CROSS-MODAL REDUNDANCY AND THE GEOMETRY OF VISION–LANGUAGE EMBEDDINGS

Grégoire Dhimoïla[a,b,c] **Thomas Fel**[d] **Victor Boutin**[e] **Agustin Picard**[c]

[a]Brown University [b]ENS Paris Saclay [c]IRT Saint Exupéry
[d]Kempner Institute, Harvard University [e]CNRS

https://github.com/Parabrele/IsoEnergy

ABSTRACT

Vision–language models (VLMs) align images and text with remarkable success, yet the geometry of their shared embedding space remains poorly understood. To probe this geometry, we begin from the Iso-Energy Assumption, which exploits cross-modal redundancy: a concept that is truly shared should exhibit the same average energy across modalities. We operationalize this assumption with an Aligned Sparse Autoencoder (SAE) that encourages energy consistency during training while preserving reconstruction. We find that this inductive bias changes the SAE solution without harming reconstruction, giving us a representation that serves as a tool for geometric analysis. Sanity checks on controlled data with known ground truth confirm that alignment improves when Iso-Energy holds and remains neutral when it does not. Applied to foundational VLMs, our framework reveals a clear structure with practical consequences: ***(i)*** sparse *bimodal* atoms carry the entire *cross-modal* alignment signal; ***(ii)*** *unimodal* atoms act as *modality-specific* biases and fully explain the modality gap; ***(iii)*** removing *unimodal* atoms collapses the gap without harming performance; ***(iv)*** restricting vector arithmetic to the *bimodal* subspace yields in-distribution edits and improved retrieval. These findings suggest that the right inductive bias can both preserve model fidelity and render the latent geometry interpretable and actionable.

1 INTRODUCTION

Vision-language models (VLMs) (Radford et al., 2021; Zhai et al., 2023; Fini et al., 2025; Tschannen et al., 2025) have become central to applications from visual question answering (Chen et al., 2024) and medical imaging (Singhal et al., 2023) to autonomous driving (Zhou et al., 2024) and embodied AI (Shukor et al., 2025), creating shared embedding spaces where visual and textual representations of similar concepts align. Despite their empirical success, we lack a principled understanding of how these models internally organize and align semantic content across modalities. This work investigates the geometric structure of cross-modal embeddings: *How do VLMs organize the alignment between visual and textual semantics, and what principles govern this shared representational space?* Understanding these mechanisms is crucial for designing more robust and interpretable vision-language architectures.

From Attributions to Concepts. To address this, the interpretability community has developed an array of tools (Gilpin et al., 2018; Bau et al., 2017) aimed at dissecting learned representations. Early efforts centered on attribution methods (Zeiler & Fergus, 2014; Sundararajan et al., 2017; Smilkov et al., 2017; Petsiuk et al., 2018; Fel et al., 2021) that highlight "where" a model focuses its attention, but these approaches often fall short (Nguyen et al., 2021; Kim et al., 2022; Colin et al., 2022; Hase & Bansal, 2020; Sixt et al., 2022) of explaining "what" abstractions and on which basis the model operates. More recent concept-based methods (Ghorbani et al., 2019; Zhang et al., 2021; Fel et al., 2023b; Elhage et al., 2022; Fel et al., 2023a) have emerged to answer this, seeking to extract meaningful features (concepts) that models implicitly compute over. Concept extraction is typically

Correspondence to: gregoire.dhimoila@ens-paris-saclay.fr framed as a dictionary learning problem (Tošić & Frossard, 2011; Rubinstein et al., 2010; Elad, 2010; Mairal et al., 2014; Dumitrescu & Irofti, 2018): identifying an overcomplete set of basis vectors that explain internal activations via sparse coding (Olshausen & Field, 1996; 1997; Lee et al., 2006; Foldiak & Endres, 2008; Rentzeperis et al., 2023). This approach is a direct response to an underlying phenomenology of the activation space: the Linear Representation Hypothesis (LRH) (Elhage et al., 2022; Wattenberg & Viégas, 2024) which posits that model activations can be viewed as sparse combinations of latent directions drawn from a high-dimensional concept basis. This is motivated not only by empirical findings, but also by geometric arguments: in high-dimensional spaces, sparse sets of nearly orthogonal directions can represent exponentially many distinct concepts with minimal interference, a principle reminiscent of Johnson-Lindenstrauss-style (Johnson et al., 1984; Larsen & Nelson, 2017) embeddings. Sparse autoencoders (SAEs) (Makhzani & Frey, 2014; Elhage et al., 2022) directly operationalize this phenomenology by learning an approximate inverse mapping from model activations to latent conceptual directions. They recover a sparse code that identifies which overcomplete basis elements (concepts) are active in a given representation, and have been effective at uncovering semantically meaningful structure in both vision (Gorton, 2024; Fel et al., 2025; Dreyer et al., 2025; Rao et al., 2024) and language models (Cunningham et al., 2023; Bricken et al., 2023; Rajamanoharan et al., 2024; Gao et al., 2025; Surkov et al., 2025).

Multimodal Interpretability. However, applying SAEs to VLMs (Bhalla et al., 2024; Pach et al., 2025) reveals a puzzling behavior: concept dictionaries often segregate by modality (Papadimitriou et al., 2025). Although VLMs are trained for cross-modal alignment, the extracted concepts tend to activate exclusively for either image or text inputs. This observation is the concept-based view of a now well-documented phenomenon called the *modality gap*. Prior work has described this separation geometrically, attributing it to a conical structure in the embedding space (Liang et al., 2022; Ethayarajh, 2019), or through the lens of training dynamics induced by the contrastive loss (Fahim et al., 2024; Shi et al., 2023; Yaras et al., 2024). Yet these accounts do not explain what such separation means at the level of shared concepts.

In this work, we introduce a framework for analyzing multimodal representations grounded in an explicit generative model. Central to this framework is the *Iso-Energy Assumption* – if a concept is truly shared across modalities, it should exhibit invariant energy, defined as the average squared activation, regardless of input domain. This assumption provides a concrete, testable criterion for identifying *bimodal* concepts. We then operationalize this simple assumption with a natural method: an alignment-penalized Matching Pursuit Sparse Autoencoder (Aligned SAE), which encourages energy consistency across modalities during training. This approach allows us to diagnose whether extracted concepts genuinely support cross-modal alignment or merely reflect modality-specific patterns. This work makes the following contributions:

- We introduce the *Iso-Energy Assumption*, which exploits cross-modal redundancy by requiring that shared concepts exhibit the same average activation energy in image and text.
- We operationalize this assumption with an Aligned Sparse Autoencoder that enforces energy consistency during training while preserving reconstruction, and validate it with ground-truth sanity checks where classical SAEs fail.
- Applied to dual-encoder vision–language foundation models, this inductive bias reveals a geometric decomposition invisible to classical SAEs: (***i***) sparse *bimodal* atoms carry the entire *cross-modal* alignment signal, while (***ii***) *unimodal* atoms carry *modality-specific* information and fully explain the modality gap, with a few high energy atoms acting as *modality-specific* biases.
- Moreover, our work reveals that cross-modal information is carried by shared atoms, as opposed to idiosyncratic ones described by Papadimitriou et al. (2025).
- This structure enables actionable interventions without loss of performance: (***iii***) removing *unimodal* atoms collapses the modality gap, and (***iv***) restricting vector arithmetic to the *bimodal* subspace yields in-distribution edits.

At the core lies a single theoretical premise: if concepts are genuinely shared across modalities, they must imprint redundant statistical traces in each domain. We formalize this intuition by modeling multimodal representations as partial inverses of a shared generative process, and introducing the *Iso-Energy Assumption* as the inductive bias that makes concept recovery feasible.

Nomenclature. We distinguish the following terms when referring to concepts based on their behavior and based on the type of information they carry. ***(i) Activation patterns:*** *(a) unimodal* concepts activate exclusively on a single modality, whereas *(b) bimodal* concepts activate on both. This distinction is made by thresholding the modality score μ, a comparison of domain-wise energy (see Appendix E.2.2). ***(ii) Information carried:*** *(c) modality-specific* concepts carry information specific to a given modality, e.g. that an image has cropping artifacts, while *(d) cross-modal* concepts carry cross-modal information and participate in the contrastive geometric alignment between modalities.

2 Related Work

Previous work extensively describes salient phenomena in the shared space of multimodal dual-encoders (Schrodi et al., 2025; Levi & Gilboa, 2025; Jiang et al., 2023; Udandarao, 2022). Most notably, Liang et al. (2022) describes what is now commonly known as the *modality-gap*: that image and text embeddings reside in disjoint cones in the latent space. This modality gap has been attributed to a cone effect (Liang et al., 2022; Ethayarajh, 2019; Schrodi et al., 2025) and training dynamics induced by the contrastive loss and by mismatched pairs of data (Fahim et al., 2024; Shi et al., 2023; Yaras et al., 2024). (Levi & Gilboa, 2025) additionally shows that embeddings are contained near the surface of ellipsoid empty shells centered near the mean of the two distributions μ_I and μ_T.

The *cone effect* naturally comes with a salient difference in modality wise means $\Delta \coloneqq \mu_I - \mu_T$ (Levi & Gilboa, 2025; Liang et al., 2022; Fahim et al., 2024), or a perfect linear separability between image and text embeddings (Schrodi et al., 2025; Levi & Gilboa, 2025; Fahim et al., 2024; Shi et al., 2023). Furthermore, Schrodi et al. (2025) shows that, surprisingly, a small subset of coordinates in the canonical latent basis accounts for most of the norm of Δ. Previous work tries to get rid of the modality gap by shifting modality-wise means (Liang et al., 2022; Levi & Gilboa, 2025), or by projecting out the few canonical directions mentioned above (Schrodi et al., 2025). In all of these works, the proposed intervention decreases cross-modal performance of embeddings, with the notable exception of Zhang et al. (2023), whose Proposition 1 precedes our more general Proposition 1. Figure 4 shows that this difference in means, while explaining the bulk of the modality gap and linear separability between the two sets of embeddings, is not enough to explain the full distributional mismatch. For that, we need to account for modality-specific information.

Jiang et al. (2023) takes an information-theoretic angle to show that modality specific variability in representations is necessary to preserve unimodal capabilities. They use this insight to design a new architecture that would expliciteley regularise for the representation of such information. We show that (*i*) unregularised foundational VLM encoders possess such representations, and that (*ii*) the underlying mechanism is to organise these types of information linearly in distinct subspaces—let us call Ω_I and Ω_T the subspaces containing modality specific information, and Γ the one containing shared information. The projection by Schrodi et al. (2025) described above, though not introduced for this purpose, is a first attempt at identifying and intervening on $\Omega_I \oplus \Omega_T$, characterised by the span of the canonical basis vector selected, while Γ would be the orthogonal complement. However, their results are negative, as their intervention is detrimental to model performance even on cross-modal focused tasks—indicating significant unintended alteration of Γ—while leaving the gap wide open (see Appendix L), thus showing that their identified directions can't describe $\Omega_I \oplus \Omega_T$.

Our novelty lies not in the description of most phenomena discussed above, but rather in the characterisation of these structures in foundational VLM encoders. Through the use of a concept-based approach, we are able to identify (*i*) high-energy unimodal features to unimodal biases, (*ii*) $\Gamma = \text{Cone}(\boldsymbol{\delta} \cdot \boldsymbol{D})$ and $\Omega_{I/T} = \text{Cone}(\boldsymbol{\delta}_{I/T} \cdot \boldsymbol{D})$. Here, $\text{Cone}(\boldsymbol{e}_1, \dots, \boldsymbol{e}_k) = \{\sum_{i=1}^{k} \lambda_i \boldsymbol{e}_i \mid \lambda_i \geq 0\}$, $\boldsymbol{D} \in \mathbb{R}^{K \times d}$ is the set of K dictionary atoms of latent dimension d, and $\boldsymbol{\delta}$ (resp. $\boldsymbol{\delta}_I$, $\boldsymbol{\delta}_T$) $\in \{0,1\}^d$ is the binary mask selecting bimodal (resp. image-, text-only) atoms. To claim this characterisation, we carefully analyse four complementary aspects of the dictionary through novel metrics. Once validated, we further test its practical value through targeted interventions on these structures.

3 Exploiting Cross-Modal Redundancy for Concept Recovery

We formalize the setting by modeling each datum (image, caption, ...) as generated from a latent concept vector and a domain-specific generator. From this perspective, encoders serve as partial inverses of a shared multimodal generative process. However, recovering the underlying concepts is in general an ill-posed inverse problem: without additional inductive biases, nonlinear ICA is provably unidentifiable (Hyvarinen & Morioka, 2016; Locatello et al., 2019; Khemakhem et al., 2020)—infinitely many mappings can account for the observed data in the absence of cross-domain constraints. Introducing Iso-Energy provides a selection principle among these solutions and yields a representation that is useful for studying the geometry observed in practice. We begin by formally

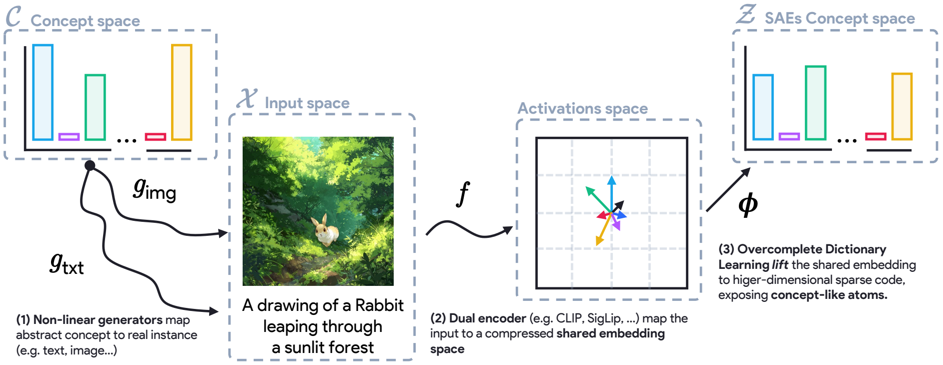
$\mathcal{C}$ Concept space
$\mathcal{X}$ Input space
Activations space
$\mathcal{Z}$ SAEs Concept space
g_{img}
g_{txt}
f
ϕ
(1) **Non-linear generators** map abstract concept to real instance (e.g. text, image...)
A drawing of a Rabbit leaping through a sunlit forest
(2) **Dual encoder** (e.g. CLIP, SigLip, ...) map the input to a compressed **shared embedding space**
(3) **Overcomplete Dictionary Learning *lift*** the shared embedding to higer-dimensional sparse code, exposing **concept-like atoms.**

Figure 1: **Multimodal data-generating process.** A latent concept vector $\boldsymbol{c} \in \mathcal{C}$ (e.g., *rabbit*, *forest*, *light*, *running*) is sampled as a sparse combination of abstract concepts and rendered through domain-specific generators $\boldsymbol{g}(\cdot)$ (e.g., image or text). Dual-encoder models (e.g., CLIP, SigLIP) map these observations to a shared activation space, which sparse autoencoders (or other overcomplete dictionary learning methods) then attempt to *lift* back to concept-like atoms. However, without additional inductive bias, encoder-decoder pairs $(\boldsymbol{f}, \boldsymbol{\phi})$ are not uniquely determined, a well-known identifiability problem in nonlinear ICA. Here we leverage cross-modal redundancy as a useful inductive bias, nudging the solution toward recovering *bimodal* concepts.

introducing our data generative process: each datum arises by (***i***) sampling a sparse concept vector and (***ii***) rendering it through a domain-specific generator. Formally, let $K \in \mathbb{N}$ be the number of latent concepts and $\mathcal{C} \subseteq \mathbb{R}^K$ the concept space. Let $\mathfrak{D} = \{\mathrm{d}_1, \dots, \mathrm{d}_{|\mathfrak{D}|}\}$ index domains with observation spaces $\mathcal{X}^{(\mathrm{d})}$.

Definition 1 (Multimodal Concept Generative Process)**.** *Sample* $\mathbf{c} \sim p = \prod_{k=1}^{K} p_k$ *with* $|\mathrm{supp}(\boldsymbol{c})| \ll K$*. For each* $d \in \mathfrak{D}$*, a deterministic generator* $\mathbf{g}^{(d)} : \mathcal{C} \to \mathcal{X}^{(d)}$ *yields* $\boldsymbol{x}^{(d)} = \boldsymbol{g}^{(d)}(\boldsymbol{c})$*. We admit that each* $\boldsymbol{g}^{(d)}$ *is* C^1 *and locally invertible.*

A VLM encoder $\boldsymbol{f}$ maps each observation $\boldsymbol{x}^{(\mathrm{d})}$ to a shared embedding, and a sparse autoencoder $\boldsymbol{\phi}$ attempts to *disentangle* this embedding back into concept coordinates, so that $\boldsymbol{\phi} \circ \boldsymbol{f}$ approximates the inverse of Definition 1. However, recovering latent concepts from nonlinear generators is not identifiable in general: without additional structure, many different dictionaries can explain the same data (Hyvarinen & Morioka, 2016; Khemakhem et al., 2020; Locatello et al., 2019). This ambiguity is visible in practice, where SAE dictionaries often vary substantially across random seeds (Fel et al., 2025). Domain membership provides auxiliary information, but by itself it is insufficient to ensure consistent recovery. We therefore introduce a cross-domain simple inductive bias, the *Iso-Energy Assumption*, which states that genuinely multimodal concepts should maintain consistent average energy across modalities. Because such concepts manifest in parallel across domains, their observations contain redundant signals that can be exploited to guide dictionary recovery toward more stable and plausible solutions. Formally,

Definition 2 (Iso-Energy Assumption)**.** *Let* $\boldsymbol{\psi} : \bigcup_{d \in \mathfrak{D}} \mathcal{X}^{(d)} \to \mathbb{R}^K$ *denote the learned encoder (VLM* $\boldsymbol{f}$ *composed with SAE* $\boldsymbol{\phi}$*). We say* $\boldsymbol{\psi}$ *satisfies* Iso-Energy *if the second moment of each coordinate is domain-invariant:*

$$\mathbb{E}_{\boldsymbol{X} \in \mathcal{X}^{(d)}} \left(\boldsymbol{\psi}(\boldsymbol{X})_k^2 \right) = \mathbb{E}_{\boldsymbol{X} \in \mathcal{X}^{(d')}} \left(\boldsymbol{\psi}(\boldsymbol{X})_k^2 \right), \quad \textit{with } k \in [\![1, K]\!], \quad (d, d') \in \mathfrak{D}^2.$$

Iso-Energy provides a testable, domain-agnostic inductive bias: *cross-modal* features should correspond to *bimodal* concepts and maintain comparable energy across domains, whereas *modality-specific* factors need not. This constraint narrows the solution set without requiring instance-level matching. It is reminiscent of the rosetta neurons (Dravid et al., 2023; Gresele et al., 2020) and the platonic representation hypothesis (Huh et al., 2024), suggesting that independent models (in our case, the vision and language encoders) converge on shared features.

Operationalization. To operationalize this assumption, we adopt as our base the Matching Pursuit (MP) sparse autoencoder, recently introduced by Costa et al. (2025) and rooted in the original framework of Mallat & Zhang (1993). MP enforces ℓ_0 sparsity through sequential residual updates, which

for any $\boldsymbol{c}$ there exists a neighborhood $U(\boldsymbol{c})$ on which $\boldsymbol{g}^{(\mathrm{d})}$ is injective with continuous inverse.

aligns with the sparse generative model introduced above and has shown strong empirical reconstruction performance in vision compared to ReLU (Bricken et al., 2023), JumpReLU (Rajamanoharan et al., 2024), or BatchTopK SAEs (Gao et al., 2025; Bussmann et al., 2024) (see Appendix A for their formal definition). We then incorporate Iso-Energy into the sparse autoencoder recipe as a soft regularizer: a small penalty encouraging the activations of the same atom to maintain similar strength across domains. Given ℓ_2-normalized codes $\boldsymbol{Z}^{(\mathrm{d})}, \boldsymbol{Z}^{(\mathrm{d}')} \in \mathbb{R}^{b \times K}$, our training loss becomes:

$$\mathcal{L}_{\mathrm{SAE-A}} = \mathcal{L}_{\mathrm{SAE}} + \beta \cdot \mathcal{L}_{\mathrm{align}} \quad \text{with} \quad \mathcal{L}_{\mathrm{align}} = -\tfrac{1}{b}\,\mathrm{Tr}\Big(\boldsymbol{Z}^{(\mathrm{d})}\, \boldsymbol{Z}^{(\mathrm{d}')\top}\Big) \tag{1}$$

with $\mathcal{L}_{\mathrm{SAE}}$ being the training loss of a standard SAE, typically an ℓ_2 norm measuring reconstruction error with a sparsity constraint on the encoder's output. $\mathcal{L}_{\mathrm{align}}$ is the soft inductive bias, where b is the batch size and K the number of latent atoms. With a small weight ($\beta \approx 10^{-4}$, see Appendix B), this regularizer gently biases the dictionary toward *bimodal* features while preserving reconstruction performance. In Section 4, we demonstrate that this bias does not force the model to create multimodal concepts that don't exist in the input data. It is important to note that the minimum of this loss function is consistent with the iso-energy principle for cross-modal concepts formalized in Definition 2: maximizing the cosine similarity of codes coming from the aligned samples from the different modalities leads to codes with the same energy across modalities. From this point onward, we denote the unregularized model as SAE, and its alignment-augmented counterpart as SAE-A.

4 Recovering Multimodal Structure with Aligned SAEs

Sanity check. Before turning to large-scale embeddings, we validate our approach on controlled toy data with known ground truth. We construct two synthetic data-generating processes that mimic CLIP-like cosine similarity statistics (Fig. 9) while giving us exact control over which atoms are *unimodal* and which are *bimodal* (Appendix C.1). Each sample is generated by drawing a sparse code $\boldsymbol{z}$ with $\|\boldsymbol{z}\|_0 = L = 20$ and producing normalized embeddings

$$\boldsymbol{x}^{(\mu)} = \boldsymbol{z}^\top \boldsymbol{D}^{(\mu)}, \qquad \boldsymbol{x}^{(\nu)} = \boldsymbol{z}^\top \boldsymbol{D}^{(\nu)}, \qquad \|\boldsymbol{x}^{(\mu)}\|_2 = \|\boldsymbol{x}^{(\nu)}\|_2 = 1.$$

Two parameters govern the process. First, τ_1 sets the cross-modal alignment of each shared (*bimodal*) atom $k \in B$ by fixing the cosine between its per-modality components. Second, τ_2 fixes the average paired image–text similarity at the embedding level:

$$\cos\angle\big(\boldsymbol{d}_k^{(\mu)}, \boldsymbol{d}_k^{(\nu)}\big) = \frac{\langle \boldsymbol{d}_k^{(\mu)}, \boldsymbol{d}_k^{(\nu)} \rangle}{\|\boldsymbol{d}_k^{(\mu)}\|_2\, \|\boldsymbol{d}_k^{(\nu)}\|_2} = \tau_1, \quad \text{and} \quad \mathbb{E}\Big[\Big\langle \boldsymbol{x}^{(\mu)}, \boldsymbol{x}^{(\nu)} \Big\rangle\Big] = \tau_2.$$

with the Iso-Energy case corresponding to $\tau_1 = 1$ (the bimodal atom is identical across modalities up to scale).

Using this approach, we can generate a distribution that matches the alignment statistics found in CLIP embeddings. We then compare a standard Sparse Autoencoder (SAE) to its ”aligned” variant trained with the loss function in Eq. 1. This new loss encourages energy consistency without compromising reconstruction quality, as indicated by the high R-squared value ($R^2 \geq 0.99$) we observed in all our experiments. Recovery is evaluated against $(\boldsymbol{D}^\star, \boldsymbol{Z}^\star)$ using the Wasserstein distance $\mathcal{W}$ between learned and true atoms (lower is better) and mean matching accuracy (mma) of usage patterns after optimal bipartite matching (higher is better). When Iso-Energy is violated ($\tau_1 \neq 1$), both SAE and SAE-A recover the dictionary equally well ($\mathcal{W} \approx 0.19,\ \mathrm{mma} \approx 0.82$), confirming that the regularizer does not hallucinate *bimodal* atoms. When Iso-Energy holds ($\tau_1 = 1$), the standard SAE fails ($\mathcal{W} = 0.396,\ \mathrm{mma} = 0.29$) while the aligned SAE-A succeeds ($\mathcal{W} = 0.184,\ \mathrm{mma} = 0.52$), showing that the inductive bias is neutral when unnecessary and decisive when appropriate. Full experimental details are provided in Appendix C.2. Having validated the principle in controlled settings, we next study its behavior on embeddings from foundation-scale VLMs.

4.1 Empirical Evaluation on Vision–Language Foundations

Setup. We train both SAE and its aligned counterpart SAE-A on activations from six representative models: CLIP (ViT-B/32, ViT-L/14) (Radford et al., 2021), OpenCLIP, OpenCLIP-L (Cherti et al., 2023), SigLIP (Zhai et al., 2023), and SigLIP2 (Tschannen et al., 2025) (see Appendix D for more details on these models). All SAE models are trained with identical hyperparameters (expansion ratio

Metric	CLIP		CLIP-L		OpenCLIP		OpenCLIP-L		SigLIP		SigLIP2	
MSE (↓)	**0.141**	0.163	**0.207**	0.213	**0.246**	0.257	**0.244**	0.253	**0.212**	0.214	**0.115**	0.115
R^2 (↑)	**0.859**	0.837	**0.793**	0.787	**0.754**	0.742	**0.755**	0.747	**0.788**	0.784	0.884	**0.885**
p_{acc} (↑)	0.847	**0.915**	0.843	**0.868**	0.849	**0.880**	0.845	**0.873**	0.897	**0.899**	0.886	**0.903**
ρ (↑)	0.327	**4.232**	1.566	**4.086**	4.072	**16.02**	8.737	**16.58**	1.370	**2.182**	0.713	**1.475**
FDA (↑)	2.630	**4.559**	3.914	**4.800**	4.369	**8.160**	9.787	**16.49**	8.831	**34.95**	8.246	**18.24**
δ_{r} (↓)	0.224	**0.125**	0.039	**0.021**	0.037	**0.018**	0.001	**-0.000**	0.023	**0.006**	0.007	**-0.006**

Table 1: **Comparison of unregularized vs. aligned sparse autoencoders across six VLMs.** We report reconstruction fidelity (MSE, R^2) and multimodality-sensitive metrics: probing accuracy (p_{acc}), functional alignment (ρ), Functional and Distributional Agreement (FDA), and interventional robustness (δ_{r}). Left values correspond to SAE, right values to SAE-A. While reconstruction is nearly identical, the aligned variant consistently improves on all multimodality metrics : ρ increases by more than an order of magnitude, FDA doubles (or triples), and δ_{r} remains near zero, demonstrating that bimodal atoms alone sustain cross-modal alignment while unimodal ones contribute little beyond modality specific bias.

8, target $\ell_0 = 20$) using a subset of 1 million LAION embeddings chosen at random, ensuring that observed differences are attributable to the Iso-Energy regularizer rather than training artifacts. Classical reconstruction metrics (MSE, R^2) reveal little difference between the two methods, confirming that the alignment penalty does not compromise fidelity. Yet reconstruction alone is uninformative: two dictionaries with identical R^2 may encode radically different concept structures. To capture these distinctions, we introduce a suite of multimodality-sensitive metrics, each designed to probe a complementary aspect of the recovered dictionary. For the full definitions of these metrics, we refer the reader to Appendix E.

Tuning β. In Equation (1), we select β via a log-sweep over $\{10^{-6}, \ldots, 10^{-1}\}$ and pick the largest value such that the difference in explained variance compared to non regularised SAE is less than 0.05. This rule reproduces our settings without hand-tuning and, under this criterion, we also observe no degenerate (always-on) features in practice; see Appendix B for details.

Metrics:

(i) Probing accuracy p_{acc} (Appendix E.2.2). This metric tests whether the geometry of dictionary atoms reflects the modality structure of the embedding space. *Unimodal* atoms should act as strong linear classifiers for domain membership, while *bimodal* atoms should remain domain-agnostic. p_{acc} gathers all these classifiers' performance in a single scalar. High p_{acc} therefore indicates that the dictionary correctly distinguishes modality-specific and shared information.

(ii) Functional alignment ρ (Appendix E.2.2). Beyond geometry, we ask which features actually drive cross-modal alignment. ρ measures the ratio of alignment explained by *bimodal* versus *unimodal* features through instance-level co-activation patterns. Values of $\rho > 1$ indicate that alignment is predominantly supported by *bimodal* concepts, consistent with the Iso-Energy Assumption.

(iii) Functional and Distributional Agreement (FDA) (Appendix E.2.2). At the population level, FDA measures whether the functional role of features is consistent with their geometric organization. While ρ is local and instance-based, FDA checks distributional alignment across large batches. Intuitively, high FDA confirms that *bimodal* atoms globally sustain the match between modalities. Neither ρ nor FDA considers the contrastive aspect of alignment, which is done by our final metric.

(iv) Interventional robustness δ_{r} (Appendix E.2.2). To test for causality, we measure how performance in retrieval tasks changes when we remove *unimodal* features. We use δ_{r} to quantify the change in recall after ablating these *unimodal* atoms. A small δ_{r} suggests that these atoms aren't essential, while a large drop indicates they are necessary for the contrastive aspect of the embeddings.

Results. As shown in Table 1, SAE-A matches SAE's MSE and R^2 scores but consistently outperforms it on all multimodality-sensitive metrics. In particular, ρ increases by more than an order of magnitude across models, indicating that alignment is almost entirely driven by *bimodal* atoms. Likewise, FDA improves substantially, confirming consistency at the distributional level. Probing accuracy improves modestly, reflecting a more distinct geometric separation, while δ_{r} remains small, demonstrating that *unimodal* atoms can be safely ablated without harming retrieval performance and therefore do not contain cross-modal information. Taken together, these results offer converging evidence that, in SAE-A, *unimodal* atoms encode modality-specific information, while *bimodal*

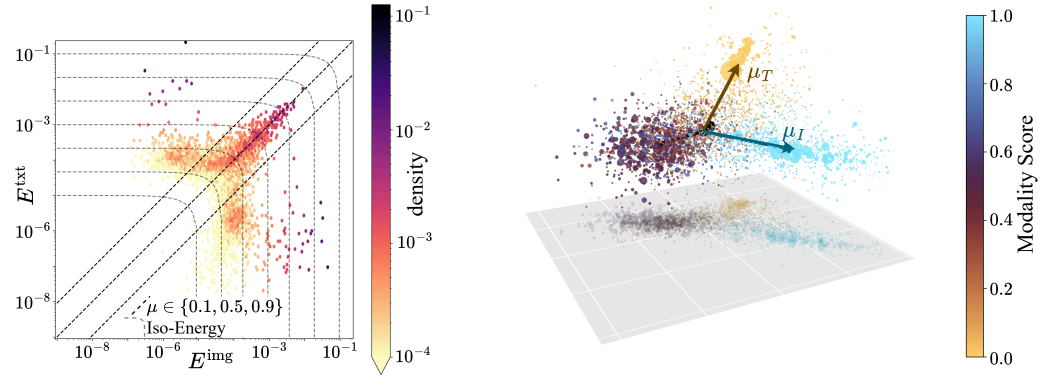

Figure 2: **(Left) Energy distribution across learned atoms.** The majority of features are *bimodal* and medium-energy (inside diagonals defined by constant modality score μ-Appendix E.2.2), while only a handful of high-energy *unimodal* features dominate modality-specific variance. These high-energy *unimodal* atoms behave like modality biases and are responsible for much of the observed modality gap. **(Right) Geometric organization of concepts.** Low-dimensional projections reveal three distinct clusters: image-only, text-only, and *bimodal*. *Unimodal* atoms align with the modality cones of the embedding space, while *bimodal* atoms occupy a modality-agnostic subspace orthogonal to these directions, thereby sustaining cross-modal alignment.

atoms, identified via Iso-Energy, constitute the principal basis for cross-modal alignment. In SAE, however, the picture is less clear, and some *unimodal* atoms carry *cross-modal* information.

These findings indicate that the Iso-Energy Assumption reveals a qualitatively distinct structure: a compact *cross-modal* subspace, entirely spanned by *bimodal* concepts, that retains the contrastive power of the original embeddings while fully supporting cross-modal alignment. As we will see in Section 5, making this structure explicit also makes it actionable: it allows us to manipulate representations directly, from closing the modality gap to performing in-distribution semantic arithmetic. But before making this structure actionable, we propose an in-depth analysis of the solution to understand some aspects of the geometry.

4.2 Concept Geometry Under Iso-Energy

Having established the effectiveness of Iso-Energy on real VLMs, we now turn to a qualitative characterization of the learned concepts. This analysis focuses on how energy is distributed, how atoms are geometrically organized, and how interpretable they are in practice.

Energy distribution. A first observation concerns how energy is distributed across modalities. As illustrated in Figure 2, the vast majority of features are *bimodal* and exhibit moderate energy levels, whereas a small subset of *unimodal* features concentrates disproportionately high energy. These high-energy *unimodal* atoms dominate modality-specific variance and act as biases, as further detailed in Appendix F.

Geometric organization. The high values of p_{acc} show a near-perfect alignment between latent structure and concept organization. This can be visualized by projecting the learned atoms into a low-dimensional space, revealing a clear separation into three clusters: image-only, text-only, and *bimodal* (Fig. 2, right). *Unimodal* atoms align tightly with the cones spanned by image and text embeddings, reproducing the geometry of the modality gap. In contrast, *bimodal* atoms occupy a modality-agnostic subspace, orthogonal to the *unimodal* directions. This geometry explains why *bimodal* (resp. *unimodal*) atoms carry cross-modal (resp. modality-specific) information.

Qualitative inspection. Finally, we examine the semantic meaning of individual atoms by inspecting their most activating examples (Appendix G). *Bimodal* atoms are semantically stable, consistently capturing the same concept across modalities (e.g., colors, objects, actions). *Unimodal* atoms, on the other hand, often reflect idiosyncratic modality-specific signals (such as poor cropping artifacts in images or “name patterns” in text) that contribute little to cross-modal semantics. Together, these

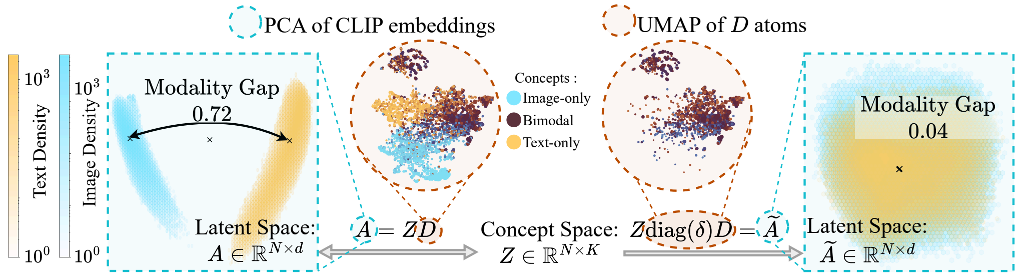

Figure 3: **The modality gap arises from multiple *unimodal* concepts, while *bimodal* concepts are sufficient to sustain cross-modal alignment. Left:** CLIP embeddings are re-expressed through a learned dictionary. A PCA projection highlights the separation between modalities, and a UMAP layout distinguishes two types of atoms: *unimodal* and *bimodal*. **Right:** Removing *unimodal* atoms with a binary mask $\boldsymbol{\delta} \in \{0, 1\}^K$ closes the gap. The reconstructed embeddings $\widetilde{\boldsymbol{A}}$ continue to support retrieval, indicating that *bimodal* atoms alone capture the structure necessary for alignment.

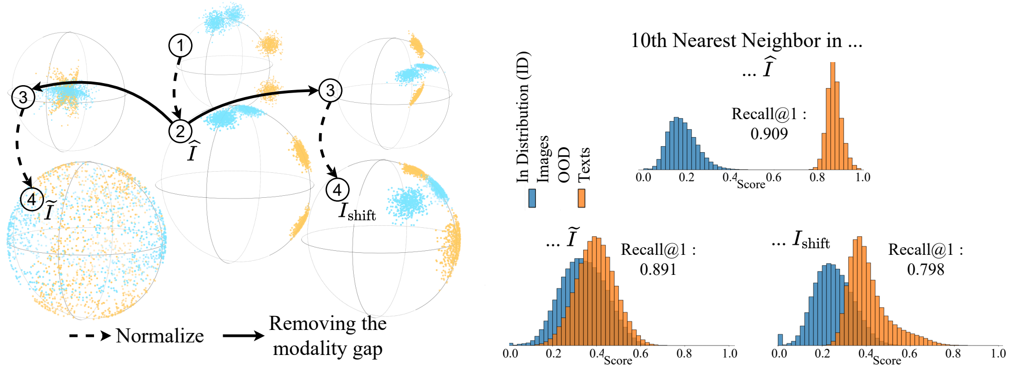

Figure 4: **Filtering *unimodal* atoms closes the modality gap without harming performance. (Left)** Synthetic illustration comparing our method with the embedding shift baseline (Liang et al., 2022). Only our approach merges image and text distributions. **(Right)** Histogram of distances from each image (ID) and caption (OOD) embedding to its 10th nearest image neighbor. The modality gap is measured as the separation between the ID and OOD histograms. Filtering *unimodal* atoms aligns the two distributions, whereas shift degrades performance and leaves the gap wide open.

three perspectives converge on the same conclusion: ***unimodal* atoms function as modality-specific biases, while *bimodal* atoms encode the shared conceptual backbone that supports cross-modal alignment.**

5 Actionable Interventions on Multimodal Embeddings

Together, these analyses demonstrate that our dictionaries yield a structured and interpretable decomposition of multimodal embeddings. Building on this foundation, we now shift from analysis to intervention: once Iso-Energy isolates the cross-modal backbone, it enables direct manipulation of embeddings in ways that were previously inaccessible. In particular, we consider the minimal intervention that removes modality information and examine its effect on two structural aspects. In fact, we show that such transformations are possible without altering ranking-related capabilities even when the modality information is non-trivial (e.g., not a bias, as proposed by Zhang et al. (2023)'s Proposition 1) and under realistic assumptions of orthogonality.

Proposition 1 (Modality information removal impact on ranking.)**.** *Consider $\boldsymbol{v} \in \mathbb{R}^d$ with decomposition $\boldsymbol{v} = \omega(\boldsymbol{x}) + \gamma(\boldsymbol{x})$ where $\omega(\boldsymbol{x}) \in \Omega$ encodes modality-specific information, $\gamma(\boldsymbol{x}) \in \Gamma$ captures cross-modal content, and $\mathbb{R}^d = \Omega \oplus \Gamma$. If visual and textual information are orthogonal, then ranking preservation is guaranteed.*

Proof. See Appendix K. □

Closing the modality gap. First, we find that by filtering out *unimodal* atoms with a binary mask (Figure 3), we nearly eliminate the modality gap while preserving retrieval and zero-shot performance.

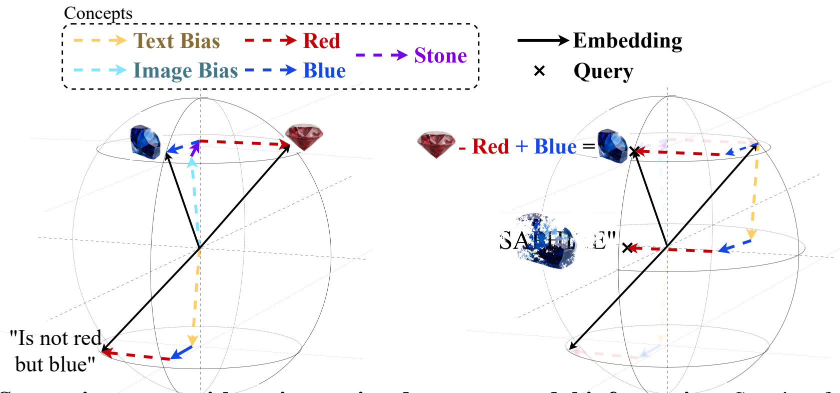

Figure 5: **Semantic vector arithmetic restricted to cross-modal information.** Starting from a Ruby (red stone), the target is a Sapphire (blue stone). The classical edit vector $\boldsymbol{\Delta} = \text{Text} + \text{Blue} - \text{Red}$ is polluted by *unimodal* directions, producing a query $\boldsymbol{Q} = \boldsymbol{I}_{\text{src}} + \boldsymbol{\Delta}$ that drifts out-of-distribution. In contrast, restricting to *bimodal* atoms yields $\boldsymbol{Q}_{\text{SAE}}$, which lies on the semantic manifold and reliably retrieves the correct target. This illustrates how *unimodal* features inject modality-specific bias into $\boldsymbol{\Delta}$, while Iso-Energy isolates the truly shared concepts that support valid semantic arithmetic.

As illustrated in Figure 4, this intervention merges the image and text distributions, unlike embedding shift baselines (Liang et al., 2022), which enforce matching means but still leave distributions well separated. Crucially, our approach preserves contrastive capabilities, showing that our dictionaries faithfully capture both contrastive and modality-specific information separately through *bimodal* and *unimodal* atoms. Measuring the modality gap is typically done by measuring the distance between the mean of text and image distributions, or by measuring their linear separability. However, these fail to account for distributional mismatches remaining post intervention. For this reason, we chose to turn to the out-of-distribution (OOD) literature to measure the modality gap, and borrow a method described by Sun et al. (2022). This method measures the separation between the blue and orange histograms in Figure 4. Our method consists in the following intervention on the activations, indicated with a tilde: $\widetilde{\boldsymbol{A}} \coloneqq (\boldsymbol{Z} \odot \boldsymbol{\delta})\boldsymbol{D}$, where $\boldsymbol{Z} \in \mathbb{R}^{N \times K}$ contains the sparse codes, $\boldsymbol{D}$ is the concept dictionary, and $\boldsymbol{\delta}$ is the binary mask filtering out *unimodal* features, broadcast to the size of $\boldsymbol{Z}$. Reconstructed activations are indicated with a hat: $\widehat{\boldsymbol{A}} \coloneqq \boldsymbol{Z}\boldsymbol{D}$. The embedding shift intervention, and variants described in the appendix, consists in adding a modality-wise constant, essentially moving the mean of each distribution $\boldsymbol{\mu}_{I/T}$. Liang et al. (2022)'s shift method transforms the images by $\boldsymbol{I}_{\text{shift}} \coloneqq \boldsymbol{I} - \boldsymbol{\mu}_I + \frac{\boldsymbol{\mu}_I + \boldsymbol{\mu}_T}{2}$, similarly for texts $\boldsymbol{T}_{\text{shift}}$.

Semantic vector arithmetic. Iso-Energy also grounds semantic manipulations. Let I_{src} be the source image, and Δ be the textual description of the difference between the source and target image – i.e., the relative caption. Restricting vector arithmetic to *bimodal* atoms $Q_{\text{SAE}} \coloneqq I_{\text{src}} + \widetilde{\Delta}$ (Figure 5) produces queries that remain in-distribution (Figure 6 and Table 2) while preserving retrieval (Appendix H). In contrast, classical arithmetic $Q = I_{\text{src}} + \Delta$ incorporates *unimodal* noise from the text embedding on top of the interesting cross-modal information, yielding degenerate queries that drift outside the embedding distribution. Our intervention consistently produces queries in-distribution without degrading performance on the FashionIQ benchmark Wu et al. (2021), demonstrating that the *bimodal* backbone revealed by Iso-Energy is practically useful.

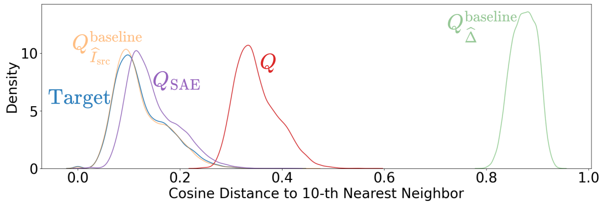

Figure 6: **Out-of-distribution behavior of semantic queries.** Histogram of distances between each query and its 10th nearest neighbor in the target image distribution. Classical arithmetic $\boldsymbol{Q} = \boldsymbol{I}_{\text{src}} + \boldsymbol{\Delta}$ drifts out-of-distribution, while our *bimodal*-restricted query $\boldsymbol{Q}_{\text{SAE}}$ remains aligned with the target space. Baselines using only the source image ($\boldsymbol{Q}^{\text{baseline}}_{\widehat{\boldsymbol{I}}_{\text{src}}}$) or only the caption difference ($\boldsymbol{Q}^{\text{baseline}}_{\widehat{\boldsymbol{\Delta}}}$) confirm the expected extremes: perfectly in-distribution and fully OOD, respectively.

OOD score (↓)	$\boldsymbol{Q}$	$\boldsymbol{Q}_{\text{SAE}}$
CLIP	0.97	**0.77**
CLIP-L	0.95	**0.76**
OpenCLIP	0.86	**0.68**
OpenCLIP-L	0.87	**0.72**
SigLIP	0.99	**0.70**
SigLIP2	0.99	**0.61**

Table 2: OOD scores for classical queries vs. concept-based queries Q_{SAE}. Lower is better, best results per line in bold.

Consider the following example. Let I_{src} be an image of a ruby, Δ be the prompt *"is not red but blue"* and the target be an image of a saphire. This is the example illustrated in Figure 5. Adding Δ to the ruby produces a query that contains both textual-only and visual-only concepts and therefore that do not correspond to any realistic embedding. Adding only the cross-modal concepts of Δ, however, produces a query that actually corresponds to an image of a saphire.

6 Conclusion

The Iso-Energy Assumption introduces a simple yet effective inductive bias for analyzing multimodal representations. In synthetic settings, it facilitates the recovery of ground-truth structure; in large-scale vision–language models, it consistently reveals a compact *bimodal* basis that supports cross-modal alignment. This basis makes multimodal concepts accessible, isolates *unimodal* concepts, closes the modality gap, and enables controlled semantic edits even in foundation-scale VLMs. In contrast, standard sparse autoencoders tend to learn a diffuse mixture of *unimodal* and *bimodal* atoms despite similar reconstruction quality, which obscures the structure that underpins alignment.

Despite its promising results, our approach has several limitations. First, the alignment penalty is sensitive to the choice of its weighting coefficient β: when too small, it becomes ineffective; when too large, it can lead to degenerate feature representations. In this work, we select β through simple sweeps, but a more principled calibration—e.g., by coupling the penalty to a performance constraint or stability criterion—remains an important direction for future work. Second, our analysis is conducted on reconstructions produced by the sparse autoencoder, rather than on the original embeddings. This constrains the reported performance to the autoencoder's reconstruction regime. Finally, our experiments are limited to dual-encoder vision–language models. Whether the same structural invariants and alignment properties hold in models with cross-attention mechanisms or generative training objectives remains an open question.

More broadly, our results support a hypothesis-driven approach to interpretability: inductive biases should be selected to reflect the structural properties relevant to downstream tasks, rather than applied indiscriminately. When properly aligned with task-relevant objectives, even simple biases can illuminate underlying mechanisms that might otherwise remain hidden—and offer actionable control without compromising core performance.

References

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

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

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.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *ArXiv e-print*, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024.

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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:2832–2845, 2022.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. *arXiv preprint arXiv:2506.03093*, 2025.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *ArXiv e-print*, 2023.

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

Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1934–1943, 2023.

Maximilian Dreyer, Jim Berend, Tobias Labarta, Johanna Vielhaben, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Mechanistic understanding and validation of large ai models with semanticlens. *Nature Machine Intelligence*, pp. 1–14, 2025.

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

Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer International Publishing, 2010.

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.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024.

Thomas Fel, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:26005–26014, 2021.

Thomas Fel, Victor Boutin, Mazda Moayeri, Remi Cadene, Louis Bethune, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:54805–54818, 2023a.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2711–2721, 2023b.

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. *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9641–9654, 2025.

Peter Foldiak and Dominik Maria Endres. Sparse coding. *ArXiv e-print*, 2008.

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

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

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

Liv Gorton. The missing curve detectors of inceptionv1: Applying sparse autoencoders to inceptionv1 early vision. *ArXiv e-print*, 2024.

Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Scholkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. *Uncertainty in Artificial Intelligence*, 2020.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.491. URL https://aclanthology.org/2020.acl-main.491/.

Sai Sumedh R Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *ArXiv e-print*, 2025.

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 Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 2016.

Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7661–7671, 2023.

William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: Evaluating the human interpretability of visual explanations. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pp. 280–298, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19774-1. doi: 10.1007/978-3-031-19775-8_17. URL https://doi.org/10.1007/978-3-031-19775-8_17.

Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th annual symposium on foundations of computer science (FOCS)*, pp. 633–638. IEEE, 2017.

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

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=QGUju9B68Z.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:17612–17625, 2022.

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. *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Scholkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 4114–4124, 2019.

Julien Mairal, Francis Bach, Jean Ponce, et al. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:26422–26436, 2021.

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.

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.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *ArXiv e-print*, 2025.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces. *ArXiv e-print*, 2025.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 151, 2018.

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, Janos Kramar, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *ArXiv e-print*, 2024.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pp. 444–461, 2024.

Ilias Rentzeperis, Luca Calatroni, Laurent U Perrinet, and Dario Prandi. Beyond l1 sparse coding in v1. *PLoS Computational Biology*, 19(9):e1011459, 2023.

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

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uAFHCZRmXk.

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 e-print*, 2021.

Peiyang Shi, Michael C Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in clip. In *ICLR 2023 workshop on multimodal representation learning: perks and pitfalls*, 2023.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. Smolvla: A vision-language-action model for affordable and efficient robotics. *ArXiv e-print*, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. Do users benefit from interpretable vision? a user study, baseline, and dataset. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 20827–20840, 2022.

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

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.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *ArXiv e-print*, 2025.

Vishaal Udandarao. Understanding and fixing the modality gap in vision-language models. *Master's thesis, University of Cambridge*, 32, 2022.

Martin Wattenberg and Fernanda Viégas. Relational composition in neural networks: A survey and call to action. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11307–11317, 2021.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 32598–32611, 2022.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. *ArXiv e-print*, 2024.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pp. 818–833, 2014.

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

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. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 35(13):11682–11690, 2021.

Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/pdf?id=D-zfUK7BR6c.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, 2024.