloneqq \sum_{i=0}^{d-1} \alpha_i \boldsymbol{e}_i$. For the latter, the probe becomes $\boldsymbol{z}^T\boldsymbol{\theta}$ instead of $\boldsymbol{x}^T\boldsymbol{\theta}$, where $\boldsymbol{z} = E(\boldsymbol{x})$ is the code of $\boldsymbol{x}$. $\boldsymbol{\theta}$ is now equal to (α_i), still with $\mathrm{supp}((\alpha_i)) \ll d$. In SAEBench, only the latter is considered. The score we report is the raw accuracy of the probe. In the vision setting, we report the sparse probing score on ImageNet.

All subsequent metrics are probe-based and can therefore be adapted in the same way to both latent space and concept space, with the same distinction as above. They all go beyond sparse probing by measuring not only the amount of downstream task-relevant information captured by dictionary atoms, but also how disentangled this information is across atoms. Due to the necessity for the following metrics to indicate that sparse probing is successful, we force them to zero when the sparse probing score is below a certain threshold.

Unlearning. Measures whether the SAE grants the ability to selectively unlearn specific information (on a *forget task*, specific) without affecting other information (measured on a *utility task*, general). We measure both (*i*) $u = \frac{\mathrm{acc}_{\mathrm{unlearned}}^{\mathrm{utility}} - \mathrm{acc}_0^{\mathrm{utility}}}{\mathrm{acc}_{\mathrm{original}}^{\mathrm{utility}} - \mathrm{acc}_0^{\mathrm{utility}}}$ the utility score, where acc_0 denotes the accuracy of a probe trained on representations where codes zs are patched with zeros, $\mathrm{acc}_{\mathrm{original}}$ the accuracy of a probe trained on original representations, and $\mathrm{acc}_{\mathrm{unlearned}}$ the accuracy of a probe trained on

representations where codes labeled as relevant for the forget task are patched with zeros. We also measure (*ii*) $f = 1 - \frac{\text{acc}_{\text{unlearned}}^{\text{forget}} - \text{acc}_0^{\text{forget}}}{\text{acc}_{\text{original}}^{\text{forget}} - \text{acc}_0^{\text{forget}}}$ the forget score. The final unlearning score is then computed as $u \cdot f$.

Other combinations of utility and forget scores are possible, such as $u + f$, where losing one point of utility is worth gaining one point of forget. We do not consider this combination, as it doesn't penalize mediocre tradeoffs where both utility and forget scores are around 0.5. More generally, the score can be the distance to the Pareto optimal $(1, 1)$: $-\|(u, f) - (1, 1)\|_p$ for some $p \in [1, +\infty]$, with easy geometric interpretations.

In the vision setting, we take ImageNet multiclass classification as the utility task, and CelebA [66] multilabel classification as the forget task.

Spurious Correlation Removal (SCR). Measures whether the SAE disentangles spurious correlations in the data. We report $\text{SCR} = \frac{\text{acc}_{\text{decorr}} - \text{acc}_{\text{original}}}{\text{acc}_{\text{oracle}} - \text{acc}_{\text{original}}}$, where $\text{acc}_{\text{original}}$ is the accuracy of a probe trained on the dataset where the spurious correlation is present (dataset A), $\text{acc}_{\text{oracle}}$ the accuracy of a probe trained on a fully de-correlated dataset (dataset B). For $\text{acc}_{\text{decorr}}$, we take the sparse probe trained on dataset A and evaluate the contribution of each atom to the spurious correlation. We zero out the atoms whose contribution is above a certain threshold, and retrain a probe on the resulting support.

In the vision setting, we select the waterbirds dataset [67], where the spurious correlation is that between the background (water vs land) and the class (waterbird vs land-bird). Dataset A is the training set, while dataset B is a balanced test set where the correlation is removed. For some reason, the models we used, i.e., DinoV2, ViT, and SigLIP, do not seem to catch the spurious correlation in the first place, rendering this metric noisy and uninformative. For this reason, we do not use it in our final evaluation. Future work should explore other known or synthetic spurious correlations in vision, such as color MNIST [68].

Targeted Probe Perturbation (TPP). Measures whether the SAE disentangles the features relevant to distinct classes across non-overlapping sets of atoms, in a multiclass classification task. Similar to unlearning, we report $u \cdot f$. In this case, we select c a class of interest, or a set of classes, and $\overline{c}$ its complement. We then consider c as the forget task, and $\overline{c}$ as the utility task.

In the vision setting, we take ImageNet multiclass classification and report the TPP score averaged across subsets c of 10 classes, selected at random.

M_{extract}. As stated before, we remove the SCR metric from our final evaluation in the vision setting. We additionally zero out the unlearning and TPP scores when the corresponding sparse probing score is below a certain threshold, typically around 0.1 in the ImageNet multiclass classification setting. We further split the metrics as follows. *InterVench-Latent* contains metrics that measure the ability of the dictionary atoms to characterize the latent space. These metrics are R^2 reconstruction and faithfulness, as well as the *latent space* version of sparse probing, unlearning, and TPP. *InterVench-Concept* contains metrics that evaluate the quality of the concept space as its own representation space. These metrics are the *concept space* version of sparse probing, unlearning, and TPP. M_{extract} is then computed as the average of all metrics in *InterVench*.

Cycle Consistency. Separately from the aggregate of the above score, we also report the R^2 associated with the cycle consistency. The only difference with the R^2 of the reconstruction is that $\widehat{\boldsymbol{x}}$ is replaced by $\widehat{\boldsymbol{x}} \coloneqq \psi_i^{-1} \circ \psi_j \circ \psi_j^{-1} \circ \psi_i(\boldsymbol{x})$.

C *InterVenchA*

The *InterVenchA* benchmark is a natural extension of the *InterVench* to the evaluation of concept alignment quality. It is based on the same metrics as the *InterVench*, but adapted to evaluate the quality of the alignment between two SAEs trained on different domains. In order to measure this, we do not rely on the ability of SAEs to capture capabilities, but to *transfer* them. Specifically, the adaptation of *InterVench-Latent* metrics measures the quality of *translation* properties, while the adaptation of *InterVench-Concept* metrics measures the quality of *concept consistency* properties.

Reconstruction. The natural extension of R^2 reconstruction and faithfulness metrics for alignment is to consider the following. Let i and j index two domains. Let $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ be joint samples from domains i and j respectively. Let E_i and D_i be the encoder and decoder of the SAE trained on domain i, similarly for E_j and D_j. We now consider the reconstruction $\widehat{\boldsymbol{x}}_j = D_j(E_i(\boldsymbol{x}_i))$ instead of $\widehat{\boldsymbol{x}}_j = D_j(E_j(\boldsymbol{x}_j))$.

Sparse Probing. Given domains i and j, we train a sparse probe exactly as before on domain i. This results in a sparse vector of parameters $\boldsymbol{\alpha} \in \mathbb{R}^K$, where K is the number of dictionary atoms and $\mathrm{supp}(\boldsymbol{\alpha}) \ll d$. In the *latent* setting, this corresponds to $\boldsymbol{\theta}_i \coloneqq \boldsymbol{\alpha}^T \boldsymbol{D}_i$, where D_i is the dictionary of domain i. We therefore evaluate the transfer capabilities of jointly trained SAEs by measuring the resulting accuracy of the probe $\boldsymbol{\theta}_j \coloneqq \boldsymbol{\alpha}^T \boldsymbol{D}_j$ on domain j. This measures alignment as the ability of dictionary atoms to jointly characterize each of their domain's geometry. In the *concept* setting, we have $\boldsymbol{\theta}_i \coloneqq \boldsymbol{\alpha}$, evaluated on domain i through the accuracy of the classifier $E_i(\boldsymbol{x}_i)^T \boldsymbol{\theta}_i$. We then evaluate the accuracy of $\boldsymbol{\theta}_j \coloneqq \boldsymbol{\theta}_i$ on domain j. This measures *concept consistency* as the ability of the same set of concepts to capture the same task-relevant information across domains.

Unlearning, SCR and TPP 's extensions naturally follow from the above. We design interventions on domain i and transfer them to domain j to evaluate the quality of the alignment.

M_{align}. All post-processing of the metrics is the same as for M_{extract}. All metrics now give a matrix of scores. The diagonal, i.e., when $i = j$, corresponds to the original *InterVench* metrics on each domain separately. The off-diagonal elements correspond to the transfer capabilities of the SAEs across domains. M_{extract} is the average across metrics of the average of the diagonal, while M_{align} is averaged across the off-diagonal elements. M_{tr} specifically targets translation properties, averaging the R^2 and faithfulness metrics as well as the *latent space* version of sparse probing, unlearning and TPP. M_{cont} on the other hand specifically targets concept consistency properties, averaging the *concept space* version of sparse probing, unlearning, and TPP.

Note. In the special setting of multimodal encoders, we select the following metrics. The R^2 is unchanged. For the faithfulness, we need to adapt it to the case of two-tower multimodal encoders. We report $\frac{\widehat{\mathcal{L}}_{ij} - \mathcal{L}_0}{\mathcal{L} - \mathcal{L}_0}$, where $\mathcal{L}$ is the original contrastive loss of the two towers, $\widehat{\mathcal{L}}_{ij}$ is the same loss but where the original representations of domain i are translated as $\widehat{x}_j = D_j(E_i(\boldsymbol{x}_i))$, and $\mathcal{L}_0$ is the same as $\widehat{\mathcal{L}}_{ij}$ but where the codes are patched with zeros. This gives a 2×2 matrix of scores. M_{extract} and M_{tr} are computed using these scores as described above. To get M_{cont}, we derive the faithfulness as follows. We replace $\widehat{\mathcal{L}}_{ij}$ with $\widehat{\mathcal{L}}$, the contrastive loss of the model composed with the encoders only. This corresponds to using the concept space as a surrogate for the latent space for contrastive learning. This gives a single score, which we use as the faithfulness metric for M_{cont}. Additionally, we include ImageNet zero-shot classification accuracy as a downstream task for all M_{extract}, M_{tr} and M_{cont} respective cases.

D Human Validation

We ran a human evaluation study for additional verification that alignment score agrees with human judgment. Most of our metrics rely on semantic categories drawn from human annotations (via ImageNet labels used in sparse probing), but do not directly query human judgment of alignment itself. This experiment checks whether alignment scores derived that way agree with direct human perceptual judgment of cross-domain feature correspondence. The protocol was approved by the IRB of an author-affiliated institution. We proceed as follows.

Participant selection. We recruited 8 voluntary participants among colleagues and relatives. Participants were not aware of the goal of the study beyond what the welcome screen described. The study was conducted online. Participants provided informed consent and did not receive monetary compensation for their time. They were instructed to play the game for as long as they desired, which averaged a duration of $\approx$ 23min per participant. We collected a total of 2063 individual trials — ≈ 688 per experimental cell.

Welcome Screen. The welcome screen displays both a short description of the task they were about to do, as well as a consent form.

Practice Session. Before the main experiment began, participants went through 10 practice runs with hand crafted features, whose purpose was to explain the task, preceded by an instruction screen. Fig. A1 shows example trials.

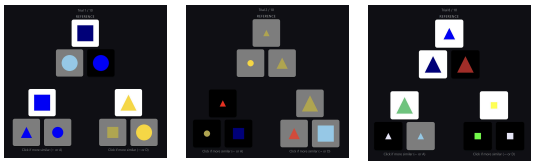


Figure A1: Example practice trials of the human evaluation study.

Main Experiment. The main experiment follows the same format as the practice sessions, but with real data. An example is shown in Fig. A2. Participants are presented with a "REF" (reference) set of maximally activating images among ImageNet for one dictionary atom from one domain (e.g. dino). Say this is atom k of dictionary D_i. Then, they are shown two sets "A" and "B" of mutually similar images. For set "A", we select atom k of dictionary D_j. For set "B", we select an atom $k' \neq k$ from D_j at random. Choices "A" or "B" are presented next to one another, with an order selected at random. Participants are asked to click on the set they find most similar to "REF".

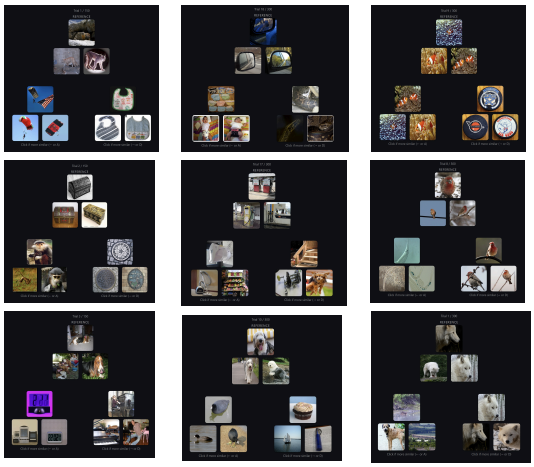


Figure A2: Example trials from the "control" (*left*), "aligned bad" (*center*) and "aligned good" (*right*) cells.

Each trial was one of three pre-defined cells: (*i*) a control, where SAEs were trained without any specific alignment regularisation across domains, and atoms are selected at random among the ones whose corresponding sparse probe is successful (accuracy > 0.9); (*ii*) an "aligned bad" condition, where features are selected from CoSAEs such that InterVenchA score is low — specifically, such that sparse probing is successful (accuracy of the sparse probe > 0.9) but the transfer accuracy is low (< 0.2). This means that this feature in particular is good quality but alignment failed (according to our metrics). (*iii*) An "aligned good" condition, where both the sparse probing is successful (same condition) and the transfer accuracy remains high (> 0.9). This means that the feature is good and the alignment is good.

The goal is to measure the probability that humans chose the set "A", and whether it is significantly higher in "aligned good" than in "aligned bad" trials. A statistically significant difference means that our measure of alignment quality tends to agree with human perception of successful alignment.

Results. The absolute differences are $+6.9$ percentage points for aligned-bad versus control, $+14.8$ points for aligned-good versus control, and $+7.9$ points for aligned-good versus aligned-bad.

A more detailed crossed random-effects binomial mixed model, accounting for participant and feature variability, gives the following contrasts. Because the fitted output reports posterior means and posterior standard deviations, we report posterior odds ratios and normal-

Cell	$P(\text{choose A})$	Cluster-bootstrap 95% CI
Control	0.517	$[0.500, 0.540]$
Aligned bad	0.586	$[0.569, 0.618]$
Aligned good	0.665	$[0.625, 0.696]$

Table A1: Probability of choosing the correct set ("A") across experimental cells.

approximation 95% credible intervals rather than frequentist p-values. We fit a model using `statsmodels.genmod.bayes_mixed_glm.BinomialBayesMixedGLM`.

Contrast	Posterior log-odds $\pm$ SD	Odds Ratio (OR)	Approx. 95% CI for OR
Aligned bad vs. control	0.621 ± 0.091	1.85	$[1.56, 2.22]$
Aligned good vs. control	1.013 ± 0.099	2.75	$[2.27, 3.34]$
Aligned good vs. aligned bad	0.405 ± 0.097	1.50	$[1.24, 1.81]$

Table A2: Pairwise contrasts between cells from the crossed random-effects binomial mixed model (participant and feature as random effects).

The feature-level random-effect standard deviation is 1.566; $[1.446, 1.701]$, whereas the participant-level standard deviation is 0.195; $[0.091, 0.444]$, indicating substantial heterogeneity across features and supporting the use of crossed random effects.

Conclusion. The human result provides graded criterion validity for the probe ranking: aligned probes are significantly preferred to control probes, and the probes labeled "aligned good" are significantly preferred to those labeled "aligned bad", even after accounting for participant and feature variability.

This result does not prove that every selected feature is semantically correct, validate every InterVenchA component, or remove other confounds common in SAE interpretability. It shows that the good-versus-bad probe distinction provided by the benchmark is reflected in human judgments.

E LINEAR CASE: ANALYTICAL SOLUTIONS

In this section, we study the analytical solutions of the proposed framework in the linear setting. By restricting all encoders and decoders to be linear maps, we can explicitly characterize the effect of each objective and identify which components of the model (encoders, decoders, or their compositions) are recovered. This analysis highlights how different combinations of losses induce different notions of coupling across and within domains.

Notation. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d_x}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times d_y}$ be centered data matrices corresponding to two representation instances. We denote their empirical covariance matrices by

$$\boldsymbol{\Sigma}_{xx} = \tfrac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}, \quad \boldsymbol{\Sigma}_{yy} = \tfrac{1}{n}\boldsymbol{Y}^\top \boldsymbol{Y}, \quad \boldsymbol{\Sigma}_{xy} = \tfrac{1}{n}\boldsymbol{X}^\top \boldsymbol{Y}. \tag{7}$$

We consider linear encoders and decoders of the form

$$\boldsymbol{Z}_x = \boldsymbol{X}\boldsymbol{W}_x, \quad \boldsymbol{Z}_y = \boldsymbol{Y}\boldsymbol{W}_y, \qquad \widehat{\boldsymbol{X}} = \boldsymbol{Z}_x \boldsymbol{V}_x, \quad \widehat{\boldsymbol{Y}} = \boldsymbol{Z}_y \boldsymbol{V}_y, \tag{8}$$

where $\boldsymbol{W} \in \mathbb{R}^{d_x \times K}$ are encoder matrices, and the corresponding decoders are given by $\boldsymbol{V} \in \mathbb{R}^{K \times d_x}$. Translation between the two domains is defined as

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{W}_x\boldsymbol{V}_x, \qquad \widehat{\boldsymbol{X}} = \boldsymbol{Y}\boldsymbol{W}_y\boldsymbol{V}_y. \tag{9}$$

When needed, we impose whitening constraints of the form

$$\frac{1}{n}\boldsymbol{Z}_x^\top \boldsymbol{Z}_x = \boldsymbol{I}_K, \qquad \frac{1}{n}\boldsymbol{Z}_y^\top \boldsymbol{Z}_y = \boldsymbol{I}_K. \tag{10}$$

All expectations in the losses are understood empirically over the dataset. Objectives are given by the losses defined in Section 3.2. We analyze these objectives and their combinations in the linear setting,

and characterize the corresponding solutions. In particular, we identify which components (encoders, decoders, or their compositions) are uniquely determined, and how different losses induce coupling across and within domains.

E.1 Reconstruction

We begin with the case where no coupling is introduced between $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e., when only self-reconstruction objectives are considered. In this setting, the two domains decouple, and each reduces to a standard approximation problem.

Lemma 1 (PCA). *Consider the self-reconstruction objectives*

$$\mathcal{L}_{\text{dcy}}{}^{x} = \frac{1}{n}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}_x\boldsymbol{V}_x\|_F^2, \qquad \mathcal{L}_{\text{dcy}}{}^{y} = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{Y}\boldsymbol{W}_y\boldsymbol{V}_y\|_F^2. \tag{11}$$

Then, for each domain independently, any minimizer satisfies:

- *The encoder $\boldsymbol{W}_x$ (resp. $\boldsymbol{W}_y$) spans the top-K principal subspace of $\boldsymbol{\Sigma}_{xx}$ (resp. $\boldsymbol{\Sigma}_{yy}$).*
- *The optimal decoder is given by $\boldsymbol{V}_x = \boldsymbol{W}_x^\top$ (resp. $\boldsymbol{V}_y = \boldsymbol{W}_y^\top$) up to a change of basis.*

Proof. See Appendix J.3.1. □

Discussion. Self-reconstruction identifies both encoders and decoders within each modality, but introduces no coupling between $\boldsymbol{X}$ and $\boldsymbol{Y}$. In particular, the solutions are invariant under invertible transformations of the latent space, i.e., for any invertible matrix $\boldsymbol{R} \in \mathbb{R}^{K\times K}$, the transformation

$$\boldsymbol{W}_x \to \boldsymbol{W}_x\boldsymbol{R}, \qquad \boldsymbol{V}_x \to \boldsymbol{R}^{-1}\boldsymbol{V}_x \tag{12}$$

leaves the objective unchanged. As a result, the latent representations are only identifiable up to a change of basis. Imposing whitening constraints selects an orthonormal basis of this subspace, corresponding to the principal directions (i.e., the right singular vectors of $\boldsymbol{X}$).

E.2 Linear Coupling

We now turn to objectives that introduce coupling between the two domains. Unlike self-reconstruction, which operates independently on each modality, these objectives enforce interactions either at the level of representations (concept consistency) or through predictive mappings (translation). We study how these different forms of coupling affect identifiability and the structure of the learned representations.

E.2.1 Concept consistency

We first consider coupling at the level of latent representations, through the concept consistency objective.

Objective. The concept consistency loss is defined as

$$\mathcal{L}_{\text{cont}} = \frac{1}{n}\|\boldsymbol{Z}_x - \boldsymbol{Z}_y\|_F^2 = \frac{1}{n}\|\boldsymbol{X}\boldsymbol{W}_x - \boldsymbol{Y}\boldsymbol{W}_y\|_F^2. \tag{13}$$

CCA. We begin by showing that the concept consistency objective alone recovers the canonical correlation analysis (CCA) for the encoders.

Lemma 2 (Degeneracy of concept consistency). *In the absence of additional constraints, minimizing $\mathcal{L}_{\text{cont}}$ is ill-posed: for any pair $(\boldsymbol{W}_x, \boldsymbol{W}_y)$, and any orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{K\times K}$, the transformation*

$$\boldsymbol{W}_x \to \boldsymbol{W}_x\boldsymbol{R}, \qquad \boldsymbol{W}_y \to \boldsymbol{W}_y\boldsymbol{R} \tag{14}$$

leaves the objective unchanged. Moreover, the objective admits degenerate solutions (e.g., collapse to zero).

Proof. See Appendix J.3.2. □

Lemma 3 (CCA from concept consistency with whitening)**.** *Consider minimizing $\mathcal{L}_{\text{cont}}$ under the whitening constraints*

$$\frac{1}{n}\boldsymbol{Z}_x^\top \boldsymbol{Z}_x = \boldsymbol{I}_K, \qquad \frac{1}{n}\boldsymbol{Z}_y^\top \boldsymbol{Z}_y = \boldsymbol{I}_K. \tag{15}$$

Then the optimal encoders $(\boldsymbol{W}_x, \boldsymbol{W}_y)$ are given by the top-K canonical directions between $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e., they solve the canonical correlation analysis (CCA) problem.

Proof. See Appendix J.3.2. □

Discussion. Concept consistency introduces coupling across modalities by enforcing alignment of latent representations. However, without additional constraints, the problem is ill-posed due to scaling and collapse. Whitening resolves these ambiguities and yields CCA, which identifies shared directions maximizing cross-correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Notably, this objective only identifies encoders: decoders remain unconstrained, and no reconstruction structure is imposed within each modality.

Identifying the Decoders. The CCA alone discards the decoders. To identify them, we now add self-reconstruction and consider the combined objective

$$\mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{dcy}}{}^{x} + \mathcal{L}_{\text{dcy}}{}^{y}. \tag{16}$$

Lemma 4 (Trade-off between CCA and PCA)**.** *Minimizing $\mathcal{L}_{\text{cont}} + \lambda \mathcal{L}_{\text{dcy}}$ yields encoders $(\boldsymbol{W}_x, \boldsymbol{W}_y)$ that solve a generalized eigenvalue problem interpolating between principal components and canonical directions. After accounting for degeneracies,*

- (i) *the solution does not depend on λ,*
- (ii) *only the pure CCA remains,*
- (iii) *the effect of the reconstruction term is to identify the decoders as optimal linear readouts*

Proof. See Appendix J.3.2. □

Discussion. While the combined objective highlights a fundamental trade-off between shared and idiosyncratic information, this trade-off disappears after solving for degenerate identification of the shared subspace. Only shared information remains, and the effect of self-reconstruction is merely to provide solutions to the decoders, as the raw CCA only considers encoders.

E.2.2 Translation

We now consider coupling through translation, which enforces predictive mappings between the two domains. Unlike concept consistency, which aligns latent representations, translation constrains the composition of encoders and decoders across modalities, and therefore couples representations through their ability to reconstruct one domain from the other.

In contrast with concept consistency, which leaves decoders entirely unconstrained and requires self-reconstruction to identify them, translation introduces a different form of degeneracy: it constrains only cross-modal compositions (e.g., $\boldsymbol{W}_x \boldsymbol{V}_y$), but does not couple encoders and decoders within each modality. As a result, the individual components remain underdetermined. As before, adding self-reconstruction restores this intra-modality coupling and leads to well-identified solutions.

RRR. First, consider the translation objective alone. This corresponds to the reduced-rank regression (RRR) and only identifies the composition of encoders from one domain to decoders of the other domain. Encoders and decoders within one domain are decoupled, and all operators are defined up to some transforms.

Lemma 5 (Translation and reduced-rank regression)**.** *Consider the (one-directional) translation objective*

$$\mathcal{L}_{\text{tr}}{}^{x\to y}(\boldsymbol{W}_x, \boldsymbol{V}_y) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}_x\boldsymbol{V}_y\|_F^2. \tag{17}$$

Then minimizing ${\mathcal{L}_{\text{tr}}}^{x\to y}$ over $(\boldsymbol{W}_x, \boldsymbol{V}_y)$ is equivalent to RRR of $\boldsymbol{Y}$ onto $\boldsymbol{X}$ with rank constraint K. The factorization $(\boldsymbol{W}_x, \boldsymbol{V}_y)$ is not identifiable: only the product $\boldsymbol{W}_x\boldsymbol{V}_y$ is. One such factorization recovers the CCA.

Proof. See Appendix J.3.3. □

Adding self-reconstruction. We now add self-reconstruction objectives, which introduce coupling within each modality:

$$\mathcal{L}_{\text{tr}} + \lambda({\mathcal{L}_{\text{dcy}}}^{x} + {\mathcal{L}_{\text{dcy}}}^{y}). \tag{18}$$

While translation alone only constrains the cross-modal compositions, self-reconstruction couples encoders and decoders within each domain. This restores intra-modality structure and yields well-defined decoders given the encoders, but does not fully resolve the non-identifiability of the encoders. Further whitening is still required, as in the concept consistency case.

Lemma 6 (Translation with reconstruction)**.** *Under this combined objective, the optimal encoders are again given by the CCA, while the decoders are given by an interpolation between the solutions of Lem. 5 and Lem. 4.*

Proof. See Appendix J.3.3. □

E.2.3 Distributional alignment

Distributional alignment enforces matching of pushforward measures rather than individual samples. Intuitively, it aligns global statistics (e.g., projections or moments) of the representations across domains. However, this does not introduce any structural coupling between X and Y: many unrelated mappings can induce identical distributions. As a result, the correlation structure may align in aggregate, but the underlying representations remain uncoupled, and no meaningful joint factorization is recovered. This is true for both translation and concept consistency.

E.2.4 Cycle consistency

Cycle consistency enforces that $\boldsymbol{X} \simeq \boldsymbol{XAB}$ and $\boldsymbol{Y} \simeq \boldsymbol{YBA}$ with $\boldsymbol{A} = \boldsymbol{W}_X\boldsymbol{V}_Y$ and $\boldsymbol{B} = \boldsymbol{W}_Y\boldsymbol{V}_X$, introducing an apparent coupling between the two domains. Without coupling, the optimal solution reduces to independent PCA decompositions, say P_x and P_y. With cycle consistency, one recovers the same PCA solution up to an arbitrary invertible transform: $\boldsymbol{A} = \boldsymbol{P}_x\boldsymbol{R}\boldsymbol{P}_y^\top$ and $\boldsymbol{B} = \boldsymbol{P}_y\boldsymbol{R}^{-1}\boldsymbol{P}_x^\top$ (Lem. 7). The factorization remains ambiguous up to an invertible transform, and the optimal solution is still equivalent to PCA, meaning the apparent coupling vanishes upon solving the problem.

F Platonic and Linear Representation Hypothesis

Platonic Representation Hypothesis (PRH). The PRH is a universal statement that any representation learning method will converge to the same shared representation, up to some transform [20]. This can be formalized in various ways. We chose to write this hypothesis under a generative framework as it allows for a formulation that is ready to be plugged into our above framework. The generative view of the PRH assumes the existence of a platonic latent space X such that all real world observations, whatever their domain $\mathcal{X}^i$, stem from this platonic latent space through a generator $g_i : X \to \mathcal{X}^i$. Then, all feature extractors $\phi_i : \mathcal{X}^i \to X^i$ invert this generative process up to a transform ψ_i. In the context laid out above, adding this extra generative step allows for intermediate observation spaces $(\mathcal{X}^i)_{i\in I}$ between the platonic latent space X and the feature extractors ϕ_i. This is especially relevant when the g_is are allowed to introduce distribution shifts or even modality shifts. For simplicity, we can further add the restriction that $X = \mathbb{R}^K$.

Hypothesis 1 (Platonic Generative Process)**.** $\exists K \in \mathbb{N}, \exists X \subseteq \mathbb{R}^K$ *s.t.* $\forall i \in I$

$$\exists g_i : X \to \mathcal{X}^i \tag{19}$$

$$\forall \phi_i : \mathcal{X}^i \to X^i, \exists \psi_i : X^i \to X, \text{ s.t. } \psi_i \circ \phi_i \circ g_i = \mathrm{id}_X \tag{20}$$

Linear Representation Hypothesis (LRH). In practice, most of the recent representation alignment and concept extraction methods implicitly assume the existence of such a generative process. In other words, they assume that representation learning inverts the generative process. In the case of representation alignment, the class of ψ_i under consideration is further constrained, allowing invariance up to translation, scaling, orthogonal transformations, etc. Dictionary learning-based methods fall under a different assumption, more realistic, and grounded in neurosciences: that platonic representations are *sparse* (Hypothesis 2), imposing different constraints on ψ_i. This sparse coding framework often comes with additional linear or close to linear constraints. The reason is (*i*) that it allows for GPU-friendly implementations through the extremely popular Sparse Autoencoder (SAE) framework, (*ii*) that representation learning is explicitly designed to produce latent spaces with meaningful Euclidean—or at least linear—geometry, with downstream operations reading linearly from X^i before optionally applying some nonlinearities.

Hypothesis 2 (Sparse Generative Process). $\boldsymbol{x} \in X \sim \mu = \prod_{k=1}^{K} p_k$ *with* $|\mathrm{supp}(\boldsymbol{x})| \ll K$.

Essentially, this generative formulation of the PRH coupled with the SAE operationalization of the LRH grounds our concept alignment framework. It does so by assuming the existence of X and of the ϕ_i, now decomposed into an additional generative step g_i and an inversion step ϕ_i, and finally, by assuming that the class of autoencoders under consideration provides the final inversion step, aligning all representations back in a shared concept space through the ψ_i. Our framework additionally provides for a natural distributional version of the PRH. We call the element-wise version the *strong* PRH, and the distributional version the *weak* PRH.

G Distributional losses

When instance-wise correspondences are unavailable, enforcing $P^{\mathrm{dist}}_{\mathrm{tr}}$ or $P^{\mathrm{dist}}_{\mathrm{cont}}$ requires matching distributions rather than individual samples. A natural approach is to minimize a discrepancy between pushforward measures, for instance, using optimal transport. However, distributional distances like the Wasserstein distance are both computationally expensive and poorly behaved in high dimensions and mini-batch settings, making them ill-suited for large-scale SAE training.

A practical alternative is provided by the Cramer-Wold theorem, which states that two distributions are equal if and only if their one-dimensional projections along all directions are equal. This motivates the use of sliced distributional objectives, which compare 1D projections of the distributions along randomly sampled directions. We take inspiration from Balestriero and LeCun [57], who use the Epps-Pulley test on random 1D slices to regularize for Gaussianity of their embedding space:

$$\mathrm{EP}(\mu_i) = b \int_{-\infty}^{\infty} |\widehat{\varphi}_{\mu_i}(t) - \varphi(t)|^2 \, w(t) \, \mathrm{d}t \tag{21}$$

where $\widehat{\varphi}_{\mu_i} = \frac{1}{b}\mathbb{E}_{\boldsymbol{x}\sim\mu_i} e^{it\langle \boldsymbol{x}, \boldsymbol{u}\rangle} = \sum_{k=1}^{b} e^{it\langle \boldsymbol{x}_{i,k}, \boldsymbol{u}\rangle}$ is the empirical characteristic function (ECF) of μ_i where a batch of b samples are projected onto a random direction $\boldsymbol{u} \in \mathcal{S}^{d-1}$; $\varphi = e^{-t^2/2}$ is the characteristic function of the normal distribution, and w is a weighting function—typically a gaussian. In the case of $P^{\mathrm{dist}}_{\mathrm{tr}}$, the target is not the normal distribution but another of our μ_j, and the source becomes $\psi^{-1} \circ \psi_i \# \mu_i$. We can therefore define the sliced translation loss as follows:

$$\mathcal{L}^{\mathrm{dist}}_{\mathrm{tr}} := \mathbb{E}_{\boldsymbol{u}\sim\mathcal{S}^{d-1}} b \int_{-\infty}^{\infty} \left| \widehat{\varphi}_{(\psi_j^{-1}\circ\psi_i)\#\mu_i}(t) - \widehat{\varphi}_{\mu_j}(t) \right|^2 w(t) \, \mathrm{d}t \tag{22}$$

$$\simeq \frac{1}{|\mathcal{U}|} \sum_{\boldsymbol{u}\in\mathcal{U}} b \int_{-\infty}^{\infty} \left| \widehat{\varphi}_{(\psi_j^{-1}\circ\psi_i)\#\mu_i}(t) - \widehat{\varphi}_{\mu_j}(t) \right|^2 w(t) \, \mathrm{d}t \tag{23}$$

where $\mathcal{U} \in (\mathcal{S}^{d-1})^m$ is a finite set of random directions, resampled at each training step. While any single slice provides only a weak constraint, aggregating over many random projections yields a consistent approximation of the target distributional alignment. Considering that $\mathcal{U}$ is resampled at each training step, the effective number of slices seen over training becomes even larger. Crucially, Balestriero and LeCun [57] show that a small coverage of $\mathcal{S}^{d-1}$ per step is sufficient to achieve good performance, with theoretical error bounds when the target distribution is gaussian. These theoretical bounds do not apply in our case since we do not control the target distribution. However, this approach is equivalent to regularizing with a sliced Maximum Mean Discrepancy (MMD) (Proposition 3),

which is widely used in practice for distributional alignment [69; 70], especially as an alternative to adversarial training [71; 72].

In the concept space $\mathbb{R}^K$, this perspective admits a particularly simple interpretation. In arbitrary latent spaces, canonical directions are meaningless, only the overall Euclidean geometry matters, which motivates arbitrary random slices for full coverage of the space. In contrast, in the concept space, canonical directions correspond to dictionary atoms, which are assumed to be meaningful, atomic units of information. Therefore, in concept space, we do not consider arbitrary random slices $u \in \mathcal{S}^{K-1}$, but rather canonical slices along the coordinate axes. This yields the following loss for distributional alignment of concepts:

$$\mathcal{L}_{\text{cont}}^{\text{dist}} := \frac{1}{|\mathcal{U}^*|} \sum_{\boldsymbol{u} \in \mathcal{U}^*} b \int_{-\infty}^{\infty} \left| \widehat{\varphi}_{\psi_i \# \mu_i}(t) - \widehat{\varphi}_{\psi_j \# \mu_j}(t) \right|^2 w(t)\, \mathrm{d}t \tag{24}$$

where $\mathcal{U}^* = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$ is the set of canonical basis vectors in $\mathbb{R}^K$. In practice, due to the sparsity of activations in concept space, we instead take $\mathcal{U}^*$ to be the set of active atoms across the batch. This formulation of $\mathcal{L}_{\text{cont}}^{\text{dist}}$ is further motivated by Dhimoïla et al. [15]'s IsoEnergy assumption.

H Additional Experimental Details

H.1 Synthetic Data Generating Process

We design a synthetic data-generating process (DGP) that follows precisely our model. First, we sample N i.i.d. samples of normally distributed latent variables $\boldsymbol{x}' \in \mathbb{R}^K$. Then, for each sample, we select its top k latent variables, set them to 1 and the rest to 0, to get the points $\boldsymbol{x} \in X$. Then, for each domain i, we construct ϕ_i as the composition of the following. First, we apply some dropout to $\boldsymbol{x}$. Then, we add a random white noise on the active latent variables, with standard deviation $\sigma = 0.1$, and final values clipped between $1 \pm \sigma$. Finally, we apply a succession of random linear transformations interleaved with ReLU nonlinearities, with a target dimensionality of $d < K$ and a bottleneck layer of dimension $d_{\text{bottleneck}} < d$.

To measure SAE quality, we monitor the explained variance (R^2) in latent space, as well as the distance of both the dictionary and codes to ground truth using a standard matching algorithm to control for permutation invariance of the codes [73; 74; 15]. To measure alignment quality, we use the R^2 of the translation and concept consistency properties, as well as the distance between the pushforward measures in both concept space and latent space using our losses. See Appendix H.2 for details on the SAE architecture and training procedure.

H.2 Hyperparameters and Training Details

SAE architecture. All variants of the SAE framework rely on specific hidden assumptions on the underlying data-generating process, on the shape of receptive fields [74], and on the type of atoms to be discovered [47; 15]. The purpose of this work is not to show which of these particular methods and assumptions best fit empirical latent spaces, but rather to focus on concept alignment properties. Thus, throughout this work, we adopt the *batchtopk* architecture [46] as implemented by *overcomplete* [47] for all SAEs across all experiments, with a dictionary size of $K = 8d$ and a sparsity level of $k = \frac{d}{20}$. These constants also hold for the synthetic setting, where we take $d = 384$, $K = 3072$ and $k = 19$.

Training. In the vision setting, we train all SAEs on $N = 1.2 \times 10^6$ unique samples from ImageNet [75], with a batch size of $b = 1024$ and a total number of tokens seen during training of $T = 5 \times 10^7$. This corresponds to approximately $E = 40$ epochs. During hyperparameter tuning, we limit the number of epochs to $E = 4$. The learning rate ($\sim 10^{-3}$) is calibrated such that the $\mathcal{L}_{\text{dcy}}$ only run converges to 1.1 times its final value in $E = 4$ epochs. We keep the same learning rate for all other runs, which all converge well within $E = 40$ epochs. We do not observe significant variance across runs with different seeds.

Our results seem not too sensitive to the regularization strengths $\alpha_{\text{tr/cont/...}}$, which are calibrated as follows. Let $\mathcal{L}^*$ be the final $\mathcal{L}_{\text{dcy}}$ loss after training on 5×10^7 tokens with no regularisation. Then, for $\text{reg} \in \{\cdot_{\text{tr}}, \cdot_{\text{cont}}, \cdot_{\text{tr}}^{\text{dist}}, \cdot_{\text{cont}}^{\text{dist}}\}$, we sweep over α_{reg} and keep the maximum order of magnitude such that the final $\mathcal{L}_{\text{dcy}}$ term of the loss is less than $0.9 \times \mathcal{L}^*$. All subsequent experiments use these

fixed values for α_{reg} if the corresponding regularisation is included, and 0 otherwise. This procedure yields $\alpha_{\text{tr}} = \alpha_{\text{cont}} = 1$, and $\alpha_{\text{tr}}^{\text{dist}} = \alpha_{\text{cont}}^{\text{dist}} = 10^3$. We find that results seem to remain consistent with α varying by one or two orders of magnitude in most cases.

In the language case, SAEs are trained on a total of 5×10^7 unique tokens extracted from the `HuggingFaceFW/fineweb` dataset, and aligned across tokenizers using the same procedure as [63]. For the models, we use `google/gemma-2-9b` and `meta-llama/Llama-3.1-8B`. All hyperparameter tuning details are the same as in the vision setting.

In the multimodal setting, we sub-sample LAION-400M [76] to get $N \sim 10^6$ unique image-text pairs, and keep the same other parameters as for ImageNet. We select CLIP (`openai/clip-vit-base-patch32`) [3] and OpenCLIP-L (`laion/CLIP-ViT-L-14-laion2B-s32B-b82K`) [77]. These models are used to extract embeddings for both images and captions. Here, we focus on SAEs coupled to the vision and text towers of the same model, as the scope of this setting is cross-modal rather than cross-model alignment. In the synthetic setting, we train all SAEs on $N = 10^6$ unique samples, with a batch size of $b = 10^3$ and a total number of tokens seen during training of $T = 10^7$, which corresponds to $E = 10$ epochs.

H.3 Mixed Training Regime

In the mixed training regime, we split the dataset into two subsets. Let $p \in [0, 1]$ be the proportion of matched pairs. We randomly select $p \cdot N$ pairs to be in the matched subset X_A^i, and the rest to be in the unmatched subset X_B^i. The token datasets are then the concatenated $X^i = X_A^i \cup X_B^i$. At this point, in all settings, all domains are perfectly aligned. We then shuffle the unmatched subsets X_B^i independently to break alignment. We keep track of matched and unmatched pairs using binary labels $\boldsymbol{l} \in \{0, 1\}^N$, where $\boldsymbol{l}_n = 1$ if the n-th pair is matched and 0 otherwise. During training, batches are randomly sampled from the whole dataset. The distributional regularization terms $\mathcal{L}_{\text{tr}}^{\text{dist}}$ and $\mathcal{L}_{\text{cont}}^{\text{dist}}$ are computed on the whole batch. The instance-wise regularization terms $\mathcal{L}_{\text{tr}}$ and $\mathcal{L}_{\text{cont}}$ are computed only on the matched pairs, i.e. those for which $\boldsymbol{l}_n = 1$.

H.4 Compute

All final experiment runs were run on a single H100 GPU. It takes about 40min to train a CoSAE on 50M precomputed CLS tokens. It takes about 7h to train a CoSAE on 50M tokens extracted on the fly from LLMs using activation buffers. Most experiments were run on an RTX3060 for early testing, where training a similar CoSAE on 5M precomputed CLS tokens takes a few minutes.

I Controlled synthetic evaluation of shared and private factor recovery

Alignment losses encourage two representations to agree, but agreement does not by itself show that they recover the same generating factor. Independently trained sparse autoencoders (SAEs) may place the same factor at different code indices, while dependent code coordinates may fail to isolate any one factor. We therefore use a controlled two-view generator to distinguish within-view factor recovery from cross-view correspondence. By varying the amount of shared structure, the experiment also tests whether alignment preserves private factors and whether equal factor marginals cause unsupported declarations of sharedness.

I.1 Design and evaluation

For each view $v \in \{1, 2\}$, observations in $\mathbb{R}^{64}$ are generated as

$$\widetilde{x}^{(v)} = A_{\text{s}}^{(v)} s + A_{\text{p}}^{(v)} p^{(v)} + \varepsilon^{(v)}, \qquad \varepsilon^{(v)} \sim \mathcal{N}(0, 0.02^2 I). \tag{25}$$

The coefficient vector s is shared exactly across views, whereas $p^{(1)}$ and $p^{(2)}$ are independent. Dictionary columns are independent normalized Gaussian vectors. Within each block, a fixed-size support is drawn uniformly, and active amplitudes follow $\exp(\mathcal{N}(0, 0.5^2))$. Thus every view has 32 factors and four active coefficients per example, while the number of shared factors changes across

Table A3: Shared/private generator. K_s and K_p are the numbers of shared and private factors, with K_p specified per view; k_s and k_p are their active support sizes. Each SAE has $M = K_s + 2K_p$ slots.

Regime	K_s	K_p	k_s	k_p	M
Fully shared	32	0	4	0	32
Partially shared	8	24	1	3	56
No shared factors	0	32	0	4	64

the three regimes in Table A3. In the no-shared regime, private coefficients have the same marginal law but independent realizations.

We generate 50,000 training, 10,000 validation, and 20,000 test examples and standardize each coordinate using training statistics. Each view has a separate affine encoder and decoder. BatchTopK [46] retains the largest $4b$ rectified activations in a batch of size b; the slot counts in Table A3 leave capacity for both shared and private factors. Models are trained with AdamW [78] for at most 30 epochs, using batches of 1024, an initial learning rate of 10^{-3}, cosine decay, and early stopping.

We compare reconstruction alone with four coupled configurations: instance consistency (coefficient 0.1), instance translation (1), distributional translation and consistency (100 and 30), and the full combination. Reconstruction is normalized by target energy, translation compares unit-norm observations, and consistency is normalized by the total code energy. Distributional translation compares random one-dimensional projections, while distributional consistency compares code coordinates using a weighted empirical characteristic-function discrepancy [79; 57]; this is a sliced maximum mean discrepancy [69]. All configurations also use a dead-feature auxiliary coefficient of 10^{-3}. The five seeds vary both the generated dictionaries and optimization, and objective configurations use matched seed identities within each regime. For results quoted with confidence intervals, we treat the five seeded experiments as independent replicates and compute a two-sided 95% Student's t confidence interval for their mean. These intervals are not adjusted for the number of reported comparisons.

Ground-truth metrics. Within each view, a one-to-one Hungarian assignment [80] matches learned slots to generating factors by absolute test-set correlation; a factor is recovered when its assigned correlation is at least 0.8. *Same-index shared recall* additionally requires both views to assign the same slot to the same shared factor above this threshold.

Choice of multiplicity correction. We use step-down maxT to identify same-index code coordinates whose absolute cross-view correlation exceeds that expected after breaking example-level correspondence. A single row permutation is applied jointly to all coordinates, preserving the dependence among slotwise statistics under the permutation reference. After ordering the observed correlations, each is compared with the permutation maximum over the hypotheses remaining at that step [81]. This can be less conservative than single-step maxT, although the first step is the same global-maximum comparison. We use 10,000 permutations and level 0.05, apply the plus-one Monte Carlo correction [82], and retain single-step results as a sensitivity analysis. A declared slot is called unsupported when it fails the ground-truth same-factor criterion. The procedure calibrates the joint row-independence null within a trained model pair; it does not establish factor identity or, without additional subset-pivotality assumptions, strong familywise control for semantic hypotheses under partial sharing.

Routing-reference excess estimators. We perturb test observations along known unit-norm generating directions, re-encode them, and cross-decode the code change. Shared responses are compared with the corresponding target direction. For each private source factor, RMS routing excess subtracts the mean normalized RMS magnitude under 5,000 fixed-point-free target-slot permutations from the same-index response, holding the source code change fixed. Strongest-target excess instead projects the response onto the ground-truth target dictionary, selects the largest target-factor RMS coordinate, and reselects that maximum within every reference permutation before subtraction. We average both factor-level excesses across private factors and translation directions. The first estimator tests for preferential total response magnitude; the second tests for preferential response along some ground-truth target factor. Neither measures absolute response, and the derangement distribution is a descriptive routing reference, not a calibrated semantic randomization null. We therefore also report

Table A4: Recovery results, reported as mean ± sample SD across five seeds. Reconstruction and private recall are averaged across views. Shared recall requires ground-truth recovery at the same code index; unsupported is the number of statistically declared slots that fail this criterion. A dash denotes an inapplicable endpoint.

Regime	Objective	Mean R^2	Private recall	Shared recall	Unsupported
Fully shared	Reconstruction	0.984 ± 0.008	—	0.025 ± 0.026	25.800 ± 3.114
Fully shared	Instance consistency	0.983 ± 0.003	—	0.075 ± 0.036	6.800 ± 1.483
Fully shared	Instance translation	0.978 ± 0.013	—	0.975 ± 0.034	0.800 ± 1.095
Fully shared	Distributional	0.862 ± 0.012	—	0.025 ± 0.014	23.200 ± 1.924
Fully shared	Full	0.963 ± 0.018	—	0.950 ± 0.047	1.400 ± 1.140
Partially shared	Reconstruction	0.994 ± 0.001	1.000 ± 0.000	0.025 ± 0.056	1.600 ± 0.894
Partially shared	Instance consistency	0.994 ± 0.000	1.000 ± 0.000	0.100 ± 0.105	1.600 ± 1.949
Partially shared	Instance translation	0.847 ± 0.005	0.988 ± 0.011	0.975 ± 0.056	1.800 ± 1.483
Partially shared	Distributional	0.954 ± 0.008	1.000 ± 0.000	0.050 ± 0.068	4.400 ± 2.966
Partially shared	Full	0.809 ± 0.006	0.988 ± 0.011	0.975 ± 0.056	2.200 ± 1.924
No shared	Reconstruction	0.994 ± 0.000	1.000 ± 0.000	—	0.200 ± 0.447
No shared	Instance consistency	0.994 ± 0.000	1.000 ± 0.000	—	0.000 ± 0.000
No shared	Instance translation	0.981 ± 0.001	1.000 ± 0.000	—	0.000 ± 0.000
No shared	Distributional	0.953 ± 0.004	1.000 ± 0.000	—	0.200 ± 0.447
No shared	Full	0.922 ± 0.009	0.997 ± 0.007	—	0.000 ± 0.000

absolute responses: a negative excess means below-reference routing, not negative causal transfer, while a near-zero excess does not establish equivalence or absent target-factor response.

I.2 Results

Reconstruction-only training shows that the generating factors are recoverable: it recovers all private factors in the partially shared and no-shared regimes, with mean reconstruction $R^2 = 0.994$. Nevertheless, same-index shared recall is only 0.025 in both regimes containing shared factors (Table A4). Accurate within-view recovery therefore does not align code indices across views.

Instance translation raises same-index shared recall to 0.975 in both shared regimes. Its reconstruction remains 0.978 when all factors are shared but falls to 0.847 under partial sharing, while private recall remains 0.988. The full objective similarly reaches shared recall of 0.950 and 0.975 in the fully and partially shared regimes, respectively, with partial-sharing private recall of 0.988. These comparisons show that shared correspondence can coexist with private recovery, although the reconstruction cost depends on the objective and the amount of private structure.

Distributional training gives the complementary dissociation. Under partial sharing it preserves reconstruction ($R^2 = 0.954$) and all private factors, but same-index shared recall remains 0.050. Matching the tested marginals is thus insufficient to identify which factors correspond. In the no-shared regime, the full and both instance-only configurations declare no shared slots across the five seeds while retaining private recall of at least 0.997.

The interventions further separate routing magnitude from destination. Under the full objective, mean shared-direction cosine is 0.370 in the fully shared regime and 0.402 under partial sharing, despite shared recall of at least 0.950. In the no-shared regime, mean private routing-magnitude excess is -0.00014 (95% interval $[-0.00133, 0.00105]$), whereas the strongest target-factor response is 0.450 and its reference excess is 0.0351 ($[0.0208, 0.0494]$). A near-zero aggregate routing excess therefore does not establish the absence of a response along a particular target factor.

I.3 Scope

The distributional coefficients were selected exploratorily using the same five seed identities reported here, so these comparisons are not an independent confirmation. Moreover, the generator uses linear mixing, independent private blocks, fixed support sizes, and sufficient slots for the union of factors; it does not cover dependent private factors, nonlinear mixing, or strongly overcomplete recovery. Within this controlled setting, the results distinguish four properties that should not be collapsed into one claim of shared geometry: reconstruction, factor recovery, same-index correspondence, and intervention destination.

J PROOFS

J.1 ALIGNMENT DUALITIES

Our first proof is that of the duality between latent and concept alignment, or translation and concept consistency, under the assumption that decoders are injective.

Proposition. *Suppose injectivity of the decoders, and suppose* P_{dcy}. *Then*

$$P_{\text{tr}} \Leftrightarrow P_{\text{cont}}$$

Proof. $P_{\text{tr}} \Leftarrow P_{\text{cont}}$: This direction does not require the injectivity assumption. Let $x \in X$ and $i, j \in I$. Suppose P_{cont}. Then:

$$\begin{aligned} P_{\text{cont}} : \psi_i(\boldsymbol{x}_i) = \psi_j(\boldsymbol{x}_j) &\Rightarrow \psi_j^{-1}(\psi_i(\boldsymbol{x}_i)) = \psi_j^{-1}(\psi_j(\boldsymbol{x}_j)) \\ &\Rightarrow \psi_j^{-1}(\psi_i(\boldsymbol{x}_i)) = \boldsymbol{x}_j : P_{\text{tr}} \qquad \text{by } P_{\text{dcy}} \end{aligned}$$

$P_{\text{tr}} \Rightarrow P_{\text{cont}}$: Let $x \in X$ and $i, j \in I$. Suppose P_{tr}. Then:

$$\begin{aligned} P_{\text{tr}} : (\psi_j^{-1} \circ \psi_i)(\boldsymbol{x}_i) = x_j &\Rightarrow \psi_j^{-1}(\psi_i(\boldsymbol{x}_i)) = \psi_j^{-1}(\psi_j(\boldsymbol{x}_j)) \qquad \text{by } P_{\text{dcy}} \\ &\Rightarrow \psi_i(\boldsymbol{x}_i) = \psi_j(\boldsymbol{x}_j) : P_{\text{cont}} \qquad \text{by injectivity of } \psi_j^{-1} \end{aligned}$$

□

Then, we prove in Proposition 2 the trivial result that pair-wise alignment implies distributional alignment.

Proposition. $P_{\text{tr}} \Rightarrow P_{\text{tr}}^{\text{dist}} \quad \wedge \quad P_{\text{cont}} \Rightarrow P_{\text{cont}}^{\text{dist}}$

Proof. $P_{\text{tr}} \Rightarrow P_{\text{tr}}^{\text{dist}}$: Assume P_{tr}, i.e., for all samples $x \in X$ and domains $i, j \in I$,

$$(\psi_j^{-1} \circ \psi_i)(\boldsymbol{x}_i) = \boldsymbol{x}_j, \quad \text{where } \boldsymbol{x}_i = \phi_i(x),\ \boldsymbol{x}_j = \phi_j(x).$$

Let $A \subseteq X^j$ be measurable. Then, by definition of the pushforward measure,

$$\begin{aligned} \big((\psi_j^{-1} \circ \psi_i)\#\mu_i\big)(A) &= \mu_i\big((\psi_j^{-1} \circ \psi_i)^{-1}(A)\big) \\ &= \mu\big(\phi_i^{-1}((\psi_j^{-1} \circ \psi_i)^{-1}(A))\big). \end{aligned}$$

For $x \in X$, using P_{tr},

$$x \in \phi_i^{-1}((\psi_j^{-1} \circ \psi_i)^{-1}(A)) \ \Leftrightarrow\ (\psi_j^{-1} \circ \psi_i)(\phi_i(x)) \in A \ \Leftrightarrow\ \phi_j(x) \in A.$$

Hence

$$\phi_i^{-1}((\psi_j^{-1} \circ \psi_i)^{-1}(A)) = \phi_j^{-1}(A),$$

and therefore

$$\begin{aligned} \big((\psi_j^{-1} \circ \psi_i)\#\mu_i\big)(A) &= \mu(\phi_j^{-1}(A)) \\ &= \mu_j(A). \end{aligned}$$

Thus $(\psi_j^{-1} \circ \psi_i)\#\mu_i = \mu_j$.

$P_{\text{cont}} \Rightarrow P_{\text{cont}}^{\text{dist}}$: Assume P_{cont}, i.e., for all samples $x \in X$ and domains $i, j \in I$,

$$\psi_i(\boldsymbol{x}_i) = \psi_j(\boldsymbol{x}_j), \quad \text{where } \boldsymbol{x}_i = \phi_i(x),\ \boldsymbol{x}_j = \phi_j(x).$$

Let $B \subseteq \mathbb{R}^K$ be measurable. Then

$$\begin{aligned} (\psi_i\#\mu_i)(B) &= \mu_i(\psi_i^{-1}(B)) \\ &= \mu(\phi_i^{-1}(\psi_i^{-1}(B))). \end{aligned}$$

For $x \in X$, using P_{cont},

$$x \in \phi_i^{-1}(\psi_i^{-1}(B)) \;\Leftrightarrow\; \psi_i(\phi_i(x)) \in B \;\Leftrightarrow\; \psi_j(\phi_j(x)) \in B \;\Leftrightarrow\; x \in \phi_j^{-1}(\psi_j^{-1}(B)).$$

Thus

$$\phi_i^{-1}(\psi_i^{-1}(B)) = \phi_j^{-1}(\psi_j^{-1}(B)),$$

and therefore

$$\begin{aligned}(\psi_i \# \mu_i)(B) &= \mu(\phi_j^{-1}(\psi_j^{-1}(B))) \\ &= \mu_j(\psi_j^{-1}(B)) \\ &= (\psi_j \# \mu_j)(B).\end{aligned}$$

Hence $\psi_i \# \mu_i = \psi_j \# \mu_j$. □

J.2 Our distribution losses are MMDs

Proposition 3. *Let μ, ν be probability measures on $\mathbb{R}^d$, and let $w : \mathbb{R} \to \mathbb{R}_+$ be an integrable weighting function. For $\boldsymbol{u} \in \mathcal{S}^{d-1}$, define*

$$\mathrm{EP}_{\boldsymbol{u}}(\mu, \nu) \coloneqq \int_{-\infty}^{\infty} \left|\varphi_{\mu,\boldsymbol{u}}(t) - \varphi_{\nu,\boldsymbol{u}}(t)\right|^2 w(t)\, dt,$$

where $\varphi_{\mu,\boldsymbol{u}}(t) \coloneqq \mathbb{E}_{x\sim\mu} e^{it\langle \boldsymbol{x}, \boldsymbol{u}\rangle}$ is the characteristic function of the one-dimensional projection of μ along $\boldsymbol{u}$. Then $\mathrm{EP}_{\boldsymbol{u}}(\mu, \nu)$ is a Maximum Mean Discrepancy:

$$\mathrm{EP}_{\boldsymbol{u}}(\mu, \nu) = \mathrm{MMD}^2_{k_{\boldsymbol{u}}}(\mu, \nu),$$

with respect to the positive definite kernel

$$k_{\boldsymbol{u}}(x, y) \coloneqq \int_{-\infty}^{\infty} e^{it\langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{u}\rangle} w(t)\, dt.$$

Proof. We expand the squared difference:

$$\begin{aligned}\mathrm{EP}_{\boldsymbol{u}}(\mu, \nu) &= \int \left|\varphi_{\mu,\boldsymbol{u}}(t) - \varphi_{\nu,\boldsymbol{u}}(t)\right|^2 w(t)\, dt \\ &= \int \Big(\varphi_{\mu,\boldsymbol{u}}(t)\overline{\varphi_{\mu,\boldsymbol{u}}(t)} + \varphi_{\nu,\boldsymbol{u}}(t)\overline{\varphi_{\nu,\boldsymbol{u}}(t)} - 2\Re\big(\varphi_{\mu,\boldsymbol{u}}(t)\overline{\varphi_{\nu,\boldsymbol{u}}(t)}\big)\Big)\, w(t)\, dt.\end{aligned}$$

Using independence of samples and Fubini's theorem,

$$\begin{aligned}\int \varphi_{\mu,\boldsymbol{u}}(t)\overline{\varphi_{\mu,\boldsymbol{u}}(t)} w(t)\, dt &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\mu} \int e^{it\langle \boldsymbol{x}-\boldsymbol{x}', \boldsymbol{u}\rangle} w(t)\, dt \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\mu} k_{\boldsymbol{u}}(\boldsymbol{x}, \boldsymbol{x}').\end{aligned}$$

Similarly,

$$\int \varphi_{\nu,\boldsymbol{u}}(t)\overline{\varphi_{\nu,\boldsymbol{u}}(t)} w(t)\, dt = \mathbb{E}_{\boldsymbol{y},\boldsymbol{y}'\sim\nu} k_{\boldsymbol{u}}(\boldsymbol{y}, \boldsymbol{y}'),$$

and

$$\int \varphi_{\mu,\boldsymbol{u}}(t)\overline{\varphi_{\nu,\boldsymbol{u}}(t)} w(t)\, dt = \mathbb{E}_{\boldsymbol{x}\sim\mu,\, \boldsymbol{y}\sim\nu} k_{\boldsymbol{u}}(\boldsymbol{x}, \boldsymbol{y}).$$

Therefore,

$$\mathrm{EP}_{\boldsymbol{u}}(\mu, \nu) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\mu} k_{\boldsymbol{u}}(\boldsymbol{x}, \boldsymbol{x}') + \mathbb{E}_{\boldsymbol{y},\boldsymbol{y}'\sim\nu} k_{\boldsymbol{u}}(\boldsymbol{y}, \boldsymbol{y}') - 2\mathbb{E}_{\boldsymbol{x}\sim\mu,\, \boldsymbol{y}\sim\nu} k_{\boldsymbol{u}}(\boldsymbol{x}, \boldsymbol{y}),$$

which is exactly $\mathrm{MMD}^2_{k_{\boldsymbol{u}}}(\mu, \nu)$.

Finally, since $k_{\boldsymbol{u}}$ is the Fourier transform of the nonnegative function w, it is positive definite by Bochner's theorem. □

J.3 Linear Alignment

This section contains the analytical solutions in the linear regression case, along with the corresponding proofs.

J.3.1 Reconstruction

Recall Lem. 1:

Lemma. *Consider the self-reconstruction objectives*

$$\mathcal{L}_{\mathrm{dcy}}{}^{x} = \frac{1}{n}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}_x\boldsymbol{V}_x\|_F^2, \qquad \mathcal{L}_{\mathrm{dcy}}{}^{y} = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{Y}\boldsymbol{W}_y\boldsymbol{V}_y\|_F^2. \tag{26}$$

Then, for each domain independently, any minimizer satisfies:

- *The encoder $\boldsymbol{W}_x$ (resp. $\boldsymbol{W}_y$) spans the top-K principal subspace of $\boldsymbol{\Sigma}_{xx}$ (resp. $\boldsymbol{\Sigma}_{yy}$).*
- *The optimal decoder is given by $\boldsymbol{V}_x = \boldsymbol{W}_x^\top$ (resp. $\boldsymbol{V}_y = \boldsymbol{W}_y^\top$) up to a change of basis.*

Proof. We treat the $\boldsymbol{X}$ domain. The $\boldsymbol{Y}$ domain is identical. We omit the x sub- and superscripts for ease of notation, and rewrite the objective

$$\min_{\boldsymbol{W},\boldsymbol{V}} \frac{1}{n}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{V}\|_F^2$$

as

$$\min_{\mathrm{rg}(\boldsymbol{M})\le K} \|\boldsymbol{X} - \boldsymbol{M}\|_F^2,$$

where $\mathrm{rg}(\boldsymbol{M})$ denotes the rank of $\boldsymbol{M}$. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the singular value decomposition of $\boldsymbol{X}$, where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{d_x\times d_x}$ are orthogonal, and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times d_x}$ is diagonal with nonnegative entries. Then, by the Eckart-Young theorem, the optimal solution is given by the truncated SVD of $\boldsymbol{X}$:

$$\boldsymbol{M} = \boldsymbol{U}_K\boldsymbol{\Sigma}_K\boldsymbol{V}_K^\top,$$

where $\boldsymbol{U}_K \in \mathbb{R}^{n\times K}$ and $\boldsymbol{V}_K \in \mathbb{R}^{d_x\times K}$ are the matrices of the top-K left and right singular vectors, and $\boldsymbol{\Sigma}_K \in \mathbb{R}^{K\times K}$ is the diagonal matrix of the top-K singular values. Therefore, any optimal factorization verifies

$$\boldsymbol{X}\boldsymbol{W}\boldsymbol{V} = \boldsymbol{U}_K\boldsymbol{\Sigma}_K\boldsymbol{V}_K^\top.$$

The general solution is given by $\boldsymbol{W} = \boldsymbol{V}_K\boldsymbol{R}$ and $\boldsymbol{V} = \boldsymbol{R}^{-1}\boldsymbol{U}_K^\top$ for any invertible matrix $\boldsymbol{R} \in \mathbb{R}^{K\times K}$. □

J.3.2 Concept Consistency

Lemma (Degeneracy of concept consistency)**.** *In the absence of additional constraints, minimizing $\mathcal{L}_{\mathrm{cont}}$ is ill-posed: for any pair $(\boldsymbol{W}_x, \boldsymbol{W}_y)$, and any orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{K\times K}$, the transformation*

$$\boldsymbol{W}_x \to \boldsymbol{W}_x\boldsymbol{R}, \qquad \boldsymbol{W}_y \to \boldsymbol{W}_y\boldsymbol{R} \tag{27}$$

leaves the objective unchanged. Moreover, the objective admits degenerate solutions (e.g., collapse to zero).

Proof. **Invariance under change of basis.** Let $\boldsymbol{R} \in \mathbb{R}^{K\times K}$ be orthogonal, and define

$$\widetilde{\boldsymbol{W}}_x = \boldsymbol{W}_x\boldsymbol{R}, \qquad \widetilde{\boldsymbol{W}}_y = \boldsymbol{W}_y\boldsymbol{R}.$$

Then the corresponding latent representations satisfy

$$\widetilde{\boldsymbol{Z}}_x = \boldsymbol{X}\widetilde{\boldsymbol{W}}_x = \boldsymbol{Z}_x\boldsymbol{R}, \qquad \widetilde{\boldsymbol{Z}}_y = \boldsymbol{Y}\widetilde{\boldsymbol{W}}_y = \boldsymbol{Z}_y\boldsymbol{R}.$$

Therefore,

$$\begin{aligned}\|\widetilde{\boldsymbol{Z}}_x - \widetilde{\boldsymbol{Z}}_y\|_F^2 &= \|(\boldsymbol{Z}_x - \boldsymbol{Z}_y)\boldsymbol{R}\|_F^2 \\ &= \operatorname{Tr}\big(\boldsymbol{R}^\top (\boldsymbol{Z}_x - \boldsymbol{Z}_y)^\top (\boldsymbol{Z}_x - \boldsymbol{Z}_y)\boldsymbol{R}\big) \\ &= \operatorname{Tr}\big((\boldsymbol{Z}_x - \boldsymbol{Z}_y)^\top (\boldsymbol{Z}_x - \boldsymbol{Z}_y)\big) \\ &= \|\boldsymbol{Z}_x - \boldsymbol{Z}_y\|_F^2.\end{aligned} \tag{28}$$

Degeneracy. Consider $\boldsymbol{W}_x = 0$ and $\boldsymbol{W}_y = 0$. Then $\boldsymbol{Z}_x = \boldsymbol{Z}_y = 0$, and

$$\mathcal{L}_{\text{cont}} = 0, \tag{29}$$

which is the global minimum. More generally, any pair $(\boldsymbol{W}_x, \boldsymbol{W}_y)$ such that $\boldsymbol{X}\boldsymbol{W}_x = \boldsymbol{Y}\boldsymbol{W}_y$ achieves zero loss, regardless of the structure or rank of the representations.

Conclusion. The objective admits infinitely many minimizers due to invariance under invertible transformations, and includes degenerate solutions such as collapse to zero. Therefore, it is ill-posed without additional constraints. □

Lemma (CCA from concept consistency with whitening)**.** *Consider minimizing $\mathcal{L}_{\text{cont}}$ under the whitening constraints*

$$\frac{1}{n}\boldsymbol{Z}_x^\top \boldsymbol{Z}_x = \boldsymbol{I}_K, \qquad \frac{1}{n}\boldsymbol{Z}_y^\top \boldsymbol{Z}_y = \boldsymbol{I}_K. \tag{30}$$

Then the optimal encoders $(\boldsymbol{W}_x, \boldsymbol{W}_y)$ are given by the top-K canonical directions between $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e., they solve the canonical correlation analysis (CCA) problem.

Proof. Recall that

$$\mathcal{L}_{\text{cont}} = \frac{1}{n}\|\boldsymbol{X}\boldsymbol{W}_x - \boldsymbol{Y}\boldsymbol{W}_y\|_F^2. \tag{31}$$

Expanding the Frobenius norm yields

$$\mathcal{L}_{\text{cont}} = \operatorname{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx}\boldsymbol{W}_x) + \operatorname{Tr}(\boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy}\boldsymbol{W}_y) - 2\operatorname{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xy}\boldsymbol{W}_y). \tag{32}$$

Under the whitening constraints

$$\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx}\boldsymbol{W}_x = \boldsymbol{I}_K, \qquad \boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy}\boldsymbol{W}_y = \boldsymbol{I}_K, \tag{33}$$

the first two terms are constant and equal to K. Therefore, minimizing $\mathcal{L}_{\text{cont}}$ is equivalent to maximizing

$$\operatorname{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xy}\boldsymbol{W}_y). \tag{34}$$

subject to the whitening constraints. This is exactly the CCA problem, whose solution is given by the top-K canonical directions between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Closed-form solution. Let $\widetilde{\boldsymbol{W}}_x = \boldsymbol{\Sigma}_{xx}^{\frac{1}{2}}\boldsymbol{W}_x$ and $\widetilde{\boldsymbol{W}}_y = \boldsymbol{\Sigma}_{yy}^{\frac{1}{2}}\boldsymbol{W}_y$. Then the whitening constraints become $\widetilde{\boldsymbol{W}}_x^\top \widetilde{\boldsymbol{W}}_x = \boldsymbol{I}_K$ and $\widetilde{\boldsymbol{W}}_y^\top \widetilde{\boldsymbol{W}}_y = \boldsymbol{I}_K$, and the objective becomes

$$\operatorname{Tr}(\widetilde{\boldsymbol{W}}_x^\top \boldsymbol{M}\widetilde{\boldsymbol{W}}_y), \tag{35}$$

where $\boldsymbol{M} = \boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}$. Let $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the singular value decomposition of $\boldsymbol{M}$. Then the optimal solution is given by $\widetilde{\boldsymbol{W}}_x = \boldsymbol{U}_K$ and $\widetilde{\boldsymbol{W}}_y = \boldsymbol{V}_K$, where $\boldsymbol{U}_K$ and $\boldsymbol{V}_K$ are the matrices of the top-K left and right singular vectors of $\boldsymbol{M}$. Therefore, the optimal encoders are given by $\boldsymbol{W}_x = \boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{U}_K$ and $\boldsymbol{W}_y = \boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}\boldsymbol{V}_K$ up to an arbitrary orthogonal transform $\boldsymbol{R} \in \mathbb{R}^{K\times K}$. □

Lemma (Concept consistency with reconstruction)**.** *Consider the objective*

$$\mathcal{L}(\boldsymbol{W}_x, \boldsymbol{W}_y, \boldsymbol{V}_x, \boldsymbol{V}_y) = \mathcal{L}_{\text{cont}} + \lambda\left(\mathcal{L}_{\text{dcy}}{}^x + \mathcal{L}_{\text{dcy}}{}^y\right). \tag{36}$$

(i) General case. *After optimizing over $(\boldsymbol{V}_x, \boldsymbol{V}_y)$, the objective depends only on the subspaces spanned by $(\boldsymbol{W}_x, \boldsymbol{W}_y)$ and combines:*

- *a consistency term favoring alignment between $\boldsymbol{X}\boldsymbol{W}_x$ and $\boldsymbol{Y}\boldsymbol{W}_y$,*
- *reconstruction terms favoring principal subspaces of $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$.*

However, the problem is ill-posed due to scaling degeneracies in the consistency term.

(ii) Whitening. *If we impose the whitening constraints*

$$\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{W}_x = \boldsymbol{I}_K, \qquad \boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy} \boldsymbol{W}_y = \boldsymbol{I}_K, \tag{37}$$

then:

- *the reconstruction terms become constant and do not influence the encoders,*
- *the objective reduces to $\max \operatorname{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xy} \boldsymbol{W}_y)$, whose solution is given by the top-K canonical directions (CCA),*
- *the optimal decoders are uniquely determined (up to invariances) as the corresponding least-squares readouts.*

In particular, under whitening, the solution is independent of λ.

Proof. We eliminate the decoders and derive the reduced objective.

Step 1: Optimal decoders. For fixed encoders, the reconstruction terms are independent least squares problems. As in the PCA case, the optimal decoders are

$$\boldsymbol{V}_x^* = (\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{W}_x)^{-1} \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx}, \qquad \boldsymbol{V}_y^* = (\boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy} \boldsymbol{W}_y)^{-1} \boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy}, \tag{38}$$

assuming invertibility.

Plugging back, the reconstructions correspond to projections onto the column spaces of $\boldsymbol{X}\boldsymbol{W}_x$ and $\boldsymbol{Y}\boldsymbol{W}_y$, and the losses reduce to

$$\frac{1}{n}\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|_F^2 = \operatorname{Tr}(\boldsymbol{\Sigma}_{xx}) - \operatorname{Tr}(\boldsymbol{P}_x \boldsymbol{\Sigma}_{xx}), \quad \frac{1}{n}\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_F^2 = \operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - \operatorname{Tr}(\boldsymbol{P}_y \boldsymbol{\Sigma}_{yy}), \tag{39}$$

where

$$\boldsymbol{P}_x = \boldsymbol{W}_x (\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{W}_x)^{-1} \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx}, \quad \boldsymbol{P}_y = \boldsymbol{W}_y (\boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy} \boldsymbol{W}_y)^{-1} \boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy}. \tag{40}$$

Step 2: Reduced objective. Up to constants independent of $(\boldsymbol{W}_x, \boldsymbol{W}_y)$, the objective becomes

$$\mathcal{L}_{\mathrm{red}}(\boldsymbol{W}_x, \boldsymbol{W}_y) = \frac{1}{n}\|\boldsymbol{X}\boldsymbol{W}_x - \boldsymbol{Y}\boldsymbol{W}_y\|_F^2 - \lambda \left(\operatorname{Tr}(\boldsymbol{P}_x \boldsymbol{\Sigma}_{xx}) + \operatorname{Tr}(\boldsymbol{P}_y \boldsymbol{\Sigma}_{yy})\right). \tag{41}$$

Expanding the first term,

$$\frac{1}{n}\|\boldsymbol{X}\boldsymbol{W}_x - \boldsymbol{Y}\boldsymbol{W}_y\|_F^2 = \operatorname{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{W}_x) + \operatorname{Tr}(\boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy} \boldsymbol{W}_y) - 2 \operatorname{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xy} \boldsymbol{W}_y). \tag{42}$$

Step 3: Variational characterization. We characterize the optimal subspaces without differentiating the reduced objective.

First, observe that the reconstruction terms depend only on the projection matrices $\boldsymbol{P}_x$ and $\boldsymbol{P}_y$, and therefore only on the column spaces of $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$. As in the PCA case, maximizing

$$\operatorname{Tr}(\boldsymbol{P}_x \boldsymbol{\Sigma}_{xx})$$

selects the top-K eigenspace of $\boldsymbol{\Sigma}_{xx}$, and similarly for $\boldsymbol{\Sigma}_{yy}$.

On the other hand, similarly to the pure CCA case, the consistency term $\|\boldsymbol{X}\boldsymbol{W}_x - \boldsymbol{Y}\boldsymbol{W}_y\|_F^2$ is ill-posed without additional constraints: it is invariant to joint rescaling of $(\boldsymbol{W}_x, \boldsymbol{W}_y)$ and admits degenerate solutions (e.g., $\boldsymbol{W}_x = \boldsymbol{W}_y = 0$).

To remove this degeneracy, we impose whitening constraints

$$\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{W}_x = \boldsymbol{I}_K, \qquad \boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy} \boldsymbol{W}_y = \boldsymbol{I}_K.$$

Under these constraints, minimizing the consistency term is equivalent to maximizing

$$\mathrm{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xy} \boldsymbol{W}_y),$$

which yields the canonical directions, as in CCA.

Importantly, under whitening, the reconstruction terms become constant:

$$\mathrm{Tr}(\boldsymbol{P}_x \boldsymbol{\Sigma}_{xx}) = \mathrm{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{W}_x) = K,$$

and similarly for $\boldsymbol{Y}$. Therefore, the reconstruction objective no longer influences the solution, regardless of the value of λ. Its only effect is to identify the decoders.

As a result, the combined objective reduces exactly to CCA, and the optimal encoders are given by the top-K canonical directions between $\boldsymbol{X}$ and $\boldsymbol{Y}$. □

J.3.3 TRANSLATION

Lemma (Translation and reduced-rank regression)**.** *Consider the (one-directional) translation objective*

$$\mathcal{L}_{\mathrm{tr}}{}^{x\to y}(\boldsymbol{W}_x, \boldsymbol{V}_y) = \frac{1}{n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}_x\boldsymbol{V}_y\|_F^2. \tag{43}$$

Then minimizing $\mathcal{L}_{\mathrm{tr}}{}^{x\to y}$ over $(\boldsymbol{W}_x, \boldsymbol{V}_y)$ is equivalent to RRR of $\boldsymbol{Y}$ onto $\boldsymbol{X}$ with rank constraint K. The factorization $(\boldsymbol{W}_x, \boldsymbol{V}_y)$ is not identifiable: only the product $\boldsymbol{W}_x\boldsymbol{V}_y$ is. One such factorization recovers the CCA.

Proof. Let $\boldsymbol{B} = \boldsymbol{W}_x\boldsymbol{V}_y$. Then the problem reduces to

$$\min_{\mathrm{rg}(\boldsymbol{B})\le K} \frac{1}{n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2, \tag{44}$$

which is exactly RRR of $\boldsymbol{Y}$ onto $\boldsymbol{X}$ with rank constraint K. The optimal unconstrained $\boldsymbol{B}$ is given by $\boldsymbol{B}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$.

We rewrite the problem as

$$\min_{\mathrm{rg}(\boldsymbol{B})\le K} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_{\mathrm{OLS}}\|_F^2 + \|\boldsymbol{X}(\boldsymbol{B} - \boldsymbol{B}_{\mathrm{OLS}})\|_F^2, \tag{45}$$

where we used the Pythagorean theorem for least squares problems. The first term is constant. Using the definition of $\boldsymbol{\Sigma}_{xx}$, we have

$$\begin{aligned}\|\boldsymbol{X}(\boldsymbol{B} - \boldsymbol{B}_{\mathrm{OLS}})\|_F^2 &= n\|\boldsymbol{\Sigma}_{xx}^{\frac{1}{2}}(\boldsymbol{B} - \boldsymbol{B}_{\mathrm{OLS}})\|_F^2 \\ &= n\|\boldsymbol{C} - \boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xy}\|_F^2.\end{aligned}$$

with $\boldsymbol{C} \coloneqq \boldsymbol{\Sigma}_{xx}^{\frac{1}{2}}\boldsymbol{B}$. We want to make the CCA appear. The only thing missing is the $\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}$ term. We introduce a whitening step: $\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xy} = \boldsymbol{M}\boldsymbol{\Sigma}_{yy}^{\frac{1}{2}}$, where $\boldsymbol{M} = \boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}$, and $\boldsymbol{C} = \boldsymbol{D}\boldsymbol{\Sigma}_{yy}^{\frac{1}{2}}$. Then, the objective becomes

$$\min_{\mathrm{rg}(\boldsymbol{D})\le K} \|\boldsymbol{D} - \boldsymbol{M}\|_{F,\boldsymbol{\Sigma}_{yy}}^2 \tag{46}$$

Conveniently, $\boldsymbol{M}$ is exactly the same as in CCA. We can express the solution in terms of the truncated SVD $\boldsymbol{U}_K\boldsymbol{\Sigma}_K\boldsymbol{V}_K^\top + \epsilon$ of $\boldsymbol{M}$:

$$\boldsymbol{B} = \boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{U}_K\boldsymbol{\Sigma}_K\boldsymbol{V}_K^\top\boldsymbol{\Sigma}_{yy}^{\frac{1}{2}} \tag{47}$$

With this decomposition, we can identify a particular factorization $\boldsymbol{W}_x = \boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{U}_K$ and $\boldsymbol{V}_y = \boldsymbol{\Sigma}_K\boldsymbol{V}_K^\top\boldsymbol{\Sigma}_{yy}^{\frac{1}{2}}$. This factorization recovers the CCA solution for the encoders and the corresponding least-squares decoder. This factorization is not unique. □

Lemma (Translation with reconstruction)**.** *Under the combined objective, the optimal encoders are again given by the CCA, while the decoders are given by an interpolation between the solutions of Lem. 5 and Lem. 4.*

Proof. As before, we start by eliminating the decoders given $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$. Due to the combined effect of self-reconstruction and translation, we get, assuming invertibility,

$$\boldsymbol{V}_x^* = (\boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy} \boldsymbol{W}_y + \lambda \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{W}_x)^{-1} (\boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{xy}^\top + \lambda \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx}) \tag{48}$$

and similarly for $\boldsymbol{V}_y^*$.

The next step is to assume whitening. We get

$$\boldsymbol{V}_x^* = \frac{1}{n(1+\lambda)} \left(\boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{xy}^\top + \lambda \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx} \right), \tag{49}$$

and similarly for $\boldsymbol{V}_y^*$. Plugging back, and thanks to whitening, the objective reduces to

$$\max_{\boldsymbol{W}_x, \boldsymbol{W}_y} \operatorname{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xy} \boldsymbol{W}_y) + \lambda \left(\operatorname{Tr}(\boldsymbol{W}_x^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{W}_x) + \operatorname{Tr}(\boldsymbol{W}_y^\top \boldsymbol{\Sigma}_{yy} \boldsymbol{W}_y) \right), \tag{50}$$

The term in λ becomes constant again. We add a whitening step to the $\boldsymbol{W}$ to make the CCA appear. Let $\widetilde{\boldsymbol{W}}_x = \boldsymbol{\Sigma}_{xx}^{\frac{1}{2}} \boldsymbol{W}_x$ and $\widetilde{\boldsymbol{W}}_y = \boldsymbol{\Sigma}_{yy}^{\frac{1}{2}} \boldsymbol{W}_y$, and define $\boldsymbol{M} = \boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}$. The objective reduces to

$$\max_{\widetilde{\boldsymbol{W}}_x, \widetilde{\boldsymbol{W}}_y} \operatorname{Tr}(\widetilde{\boldsymbol{W}}_x^\top \boldsymbol{M} \widetilde{\boldsymbol{W}}_y), \tag{51}$$

and we again get the CCA solution for the encoders. The difference is that now, the decoders interpolate between self-projections and cross-projections, with a weighting controlled by λ. □

J.3.4 Cycle consistency

Lemma 7. *Under the $\mathcal{L}_{\text{cy}}$ objective, the encoders and decoders recover the PCA.*

Proof. Let $\boldsymbol{A} \coloneqq \boldsymbol{W}_x \boldsymbol{V}_y$ and $\boldsymbol{B} \coloneqq \boldsymbol{W}_y \boldsymbol{V}_x$. Let $\boldsymbol{M} \coloneqq \boldsymbol{A}\boldsymbol{B}$ and $\boldsymbol{N} \coloneqq \boldsymbol{B}\boldsymbol{A}$. Then the cycle consistency objective can be rewritten as

$$\min_{\operatorname{rg}(\boldsymbol{A}) \leq K, \operatorname{rg}(\boldsymbol{B}) \leq K} \frac{1}{n} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{M}\|_F^2 + \frac{1}{n} \|\boldsymbol{Y} - \boldsymbol{Y}\boldsymbol{N}\|_F^2. \tag{52}$$

Without the coupling constraint between $\boldsymbol{M}$ and $\boldsymbol{N}$, the problem would reduce to two independent PCA problems, whose solutions are given by the top-K principal subspaces of $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$: $\boldsymbol{M}^* = \boldsymbol{P}_x \boldsymbol{P}_x^\top$ and $\boldsymbol{N}^* = \boldsymbol{P}_y \boldsymbol{P}_y^\top$. This gives a lower bound on the objective. We now show that this lower bound is achievable, and therefore optimal. Let $\boldsymbol{R} \in \mathbb{R}^{K \times K}$ be an arbitrary invertible matrix, and define

$$\boldsymbol{A} = \boldsymbol{P}_x \boldsymbol{R} \boldsymbol{P}_y^\top, \qquad \boldsymbol{B} = \boldsymbol{P}_y \boldsymbol{R}^{-1} \boldsymbol{P}_x^\top. \tag{53}$$

Then $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{B} = \boldsymbol{P}_x \boldsymbol{R} \boldsymbol{P}_y^\top \boldsymbol{P}_y \boldsymbol{R}^{-1} \boldsymbol{P}_x^\top = \boldsymbol{P}_x \boldsymbol{P}_x^\top = \boldsymbol{M}^*$, and similarly $\boldsymbol{N} = \boldsymbol{P}_y \boldsymbol{P}_y^\top = \boldsymbol{N}^*$. □

K Additional results and ablations

Note that due to the normalisation of some metrics in Appendices B and C, some values can go above 1.

Regularization	M_{extract} (↑)	M_{tr} (↑)	M_{cont} (↑)
$\mathcal{L}_{\text{dcy}}$	0.99	0.00	0.00
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}$	0.99	0.99	0.00
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}$	1.00	0.98	0.99
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}}$	0.99	0.99	0.00
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	1.00	0.98	0.99

Table A5: In the **synthetic DGP experiment**, the P_{cont} and P_{tr} duality **does not** hold. The instance-wise vs distributional duality **does** hold.

	InterVench— Language								
	Latent Space					Concept Space			
Regularization	R^2	Faith	SP	Unlearn	TPP	SP	Unlearn	TPP	M_{extract}
$\mathcal{L}_{\text{dcy}}$	0.6243	0.9532	0.4663	0.9498	0.9827	0.4133	0.9075	0.5030	0.7250
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}$	0.6746	0.9606	0.4631	0.9733	1.0026	0.3914	0.9897	0.9954	0.8063
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}}$	0.0003	0.1701	0.4645	0.9595	0.9893	0.0000	0.0000	0.0000	0.3230
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}$	0.7298	0.9854	0.4589	0.9806	1.0061	0.1850	0.9513	1.3475	0.8306
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.7166	0.9825	0.4683	0.9865	1.0048	0.2104	0.8391	0.9188	0.7659
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}} + \mathcal{L}_{\text{cont}}$	0.6983	0.9621	0.4570	0.9736	0.9984	0.3138	0.9749	1.1053	0.8104
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.4412	0.9094	0.4265	0.9689	0.9950	0.2226	1.1094	1.3087	0.7977
mixed (1 in 10)	0.7111	0.9868	0.4437	0.9840	1.0096	0.3593	0.9399	0.9942	0.8036
mixed (1 in 100)	0.7078	0.9868	0.4521	0.9748	0.9990	0.3591	0.9276	0.8923	0.7874
mixed (1 in 1000)	0.7120	0.9841	0.4507	0.9676	0.9905	0.3562	0.9903	1.0059	0.8072

Table A6: Detailed *InterVench* metrics on **language embeddings** under different regularization regimes.

	InterVenchA— Language								
	Translation						Concept Consistency		
Regularization	R^2	Faith	SP	Unlearn	TPP	M_{tr}	SP	Unlearn	TPP	M_{cont}
$\mathcal{L}_{\text{dcy}}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}$	0.4897	0.9051	0.4348	0.7009	0.7718	0.6605	0.1821	0.3131	0.4306	0.3086
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}}$	0.0000	0.0000	0.1891	0.3409	0.3864	0.1833	0.0000	0.0000	0.0000	0.0000
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}$	0.0000	0.0000	0.2233	0.4710	0.4660	0.2321	0.2783	0.8061	0.9278	0.6707
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1177	0.3549	0.4103	0.2943
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}} + \mathcal{L}_{\text{cont}}$	0.5863	0.9410	0.4150	0.8174	0.9075	0.7334	0.2839	0.8740	0.9548	0.7042
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.0000	0.0000	0.1373	0.1278	0.1388	0.0808	0.0000	0.0000	0.0000	0.0000
mixed (1 in 10)	0.6307	0.9528	0.3929	0.7120	0.7611	0.6899	0.3569	0.9252	0.9615	0.7479
mixed (1 in 100)	0.6272	0.9546	0.4068	0.7366	0.7961	0.7043	0.3548	0.9350	0.9411	0.7436
mixed (1 in 1000)	0.6310	0.9570	0.3895	0.8033	0.8425	0.7247	0.3541	0.9136	0.9858	0.7512

Table A7: Detailed *InterVenchA* metrics on **language embeddings** under different regularization regimes.

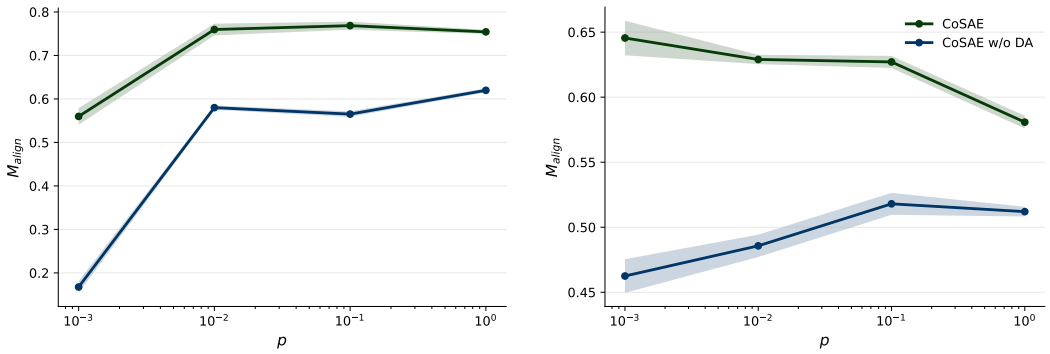


Figure A3: Ablation of the effect of distributional alignment when computing element-wise alignment losses on a proportion p of the training data—used as anchor points. CoSAE is the mixed regime described in Section 4.3. "CoSAE w/o DA" is the same procedure but without Distributional Alignment regularization. This plot shows results for SAEs trained on visual (*left*) and textual (*right*) embeddings. Plain dots report the mean value over 10 runs, shaded regions indicate $\pm$std.

	InterVench— Vision								
	Latent Space					Concept Space			
Regularization	R^2	Faith	SP	Unlearn	TPP	SP	Unlearn	TPP	M_{extract}
$\mathcal{L}_{\text{dcy}}$	0.8385	1.0707	0.7788	0.9438	0.9820	0.7572	0.9442	0.9872	0.9128
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}$	0.7228	1.0896	0.7715	0.9237	0.9945	0.7772	0.9475	0.9859	0.9016
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}}$	0.8222	1.0426	0.7282	0.9570	0.9931	0.6843	0.9116	0.9751	0.8893
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}$	0.8312	1.0598	0.7799	0.9417	0.9954	0.0000	0.0000	0.0000	0.5760
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.8228	1.0605	0.7776	0.9095	0.9932	0.7641	0.9185	0.9921	0.9048
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}} + \mathcal{L}_{\text{cont}}$	0.7160	1.0945	0.7721	0.9793	0.9964	0.3230	0.8885	0.9179	0.8360
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.7352	1.0582	0.7737	0.9459	0.9853	0.7604	0.8935	0.9889	0.8926
mixed (1 in 10)	0.7124	1.1245	0.7623	0.9615	0.9951	0.7630	0.9555	1.0011	0.9094
mixed (1 in 100)	0.6734	1.0819	0.7675	0.9385	0.9944	0.7541	0.9549	0.9815	0.8933
mixed (1 in 1000)	0.6971	1.0931	0.7698	0.9460	0.9988	0.7544	0.9532	0.9901	0.9003

Table A8: Detailed *InterVench* metrics on **vision embeddings** under different regularization regimes.

	InterVenchA— Vision									
	Translation						Concept Consistency			
Regularization	R^2	Faith	SP	Unlearn	TPP	M_{tr}	SP	Unlearn	TPP	M_{cont}
$\mathcal{L}_{\text{dcy}}$	0.0000	0.6862	0.0000	0.0000	0.0000	0.1372	0.0000	0.0000	0.0000	0.0000
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}$	0.4660	1.1853	0.5350	0.8804	0.5750	0.7283	0.1896	0.2415	0.2211	0.2174
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}}$	0.0000	0.7813	0.2952	0.8119	0.7327	0.5242	0.0000	0.0000	0.0000	0.0000
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}$	0.0000	0.6998	0.0000	0.0000	0.0000	0.1400	0.4598	0.8263	0.5209	0.6023
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.0000	0.7010	0.0000	0.0000	0.0000	0.1402	0.0000	0.0000	0.0000	0.0000
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}} + \mathcal{L}_{\text{cont}}$	0.4705	1.1824	0.5451	0.9235	0.5440	0.7331	0.4694	0.8765	0.5102	0.6187
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.0000	0.7141	0.0000	0.0000	0.0000	0.1428	0.0000	0.0000	0.0000	0.0000
mixed (1 in 10)	0.5243	1.1509	0.5221	0.9225	0.5228	0.7285	0.7267	0.9349	0.7786	0.8134
mixed (1 in 100)	0.4224	1.0931	0.6267	0.8847	0.6318	0.7317	0.7152	0.9312	0.7373	0.7946
mixed (1 in 1000)	0.4056	0.9455	0.4667	0.6016	0.5905	0.6020	0.3528	0.8211	0.4584	0.5442

Table A9: Detailed *InterVenchA* metrics on **vision embeddings** under different regularization regimes.

Regularization	M_{extract} (↑)	M_{tr} (↑)	M_{cont} (↑)
$\mathcal{L}_{\text{dcy}}$	**0.8956**	0.0000	0.0000
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}$	0.7391	0.2974	0.2377
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}$	0.8786	0.0000	0.3271
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}}$	0.7658	0.000	0.0000
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.8879	0.000	0.0000
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}} + \mathcal{L}_{\text{cont}}$	0.7496	0.2765	0.3454
$\mathcal{L}_{\text{dcy}} + \mathcal{L}_{\text{tr}}^{\text{dist}} + \mathcal{L}_{\text{cont}}^{\text{dist}}$	0.8889	0.000	0.0000
mixed (1 in 1000)	0.7814	**0.5053**	**0.5535**

Table A10: **Ablation studies on the effect of regularization terms.** SAE quality on **multimodal embeddings** under different regularization regimes.