REVEALING AND EMPLOYING OBLIQUE SEMANTIC GEOMETRY IN VISION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision foundation models learn rich representation spaces that encode a wide range of visual semantics. Yet, understanding the organization of this semantic structure often relies on predefined concepts, supervision, or training additional models. Thus, it remains unclear which semantic factors emerge directly from the representation distribution itself, and how these factors relate to one another. In this work, we first examine, in a controlled manner, the relations among semantic attributes in known datasets. We observe that similar geometry emerges across both vision-only and vision-language encoders, suggesting that this structure is largely data-driven. Since semantic directions are generally non-orthogonal, we develop a geometric framework for analyzing their organization. To recover the underlying factors without labels or semantic supervision, we seek a complete linear decomposition that separates distinct sources. Independent Component Analysis (ICA) provides a natural construction for this purpose by separating statistically independent, non-Gaussian factors through a complete and invertible transformation. Its dual representations form a biorthogonal pair, enabling precise feature disentanglement while exposing the geometry between semantic factors. We show across several measures and datasets that ICA recovers semantic factors more effectively than PCA and SAEs, with less identifiable sources being harder to recover. We further find that geometrically aligned factors form coherent semantic groups, revealing structure from fine-grained attributes to broader concepts. Finally, we employ this structure through controlled interventions, demonstrating its utility for image retrieval and mitigating spurious correlations.

1 INTRODUCTION

Vision foundation models map images to compact representations that support a wide range of downstream tasks (Radford et al., 2021; Oquab et al., 2023). Their success suggests that these spaces contain rich semantic structure, but the form of this structure remains only partly understood. A common view is that individual semantic properties correspond to linear directions in representation space. This idea is formalized by the *Linear Representation Hypothesis* (LRH) (Park et al., 2024), and related evidence has been found across language and vision representations (Zimmermann et al., 2021; Sonthalia et al., 2026). Recent work has also shown that richer semantic relations, such as categorical and hierarchical structure, can have a geometric form (Park et al., 2025). However, it remains unclear whether the geometry between semantic factors carries meaningful structure.

Most existing approaches for identifying semantic directions follow one of two routes. Some begin with predefined concepts, specified through labels, attributes, text descriptions, or semantic dictionaries (Kim et al., 2018; Koh et al., 2020; Bhalla et al., 2024). These methods can locate known concepts in the representation space, but cannot reveal semantic factors that were not specified in advance. Other approaches, such as sparse dictionary learning, discover concepts directly from model activations (Li et al., 2025; Fel et al., 2026b). While these methods remove the need for predefined semantics, they typically learn overcomplete dictionaries, as insufficient dictionary width can distort correlated features, while their explicit sparsity prior can cause feature absorption when the underlying features are hierarchical (Chanin et al., 2025a;b).

This motivates a direct analysis of the pretrained representation distribution without imposing an explicit sparsity constraint. ICA (Comon, 1994; Hyvärinen & Oja, 2000) provides a natural construc-

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.

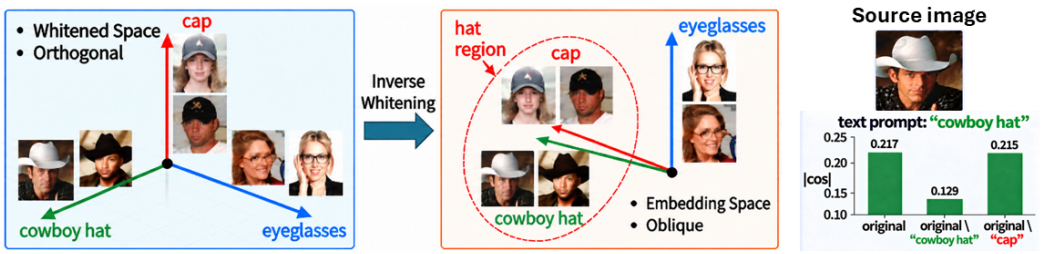


Figure 1: **Oblique biorthogonal structure and precise semantic intervention.** ICA yields orthogonal semantic coordinates in the whitened space and inverse whitening maps them to an oblique basis in the original embedding space. In this native space, semantically related factors such as *cap* and *cowboy hat* become geometrically close. Despite this proximity, the coordinates remain independently controllable: removing the *cowboy hat* coordinate strongly reduces $\cos(\text{image}, \text{“cowboy hat”})$, while removing the nearby *cap* coordinate leaves it nearly unchanged.

tion by recovering statistically independent, non-Gaussian components through an invertible linear transformation, yielding a complete coordinate system without training an auxiliary dictionary. Prior work has shown that ICA components can correspond to interpretable semantics in pretrained word and image embeddings (Yamagiwa et al., 2023), as well as in language-model activations (Liu & Han, 2026). We build on these observations, but move beyond individual components to study how the recovered semantic factors are geometrically organized in the original embedding space.

ICA naturally admits a biorthogonal analysis-synthesis view: semantic factors are separated along orthogonal coordinates in the whitened space, while their corresponding directions in the original embedding space retain geometric correlations (see Figure 1). These native-space directions therefore form an *oblique coordinate system*, whose pairwise geometry provides a natural way to characterize relations between the recovered factors. Crucially, we find that this geometry is semantically structured: directions associated with related concepts become more aligned and organize into coherent semantic groups. These findings suggest a simple view of semantic organization. Individual coordinates separate factors, while their pairwise geometry captures the relationships among them.

This oblique system is also operational. The invertible transformation allows modified ICA coordinates to be mapped back exactly to the original representation space. The recovered structure naturally supports interventions at multiple semantic scales, from individual factors to coherent groups of related concepts. We study these interventions across vision and vision-language foundation models on several distinct datasets. We demonstrate their utility for controlled image retrieval and mitigation of spurious correlations, and further show that the recovered semantic structure reveals fine-grained factors beyond those captured by existing annotations.

Our main contributions are threefold:

- **An oblique representation framework.** We propose an oblique geometric model of representation spaces through a biorthogonal coordinate system. This formulation separates distinct semantic factors while preserving meaningful structure in their pairwise geometry.
- **A complete and data-driven semantic system.** We show that ICA recovers a complete and invertible semantic coordinate system directly from frozen representations. The geometry of the directions is itself semantically structured, organizing related factors into coherent higher-level groups.
- **Operational fine-to-coarse semantic control.** The recovered system supports direct interventions on individual and groups of factors, enabling semantic control at multiple levels of granularity. We demonstrate its utility for controlled image retrieval, mitigation of spurious correlations, and for finding fine-grained concepts beyond existing annotations.

2 Related Work

Linear semantic structure of representation spaces. A large body of work studies whether semantic information is represented linearly in learned representations. In language models, the *Linear*

Representation Hypothesis (LRH) formalizes the view that high-level concepts are represented as linear directions (Park et al., 2024). Related theoretical work on contrastive representation learning shows that learned representations can recover latent factors of the data-generating process under suitable assumptions (Zimmermann et al., 2021). More recent work extends the linear view beyond binary concepts, showing that categorical concepts can form structured polytopes and that semantic hierarchies can be reflected geometrically (Park et al., 2025; Fel et al., 2026b). In vision, Sonthalia et al. (2026) show linear structure for continuous properties such as age, pose, and object count. Together, these works show that semantic concepts can exhibit meaningful linear structure across language and vision. In this work we focus on whether semantic factors recovered from the representation distribution form a coherent geometric organization.

Concept discovery. A broad line of work seeks to expose semantic concepts in learned representations. In vision, concept-based interpretability methods relate internal representations to human-defined semantics through labeled units, concept directions, or explicit concept representations (Bau et al., 2017; Kim et al., 2018; Koh et al., 2020; Espinosa Zarlenga et al., 2022). Other vision methods discover concepts directly from activations through clustering or factorization (Ghorbani et al., 2019; Fel et al., 2023). In vision-language models, semantic decomposition of CLIP (Radford et al., 2021) representations has been performed using predefined textual concept dictionaries (Bhalla et al., 2024). More recent concept-based models recover finer-grained and hierarchical concepts from coarse semantic supervision (Hill et al., 2026b;a). These methods either rely on human-defined semantics or use them to organize and interpret the discovered concepts. We instead recover semantic factors without supervision, directly from the pre-trained embedding distribution.

Sparse autoencoders and learned feature dictionaries. Sparse autoencoders (SAEs) have become a prominent approach for unsupervised feature discovery in language models, reconstructing activations through sparse, typically overcomplete dictionaries (Huben et al., 2024). More recently, these ideas have been extended to vision and vision-language representations, where learned features reveal interpretable concepts and non-trivial geometric organization (Papadimitriou et al., 2025; Li et al., 2025). In particular, it has been shown that visual concepts may be represented by non-orthogonal directions or low-dimensional manifolds rather than isolated axes (Fel et al., 2026b;a). At the same time, feature absorption and correlations between underlying factors can affect the recovered decomposition (Chanin et al., 2025b;a). Unlike these learned, typically overcomplete representations, our approach requires no additional training and yields a complete, invertible coordinate system without imposing an explicit sparsity constraint.

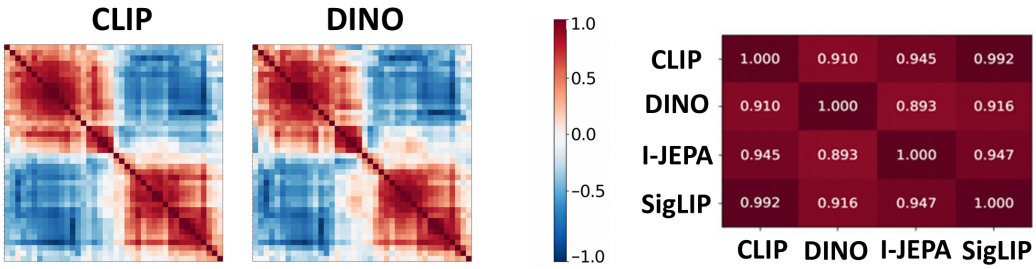
ICA for representation interpretability. Independent Component Analysis (ICA) provides a classical framework for recovering statistically independent factors through a linear transform (Comon, 1994; Hyvärinen & Oja, 2000). In pretrained text and image embeddings, Yamagiwa et al. (2023) show that ICA recovers interpretable semantic components that recur across models and modalities. Follow-up work on text embeddings studies relations among these components, either by organizing them according to semantic continuity (Yamagiwa et al., 2024) or by showing that residual higher-order dependencies between components carry semantic information (Oyama et al., 2024). More recently, ICA has also been applied to language-model activations, recovering interpretable directions without training any dictionary (Liu & Han, 2026). These works establish the semantic meaning of ICA components and relations between their activations. We study the geometry of the corresponding directions after mapping them back to the model's native embedding space, and analyze the geometry between the independently recovered semantic factors.

3 Oblique Semantic Coordinate System

We construct a complete semantic coordinate system from existing representation spaces. Our starting point is the observation that different foundation models often organize semantic properties in similar ways (Figure 2). We use this structure to motivate the recovery of latent factors directly from the embedding distribution and to analyze how they are organized in the representation space.

3.1 Shared Semantic Structure Across Encoders

A growing body of work suggests that independently trained models can develop similar representational structure, an idea formalized by the *Platonic Representation Hypothesis* (Huh et al.,



(a) **CelebA:** Representational Similarity Matrices (RSMs). (b) **MSCOCO:** Correlation between RSMs.

Figure 2: **Shared semantic structure across encoders.** Pairwise relations between feature directions (CelebA attributes and MSCOCO classes) are computed independently for each encoder using Equation (1). Despite differences in training objectives the relative organization of the features is highly consistent across encoders. Complementary results are presented in Figure 8.

2024). The extent of this convergence remains debated, particularly under broader, large-scale evaluations (Koepke et al., 2026). Here, we begin with a simple observation: across different vision encoders, known semantic attributes exhibit remarkably similar relative geometry.

For each semantic factor k and encoder m, let $v_k^{(m)}$ denote a linear reference direction in that representation space. We use the difference between the mean embeddings of positive and negative examples (Sonthalia et al., 2026). We then characterize the relations among factors within each encoder by their pairwise cosine similarities:

$$v_k^{(m)} = \mu_{k,+}^{(m)} - \mu_{k,-}^{(m)}, \qquad G_{ij}^{(m)} = \cos\left(v_i^{(m)}, v_j^{(m)}\right). \tag{1}$$

The directions $v_k^{(m)}$ cannot be compared directly across encoders because they live in different representation spaces. The relation matrices $G^{(m)}$, however, capture how the same semantic factors are organized within each space and can therefore be compared across models. As illustrated in Figure 2, these relation patterns are strongly preserved across different vision encoders; full experimental details and quantitative results are provided in Appendix D.1.

We propose that semantic structure in representation space is organized by a **complete set of latent factors whose directions are generally non-orthogonal**. The factors themselves are distinct, while the geometry between their directions captures semantic relations. Motivated by prior results showing that learned representations can recover latent factors of the data-generating process (Zimmermann et al., 2021), we next seek such a coordinate system directly from the embedding distribution.

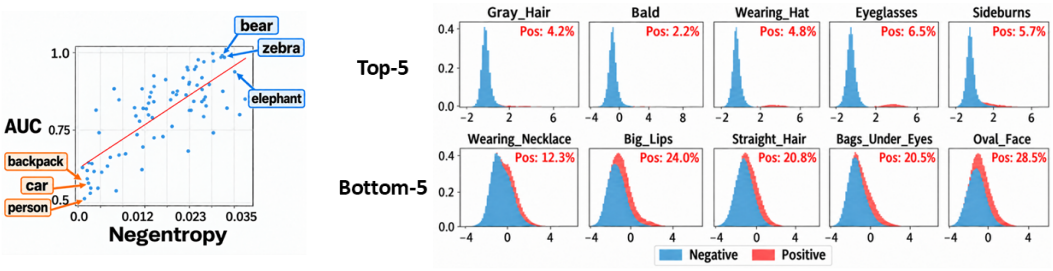
3.2 Recovering Semantic Factors from the Embedding Distribution

Recovering the latent factors directly from the embedding distribution requires relying solely on their statistical signatures. Semantic features can manifest differently in this distribution. Some vary smoothly across the population, such as age or pose, while others are more *presence-like*: an object, accessory, background, or artifact that appears only in a subset of images. Such selective factors can induce strongly non-Gaussian projections, making non-Gaussianity a useful signal for their discovery. We quantify this relationship in Figure 3, where stronger non-Gaussianity is associated with more reliable factor recovery.

ICA recovers statistically independent, non-Gaussian factors. Given embeddings $z_n \in \mathbb{R}^D$, we whiten and rotate them to obtain ICA coordinates c_n:

$$z_n = f(x_n) \in \mathbb{R}^D, \qquad y_n = P(z_n - \mu), \qquad c_n = Ry_n. \tag{2}$$

Here, μ is the empirical mean, P whitens the embeddings, and R is the ICA rotation. The exact whitening and ICA construction is detailed in Appendix A. We retain all D components, so the transformation is complete and invertible. No labels, captions, semantic vocabulary, or additional model training are required. Prior work has shown that individual ICA components can capture interpretable semantics in text and image embeddings (Yamagiwa et al., 2023), as well as in



(a) **MSCOCO** (b) **CelebA:** projection distributions of representative attributes.

Figure 3: **Statistical properties of ICA components.** Selective, presence-like factors tend to induce more non-Gaussian projections and are better captured by ICA, while continuous or entangled factors are more Gaussian-like and harder to isolate. (a) Higher-negentropy factors achieve higher classification AUC along their closest ICA direction. (b) Projection distributions for representative high- and low-negentropy attributes. Complementary results are shown in Figure 9.

language-model activations (Liu & Han, 2026). We similarly find that many recovered components in vision foundation models correspond to coherent semantic factors, including objects, attributes, backgrounds, and visual styles; see Figure 4 for qualitative examples. Beyond the semantics of individual components, we are also interested in whether the recovered factors exhibit meaningful geometric relations in the original embedding space.

3.3 The Oblique Geometry of Semantic Factors

ICA recovers distinct factors in the whitened space, but this space does not reveal geometric relations between them: by construction, the ICA directions are orthogonal. To study how the recovered factors relate, we map them back to the model's native representation space. Inverting Equation (2):

$$z_n = \mu + Ac_n, \qquad A = P^{-1}R^\top = [a_1, \ldots, a_D], \qquad g_{ij} = \cos(a_i, a_j). \tag{3}$$

Each a_i is the native-space direction associated with coordinate c_i; changing that coordinate moves the representation along a_i. This yields the dual geometry illustrated in Figure 1: the ICA coordinates are orthogonal in the whitened space, while their corresponding native-space directions a_i are generally non-orthogonal. These directions therefore form an *oblique coordinate system*, whose pairwise similarities g_{ij} characterize the geometry between recovered factors.

Importantly, obliqueness itself is expected: any orthogonal basis can become non-orthogonal after inverse whitening. Our key finding is that the resulting ICA geometry is not arbitrary. Semantically related factors become more aligned in the native space and form coherent, distinct groups. We quantify this structure and compare it against alternative controls in Section 4. The full algebraic form of this dual geometry, including the native Gram matrix and the distinction between measurement and intervention directions, is given in Appendix B.

3.4 Operational Fine-to-Coarse Semantics

The recovered coordinate system is not only informative, it can also be operated on directly. Because the transformation is complete and invertible, we can modify selected factors in the whitened space and map the result exactly back to the native representation (Δ specifies which semantic coordinates are modified and by how much):

$$c' = c + \Delta, \qquad z' = \mu + Ac' = z + A\Delta. \tag{4}$$

This provides two natural levels of intervention. When Δ is supported on a single coordinate, the representation is changed along one recovered semantic factor. When it is supported on a group of geometrically related coordinates, the intervention acts jointly on a broader semantic family. The oblique geometry therefore provides not only relations between factors, but also a principled way to move from fine-grained to broader semantic control. We evaluate these interventions in the next section and demonstrate their effectiveness across several practical applications.

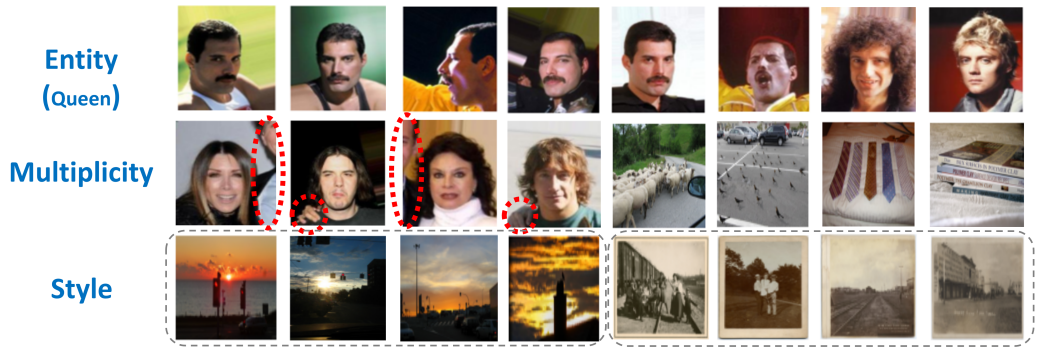


Figure 4: **Examples of semantic sources revealed by ICA.** Extreme activations along selected ICA coordinates expose coherent sources at different semantic levels, including *entity*, *multiplicity*, and *style*. Images within each group share the corresponding recovered factor, with red markers highlighting the additional person in the multiplicity examples. The displayed images are from CLIP and DINO, see more examples in Figure 11.

4 Experiments

We organize the experiments to progressively examine the semantic structure of the representation spaces. We first show that semantic factors exhibit a shared, non-orthogonal structure across vision foundation models. We then examine how these factors can be recovered directly from the unlabeled embedding distribution and study the structural relations among them. Finally, we demonstrate the practical use of the resulting coordinate system through controlled interventions.

Datasets. We primarily use two complementary datasets. **CelebA** (Liu et al., 2015) provides a controlled setting with 40 binary facial attributes and repeated identities. **MSCOCO** (Lin et al., 2014) provides a substantially more diverse setting, containing natural scenes with multiple objects, backgrounds, interactions, and visual compositions. Together, the two datasets allow us to study semantic structure both in a relatively controlled domain and in general natural imagery.

Encoders. We study frozen representations from three vision foundation models with different training objectives: **CLIP** (Radford et al., 2021), trained with image-text contrastive supervision; **DINOv2** (Oquab et al., 2023), trained using image-only self-supervision; and **I-JEPA** (Assran et al., 2023), trained using predictive representation learning. For the cross-encoder analysis in Section 4.1, we additionally include SigLIP (Zhai et al., 2023).

Implementation details. For each dataset-encoder pair, we apply FastICA (Hyvärinen & Oja, 2000) to the whitened embeddings and retain all dimensions, yielding a complete and invertible coordinate system. The ICA decomposition is fully unsupervised and does not use semantic labels or annotations. After discovery, we use LLaVA (Liu et al., 2023) to assign a semantic description to each ICA coordinate by contrasting images with high and low activations along that direction. Further details on ICA construction and direction annotation are provided in Appendices A and C.

Table 1: Semantic coverage and mean AUC of directions discovered in CelebA and MSCOCO. Coverage denotes the number of unique discovered directions matched to the annotated factors, divided by the total number of attributes or classes. Higher is better for both metrics.

	CelebA						MSCOCO					
	CLIP		DINO		I-JEPA		CLIP		DINO		I-JEPA	
Method	Cov. (%)	AUC	Cov. (%)	AUC	Cov. (%)	AUC	Cov. (%)	AUC	Cov. (%)	AUC	Cov. (%)	AUC
ICA	67.5	**0.80**	**65.0**	**0.776**	**57.5**	**0.803**	**76.2**	**0.932**	80.0	**0.903**	**76.2**	**0.881**
PCA	15.0	0.734	7.5	0.741	7.5	0.760	15.0	0.825	37.5	0.852	13.8	0.782
SAE-D	60.0	0.753	27.5	0.765	15.0	0.734	66.2	0.911	80.0	0.860	32.5	0.793
SAE-$2D$	65.0	0.744	40.0	0.749	20.0	0.738	67.5	0.919	78.8	0.877	33.8	0.834
SAE-$4D$	57.5	0.730	32.5	0.736	17.5	0.741	66.2	0.918	**86.2**	0.877	38.8	0.814
SAE-$8D$	57.5	0.750	32.5	0.747	27.5	0.768	71.2	0.916	76.2	0.877	38.8	0.833
SAE-$16D$	**70.0**	0.764	47.5	0.742	22.5	0.760	72.5	0.922	76.2	0.875	51.2	0.823

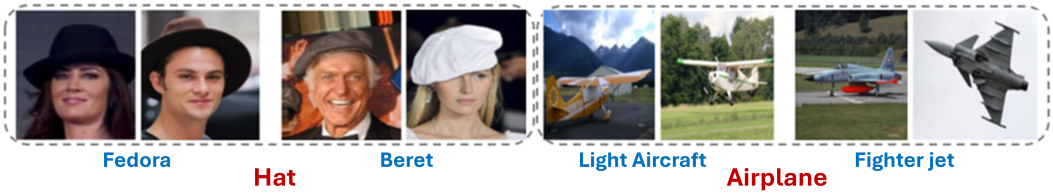


Figure 5: **Subclass separability.** The samples most aligned with different discovered ICs can serve as labeled examples of previously unknown semantic groups, such as subclasses within an existing category (e.g., *beret* and *fedora* hats). Training a simple linear classifier on these automatically discovered groups produces high subclass separation performance, as detailed in Appendix D.4.

Table 2: **Hierarchical organization of discovered directions on ImageNet-100.** We report the ratio of mean cosine similarity for class directions sharing the same semantic parent to those from different parents (Within/Across). Higher ratios indicate stronger hierarchical organization. Since PCA's directions are orthogonal by construction, the measure is inapplicable.

Model	ICA	PCA	SAE-D	SAE-$2D$	SAE-$4D$	SAE-$8D$	SAE-$16D$
CLIP	**2.821**	–	1.504	1.467	1.326	1.243	1.203
DINO	**5.721**	–	1.279	1.439	1.454	1.330	1.220
I-JEPA	**4.046**	–	1.139	1.081	1.074	1.044	1.029

Baselines. For unsupervised factor discovery, we compare ICA with PCA and sparse autoencoders (SAEs) with dictionary sizes ranging from D to $16D$, where D is the embedding dimension. These alternatives represent two common forms of representation decomposition: orthogonal variance-based directions and learned sparse, potentially overcomplete feature dictionaries. We additionally include two supervised oracle baselines that use attribute labels directly to construct semantic directions: the mean-difference (*mean-diff*) direction (Sonthalia et al., 2026) and logistic-regression classifiers. For the unsupervised methods, semantic annotations may be used only *after* factor discovery for component-to-attribute matching and evaluation. Full details in Appendix C.

4.1 Shared Semantic Structure Across Representations

We start from the observation that semantic factors exhibit a consistent geometric organization across independently trained vision encoders. For each annotated factor (CelebA attribute or MSCOCO class), we use the dataset labels to construct a reference semantic direction as the difference between the mean embeddings of positive and negative samples (*mean-diff*). For each encoder, we then compute the pairwise cosine similarities between these directions, obtaining a semantic relation matrix that describes the relative organization within each representation space.

As shown in Figure 2, these relation matrices are remarkably similar across CLIP, DINOv2, I-JEPA, and SigLIP. The correlations between matrices are consistently high across different encoders. At the same time, the off-diagonal structure of the matrices shows that semantic factors are generally non-orthogonal, supporting a semantic organization of directions. Complementary results on both datasets are provided in Figure 8 and full experimental details are in Appendix D.1.

4.2 Recovering Semantic Factors from the Data

We next examine the semantic factors revealed by the unlabeled embedding distribution using known annotations. We then compare well-recovered and missed concepts to understand which factors are more easily recovered, and identify additional factors beyond the annotated vocabulary.

Semantic coverage and separability. We first compare ICA with PCA and SAEs using the annotated CelebA attributes and MSCOCO categories as post-hoc references. For each semantic factor, we match the discovered direction that best represents it in terms of angular alignment (different factors can be mapped to the same discovered direction). We measure both classification AUC and semantic coverage, defined as the number of unique matched directions divided by the number of annotated attributes or classes. Low coverage indicates that multiple semantic factors collapse onto

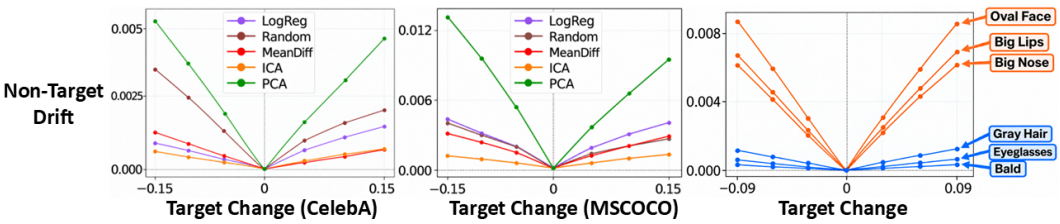


Figure 6: **Collateral damage.** We intervene along a target factor until its angular alignment with the corresponding *mean-diff* reaches a desired change (horizontal axis), and measure the mean angular drift of other, non-target factors (vertical axis). **Left, Middle:** ICA yields the lowest non-target drift, highlighting stronger disentanglement from its oblique biorthogonal framework on CelebA (left) and MSCOCO (middle). **Right:** Within ICA, presence-like factors show lower non-target drift, indicating stronger disentanglement (CelebA).

the same discovered direction. As shown in Table 1, ICA consistently achieves strong coverage and the highest separability across datasets and encoders. These results show that ICA consistently recovers semantically meaningful individual factors across datasets and encoders.

Characteristics of the recovered factors. We next examine what distinguishes well-recovered concepts from those that are missed by ICA. For each annotated factor, we measure the non-Gaussianity of its projection in the whitened representation using negentropy (details in Appendix D.2), and relate it to both its separability along the closest ICA coordinate (AUC) and its angular alignment with that coordinate (Cosine similarity). As shown in Figure 3, selective, presence-like factors tend to have higher negentropy and are recovered more reliably. The complementary results in Figure 9 show the same trend geometrically, with higher-negentropy factors forming smaller angles with their closest ICA direction.

Factors beyond the annotated vocabulary. We also demonstrate that the recovered coordinates capture coherent semantic factors that are not represented by the available dataset annotations. As shown in Figures 4 and 11, ICA reveals distinct factors related to identity, background, lighting, multiplicity, and global style. These examples show that the discovered coordinate system extends beyond the predefined CelebA attributes and MSCOCO object categories. Additional ablations and sensitivity analyses are provided in Appendix D.3.

4.3 Semantic Relations Between Recovered Factors

Having established that ICA recovers meaningful semantic factors, we next examine whether the relations among the recovered directions reflect semantic structure. As illustrated in Figure 1, semantically related factors such as *cowboy hat* and *cap* are represented by distinct ICA coordinates, and their corresponding representation space directions remain geometrically close. A similar effect is observed in Table 4: different ICA directions capture distinct subclasses within the broader *Wearing Hat* and *airplane* categories, with high classification accuracy across CLIP and DINO (see full details in Appendix D.4). These results suggest that related semantic factors remain separable while retaining a meaningful geometric relation.

To examine this effect quantitatively at a broader scale, we use ImageNet-100 (Deng et al., 2009), where classes are grouped under higher-level semantic parents in the ImageNet hierarchy. We compare the cosine similarity between directions associated with classes sharing the same parent and classes belonging to different parents. As shown in Table 2, ICA yields substantially larger within-parent than across-parent similarity across CLIP, DINO, and I-JEPA, with markedly stronger ratios than the SAE baselines. These results show that the recovered factors are not only individually meaningful: their native-space geometry also reflects semantic organization.

4.4 Operating on the Semantic Coordinate System

The recovered geometry is not only descriptive; it also provides a structured mechanism for semantic control in the representation space. Because the ICA transformation is complete and invertible, indi-

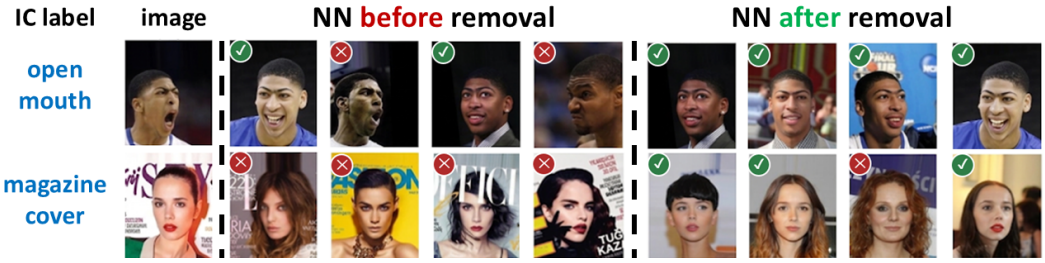


Figure 7: **Controlled image retrieval for identity preservation.** We suppress ICA coordinates corresponding to nuisance factors in the query image, such as *open mouth* (top) or *magazine cover* (bottom), before retrieval. Suppressing these factors improves identity preservation among retrieved neighbors: green checks denote the query identity; red crosses denote different identities. ICA directions are captioned using LLaVA (Liu et al., 2023). Additional examples are in Figure 13.

Table 3: **Spurious mitigation: worst-group accuracy (WG).** We report WG with and without removing spurious ICA directions. STDs are omitted when negligible. Overall results in Table 6.

	CLIP			DINO		I-JEPA	
	ERM	DFR	CA	ERM	DFR	ERM	DFR
Original	69.9	85.6±2.1	74.4±1.9	92.1	92.1±0.2	71.2	88.1±1.3
With ICA	**73.7**	**88.3±0.1**	**76.8±1.1**	**93.5**	**92.6±0.3**	**77.3**	**88.9±1.0**

vidual coordinates can be modified independently and then mapped back to the original embedding space. The technical details of all experiments are provided in Appendix D.5.

Semantic intervention. We first evaluate whether individual semantic factors can be manipulated without substantially affecting related concepts. As illustrated in Figure 1, the *cowboy hat* and *cap* directions are geometrically close in the native space, yet projecting out the corresponding ICA coordinates produces distinct semantic effects. We quantify this behavior on CLIP in Figure 6, where ICA yields the smallest non-target drift for a matched target change compared with the baselines. These results show that semantically related factors can remain geometrically close while still being independently controlled in the ICA coordinate system. Ablations and robustness experiments are in Appendix D.5.

Controlled retrieval. We next use coordinate interventions to modify nearest-neighbor retrieval. In Figure 7 and Figure 13, suppressing nuisance factors such as *open mouth* or *magazine cover* shifts retrieval toward images with the same identity. Conversely, Appendix Figure 14 shows that increasing coordinates associated with properties such as lighting or image style retrieves semantically similar images exhibiting the desired property.

Mitigating spurious correlations. We next evaluate the same idea quantitatively by removing semantic factors that act as spurious cues for downstream classification, on the Waterbirds dataset (Sagawa et al., 2020). Tables 3 and 6 compare several classifiers before and after suppressing the corresponding ICA directions. Across CLIP, DINO, and I-JEPA, removing these factors consistently improves worst-group accuracy while largely preserving average accuracy.

5 Conclusions

We introduced an oblique, biorthogonal view of semantic structure in vision foundation models. Our results first show that semantic factors exhibit a shared, non-orthogonal organization across vision encoders. We then show that these factors can be revealed directly from embedding distributions, and that the geometry among the recovered directions reflects semantic relations. Finally, we demonstrate that the resulting coordinate system is operational, enabling selective semantic intervention, controlled retrieval, and mitigation of spurious correlations. Together, these findings provide a simple framework for both understanding and manipulating semantic structure in pretrained vision representations.

AI USE STATEMENT

Large language models were used for writing assistance, including improving language clarity, correcting grammar, generating illustrations, and for code generation limited to implementing specified functions or refactoring existing code according to explicit instructions. In addition, LLaVA (Liu et al., 2023), a vision-language model, was used as part of the proposed pipeline to assign semantic descriptions to discovered directions, as described in the method. All scientific decisions, analyses, interpretations, and conclusions were independently determined by the authors, who take full responsibility for the content and claims of this work.

REPRODUCIBILITY STATEMENT

The Appendix provides full details of the mathematical formulation, datasets, experimental settings, model configurations, hyperparameters, and evaluation protocols used throughout the paper. We rely on standard publicly available datasets and pretrained models, and specify the settings required to reproduce all reported experiments and figures. Upon acceptance, we will release the complete codebase for feature extraction, ICA fitting, evaluation, and downstream interventions.

REFERENCES

Mahmoud Assran et al. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pp. 15619–15629, 2023.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3319–3327. IEEE, 2017.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

David Chanin, Tomáš Dulka, and Adrià Garriga-Alonso. Feature hedging: Correlated features break narrow sparse autoencoders. *arXiv preprint arXiv:2505.11756*, 2025a.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Isaac Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, 2025b.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

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

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in neural information processing systems*, 35:21400–21413, 2022.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2711–2721, 2023.

Thomas Fel, Matthew Kowal, Mozes Jacobs, Dron Hazra, Usha Bhalla, Lee Sharkey, Lucius Bushnaq, Satchel Grant, Tal Haklay, Thomas Icard, et al. Structuring sparsity: Block-sparse featurizers capture visual concept manifolds. *arXiv preprint arXiv:2606.25234*, 2026a.

Thomas Fel, Binxu Wang, Michael Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep Singh Lubana, Talia Konkle, Demba Ba, et al. Into the rabbit hull: From task-relevant concepts in dino to minkowski geometry. In *International Conference on Learning Representations*, volume 2026, pp. 154538–154584, 2026b.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Oscar Hill, Mateo Espinosa Zarlenga, and Mateja Jamnik. Hierarchical concept-based interpretable models. In *International Conference on Learning Representations*, volume 2026, pp. 11581–11609, 2026a.

Oscar Hill, Mateo Espinosa Zarlenga, and Mateja Jamnik. Digging deeper: Learning multi-level concept hierarchies, 2026b. URL https://arxiv.org/abs/2603.10084.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

A 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.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. Pmlr, 2020.

Yuxiao Li, Eric J Michaud, David D Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Sida Liu and Feijiang Han. Ica lens: Interpreting language models without training another dictionary. *ArXiv*, abs/2606.11722, 2026. URL https://api.semanticscholar.org/CorpusID:289167517.

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*, pp. 3730–3738, 2015.

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. *arXiv preprint arXiv:2304.07193*, 2023.

Momose Oyama, Hiroaki Yamagiwa, and Hidetoshi Shimodaira. Understanding higher-order correlations among semantic components in embeddings. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2883–2899, 2024.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham Kakade, and Stephanie Gil. Interpreting the Linear Structure of Vision-Language Model Embedding Spaces. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=qPsmGjpq1j.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 76441–76463, 2025.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Ankit Sonthalia, Arnas Uselis, and Seong Joon Oh. On the rankability of visual embeddings. *Advances in Neural Information Processing Systems*, 38:66169–66203, 2026.

Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. Discovering universal geometry in embeddings with ica. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4647–4675, 2023.

Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. Axis tour: Word tour determines the order of axes in ica-transformed embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 477–506, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35:21682–21697, 2022.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International conference on machine learning*, pp. 12979–12990. PMLR, 2021.

OVERVIEW

This appendix complements the main paper with the full mathematical derivations, implementation details, and extended experimental results. We first detail the ICA and whitening construction and derive the associated oblique, biorthogonal geometry. We then provide the models, datasets, baselines, and experimental protocols used throughout the evaluation. Finally, we present additional analyses of the recovered semantic structure, together with extended intervention results, sensitivity studies, and ablations.

A ICA CONSTRUCTION AND WHITENING DETAILS

This appendix provides the exact linear transformation used to obtain the ICA coordinates introduced in Equation (2).

Let the centered embedding of image x_n be

$$\tilde{z}_n = z_n - \mu, \qquad \mu = \frac{1}{N}\sum_{n=1}^{N} z_n. \tag{5}$$

The empirical covariance of the centered embeddings is

$$\Sigma = \frac{1}{N}\sum_{n=1}^{N} \tilde{z}_n \tilde{z}_n^\top. \tag{6}$$

We use the eigendecomposition

$$\Sigma = U\Lambda U^\top, \tag{7}$$

where $U \in \mathbb{R}^{d\times d}$ is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ contains the covariance eigenvalues.

We define the whitening transform as

$$P = \Lambda^{-1/2} U^\top, \tag{8}$$

so that the whitened representation is

$$y_n = P\tilde{z}_n = \Lambda^{-1/2} U^\top (z_n - \mu). \tag{9}$$

By construction, the whitened embeddings have identity covariance:

$$\mathrm{Cov}(y) = P\Sigma P^\top = I. \tag{10}$$

We then apply FastICA to the whitened embeddings. The recovered coordinates are

$$c_n = R y_n, \tag{11}$$

where $R \in \mathbb{R}^{d\times d}$ is the ICA rotation. Because the input has been whitened, the search for independent components can be restricted to orthogonal rotations, so

$$RR^\top = R^\top R = I. \tag{12}$$

FastICA estimates R by maximizing the non-Gaussianity of the projected coordinates. Equivalently, for a unit vector r, the algorithm searches for projections

$$c = r^\top y \tag{13}$$

that maximize a contrast function measuring deviation from Gaussianity, subject to orthogonality between the recovered directions. In practice, FastICA uses a fixed-point approximation to this objective (Hyvärinen & Oja, 2000).

Combining whitening and ICA gives the complete mapping from the native embedding space to the ICA coordinate system:

$$c_n = RP(z_n - \mu) = R\Lambda^{-1/2} U^\top (z_n - \mu). \tag{14}$$

Because both P and R are invertible, the mapping is bijective when all d components are retained. The inverse transformation is

$$z_n - \mu = P^{-1} R^\top c_n = U\Lambda^{1/2} R^\top c_n. \tag{15}$$

Defining

$$A = P^{-1} R^\top = U\Lambda^{1/2} R^\top, \tag{16}$$

we obtain the native-space reconstruction

$$z_n = \mu + Ac_n. \tag{17}$$

Writing

$$A = [a_1, \ldots, a_d], \tag{18}$$

the representation can equivalently be expressed as

$$z_n = \mu + \sum_{i=1}^{d} c_{n,i} a_i. \tag{19}$$

Thus, each ICA coordinate $c_{n,i}$ is associated with a native-space direction a_i. The geometry of these directions is analyzed in Appendix B.

B Geometry of the Oblique Coordinate System

Let the empirical covariance of the centered embeddings be

$$\Sigma = U\Lambda U^\top,$$

and choose the whitening transform

$$P = \Lambda^{-1/2} U^\top.$$

After ICA,

$$c = RP(z - \mu),$$

where R is orthogonal in the whitened space. The inverse transform is

$$z - \mu = U\Lambda^{1/2} R^\top c = Ac.$$

The native-space Gram matrix of the basis vectors is therefore

$$A^\top A = R\Lambda R^\top.$$

Unless Λ is proportional to the identity, or R aligns with its eigenspaces, the off-diagonal entries are generally non-zero. Thus, orthogonality in the whitened ICA space does not imply orthogonality in the original embedding space.

B.1 Measurement and intervention directions

Let

$$W = RP = A^{-1}.$$

The i-th coordinate is measured by the i-th row $w_i^\top$ of W,

$$c_i = w_i^\top (z - \mu).$$

By contrast, changing only c_i changes the original embedding along the i-th column a_i of A.

Since

$$WA = I,$$

the measurement and intervention directions satisfy

$$w_i^\top a_j = \delta_{ij}.$$

They form a dual pair and coincide only when the coordinate basis is orthogonal. This distinction is useful when interpreting linear probes and intervention directions in an oblique representation space.

B.2 BIORTHOGONAL BASES

Let $\{\mathbf{a}_i\}_{i=1}^D$ be a basis of a D-dimensional vector space. Any vector $\mathbf{x}$ admits a unique representation

$$\mathbf{x} = \sum_{i=1}^{D} c_i \mathbf{a}_i, \tag{20}$$

where c_i denotes the coordinate of $\mathbf{x}$ along the i-th basis axis.

For an orthonormal basis, the coordinates can be obtained directly by projection onto the basis vectors,

$$c_i = \mathbf{a}_i^\top \mathbf{x}, \tag{21}$$

since $\mathbf{a}_i^\top \mathbf{a}_j = \delta_{ij}$. In this case, the same basis is used both to *analyze* a vector into its coordinates and to *synthesize* the vector from these coordinates.

More generally, the basis vectors need not be orthogonal. A set of linearly independent, non-orthogonal vectors $\{\mathbf{a}_i\}_{i=1}^D$ defines an *oblique coordinate system*. In this case, projecting $\mathbf{x}$ directly onto $\mathbf{a}_i$ does not, in general, recover its i-th coordinate, since

$$\mathbf{a}_i^\top \mathbf{x} = \sum_{j=1}^{D} c_j \mathbf{a}_i^\top \mathbf{a}_j, \tag{22}$$

which contains contributions from other coordinates whenever $\mathbf{a}_i^\top \mathbf{a}_j \neq 0$.

Associated with every basis $\{\mathbf{a}_i\}_{i=1}^D$ is a unique *dual basis* $\{\mathbf{b}_i\}_{i=1}^D$ satisfying

$$\mathbf{b}_i^\top \mathbf{a}_j = \delta_{ij}. \tag{23}$$

The two bases $\{\mathbf{a}_i\}$ and $\{\mathbf{b}_i\}$ are therefore said to be *biorthogonal*. The dual vectors provide the analysis operation,

$$c_i = \mathbf{b}_i^\top \mathbf{x}, \tag{24}$$

while the primal vectors provide the corresponding synthesis operation,

$$\mathbf{x} = \sum_{i=1}^{D} \left(\mathbf{b}_i^\top \mathbf{x}\right) \mathbf{a}_i. \tag{25}$$

Equivalently, defining

$$A = [\mathbf{a}_1 \quad \cdots \quad \mathbf{a}_D], \qquad B = [\mathbf{b}_1 \quad \cdots \quad \mathbf{b}_D], \tag{26}$$

biorthogonality can be expressed as

$$B^\top A = I, \tag{27}$$

and hence, for an invertible A,

$$B = (A^{-1})^\top. \tag{28}$$

This reveals the distinction between orthonormal and oblique coordinate systems. For an orthonormal basis, $A^\top A = I$, and therefore $B = A$: the primal and dual bases coincide. For an oblique basis, $A^\top A \neq I$ in general, and the analysis and synthesis directions are distinct. Thus, an oblique coordinate system naturally admits a biorthogonal analysis-synthesis representation: the dual basis $\{\mathbf{b}_i\}$ extracts the coordinates, while the primal basis $\{\mathbf{a}_i\}$ reconstructs their contributions to the original space.

C IMPLEMENTATION DETAILS

C.1 MODELS AND HARDWARE

- **Computing hardware:** Experiments were conducted on an NVIDIA GeForce RTX 3090 GPU.

- **Software libraries:** We use `PyTorch` as the main deep-learning framework; `scikit-learn` for the FastICA implementation; `matplotlib` for scientific visualization; `transformers` for loading CLIP, LLaVA, DINOv2, and SigLIP models; and `PIL` for image loading and preprocessing.
- **Models:**
  - **Vision encoders:** DINOv2 ViT-L/14 with a 1024-dimensional embedding space, and I-JEPA ViT-H/14 with a 1280-dimensional embedding space.
  - **Vision-language models:** CLIP ViT-L/14 with a 768-dimensional embedding space, and SigLIP ViT-SO400M with a 1152-dimensional embedding space.
  - **Vision-language assistant:** LLaVA-v1.6-Mistral-7B-HF.
- **Experimental details:** We fit a separate ICA model for each dataset, using $150{,}000$ training samples for CelebA and the full training set for each of the remaining datasets (MSCOCO, ImageNet-100, and Waterbirds).
- **LLaVA prompt example (MSCOCO):**

> `You are given` N `images that were selected because they share a common visual property. Identify the single most salient visual concept that is consistently shared across these images and would best distinguish this set from typical, randomly selected images.`
> `The concept may involve an object, object category, animal, person, action, activity, scene, environment, spatial arrangement, object interaction, number of objects, shape, texture, material, color, pose, viewpoint, lighting, background, weather, photographic style, image quality, or any other visible property.`
> `Do not force the answer into one of these categories; they are only examples. Focus on the most specific and coherent property supported by the images as a group, rather than properties that appear in only a few images.`
> `Return only a concise description of the shared concept, using at most` K `words.`

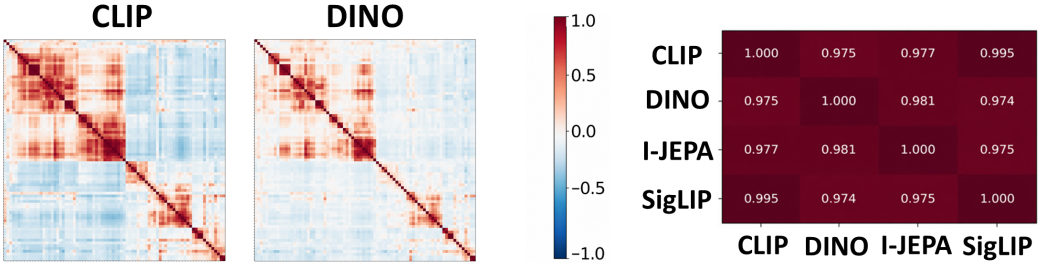
C.2 Methods

- **Mean Difference (MeanDiff):** For a label of interest, we compute the mean embedding of images containing the label and the mean embedding of images not containing it. Their difference defines a linear direction in the embedding space that represents the label.
- **Logistic regression (LogReg):** For a label of interest, we train a linear classifier to separate embeddings of images with and without the label. The normal vector to the resulting decision hyperplane is then used as the representative direction of the label in the embedding space.
- **SAE:** We enforce sparsity using a Top-K SAE variant, which retains at most K active latent entries for each input. We perform an ablation over $K \in 16, 32, 64, 128, 256$ for all experiments involving the SAE. In most cases, performance is relatively insensitive to the choice of K. In Table 1, however, $K = 32$ provides the best trade-off between coverage and AUC, and we therefore use this value in our experiments. We also evaluated SAEs with explicit ℓ_1 regularization, which produced worse performance than the Top-K variant.

D Additional experimental details and Results

D.1 Shared semantic structure

In Figure 2 and Figure 8, the RSMs are constructed from the *mean-diff* directions by computing the pairwise cosine similarity between every pair of directions. The similarity between two RSMs is then quantified using Pearson's correlation coefficient, computed between the vectorized upper-triangular entries of the two matrices.



(a) **MSCOCO:** Representational Similarity Matrices (RSMs). (b) **CelebA:** Correlation between RSMs.

Figure 8: **Shared semantic structure across encoders.** Complementary results to Figure 2.

D.2 Statistical properties of components

We measure non-Gaussianity using *negentropy*, the criterion employed by FastICA to identify non-Gaussian directions in the whitened space. Since exact negentropy is difficult to estimate, FastICA uses the approximation

$$J(y) \propto \left(\mathbb{E}[G(y)] - \mathbb{E}[G(v)]\right)^2, \tag{29}$$

where $y = \mathbf{w}^\top \mathbf{x}$ is the projection of the whitened representation $\mathbf{x}$ onto a candidate direction $\mathbf{w}$, $v \sim \mathcal{N}(0,1)$ is a standard Gaussian variable, and G is a non-quadratic contrast function. We use the default FastICA choice,

$$G(u) = \log\cosh(u). \tag{30}$$

FastICA searches for directions $\mathbf{w}$ that maximize this approximation, thereby favoring projections with maximal deviation from Gaussianity. For analysis, we use the same approximation as a post-hoc non-Gaussianity score for an already fixed projection. Given projected values $\{y_i\}_{i=1}^N$, we first standardize them to zero mean and unit variance,

$$\tilde{y}_i = \frac{y_i - \mu_y}{\sigma_y}, \tag{31}$$

and compute

$$\widehat{J}(\tilde{y}) = \left[\frac{1}{N}\sum_{i=1}^{N} \log\cosh(\tilde{y}_i) - c_G\right]^2, \qquad c_G = \mathbb{E}_{v\sim\mathcal{N}(0,1)}[\log\cosh(v)]. \tag{32}$$

In our implementation, c_G is estimated once using 10^6 samples from a standard Gaussian distribution ($c_G \approx 0.3745$). Thus, Gaussian-like projections obtain scores close to zero, while increasingly non-Gaussian projections obtain larger scores. The squared difference makes the measure insensitive to the sign of the deviation from the Gaussian reference.

D.3 Granular annotations

Below, we present the semantic labels assigned by LLaVA to the IC directions extracted from CLIP representations on the MSCOCO dataset:

0: video game; **1**: rowing boat; **2**: baseball game; **3**: umbrellas; **4**: tennis court; **5**: camera phone; **6**: lighting; **7**: birds; **8**: baseball; **9**: apple; **10**: birthday; **11**: man in suit; **12**: teddy bears; **13**: toothbrush; **14**: clock; **15**: laptop; **16**: books; **17**: coffee cup; **18**: rainy day; **19**: umbrella; **20**: cat looking out; **21**: bicycle; **22**: skateboard; **23**: foggy street; **24**: birthday cake; **25**: teddy bear; **26**: beach; **27**: antique furniture; **28**: father and child; **29**: teddy bear; **30**: remote controls; **31**: pizza; **32**: food and drink; **33**: giraffe; **34**: bananas; **35**: baseball game; **36**: banana bunch; **37**: detour; **38**: political campaign; **39**: bananas; **40**: train; **41**: stop sign; **42**: elephant; **43**: umbrella; **44**: people; **45**: duck; **46**: baseball bat; **47**: food on plate; **48**: tennis racket; **49**: bear; **50**: pizza; **51**: elephant; **52**: bookshelf; **53**: dolls; **54**: skateboard; **55**: heart; **56**: white kitchen; **57**: toilet tricks; **58**: cat; **59**: bus; **60**: birthday cake; **61**: direction; **62**: train; **63**: old fashioned; **64**: baseball player; **65**: blue sky; **66**: nathan homestead; **67**: skiing; **68**: fruit; **69**: reflection; **70**: fish; **71**: clock; **72**: umbrella; **73**: plane; **74**: water; **75**: toys on floor; **76**: stop sign; **77**: elephant in refrigerator; **78**: food; **79**: breakfast; **80**: baseball field; **81**: pitcher throwing ball; **82**: train; **83**: oranges in bowl; **84**: white shower curtain; **85**: paddle surfing; **86**: stairs; **87**: man holding wii; **88**: man in suit; **89**: people; **90**: bird; **91**: plane; **92**: baseball bat; **93**: scissors; **94**: baseball player; **95**: scissors; **96**: welcome; **97**: food truck; **98**: rowing team; **99**: clock; **100**: cows; **101**: baseball player; **102**: giraffe; **103**: toilet; **104**: red bus; **105**: cows; **106**: graffiti; **107**: tennis player; **108**: smiling couple; **109**: laptop; **110**: baseball player; **111**: kite flying; **112**: bird on branch; **113**: reflection; **114**: laptop on desk; **115**: clock; **116**: umbrellas; **117**: red fire hydrant; **118**: sleeping; **119**: street sign; **120**: old photo;

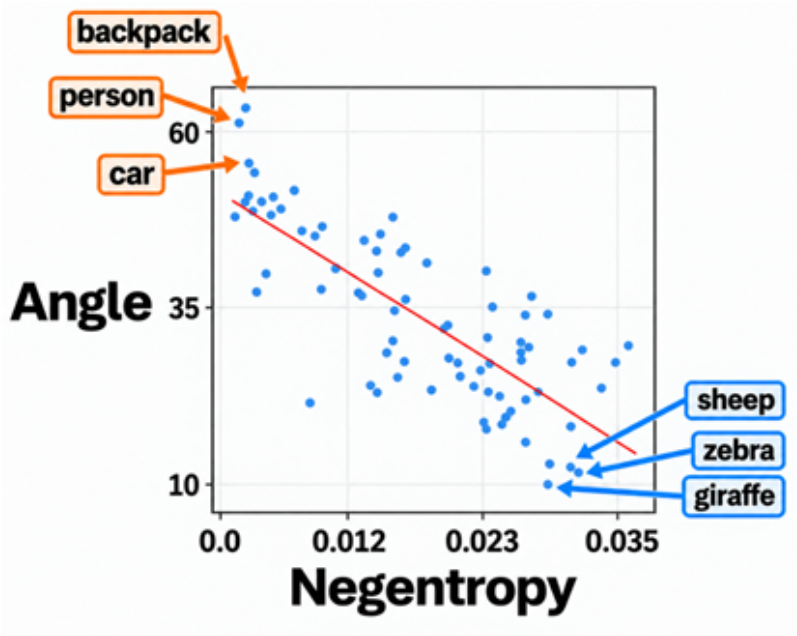


Figure 9: **Statistical properties of ICA components.** Complementary results to Figure 3. Features with higher negentropy tend to create smaller angle with the closest IC to their representative *mean-diff* direction, as suggested by the negative relationship between negentropy and angle.

121: arrow; **122**: construction; **123**: bed; **124**: garden; **125**: baseball player; **126**: man holding wine; **127**: toothbrush; **128**: tile floor; **129**: cat watching tv; **130**: train; **131**: doughnut; **132**: art project; **133**: baseball bat; **134**: jumping; **135**: tie; **136**: cell phone; **137**: man holding laptop; **138**: one way; **139**: water skiing; **140**: wii controllers; **141**: donuts; **142**: object; **143**: sunset; **144**: pizza; **145**: soccer ball; **146**: stop sign; **147**: pasta and broccoli; **148**: bicycles; **149**: vase; **150**: stop harper; **151**: train; **152**: church; **153**: clock; **154**: surfing; **155**: boxes; **156**: pizza; **157**: man and woman; **158**: clock; **159**: smiling people; **160**: red and white; **161**: texting; **162**: accident; **163**: stop sign; **164**: color; **165**: white horse; **166**: food on plate; **167**: orange bus; **168**: cat on laptop; **169**: fruit salad; **170**: cutting cake; **171**: doughnut; **172**: baseball player; **173**: street sign; **174**: tennis ball; **175**: giraffe; **176**: food; **177**: baseball player; **178**: smiling man; **179**: propeller; **180**: wakeboarding; **181**: surfing; **182**: cell phone; **183**: ban park; **184**: truck; **185**: woman; **186**: food; **187**: refrigerator; **188**: luggage; **189**: kitchen; **190**: living room; **191**: bench; **192**: selfie; **193**: white bed; **194**: motorcycle; **195**: street sign; **196**: man holding umbrella; **197**: north beverly; **198**: direction; **199**: crochet; **200**: curved perspective; **201**: baby with teddy; **202**: train; **203**: suitcase; **204**: snowboarding; **205**: cat; **206**: speed limit; **207**: clouds; **208**: street sign; **209**: train; **210**: boat; **211**: tennis player; **212**: man wearing hat; **213**: horse and carriage; **214**: wooden bench; **215**: cat; **216**: bus; **217**: wall st; **218**: traffic calmed; **219**: cutting cake; **220**: cell phone; **221**: plane; **222**: airplane; **223**: muffin; **224**: happy birthday emmett; **225**: snow; **226**: no; **227**: rugby ball; **228**: no; **229**: tennis player; **230**: man wearing glasses; **231**: cutting money; **232**: clock tower; **233**: baseball; **234**: golf; **235**: propeller plane; **236**: jumping horse; **237**: smiling children; **238**: man holding phone; **239**: clock tower; **240**: tennis ball; **241**: police motorcycle; **242**: direction; **243**: birds flying; **244**: street sign; **245**: love trees; **246**: plane; **247**: soccer field; **248**: messy; **249**: motorcycle; **250**: camera; **251**: birthday cake; **252**: stop sign; **253**: tennis ball; **254**: man standing; **255**: street sign; **256**: bear; **257**: beer bottle; **258**: sidewalk; **259**: frisbee; **260**: pizza; **261**: train; **262**: elephant; **263**: strawberry; **264**: bridge; **265**: sheep grazing; **266**: broccoli; **267**: train; **268**: zombie; **269**: truck; **270**: food on plate; **271**: baseball player; **272**: tennis player; **273**: dirt ground; **274**: plane; **275**: train; **276**: plane; **277**: bat; **278**: baseball player; **279**: white bear; **280**: french fries; **281**: haircut; **282**: white tile; **283**: grill; **284**: zebra fight; **285**: food on plate; **286**: one way; **287**: surfing; **288**: kite surfing; **289**: birds; **290**: frisbee; **291**: nude woman; **292**: news set; **293**: bench; **294**: color; **295**: zebra; **296**: birthday party; **297**: surfing; **298**: man holding frisbees; **299**: clock tower; **300**: baseball; **301**: camping; **302**: busy street; **303**: dunkin donuts; **304**: office cubicle; **305**: bench; **306**: truck; **307**: bus; **308**: bench; **309**: tie; **310**: sleeping on bench; **311**: bus; **312**: man in suit; **313**: planes in sky; **314**: black bird; **315**: star wars; **316**: bus; **317**: american flag; **318**: harry potter; **319**: cat; **320**: people standing; **321**: surfboards; **322**: zebra; **323**: snow; **324**: zebra; **325**: umbrellas; **326**: woman sitting; **327**: beach; **328**: throwing ball; **329**: bars; **330**: scissors; **331**: stop sign; **332**: smiley faces; **333**: fruit; **334**: military vehicle; **335**: tennis player; **336**: carrots; **337**: giraffe; **338**: wedding; **339**: cat; **340**: street sign; **341**: scissors; **342**: skiing; **343**: abandoned; **344**: color; **345**: cat; **346**: cup; **347**: skateboarding; **348**: sailboat; **349**: smiling men; **350**: dog catching frisbee; **351**: traffic light; **352**: birds; **353**: no; **354**: tennis ball; **355**: stop sign; **356**: baseball; **357**: miniature food; **358**: ping pong; **359**: kitchen; **360**: skateboarding; **361**: street sign; **362**: baby; **363**: donut making; **364**: clock; **365**: street sign; **366**: luggage; **367**: bananas; **368**: umbrella; **369**: teddy bears; **370**: red sign; **371**: scissors; **372**: umbrella; **373**: living room; **374**: cat in suitcase; **375**: cat; **376**: man in suit; **377**: man in suit; **378**: pizza; **379**: birds on wire; **380**: man playing tennis; **381**: purple shirts; **382**: kite; **383**: toothbrush; **384**: carrot; **385**: dinner party; **386**: birds in water; **387**: clock; **388**: umbrella; **389**: soccer ball; **390**: red truck; **391**: hand; **392**: electrical cord; **393**: baseball player; **394**: laptop; **395**: dessert; **396**: bread; **397**: stairs; **398**: clock tower; **399**: food on plate; **400**: no hits; **401**: team photo; **402**: motorcycle; **403**: donuts; **404**: church; **405**: tennis court; **406**: parking meter; **407**: imminent danger; **408**: orange table; **409**: dog jumping; **410**: reflection; **411**: snow; **412**: train; **413**: direction; **414**: breakfast; **415**: train; **416**: orange; **417**: zebra grazing; **418**: parade; **419**: white truck; **420**: sandwich; **421**: cutting cake; **422**: ride elephant; **423**: train station; **424**: cows in field; **425**: street light; **426**: color; **427**: motion blur; **428**: giraffe; **429**: overturned bus; **430**: no hits, four; **431**: cake; **432**: zebra; **433**: toast; **434**: pizza; **435**: baseball team; **436**: bench; **437**: stop sign; **438**: cow; **439**: elephant; **440**: food on plate; **441**: object; **442**: kitten drinking milk; **443**: street sign; **444**: white wii remote; **445**: open oven; **446**: food; **447**: stone wall; **448**: baseball field; **449**: bicycle; **450**: herd of sheep;

451: boy; **452**: broccoli; **453**: bicycle; **454**: baseball; **455**: dog on chair; **456**: surfboard; **457**: baseball game; **458**: horse; **459**: crowd; **460**: stop sign; **461**: computer desk; **462**: no to bike; **463**: snowboarder in air; **464**: business meeting; **465**: horse jumping; **466**: stuffed animals; **467**: dirt road; **468**: cell phone; **469**: wedding cake; **470**: roses are red; **471**: tennis player; **472**: baseball; **473**: vase; **474**: zombie; **475**: baseball player; **476**: woman holding knife; **477**: color; **478**: artichoke; **479**: large pizza; **480**: man with hat; **481**: eating; **482**: ice cream; **483**: train; **484**: under construction; **485**: train; **486**: birthday cake; **487**: stop sign; **488**: scissors; **489**: white truck; **490**: snow; **491**: smiling children; **492**: street sign; **493**: birds; **494**: wii remote; **495**: cat in sink; **496**: laptop on table; **497**: low gear now; **498**: horse jumping; **499**: birthday cake; **500**: snowboarding; **501**: direction; **502**: scissors; **503**: open refrigerator; **504**: cat hiding; **505**: baseball bat; **506**: tennis ball; **507**: bus; **508**: umbrella; **509**: hot dog; **510**: man eating; **511**: fire truck; **512**: bananas; **513**: train; **514**: cat; **515**: sandwich; **516**: bench; **517**: man sitting; **518**: orange, bowl, blurry; **519**: united glass; **520**: lighting; **521**: woman sitting; **522**: tennis player; **523**: bananas; **524**: skateboarding; **525**: man smiling; **526**: statue; **527**: cheese pizza; **528**: blender; **529**: motorcycle; **530**: toy; **531**: man jumping; **532**: plane; **533**: flowers; **534**: train; **535**: cat; **536**: surfboard; **537**: clock tower; **538**: attention dog owners; **539**: surfing; **540**: bananas; **541**: rope; **542**: cat; **543**: bus; **544**: magnet; **545**: kitchen; **546**: 100; **547**: white cabinets; **548**: food; **549**: elephant; **550**: plane; **551**: snowy mountains; **552**: people standing; **553**: paintings; **554**: snow; **555**: baseball; **556**: surfing; **557**: bird on laptop; **558**: umbrella; **559**: one way; **560**: apples; **561**: woman, bikini, motorcycle; **562**: sugar; **563**: bird; **564**: color; **565**: woman sitting; **566**: wine glass; **567**: flowers; **568**: scooters; **569**: street sign; **570**: wooden ceiling; **571**: street; **572**: bird; **573**: direction; **574**: street sign; **575**: corned beef sandwich; **576**: man with cat; **577**: skateboarding; **578**: school bus; **579**: signing; **580**: umbrella; **581**: sandwich; **582**: kites; **583**: lunch box; **584**: bicycles; **585**: giraffe; **586**: horse; **587**: blue truck; **588**: clock tower; **589**: surfing; **590**: microwave; **591**: double decker bus; **592**: zebra; **593**: computer; **594**: baseball bat; **595**: traffic; **596**: orange hats; **597**: white bed; **598**: surfing; **599**: tennis ball; **600**: birthday cake; **601**: soup; **602**: cutting emissions; **603**: wii controller; **604**: orange leaves; **605**: food; **606**: truck; **607**: green motorcycle; **608**: cutting board; **609**: clock tower; **610**: keep calm and; **611**: fruit stand; **612**: tennis player; **613**: vegetables; **614**: food in oven; **615**: food on plate; **616**: toilet; **617**: street light; **618**: gold; **619**: feeding giraffe; **620**: congratulations; **621**: bicycle; **622**: street sign; **623**: rain; **624**: kitchen utensils; **625**: motorcycle; **626**: cutting cake; **627**: snow; **628**: dirt bike; **629**: red and white; **630**: airplane; **631**: clock tower; **632**: street sign; **633**: stop sign; **634**: travel items; **635**: runway; **636**: bus; **637**: sausage; **638**: clock tower; **639**: train; **640**: ski lift; **641**: laptop; **642**: bottles; **643**: urinal; **644**: candles; **645**: medical; **646**: laptop; **647**: surfboards; **648**: wooden cabinets; **649**: food; **650**: skateboarder; **651**: orange train; **652**: colorful, blurry, people; **653**: keyboard; **654**: bundle busted; **655**: exit sign; **656**: baseball field; **657**: toothbrush; **658**: kites; **659**: happy birthday emmett; **660**: newspaper; **661**: watch for senior; **662**: train; **663**: colorful umbrella; **664**: sadness; **665**: laptop; **666**: cat and dog; **667**: cat; **668**: birthday cake; **669**: color; **670**: man with mustache; **671**: onion rings; **672**: music studio; **673**: snow; **674**: baseball player; **675**: laptop; **676**: airplane; **677**: man in hat; **678**: fruit; **679**: umbrella; **680**: rodeo; **681**: head south; **682**: beach; **683**: umbrellas; **684**: banana; **685**: baseball bat; **686**: feeding cat; **687**: cat; **688**: umbrellas; **689**: mobile webcasting; **690**: clock; **691**: rainbow flags; **692**: baseball player; **693**: sheep; **694**: pizza; **695**: welcome to golden; **696**: man in suit; **697**: baseball; **698**: people eating pizza; **699**: one way; **700**: street sign; **701**: window; **702**: highway; **703**: light; **704**: boats; **705**: motorcycle; **706**: dirty bathroom; **707**: basketball; **708**: desserts; **709**: garbage truck; **710**: umbrella; **711**: flood; **712**: white bear; **713**: kitchen; **714**: laptop; **715**: train; **716**: old town; **717**: cat; **718**: kitchen; **719**: cat on laptop; **720**: broccoli; **721**: pare; **722**: skateboarding; **723**: traffic light; **724**: bus; **725**: double decker bus; **726**: white sink; **727**: crosswalk; **728**: baseball player; **729**: airplane; **730**: stop sign; **731**: stop sign; **732**: glass; **733**: food on plate; **734**: bird feeder; **735**: suitcases; **736**: santa claus; **737**: man and woman; **738**: social gathering; **739**: lego; **740**: cat; **741**: group photo; **742**: food; **743**: baseball field; **744**: bus; **745**: teddy bear; **746**: toilet; **747**: sheep; **748**: computer; **749**: pink umbrellas; **750**: zebra; **751**: wine glass; **752**: pizza; **753**: skateboarding; **754**: open mouth; **755**: person; **756**: plane; **757**: bottle; **758**: elephant; **759**: banana; **760**: umbrellas; **761**: rain; **762**: baseball player; **763**: bed; **764**: stop sign; **765**: train; **766**: apple; **767**: tennis court.

Animals (n=12)

IC	Label
33	giraffe
100	cows
256	bear
262	elephant
337	giraffe
428	giraffe
439	elephant
585	giraffe
592	zebra
693	sheep
747	sheep
758	elephant

Fruits & vegetables (n=7)

IC	Label
83	oranges in bowl
333	fruit
336	carrots
367	bananas
560	apples
611	fruit stand
613	vegetables

Vehicles (n=11)

IC	Label
40	train
62	train
167	orange bus
202	train
415	train
483	train
485	train
591	double decker bus
639	train
724	bus
725	double decker bus

Figure 10: **IC semantic groups.**

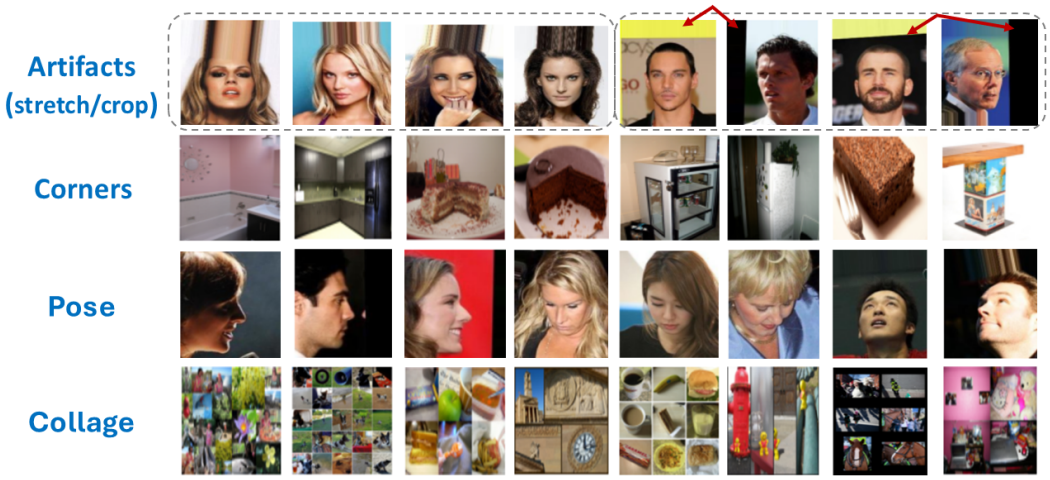


Figure 11: **Additional semantic sources revealed by ICA.** Extreme activations along selected ICA coordinates expose recurring visual factors beyond object identity or class labels. The examples include image artifacts such as stretching and cropping, as well as spatial composition patterns such as corners. Red arrows highlight representative crop-related artifacts. The displayed images are from CLIP and I-JEPA.

D.4 Relations between factors

- **Subclass separability.** In the class separability experiment (Figure 5, Table 4), each subclass is associated with an IC direction. We select the n most extreme training samples along the corresponding IC and manually verify that they belong to the target subclass (e.g., 100 *fedora* and 100 *beret* images in CelebA). The same procedure is used to select k test samples per subclass. A linear binary classifier is then trained on the selected training samples and evaluated on the corresponding test samples.
- **Hierarchical organization.** In the hierarchical organization experiment (Table 2), each discovered direction is assigned to a class by majority vote over the labels of its n most extreme samples. We then compute pairwise angular alignment between the assigned di-

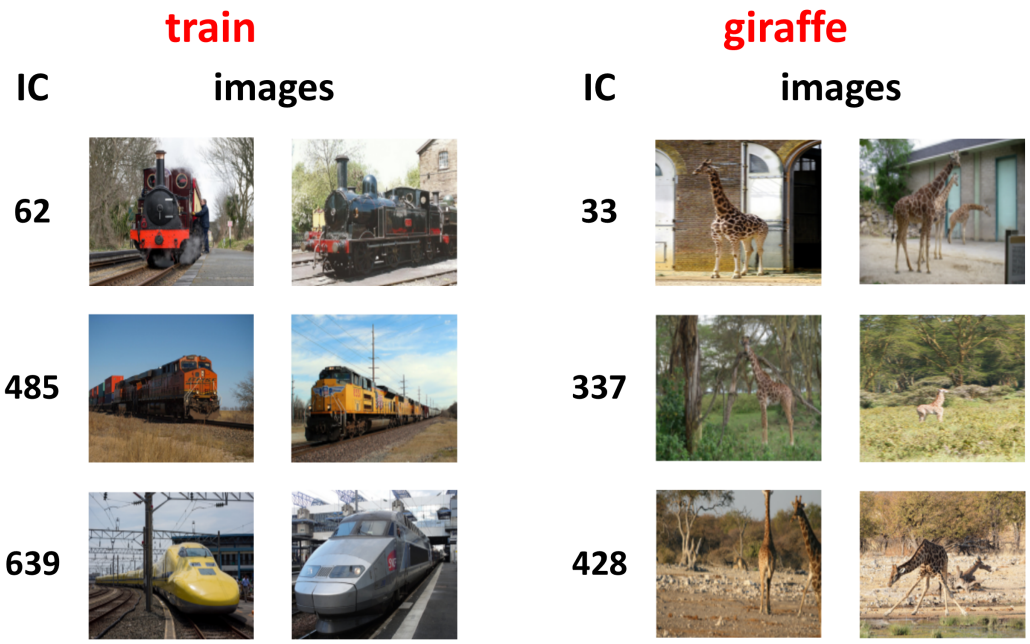


Figure 12: **Fine-grained semantic structure of ICs.** ICs assigned the same semantic label by LLaVA (Figure 10) can nevertheless capture distinct sub-concepts. Different *train* ICs correspond to different types of trains, whereas different *giraffe* ICs capture variations in their surrounding environments.

Table 4: **Subclass separability.** For each parent concept, two ICA directions corresponding to distinct subclasses are selected. We report the classification accuracy for the recovered subclasses.

Model	Dataset	Hat-1	Hat-2	Dataset	Airplane-1	Airplane-2
CLIP	CelebA	0.98	0.98	MSCOCO	1.00	1.00
DINO		0.98	0.98		1.00	1.00

rections and use the ImageNet-100 parent labels to test whether directions representing classes under the same parent are more closely aligned than those from different parents. For computational efficiency, we retain only directions for which at least $n/2$ of the extreme samples belong to the assigned class. The same trend is observed without this constraint, although the filtering makes it more pronounced. For example, for DINO, the mean cosine similarity of within-parent pairs is approximately $5.7\times$ that of across-parent pairs, compared with only $1.45\times$ for the best-performing SAE (SAE-4D). This indicates that hierarchical relatedness is much more strongly reflected in directional proximity for ICA than for SAEs.

In Figure 10, we group ICs assigned similar semantic labels by LLaVA (out of the captions list presented above) into three broad categories: *Animals*, *Fruits & vegetables*, and *Vehicles*. Representative samples from several of these groups are shown in Figure 12. We then compute the pairwise angular alignment between ICs within the same group and across different groups. The within-group alignment is substantially higher than the across-group alignment, indicating that semantically related ICs also exhibit stronger geometric similarity (Table 5).

Table 5: **Semantic grouping of IC directions.** Mean absolute pairwise cosine similarity ($|\cos|$) between IC directions within and across semantic groups. Diagonal entries denote within-group similarity, while off-diagonal entries denote between-group similarity.

	Animals	Fruits & vegetables	Vehicles
Animals	**0.3602**	0.0412	0.0394
Fruits & vegetables	0.0412	**0.3467**	0.0752
Vehicles	0.0394	0.0752	**0.3537**

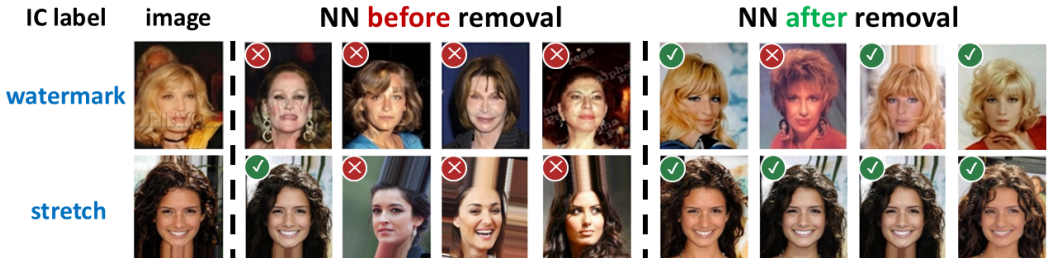


Figure 13: **Controlled image retrieval for identity preservation.** Complementary results to Figure 7, on the factors *watermark* (top) and *stretch* (bottom).

D.5 Intervention and applications

Collateral damage. In the collateral damage experiment (Figure 6), we use the *mean-diff* directions of CelebA attributes and MSCOCO categories as ground-truth semantic directions. For each semantic factor, we assign the direction from each method with the closest angular alignment to its corresponding *mean-diff*. To evaluate disentanglement, we intervene along a target factor until reaching a prescribed change in its alignment with the corresponding *mean-diff*, and then measure the induced drift in all non-target factors. Better disentanglement should yield smaller non-target drift, as modifying one factor should minimally affect the others. We report results averaged over all target factors; for example, on CelebA, each of the 40 attributes is treated in turn as the target, while the drift over the remaining 39 attributes is measured and then averaged across all targets. For a fair comparison, we include only samples for which all methods successfully reach every prescribed target change. The intervention procedure is analogous across ICA, PCA, and SAE, using IC directions, PC directions, and SAE atoms, respectively. For LogReg, we intervene along the normal to the learned separating hyperplane, while for MeanDiff we intervene directly along the *mean-diff* direction as a baseline. As shown in the left and middle panels of Figure 6, ICA exhibits lower non-target drift than the competing methods on both CelebA and MSCOCO using CLIP representations. The right panel further shows that *existence-like* attributes, which exhibit higher negentropy and are better captured by ICA, incur less collateral damage than more continuous or entangled attributes.

Controlled image retrieval. In the controlled image retrieval experiments (Figure 7, Figure 13, Figure 14), we intervene on selected ICA coordinates similarly to the collateral damage experiment. To suppress an undesired semantic factor (for instance, a global feature that strongly influences nearest neighbor retrieval) we zero its corresponding IC coordinate in the whitened space, allowing other important properties, such as identity, to be better preserved (Figure 7, Figure 13). The identity-preservation experiment is conducted primarily on CLIP, since its text aligned representation tends to better preserve identity related nearest neighbors than purely visual encoders such as DINOv2 and I-JEPA. Conversely, a semantic factor can be injected by increasing the magnitude of its corresponding IC coordinate, after which we retrieve the nearest database image that remains similar to the original while exhibiting the desired added property (Figure 14).

Spurious mitigation. In the spurious mitigation experiment (Table 3, Table 6), we use the Waterbirds dataset to evaluate whether ICA can improve worst group accuracy (WG) by identifying and removing directions associated with spurious background features, namely land and water. We identify spurious ICs using the background labels of their most extreme training samples: directions

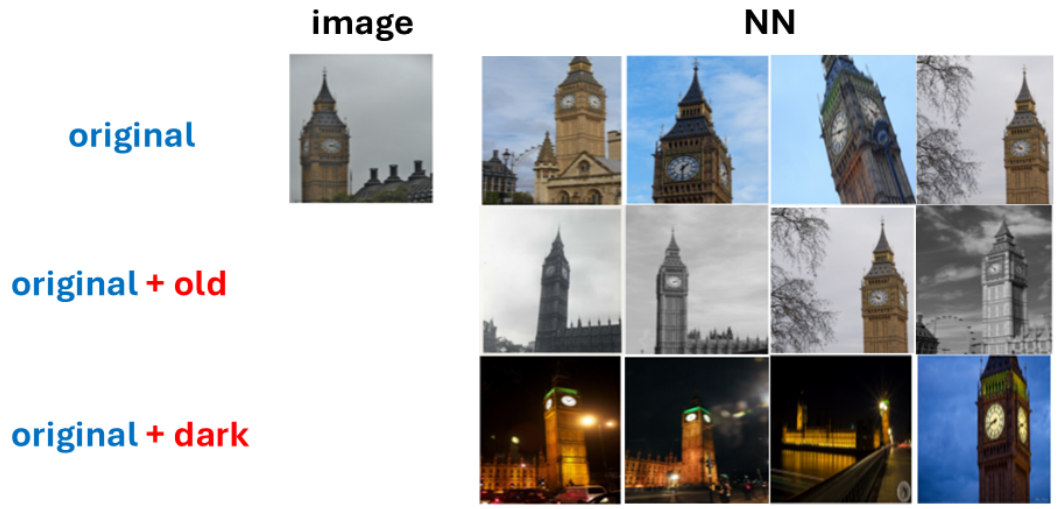


Figure 14: **Controlled image retrieval.** We identify IC directions that represent visual properties such as style (e.g., old-fashioned images in the second row) or lighting (e.g., images captured in the dark in the third row). We then enhance the corresponding coordinates of a specific target image and retrieve its nearest neighbors. This enables us to extract from a database of interest images that are semantically close to the target while exhibiting the desired visual change; for instance, an image of a clock tower in the dark (third row).

whose extremes are highly concentrated on a single background type are ranked as more spurious. We then remove the highest-ranked ICs and apply established spurious-mitigation methods directly to the resulting embeddings, without otherwise modifying their training procedures. We performed an ablation over the number of highest-ranked ICs removed and the number of extreme samples used to rank each IC, and report the hyperparameter combination that achieved the best performance for each setting. For CLIP, the selected configurations were ERM (13 removed ICs, 6 extreme samples), DFR (4 removed ICs, 11 extreme samples), and CA (8 removed ICs, 13 extreme samples). For DINO, the selected configurations were ERM (15 removed ICs, 8 extreme samples) and DFR (3 removed ICs, 7 extreme samples). For I-JEPA, they were ERM (30 removed ICs, 13 extreme samples) and DFR (10 removed ICs, 7 extreme samples). Each experiment was repeated over five independent runs, and we report the mean and standard deviation across runs. Across methods, removing spurious ICs consistently improves WG accuracy while largely preserving average accuracy (AVG). We further extend ERM and DFR from CLIP to DINOv2 and I-JEPA representations and observe the same general improvement after removing the identified spurious ICs.

For the competing methods, **ERM** trains a logistic regression classifier on the full training set over frozen embeddings, without explicitly accounting for group imbalance. **Deep Feature Reweighting (DFR)** (Kirichenko et al., 2023) keeps the representation frozen and retrains only the linear classification head using group balanced validation data, thereby reducing reliance on spurious correlations. In our implementation, balancing is performed through group-based sample weighting, while the logistic regression regularization is selected according to validation WG accuracy. **Contrastive Adapters (CA)** (Zhang & Ré, 2022) operates on frozen CLIP embeddings and learns a lightweight two layer adapter using a combination of classification and supervised contrastive objectives, encouraging difficult samples to align with same class examples while separating them from other classes. We evaluate ERM, DFR, and CA with CLIP, and additionally apply the same ERM and DFR procedures to frozen DINOv2 and I-JEPA embeddings.

Table 6: **Spurious mitigation: average accuracy (AVG).** We report AVG with and without removing spurious ICA directions. STDs are omitted when negligible.

	CLIP			DINO		I-JEPA	
	ERM	DFR	CA	ERM	DFR	ERM	DFR
Original	**89.9**	90.6±1.0	**90.0±0.7**	97.2	95.6±0.4	90.7	**92.3±0.5**
With ICA	89.6	**91.7±0.3**	89.9±0.7	**97.5**	**96.2±0.8**	**91.6**	92.0±0.5

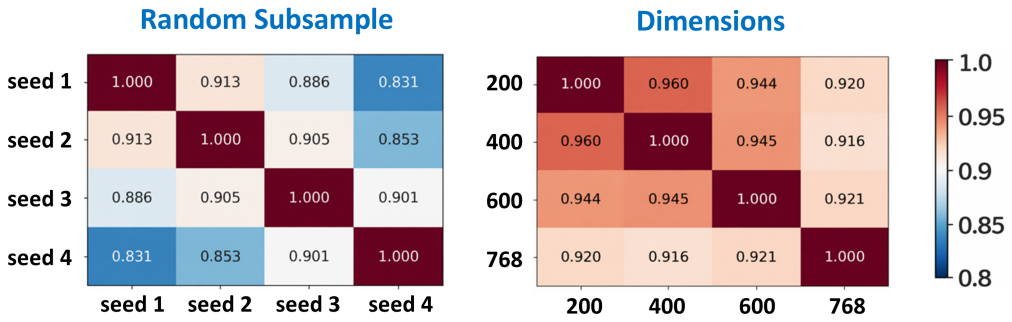


Figure 15: **Sensitivity tests.**

Sensitivity analysis. In Figure 15, we follow a framework similar to Figure 2 and Figure 8, but use the IC direction closest to each *mean-diff* rather than the *mean-diff* directions themselves. The *mean-diff* directions are kept fixed throughout and are computed once from the full image set. We evaluate the stability of the resulting ICA directions under two settings: different random subsamples of the training set (75% of the data), and different numbers of PCA dimensions used before ICA. In both cases, the resulting directions remain highly correlated, indicating strong stability. These experiments are conducted on CelebA using CLIP representations.

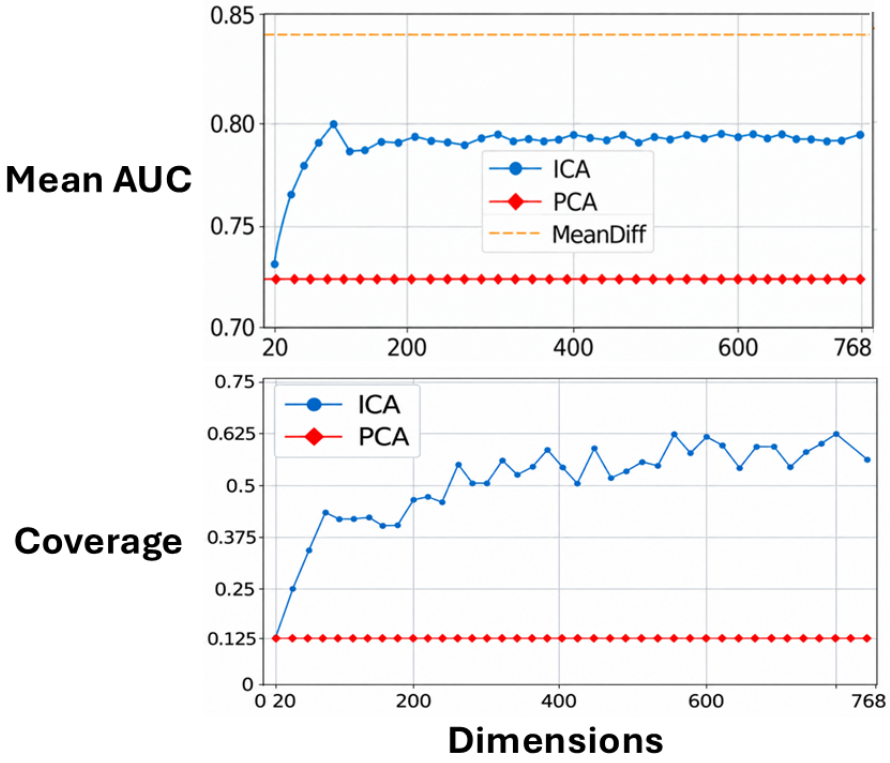


Figure 16: **Ablation study.**

Ablation study. In Figure 16, we follow a framework similar to Table 1 and evaluate the mean AUC and coverage of ICA and PCA as a function of the number of retained PCA components. While PCA performance remains consistently low across dimensionalities, ICA improves rapidly even at relatively low dimensions and then saturates to a stable plateau. Together with the sensitivity analysis, these results indicate that ICA preserves much of its performance even in reduced dimensional representations. This observation is consistent with the low effective dimensionality often exhibited by learned representations, where much of the meaningful variation is concentrated within a relatively small number of principal directions despite the higher nominal dimensionality of the embedding space. Although this provides motivation for operating at reduced dimensionality, we use the full embedding dimension in our main experiments in order to retain a complete oblique coordinate system.