REFRACTION: BEYOND AXIS-ALIGNED INTERPRETABLE SUBSPACES IN VISUAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neuron explanation (NE) is a crucial task for opening the "black box" of deep neural networks, linking neurons to human-understandable textual concepts. Existing research predominantly focuses on single-neuron explanation (SNE), while multi-neuron explanation (MNE) has received less attention, despite evidence that neurons jointly construct various subspaces to encode concepts. In this paper, we aim to discover interpretable subspaces more comprehensively. We first show that axis-aligned discovery in existing MNE methods can overlook interpretable subspaces. These methods rely on highly co-activated neurons, whereas meaningful concepts can also be encoded by weighted combinations of neurons with individually moderate activations, a phenomenon we term **mesosemanticity**. Motivated by this finding, we propose **Refraction**, which adaptively constructs new axes as weighted combinations of the original neuron axes to discover and explain multi-neuron subspaces. Specifically, Refraction uses generalized eigendecomposition to contrast representative images within a neuron-activation cluster with similar images outside it, identifying directions with high target-to-background response energy. We also prove that the classical MNE method FALCON is a special case of Refraction. To evaluate MNE, we extend Concept Synthesis (CoSy) to multi-neuron subspaces, termed MN-CoSy, which measures the alignment between subspaces and their textual explanations. We also introduce Effective Interpretable Subspace Count (EISC) to measure the quality and efficiency of discovered subspaces. Extensive experiments demonstrate substantial improvements over existing NE methods. Further analyses provide evidence for mesosemanticity by revealing interpretable subspaces missed by axis-aligned discovery and showing how the learned axes combine contributions from multiple neurons to represent visual concepts.

1 INTRODUCTION

The rise of foundation models has made neural networks larger and more complex. Understanding their behavior is essential for building trustworthy machine learning systems (Dwivedi et al., 2023). Neuron explanation (NE) aims to describe the function of internal neurons using human-understandable concepts (Oikarinen et al., 2025). These descriptions can help identify the concepts represented by a model and diagnose errors in downstream tasks.

NE methods can be broadly divided into single-neuron explanation (SNE) and multi-neuron explanation (MNE). Most existing NE work has focused on SNE. For example, Network Dissection (Bau et al., 2017) matches the spatial activations of individual neurons to annotated visual concepts. MILAN (Hernandez et al., 2022) generates natural-language descriptions from images that highly activate a neuron. CLIP-Dissect (Oikarinen & Weng, 2022) uses a pretrained vision-language model to match neurons with concepts without requiring concept annotations. DnD (Bai et al., 2025) uses language models to propose descriptions and generated images to select among them. SIEVE (Ji et al., 2026) selects neurons with discriminative responses and verifies their proposed concepts using generated images.

Despite substantial progress in SNE, MNE has received less attention: it seeks to identify subspaces formed by multiple neurons and assign them human-understandable textual concepts. An early MNE method, HINT (Wang et al., 2022), identifies neurons that collaborate to represent concepts

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.

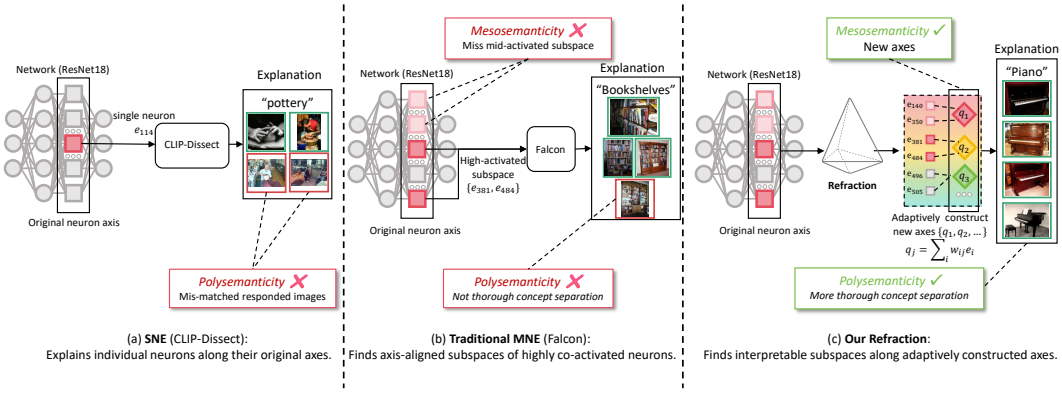


Figure 1: Illustrative comparison of Refraction with single-neuron explanation (SNE) and conventional multi-neuron explanation (MNE). SNE assigns explanations to individual neurons and is therefore vulnerable to polysemanticity. FALCON partly alleviates this problem by explaining jointly highly activated neuron groups, but its axis-aligned search can overlook subspaces encoded by moderately activated neurons (mesosemanticity) and remains susceptible to polysemanticity. Refraction instead adaptively constructs new axes as weighted combinations of the original neuron axes, extending subspace discovery beyond axis-aligned spaces. This broader search recovers the overlooked subspaces associated with mesosemanticity, and more thoroughly separates concepts that share neurons, further mitigating polysemanticity. Best viewed in color on a computer.

in a predefined concept hierarchy. FALCON (Kalibhat et al., 2023) discovers neuron subspaces that are jointly highly activated by probe images. It names these groups using captions of highly activating image regions and uses low-activation images to filter out spurious concepts. More recently, NeurFlow (Cao et al., 2025) studies functional interactions between neuron groups across network layers.

Although SNE has received far more attention than MNE, many SNE methods assign a single explanation to each neuron, implicitly treating its behavior as an individual unit, as shown in Fig. 1(a). This assumption is challenged by **polysemanticity** (Kalibhat et al., 2023; Elhage et al., 2022): the same neuron can participate in different multi-neuron subspaces that encode different concepts. Recent SNE methods such as SPICE (Lee et al., 2026) address polysemanticity by clustering images that activate the same neuron and assigning a separate explanation to each cluster. However, these explanations remain centered on individual neurons and do not identify the multi-neuron subspaces associated with the different concepts. MNE offers a more direct route to this goal: discovering such subspaces and associating them with human-understandable textual concepts.

In this paper, we identify a second, distinct phenomenon in neural representations. In our experiments, neurons with individually moderate responses can jointly encode a visual concept: a weighted combination of their activations distinguishes concept-relevant images from similar background images. We call this phenomenon **mesosemanticity**, as shown in Fig. 1(b). Polysemanticity concerns how one neuron can contribute to multiple concepts; mesosemanticity concerns how multiple moderately responsive neurons can contribute to one concept. The former challenges assigning a single explanation to each neuron, while the latter can cause MNE methods to miss interpretable subspaces when they search only for axis-aligned subspaces spanned by highly co-activated neurons. This axis-aligned search may also leave concepts that share neurons insufficiently separated, so these methods remain susceptible to polysemanticity.

Motivated by mesosemanticity, we propose **Refraction**, as shown in Fig. 1(c), to discover and explain multi-neuron subspaces beyond those spanned by selected original neurons. Refraction changes the axes along which visual representations are examined, constructing new axes as weighted combinations of the original neuron axes. For each cluster of probe-image features, it contrasts representative images within the cluster with similar images outside it. Generalized eigendecomposition identifies directions with high response energy in the target images relative to this background. The resulting directions form a subspace, which Refraction names using its highly re-

sponsive images. By searching along these new directions, Refraction can uncover concepts jointly encoded by neurons whose individual responses are only moderate. We further theoretically demonstrate that FALCON's axis-aligned subspaces are a special case of the Refraction subspaces.

A further challenge for MNE is evaluation: it lacks a widely adopted automatic measure for comparing the subspaces discovered by different methods. For SNE, Concept Synthesis (CoSy) (Kopf et al., 2024), referred to here as SN-CoSy, has made quantitative comparisons easier. MNE evaluations, in contrast, often rely on human judgments, which are costly and difficult to scale. We therefore extend SN-CoSy to **MN-CoSy**, which tests whether images generated from a textual explanation elicit stronger responses from its associated subspace than control images. We also introduce **Effective Interpretable Subspace Count (EISC)**, which measures how many distinct, well-explained subspaces a method discovers, and EISC-Rate, which normalizes this count by the number of outputs. Experiments show that Refraction performs strongly MN-CoSy and EISC. They also provide empirical evidence for mesosemanticity and reveal interpretable subspaces missed by axis-aligned discovery.

Our contributions are summarized as follows:

- We identify and empirically validate **mesosemanticity**: visual concepts can be encoded by weighted combinations of neurons with individually moderate responses. We show that restricting discovery to highly co-activated neurons can consequently miss interpretable subspaces.
- We propose **Refraction**, which uses target–background generalized eigendecomposition to construct new feature axes and discover multi-neuron subspaces beyond axis-aligned groups. We prove that FALCON's axis-aligned subspaces are a special case of Refraction's subspace representation.
- We introduce **MN-CoSy** and **EISC**, together with EISC-Rate, to evaluate explanation alignment and effective subspace discovery. Experiments demonstrate Refraction's effectiveness and further examine the presence and impact of mesosemanticity.

2 Related Work

2.1 Single-Neuron Explanation

Neuron explanation (NE) links internal neurons to human-understandable textual concepts. Early studies mainly explained individual neurons, and this single-neuron explanation (SNE) setting remains common today. Network Dissection (Bau et al., 2017) matches activation maps with pixel-level concept annotations, while CompExp (Mu & Andreas, 2020) combines annotated concepts using logical operators. MILAN (Hernandez et al., 2022) describes neurons from highly activating image regions, whereas CLIP-Dissect (Oikarinen & Weng, 2022) selects concepts using a pretrained vision-language model Radford et al. (2021) without concept annotations. INVERT (Bykov et al., 2023) contrasts high- and low-response inputs to find compositional concepts; its formulation also applies to other scalar neural representations. More recently, DnD (Bai et al., 2025) uses language models to propose descriptions and generated images to select among them, while SIEVE (Ji et al., 2026) selects discriminative neurons and verifies their concepts using generated images. CoSy (Kopf et al., 2024) evaluates neuron explanations by comparing responses to concept-conditioned synthetic images and control images.

However, a single explanation may not capture all the concepts associated with a polysemantic neuron. SPICE (Lee et al., 2026) clusters images that activate the same neuron and assigns a separate explanation to each cluster. Moreover, our observation of mesosemanticity points to another limitation of studying neurons individually: even a moderately responsive neuron can contribute substantially to a concept when combined with other neurons.

2.2 Multi-Neuron Explanation

Multi-neuron explanation (MNE) seeks to associate concepts with the joint activity of neurons. HINT (Wang et al., 2022) identifies neurons that collaborate to represent concepts in a predefined hierarchy. FALCON (Kalibhat et al., 2023) discovers groups of jointly highly activated neurons, names

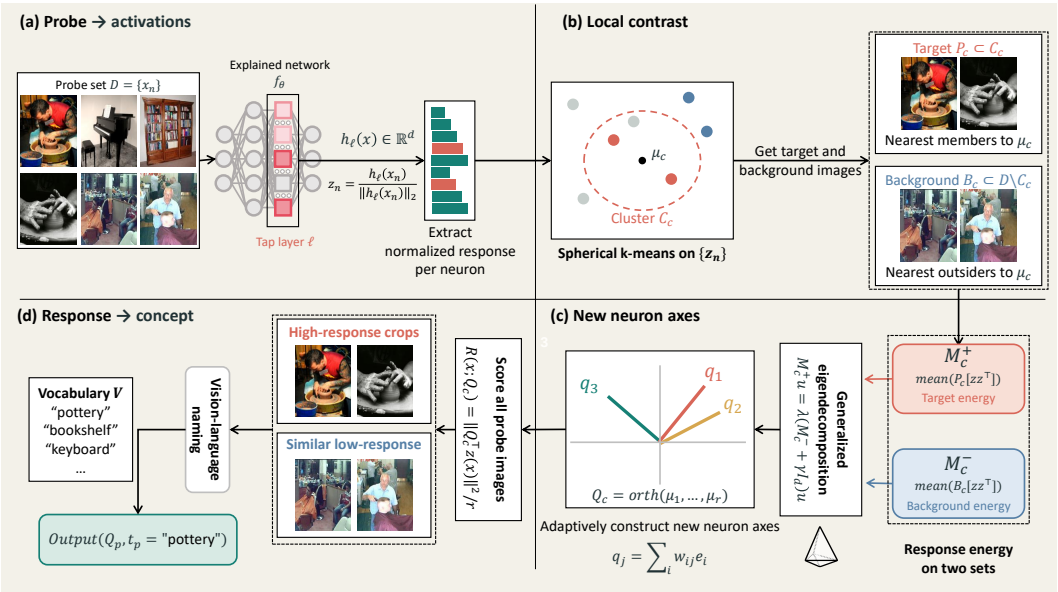


Figure 2: Overview of Refraction. (a) Probe images are passed through the network, and activations at the explained layer are normalized into neuron-response vectors. (b) Spherical k-means defines a cluster; representative members form the target set, while images outside the cluster but near its center form the background set. (c) Generalized eigendecomposition contrasts the target and background second-moment matrices to construct new axes, which are orthonormalized into a subspace basis. (d) The subspace scores probe images, and a vision-language naming procedure uses high-response crops, similar low-response images, and the vocabulary to assign a textual concept. Refinement and candidate filtering are omitted for clarity.

them using captions of activating image regions, and filters proposed concepts using low-activation images. NeurFlow (Cao et al., 2025) studies functional interactions between neuron groups across network layers. A related line of work uses sparse autoencoders to learn interpretable features from model activations (Huben et al., 2024; Pach et al., 2026); these methods introduce an additional trained representation rather than directly naming subspaces discovered in the original network.

FALCON is closest to our work, but it searches only axis-aligned subspaces formed by jointly highly activated neurons. This can miss combinations of moderately activated neurons, as highlighted by mesosemanticity, and leave concepts that share neurons insufficiently separated, so polysemanticity remains a problem. In contrast, Refraction constructs new axes from weighted neuron combinations, extending the search beyond axis-aligned subspaces. It thereby recovers overlooked subspaces and explains concepts that share neurons more accurately, mitigating both limitations.

3 Methodology

Given a pretrained vision model and a probe set, we seek visual concepts represented jointly by neurons in a chosen layer. We first describe a case in which a concept is carried by a combination of neurons without requiring any one of them to be highly activated (Section 3.1). We then introduce Refraction, which finds such combinations by contrasting images within a feature cluster against similar images outside it (Section 3.2). Finally, we extend Concept Synthesis to multi-neuron subspaces and define a quality-weighted count of distinct interpretable subspaces (Section 3.3). We also provide a detailed algorithm in Alg. 1.

Notation and tasks. Let f_θ be a fixed vision model, ℓ the layer to be explained, and $\mathcal{D} = \{x_n\}_{n=1}^N$ a probe set. We pool the spatial locations of a convolutional feature map, or the patch tokens of a

vision transformer, to obtain $h_\ell(x) \in \mathbb{R}^d$. Write

$$z(x) = \frac{h_\ell(x)}{\|h_\ell(x)\|_2}, \qquad z_n = z(x_n), \tag{1}$$

where d is the number of neurons in the chosen representation. The single-neuron explanation (SNE) task takes a specified neuron $i \in \{1, \ldots, d\}$ and returns a textual concept t_i describing its response. Multi-neuron explanation (MNE) does not prespecify a group: it returns a set of pairs $\{(Q_p, t_p)\}_p$, where the orthonormal columns of $Q_p \in \mathbb{R}^{d \times r_p}$ span an r_p-dimensional subspace and t_p is its textual explanation. A fixed vocabulary $\mathcal{V}$ supplies candidate concepts during naming. For a subspace Q, we use the mean squared projection

$$R(x; Q) = \frac{1}{r}\|Q^\top z(x)\|_2^2, \qquad Q^\top Q = I_r, \tag{2}$$

as its response. This definition also covers a single original neuron by setting $Q = e_i$, the i-th coordinate vector, for which $R(x; e_i) = z_i(x)^2$.

3.1 Mesosemanticity

Selecting neurons whose individual responses exceed a high threshold is a natural way to find a jointly active group, as in FALCON (Kalibhat et al., 2023). It does not, however, cover every combination that can distinguish a visual concept. For a unit direction $u \in \mathbb{R}^d$,

$$u^\top z(x) = \sum_{i=1}^{d} u_i z_i(x). \tag{3}$$

Several terms in this sum may be moderate when viewed separately, yet their combination can respond consistently to one set of images and weakly to visually similar images outside that set. We call this phenomenon *mesosemanticity*: concept-discriminating activity can arise from the weighted combination of neurons that need not be individually highly activated. The term describes an observed form of distributed response, not a new type of neuron or a claim that moderate activation alone implies interpretability. It motivates searching over directions rather than only over subsets of highly activated original neurons.

3.2 Refraction

Refraction changes the directions along which a representation is examined. It first uses the probe features to identify local groups of related images. For each group, it then finds directions that respond more strongly to representative images in that group than to similar images outside it. The resulting directions are weighted combinations of the original neuron axes. The procedure requires no change to the target model.

Target and background images. As shown in Fig.2(b), Refraction partitions $\{z_n\}_{n=1}^N$ into C clusters by spherical k-means. For cluster c, let μ_c be its unit-normalized center and $\mathcal{C}_c$ its member indices. In our experiments, we fix $n_+ = 32$ and $n_- = 128$. The initial target set P_c consists of the n_+ members with the largest $z_n^\top \mu_c$. To construct a difficult comparison, the background set B_c contains the n_- images outside $\mathcal{C}_c$ that are closest to the same center. Thus, the target and background are both near μ_c, but only the target images belong to the cluster initially. These sets are constructed separately for each cluster; they are not object-class partitions.

Contrastive directions. As shown in Fig.2(c), for each pair (P_c, B_c), form the uncentered second-moment matrices

$$M_c^+ = \frac{1}{|P_c|}\sum_{n \in P_c} z_n z_n^\top, \qquad M_c^- = \frac{1}{|B_c|}\sum_{n \in B_c} z_n z_n^\top. \tag{4}$$

We solve the regularized generalized eigenvalue problem

$$M_c^+ u = \lambda(M_c^- + \gamma I_d)u, \qquad \gamma > 0. \tag{5}$$

For a direction u, the associated generalized Rayleigh quotient is

$$\lambda(u) = \frac{u^\top M_c^+ u}{u^\top (M_c^- + \gamma I_d) u}. \tag{6}$$

A large value therefore indicates high mean squared response on the target images relative to the similar background, with γ stabilizing the denominator. We retain the r leading generalized eigenvectors and orthonormalize them in the Euclidean metric to obtain Q_c. Their span, rather than any particular ordering or sign of its basis vectors, defines the discovered subspace. Its response is given by Eq. 2.

Refinement and naming. After the initial decomposition, we update the target and background sets once:

$$\widetilde{P}_c = \mathrm{TopK}_{32}\big(R(x_n; Q_c) : n \in [N]\big), \tag{7}$$

$$\widetilde{B}_c = \mathrm{TopK}_{128}\big(z_n^\top \mu_c : n \in [N] \setminus (\mathcal{C}_c \cup \widetilde{P}_c)\big), \tag{8}$$

where TopK_k returns the indices of the k largest scores and $[N] = \{1, \ldots, N\}$. Thus, the updated target may lie outside the original cluster, whereas the background remains outside it and excludes the updated target. We recompute the subspace from $\widetilde{P}_c$ and $\widetilde{B}_c$. Then, we reselect background images from outside the original cluster, excluding the final high-response images. Candidate subspaces are prioritized using the fraction of their final high-response images that belong to the source cluster, their response gap against the background, split-half stability, and the mean generalized eigenvalue. Near-duplicates are removed using both subspace overlap and overlap of their high-response image sets. The remaining subspaces are named from their high-response image regions using a fixed vision-language naming procedure; low-response, visually similar images help reject concepts shared with the background. As shown in Fig.2(d), given a concept t_p, we denote by $\mathrm{conceptSupport}(p)$ the number of high-response crops for which t_p appears among the top-ranked text matches. This support ranks the discovered subspaces for Top-K analysis; evaluation scores are not used to rank them.

Relation to FALCON. FALCON's neuron groups span axis-aligned subspaces of the original representation. We show in Appendix A.1 that these subspaces are a special case of Refraction's subspace representation, with strict inclusion when $r < d$.

3.3 Two New Metrics for Multi-Neuron Explanation

To facilitate the automated evaluation of MNE methods, we propose two new metrics: MN-CoSy and EISC. MN-CoSy asks whether an explanation matches the response of a discovered subspace. EISC asks how many distinct, well-explained subspaces a method discovers under a fixed output budget. The two measures therefore separate explanation alignment from the effective size of the discovered set.

3.3.1 MN-CoSy

Concept Synthesis (CoSy) (Kopf et al., 2024) evaluates a textual explanation by generating images from Stable Diffusion (SD) model (Rombach et al., 2022) and comparing a neuron's response on those images with its response on a natural-image control set. We extend the evaluation object from a neuron to a subspace Q_p. Let $X_p^+ = \{\tilde{x}_j\}_{j=1}^m$ be images generated from the text t_p, and let $X^0 = \{x_j^0\}_{j=1}^n$ be the shared control set. With $a_p(x) = R(x; Q_p)$, MN-CoSy reports

$$\mathrm{AUC}_p = \Pr\big[a_p(\tilde{x}) > a_p(x^0)\big] + \tfrac{1}{2}\Pr\big[a_p(\tilde{x}) = a_p(x^0)\big], \tag{9}$$

$$\mathrm{MAD}_p = \frac{\mathrm{mean}_{\tilde{x} \in X_p^+}\, a_p(\tilde{x}) - \mathrm{mean}_{x \in X^0}\, a_p(x)}{\mathrm{std}_{x \in X^0}\, a_p(x)}. \tag{10}$$

Here the probabilities are empirical over synthetic–control image pairs, and the standard deviation uses the control population. For a FALCON group we set $Q_p = E_I$; for Refraction we use its learned Q_p. All methods use the same generator, prompt protocol, and control set. In the separate SNE evaluation, the neuron i is fixed across methods and both are scored on $a_i(x) = z_i(x)^2$, irrespective of any subspace used to find its explanation.

3.3.2 EISC

Comparing only mean MN-CoSy scores can favor a method that returns a few easy-to-name or redundant subspaces. We therefore define the *Effective Interpretable Subspace Count* (EISC) for the

first K subspaces selected without consulting their MN-CoSy scores. Each subspace receives the quality weight

$$w_p = \max\{0,\, 2\,\mathrm{AUC}_p - 1\} \in [0, 1]. \tag{11}$$

To identify near-duplicate responses, let T_p contain the indices of the m_s control images with the largest $R(x; Q_p)$. We compare two signatures by their Jaccard similarity $J(p, q) = |T_p \cap T_q|/|T_p \cup T_q|$ and use complete-link clustering at a fixed threshold τ_J to obtain groups $\mathcal{G}_1, \ldots, \mathcal{G}_L$. In our experiments, $m_s = 100$ and $\tau_J = 0.5$. The clustering is constructed once from the joint set of subspaces being compared at a given layer and is held fixed for all values of K. For a method's selected set $\mathcal{S}_K$, define

$$\mathrm{EISC}(\mathcal{S}_K) = \sum_{j:\, \mathcal{G}_j \cap \mathcal{S}_K \neq \varnothing} \frac{\sum_{p \in \mathcal{G}_j \cap \mathcal{S}_K} w_p}{|\mathcal{G}_j \cap \mathcal{S}_K|}, \tag{12}$$

$$\text{EISC-Rate}(\mathcal{S}_K) = \frac{\mathrm{EISC}(\mathcal{S}_K)}{K}, \qquad K > 0. \tag{13}$$

Each represented response group contributes at most one quality-weighted count; repeated subspaces do not increase the count simply by being repeated. EISC-Rate gives the effective count per selected subspace. Both are proxies based on response and synthetic-image alignment, rather than a count of manually verified semantic concepts. We compare methods at the same K: FALCON groups are ordered by occurrence support and Refraction subspaces by conceptSupport. Neither ordering uses the evaluation scores.

4 Experiments

4.1 Experimental Settings

We evaluate ImageNet-1K-pretrained ResNet-18 and ResNet-50 (He et al., 2016) at layers 2–4, and ViT-B/16 (Liu et al., 2021) at blocks 4, 8, and 12. Both methods use the same 50,000 ImageNet training images as probes. Following CoSy (Kopf et al., 2024), MN-CoSy uses the 50,000 ImageNet validation images as controls.

We evaluate each method with two independent vocabularies, CLIP-Dissect-20k and LAION-NP-20k, constructed as described in Appendix A.3.2. Within each vocabulary, both methods use the same contrastive text-selection rule adapted from FALCON (Kalibhat et al., 2023); response localization and image-budget differences are detailed in Appendix A.3.3. We compare their outputs at matched Top-K budgets.

Refraction uses $(C, r, \gamma) = (486, 2, 0.01)$ on ResNet-50 and ViT-B/16, while its ResNet-18 settings vary by layer and vocabulary. All configurations are listed in Appendix A.3.1.

4.2 Quantitative Results

Explanation alignment. Tables 1, 2, and 3 report the average results for the three backbones. Refraction obtains higher mean MN-CoSy AUC and MAD than FALCON under both vocabularies on each backbone. The per-layer results provide a more qualified picture: FALCON has higher AUC for ResNet-18 layer 4 with CLIP-Dissect-20k and for ResNet-50 layer 4 with either vocabulary. Refraction nevertheless has higher MAD in every reported layer–vocabulary and block–vocabulary comparison.

Effective subspace discovery. Refraction also achieves higher average EISC and EISC-Rate on all three backbones under both vocabularies. The advantage is particularly clear in the intermediate ResNet-50 layer and ViT-B/16 block 8. It is not uniform across individual layers: FALCON has slightly higher EISC on ResNet-50 layer 2 with LAION-NP-20k and on layer 4 with the same vocabulary. These comparisons indicate that Refraction generally returns more distinct subspaces with well-aligned explanations at matched output budgets, rather than merely increasing the average score of a small set of subspaces.

Table 1: Average MN-CoSy and EISC results across ResNet-18 layers 2–4. Per-layer results are reported in Table 5. Bold indicates the higher score within each vocabulary.

Method	Vocabulary	MN-CoSy AUC	MN-CoSy MAD	EISC	EISC-Rate
FALCON	CLIP-Dissect-20k	0.7868	3.7205	23.7245	0.5931
Refraction (Ours)		**0.9051**	**6.9890**	**29.8619**	**0.7465**
FALCON	LAION-NP-20k	0.7763	3.8451	22.8214	0.5705
Refraction (Ours)		**0.9100**	**7.6748**	**32.1664**	**0.8042**

Table 2: Average MN-CoSy and EISC results across ResNet-50 layers 2–4. Per-layer results are reported in Table 6. Bold indicates the higher score within each vocabulary.

Method	Vocabulary	MN-CoSy AUC	MN-CoSy MAD	EISC	EISC-Rate
FALCON	CLIP-Dissect-20k	0.6204	1.3716	13.2416	0.3464
Refraction (Ours)		**0.7755**	**6.2210**	**21.7096**	**0.5585**
FALCON	LAION-NP-20k	0.6724	1.8955	16.6824	0.4357
Refraction (Ours)		**0.8010**	**8.3004**	**23.9938**	**0.6175**

4.3 Ablation Study

We examine whether Refraction's gains depend on a particular parameter choice. On ResNet-18 layers 2 and 3 with two independent vocabularies, we vary the cluster count C, subspace dimension r, and regularization coefficient γ one at a time, yielding 17 fixed configurations for each layer–vocabulary setting.

Results could found in Appendix A.3.5. The results show that **Refraction's advantage is not limited to a single parameter choice.** At $K = 40$, 63 of the 68 configuration–layer–vocabulary combinations outperform FALCON on all four metrics. In three of the four settings, all 17 configurations do so. For each setting, we highlight the configuration with the highest Top-40 AUC and keep it fixed across the entire Top-K curve. All four highlighted configurations outperform FALCON on all four metrics at every K from 1 to 40. In particular, the EISC and EISC-Rate gains show that Refraction discovers more effective interpretable subspaces after accounting for explanation quality and redundancy. Although the preferred configuration varies across layers and vocabularies, the advantage persists across most tested configurations.

4.4 Qualitative Results

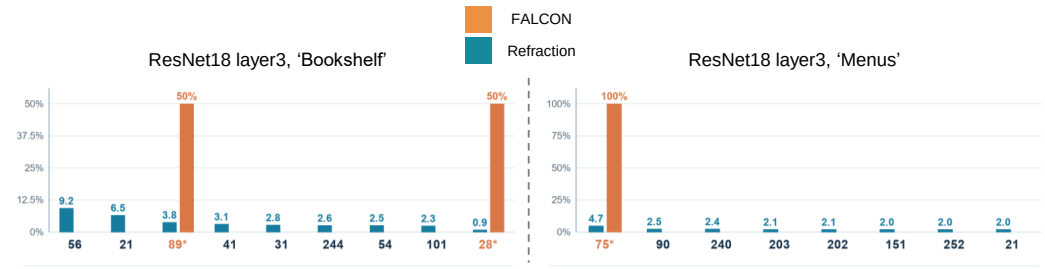
Evidence of Mesosemanticity. Figure 3 illustrates mesosemanticity: a concept-specific response can emerge from combining neurons whose individual activations are not dominant. The *menus* example shows this most clearly. FALCON assigns all coordinate mass to one neuron, whose median activation is 0.188. Refraction instead spreads its mass across multiple coordinates; one of its largest contributors has a median activation of only 0.054. Nevertheless, Refraction achieves higher MN-CoSy AUC (0.9784 versus 0.7510) and MAD (3.995 versus 1.369), and all six displayed high-response images match the concept, compared with one of six for FALCON.

The *bookshelf* example shows a similar difference in subspace structure: FALCON divides its mass equally between two axes, whereas Refraction's largest coordinate accounts for only 9.2% of its mass. Their AUC scores are nearly identical (0.9995 and 1.0000), but Refraction has a higher MAD (10.406 versus 8.074) and more concept-matching images among its top responses. Together, these examples illustrate why restricting discovery to a few individually prominent neurons can overlook useful multi-neuron combinations. More examples could be found in Appendix A.3.6.

More Subspaces found by Refraction. We also visually compared the interpretable subspaces identified by Falcon and Refraction; specific data and discussion can be found in the Appendix A.3.6. In general, Refraction discovers more effective interpretable subspaces and achieves higher image matching scores.

Table 3: Average MN-CoSy and EISC results across ViT-B/16 blocks 4, 8, and 12. Per-block results are reported in Table 7. Bold indicates the higher score within each vocabulary.

Method	Vocabulary	MN-CoSy AUC	MN-CoSy MAD	EISC	EISC-Rate
FALCON	CLIP-Dissect-20k	0.5847	0.4193	7.8836	0.2499
Refraction (Ours)		**0.7910**	**5.5079**	**14.9783**	**0.6164**
FALCON	LAION-NP-20k	0.6161	0.6451	9.3361	0.3094
Refraction (Ours)		**0.8587**	**6.8360**	**18.9978**	**0.7374**



(a) The largest coordinate masses of the subspaces associated with 'Bookshelf' and 'Menus.' Refraction draws on several original neuron axes, including axes with moderate individual activations, whereas FALCON concentrates its mass on one or two axes.



(b) Six highest-response real probe images for each method.
Green borders mark images matching the assigned concept; Red borders mark semantic mismatches.

Figure 3: Illustrations of the mesosemanticity phenomenon identified in this work, using ResNet-18 layer-3 representations.

5 Conclusion

We identified mesosemanticity: a visual concept can be encoded by neurons whose individual responses are only moderate, yet whose weighted combination is concept-discriminating. This exposes a limitation of searching only for highly co-activated, axis-aligned neuron groups. To search beyond these groups, Refraction constructs new axes through target–background generalized eigendecomposition; its subspace representation includes FALCON's axis-aligned subspaces as a special case. We also introduced MN-CoSy and EISC to assess explanation alignment and effective subspace discovery, respectively. Across ResNet-18, ResNet-50, and ViT-B/16, Refraction improves the average results under both concept vocabularies. Qualitative examples further show interpretable subspaces missed by axis-aligned discovery and more concept-consistent high-response images. These results suggest that examining weighted neuron directions can reveal interpretable structure that individual neurons and coordinate-aligned groups leave hidden.

AI USE STATEMENT

We used OpenAI conversational assistants (ChatGPT and Codex) for feedback on hypotheses, methodology, experimental design, and mathematical formulations; assistance with proof drafts, code implementation and debugging, parameter exploration, experiment orchestration, and result interpretation; and drafting and editing manuscript text, creating quantitative plots, and preparing preliminary visualization drafts. We used SDXL 1.0 to generate concept-conditioned evaluation images under the CoSy protocol.

The authors originated the research question and its neuron-explanation framing, identified and verified the cited literature, proposed generalized eigendecomposition for discovering latent subspaces, recognized the representational special-case relationship between FALCON and Refraction, and designed and selected the experiments. The authors determined the paper's argument and structure, drew the final conceptual and methodological illustrations, revised and finalized the quantitative visualizations, and edited the manuscript sentence by sentence. Generative AI did not originate these central ideas or independently determine the final methods, experiments, or claims.

We checked AI-assisted code and analyses against the original experimental artifacts and reviewed the mathematical arguments. AI-assisted image-match judgments were used for preliminary qualitative screening, not as an independent human evaluation. The authors take responsibility for the final text, claims, code, analyses, and figures.

REFERENCES

Nicholas Bai, Rahul Ajay Iyer, Tuomas Oikarinen, Akshay R. Kulkarni, and Tsui-Wei Weng. Interpreting neurons in deep vision networks with language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

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.

Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, and Marina Höhne. Labeling neural representations with inverse recognition. *Advances in Neural Information Processing Systems*, 36:24804–24828, 2023.

Tue Cao, Nhat Hoang-Xuan, Hieu Pham, Phi Le Nguyen, and My Thai. Neurflow: Interpreting neural networks through neuron groups and functional interactions. In *International Conference on Learning Representations*, volume 2025, pp. 73329–73352, 2025.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM computing surveys*, 55(9):1–33, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2022.

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.

ZeBin Ji, Yang Hu, Xiuli Bi, Bo Liu, and Bin Xiao. Select, hypothesize and verify: Towards verified neuron concept interpretation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31316–31325, June 2026.

Neha Kalibhat, Shweta Bhardwaj, C Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. In *International Conference on Machine Learning*, pp. 15623–15638. PMLR, 2023.

Laura Kopf, Philine L Bommer, Anna Hedström, Sebastian Lapuschkin, Marina M Höhne, and Kirill Bykov. Cosy: Evaluating textual explanations of neurons. *Advances in Neural Information Processing Systems*, 37:34656–34685, 2024.

Sehyun Lee, Dahee Kwon, Damin Lee, and Jaesik Choi. Spice: Simple polysemantic feature interpretation via clustering-based explanation. *arXiv preprint arXiv:2609.13198*, 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163, 2020.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. 2022.

Tuomas Oikarinen, Ge Yan, and Tsui-Wei Weng. Evaluating neuron explanations: A unified framework with sanity checks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47090–47131. PMLR, 13–19 Jul 2025.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *Advances in Neural Information Processing Systems*, 38:95706–95742, 2026.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Andong Wang, Wei-Ning Lee, and Xiaojuan Qi. Hint: Hierarchical neuron concept explainer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10244–10254. IEEE, 2022.

A Appendix

A.1 FALCON and Refraction: Subspace Representations

Theorem 1 (FALCON as an axis-aligned special case)**.** *Fix a subspace dimension $1 \le r \le d$. For an index set $I \subseteq \{1, \dots, d\}$ with $|I| = r$, let $E_I = [e_i]_{i \in I}$. Define the sets of subspace projectors*

$$\mathcal{P}_{\mathrm{F}}^{(r)} = \{E_I E_I^\top : I \subseteq \{1, \dots, d\},\ |I| = r\}, \tag{14}$$

$$\mathcal{P}_{\mathrm{R}}^{(r)} = \{QQ^\top : Q \in \mathbb{R}^{d \times r},\ Q^\top Q = I_r\}. \tag{15}$$

Then $\mathcal{P}_{\mathrm{F}}^{(r)} \subseteq \mathcal{P}_{\mathrm{R}}^{(r)}$, with strict inclusion when $r < d$. Moreover, under the response in Eq. 2, a basis Q has the same response as a FALCON coordinate group I for every $z \in \mathbb{R}^d$ if and only if $Q = E_I O$ for some $r \times r$ orthogonal matrix O.

Proof. The columns of E_I are orthonormal, so $E_I E_I^\top$ belongs to $\mathcal{P}_{\mathrm{R}}^{(r)}$. If $r < d$, let

$$Q_* = \left[\frac{e_1 + e_{r+1}}{\sqrt{2}}, e_2, \dots, e_r\right].$$

Its columns are orthonormal, but $Q_* Q_*^\top$ has a nonzero off-diagonal entry, whereas every projector $E_I E_I^\top$ is diagonal. Thus the inclusion is strict.

If $Q = E_I O$ with $O^\top O = I_r$, then $QQ^\top = E_I E_I^\top$, and Eq. 2 gives identical responses for every z. Conversely, equality of the responses for every z implies

$$z^\top (QQ^\top - E_I E_I^\top) z = 0 \qquad \text{for all } z \in \mathbb{R}^d.$$

The matrix in parentheses is symmetric, so it must be zero. The two projectors therefore have the same range. Since Q and E_I are orthonormal bases of that range, $O = E_I^\top Q$ is orthogonal and $Q = E_I O$. □

Discussion. In particular, choosing $Q = E_I$ gives

$$R(x; E_I) = \frac{1}{r} \sum_{i \in I} z_i(x)^2.$$

This result concerns subspace representation and the common projection-based response. It does not imply that Refraction's generalized eigendecomposition reproduces FALCON's co-activation thresholds, group discovery, or concept naming.

A.2 Algorithm for Refraction

To better help readers understand our algorithmic workflow, we provide the complete PyTorch-style pseudocode in Alg. 1.

A.3 Additional Experiments

A.3.1 Hyperparameter Configurations

Table 4 lists the (C, r, γ) settings used for each model, layer, and vocabulary. The target size, background size, and refinement procedure are specified in Section 3.

A.3.2 Implementation of Vocabulary

FALCON originally uses LAION-400M captions as candidate descriptions, whereas CLIP-Dissect uses a fixed 20,000-word vocabulary. This difference matters for CoSy-based evaluation: free-form captions may contain overly specific, compositional, or non-visual details that an image generator cannot reliably reproduce. Comparing methods under their original, different vocabularies can

Algorithm 1 PyTorch-like pseudo-code for Refraction

Require: Probe set $\mathcal{D}$, fixed model f_θ, layer ℓ, vocabulary $\mathcal{V}$; cluster count C, rank r, regularizer γ
Ensure: Named subspaces $\{(Q_p, t_p)\}_p$: subspace bases Q_p and their textual concepts t_p

Compute normalized neuron activations z_n using Eq. 1
Run spherical k-means to obtain $(\mathcal{C}_c, \mu_c)$ for $c = 1, \ldots, C$
Fix target size $n_+ = 32$ and background size $n_- = 128$
for $c = 1, \ldots, C$ **do**
    $P_c \leftarrow$ the n_+ members of $\mathcal{C}_c$ nearest to μ_c
    $B_c \leftarrow$ the n_- images outside $\mathcal{C}_c$, excluding P_c, nearest to μ_c
    Form M_c^+ and M_c^- from P_c and B_c using Eq. 4
    Solve Eq. 5 for the leading r directions and orthonormalize them as Q_c
    Compute $R(x_n; Q_c)$ for all $x_n \in \mathcal{D}$
    $P_c \leftarrow$ the n_+ highest-response images in $\mathcal{D}$
    $B_c \leftarrow$ the n_- images outside $\mathcal{C}_c$, excluding the updated P_c, nearest to μ_c
    Re-form M_c^+ and M_c^- and re-solve Eq. 5 to update Q_c
    Compute $R(x_n; Q_c)$ for all $x_n \in \mathcal{D}$
    $H_c \leftarrow$ the n_+ highest-response images in $\mathcal{D}$
    Re-select B_c outside $\mathcal{C}_c$, excluding H_c, for the final diagnostics
    Record Q_c, H_c, response gap, and split-half stability
end for
Prioritize candidates by the recorded diagnostics and remove near-duplicates
for all retained subspaces Q_p **do**
    Select high-response regions and a similar low-response background
    Assign textual concept t_p using the fixed contrastive naming procedure and vocabulary $\mathcal{V}$
end for
return $\{(Q_p, t_p)\}_p$

Table 4: Hyper-parameter configurations used in the multi-neuron experiments. C is the number of clusters, r the subspace rank, and γ the regularization coefficient. Each entry reports (C, r, γ).

Model	Layer / Block	CLIP-Dissect-20k	LAION-NP-20k
ResNet-18	Layer 2	$(486, 8, 0.01)$	$(486, 2, 0.03)$
	Layer 3	$(486, 2, 0.001)$	$(486, 8, 0.01)$
	Layer 4	$(128, 2, 0.01)$	$(486, 2, 0.1)$
ResNet-50	Layers 2–4	$(486, 2, 0.01)$	$(486, 2, 0.01)$
ViT-B/16	Blocks 4, 8, 12	$(486, 2, 0.01)$	$(486, 2, 0.01)$

therefore conflate explanation quality with the difficulty of generating images from their concepts. To retain a LAION-derived concept source while making the candidate set more comparable in size and granularity, we construct LAION-NP-20k, a fixed vocabulary of frequent noun phrases.

We sample 4,096 captions from each LAION-400M metadata shard with a fixed seed, extract adjective–noun sequences containing at least one noun, and retain phrases of at most five tokens. After removing generic image and web-metadata terms, we select the 20,000 phrases with the highest caption frequency. CLIP-Dissect-20k uses CLIP-Dissect's 20,000-word list. We evaluate both FALCON and Refraction separately under each vocabulary, so their comparisons within a vocabulary use identical candidate concepts.

A.3.3 Implementation of Concept Naming

Response-conditioned crops. For each discovered subspace, we select its highest-response probe images and crop the regions associated with its response. For a ResNet, Refraction obtains the response map by applying Grad-CAM to $R(x; Q)$; if this map is constant, we use the squared projection of the spatial activations onto Q instead. For a FALCON group, the response map is the sum of its selected post-ReLU channel maps. We min–max normalize each map and crop the smallest

Table 5: Per-layer Results of MN-CoSy and EISC on ResNet-18 layers 2–4. Results are compared at matched Top-K (All $K = 40$). Bold indicates the higher score within each layer and vocabulary.

Layer	Method	Vocabulary	MN-CoSy AUC	MN-CoSy MAD	EISC	EISC-Rate
Layer 2	FALCON	CLIP-Dissect-20k	0.6591	1.6082	15.9904	0.3998
	Refraction (Ours)		**0.9280**	**6.2176**	**25.7370**	**0.6434**
	FALCON	LAION-NP-20k	0.6591	1.4799	16.5066	0.4127
	Refraction (Ours)		**0.8452**	**4.9223**	**25.9893**	**0.6497**
Layer 3	FALCON	CLIP-Dissect-20k	0.8308	3.4221	27.5947	0.6899
	Refraction (Ours)		**0.9409**	**4.6293**	**34.3128**	**0.8578**
	FALCON	LAION-NP-20k	0.7883	3.7025	24.6635	0.6166
	Refraction (Ours)		**0.9568**	**6.3447**	**35.5441**	**0.8886**
Layer 4	FALCON	CLIP-Dissect-20k	**0.8704**	6.1312	27.5883	0.6897
	Refraction (Ours)		0.8465	**10.1202**	**29.5359**	**0.7384**
	FALCON	LAION-NP-20k	0.8814	6.3528	27.2941	0.6824
	Refraction (Ours)		**0.9282**	**11.7572**	**34.9658**	**0.8741**

image region containing locations with normalized values of at least 0.6. If a valid region cannot be obtained, we retain the whole image.

ViT-B/16 has no CNN activation map. For a subspace at block j, we instead compute a response-conditioned map from the patch tokens of block $j-1$. For patch s, its value is

$$A_s(x;Q) = \left| \sum_k H_{s,k}^{j-1}(x) \frac{\partial R(x;Q)}{\partial H_{s,k}^{j-1}(x)} \right|, \tag{16}$$

where $H_{s,k}^{j-1}(x)$ is channel k of patch token s. The resulting values form a 14×14 map. We use this same localization rule for FALCON and Refraction on ViT, with Q set to the corresponding group's coordinate basis for FALCON.

Text selection. We encode the high-response crops, the selected background images, and each candidate concept with CLIP ViT-B/32. The background images are encoded as whole images. Within each vocabulary, we retrieve the top five concepts for each high-response crop and aggregate their softmax scores across crops. Concepts also strongly retrieved from the background images are excluded; we select the highest-scoring remaining concept. If none remains, we select the concept with the largest mean high-minus-background CLIP similarity. We process CLIP-Dissect-20k and LAION-NP-20k independently. The $\mathrm{conceptSupport}$ of a selected concept is the number of high-response crops for which it appears among the top five matches.

Image budgets and filtering. Refraction uses at most 32 high-response images per candidate. In the ResNet-50 and ViT-B/16 extensions, FALCON is also limited to 32 high-response images per group; the ResNet-18 FALCON reference uses all images supporting a group. We record the squared norm of the mean normalized CLIP embedding of the high-response crops as a consistency diagnostic. Candidates with a value no greater than 0.8 are excluded at ResNet layer 4. We do not impose this hard filter at ResNet layers 2–3 or ViT blocks 4, 8, and 12.

A.3.4 Additional Qualitative Results

Re-evaluating FALCON on SN-CoSy. In the original CoSy paper (Kopf et al., 2024), FALCON was considered inferior to other methods (such as CLIP-dissect) in single-neuron explanation tasks. However, our analysis revealed that the CoSy evaluation did not decouple the method itself from the vocabulary used. When we re-evaluated CLIP-dissect and FALCON under standardized conditions, we found that FALCON actually performs no worse than CLIP-dissect—and in most cases, even better; the experimental results are presented in the Table 8.

Table 6: Per-layer Results of MN-CoSy and EISC on ResNet-50 layers 2–4. FALCON and Refraction are compared at the same output budget within each layer and vocabulary. Bold indicates the better score in each pair.

Layer	Method	Vocabulary	MN-CoSy AUC	MN-CoSy MAD	EISC	EISC-Rate
Layer 2 ($K = 40$)	FALCON	CLIP-Dissect-20k	0.5874	0.6363	10.8485	0.2712
	Refraction (Ours)		**0.7372**	**2.6342**	**16.2503**	**0.4063**
	FALCON	LAION-NP-20k	0.6587	0.9205	**15.8724**	**0.3968**
	Refraction (Ours)		**0.7182**	**2.4793**	15.6740	0.3918
Layer 3 ($K = 40$)	FALCON	CLIP-Dissect-20k	0.4737	-0.0400	6.1452	0.1536
	Refraction (Ours)		**0.8096**	**6.2896**	**25.5589**	**0.6390**
	FALCON	LAION-NP-20k	0.4850	0.0026	6.6533	0.1663
	Refraction (Ours)		**0.8604**	**8.6144**	**30.1302**	**0.7533**
Layer 4 ($K = 37$)	FALCON	CLIP-Dissect-20k	**0.8001**	3.5185	22.7310	0.6144
	Refraction (Ours)		0.7796	**9.7393**	**23.3197**	**0.6303**
	FALCON	LAION-NP-20k	**0.8735**	4.7633	**27.5213**	**0.7438**
	Refraction (Ours)		0.8245	**13.8077**	26.1771	0.7075

Table 7: Per-block Results of MN-CoSy and EISC on ViT-B/16 using two independent concept vocabularies. Results are compared at matched Top-K. Bold indicates the higher score within each block and vocabulary.

Block	Method	Vocabulary	MN-CoSy AUC	MN-CoSy MAD	EISC	EISC-Rate
4 ($K = 40$)	FALCON	CLIP-Dissect-20k	0.6910	0.9658	16.3729	0.4093
	Refraction (Ours)		**0.8030**	**1.6747**	**24.8989**	**0.6225**
	FALCON	LAION-NP-20k	0.7357	1.3231	18.2919	0.4573
	Refraction (Ours)		**0.8408**	**1.7806**	**27.6991**	**0.6925**
8 ($K = 40$)	FALCON	CLIP-Dissect-20k	0.5065	0.1527	6.1594	0.1540
	Refraction (Ours)		**0.6430**	**4.7529**	**14.9122**	**0.3728**
	FALCON	LAION-NP-20k	0.5558	0.3495	8.1065	0.2027
	Refraction (Ours)		**0.7721**	**8.8500**	**23.7356**	**0.5934**
12 ($K = 6$)	FALCON	CLIP-Dissect-20k	0.5567	0.1394	1.1184	0.1864
	Refraction (Ours)		**0.9270**	**10.0961**	**5.1237**	**0.8540**
	FALCON	LAION-NP-20k	0.5570	0.2627	1.6098	0.2683
	Refraction (Ours)		**0.9632**	**9.8774**	**5.5586**	**0.9264**

A.3.5 Additional Ablation Results

Figures 4–7 show the four layer–vocabulary settings from the parameter ablation in Fig. 4. All settings use the same 17 fixed configurations and Top-K evaluation protocol.

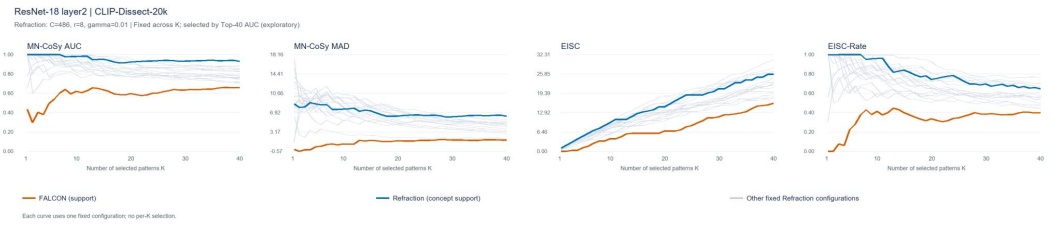


Figure 4: Parameter sensitivity on ResNet-18 layers 2 with CLIP-dissect-20k vocabulary. Blue shows the fixed Refraction configuration selected by Top-40 MN-CoSy AUC, orange shows FALCON, and gray shows the other tested Refraction configurations.

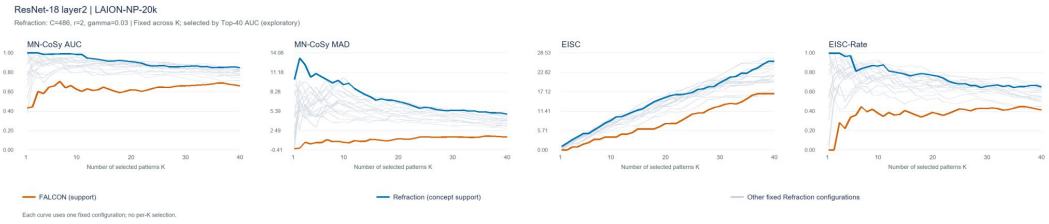


Figure 5: Parameter sensitivity on ResNet-18 layer 2 with LAION-NP-20k. The panels report MN-CoSy AUC, MN-CoSy MAD, EISC, and EISC-Rate. Colors and configuration selection follow Fig. 4.

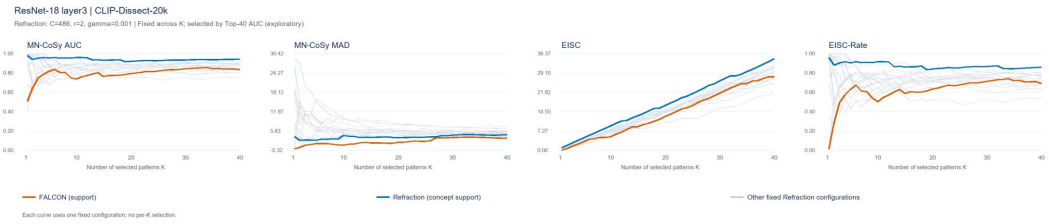


Figure 6: Parameter sensitivity on ResNet-18 layer 3 with CLIP-Dissect-20k. The panels and plotting conventions follow Fig. 4.

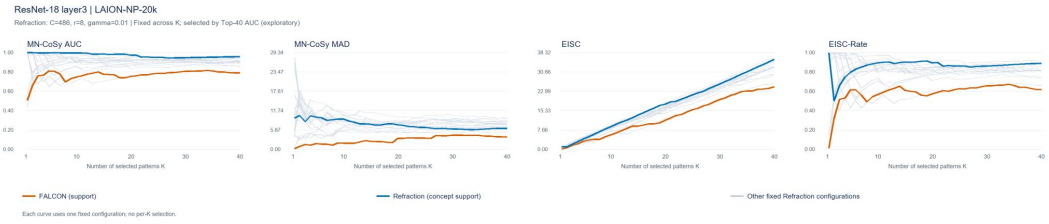


Figure 7: Parameter sensitivity on ResNet-18 layer 3 with LAION-NP-20k. The panels and plotting conventions follow Fig. 4.

Table 8: SN-CoSy results on 50 ResNet-18 layer-4 neurons. All methods use the same normalized neuron response z_i^2; FALCON uses its single-neuron adaptation from Kopf et al. (2024).

Method	Vocabulary	SN-CoSy AUC	SN-CoSy MAD
CLIP-Dissect	CLIP-Dissect-20k	0.8142	3.7952
FALCON		0.8073	3.7950
CLIP-Dissect	LAION-NP-20k	0.8622	4.9384
FALCON		0.8760	5.1361

A.3.6 Additional Quantitative Results

Comparison of Discovered Subspaces. Figures 8–10 compare subspace discovery at matched Top-K output budgets. Each bar represents a non-redundant response group with positive EISC contribution, and its length shows that contribution. Refraction yields more such groups than FALCON in all six ResNet-18 settings and all six ViT-B/16 settings. On ResNet-50, Refraction leads in both layer-3 settings, whereas FALCON yields slightly more groups in layers 2 and 4. Overall, Refraction leads in group count in 14 of the 18 model–layer–vocabulary settings and achieves higher EISC in 16. Thus, the results indicate broader discovery in most settings, while the EISC gains also reflect the quality of the discovered groups.

Additional Evidence of Mesosemanticity. Figure 11– 15 illustrates more examples for mesosemanticity.

Independently discovered interpretable subspaces by Refraction. Figures 16–19 show additional patterns discovered by Refraction but not covered by FALCON's reported results. For each pattern, we show six high-response images, their response ranks, the assigned textual concept, and its MN-CoSy AUC.

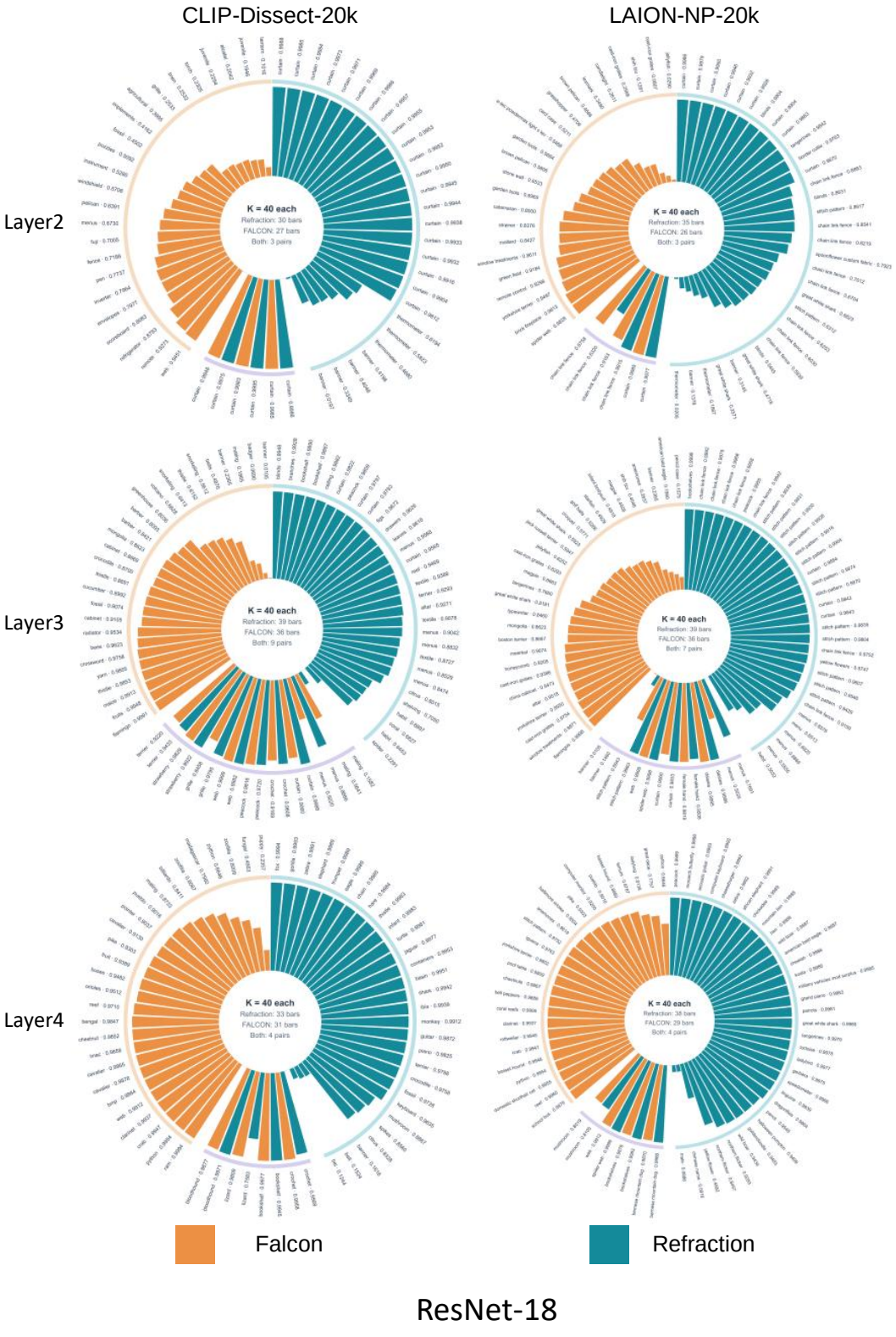


Figure 8: Comparison of discovered subspaces on ResNet-18. Rows show layers 2–4 and columns show the two independent vocabularies. Each bar represents a deduplicated group with positive EISC contribution; its length and adjacent value show that contribution. Center labels report the number of bars for each method and the number of shared pairs. Groups are arranged clockwise as Refraction-only, shared, and FALCON-only. A pair is shared when its explanations are semantically equivalent and its Top-200 response sets have at least 10 images in common.

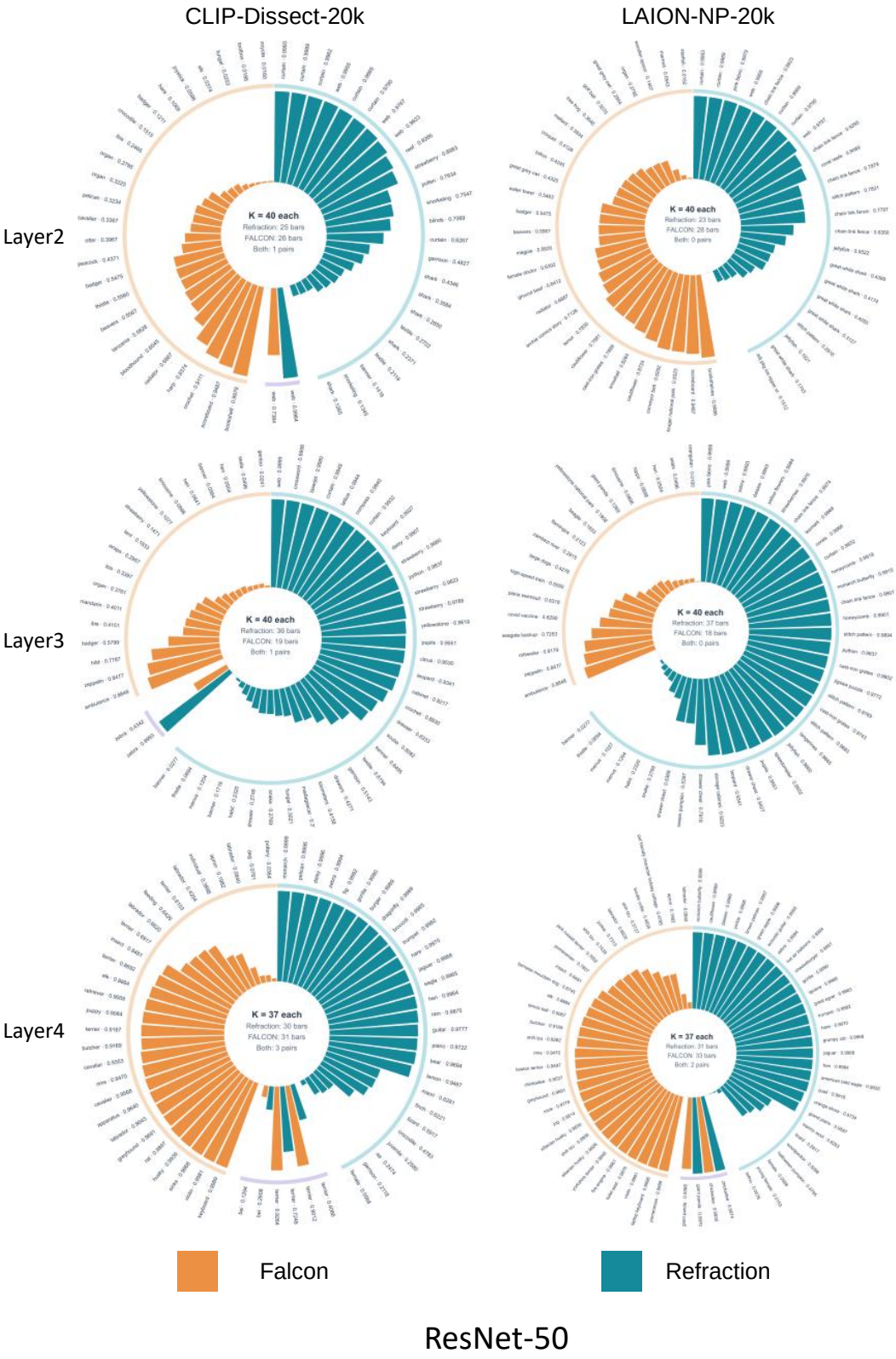


Figure 9: Comparison of discovered subspaces on ResNet-50. Rows show layers 2–4 and columns show the two independent vocabularies. Bar length gives each deduplicated group's positive EISC contribution; center labels give the method-wise bar counts and the number of shared pairs. Shared pairs follow the name-and-response-overlap criterion in Fig. 8.

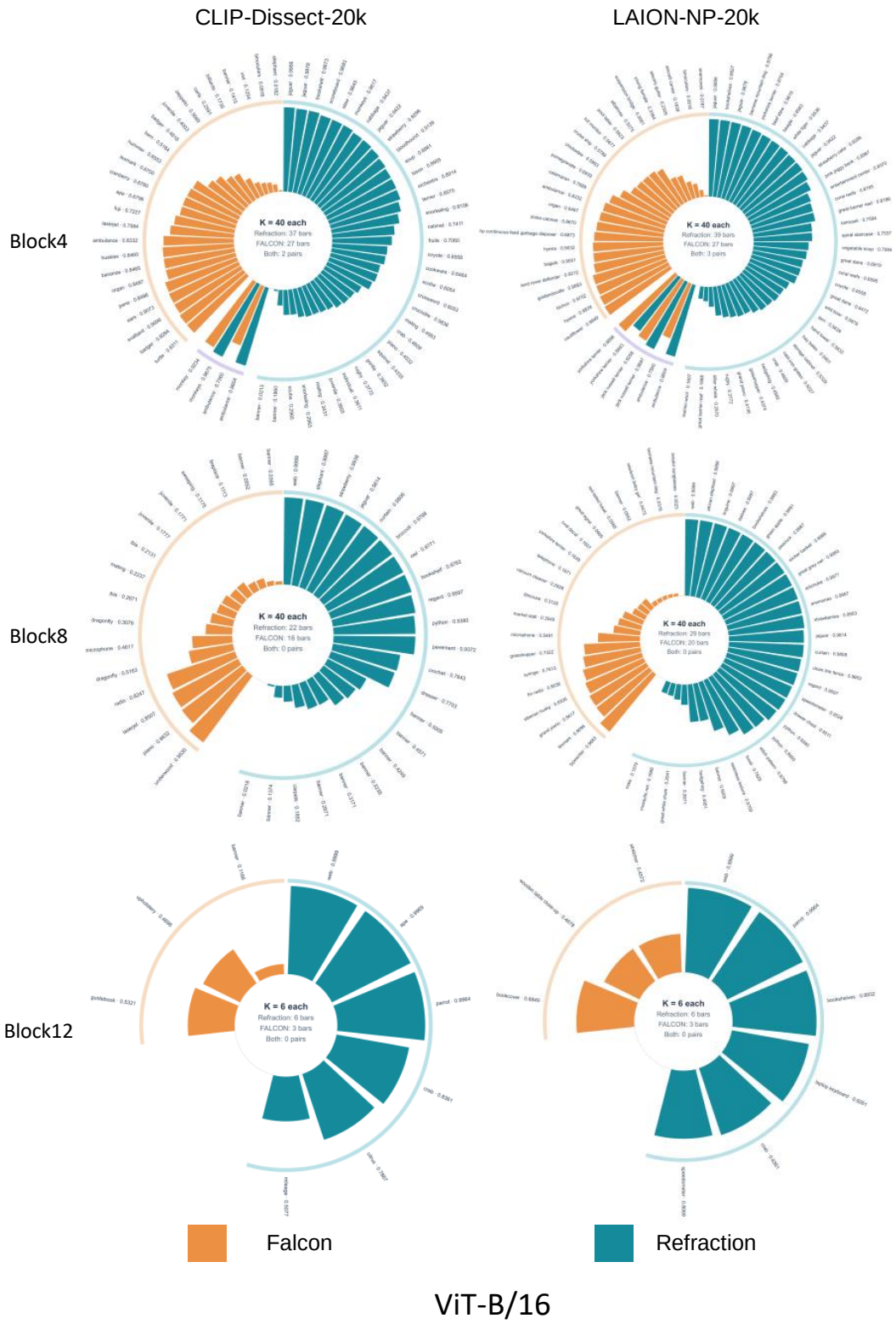


Figure 10: Comparison of discovered subspaces on ViT-B/16. Rows show blocks 4, 8, and 12; columns show the two independent vocabularies. Bar length gives each deduplicated group's positive EISC contribution; center labels give the method-wise bar counts and the number of shared pairs. Shared pairs follow the criterion in Fig. 8.

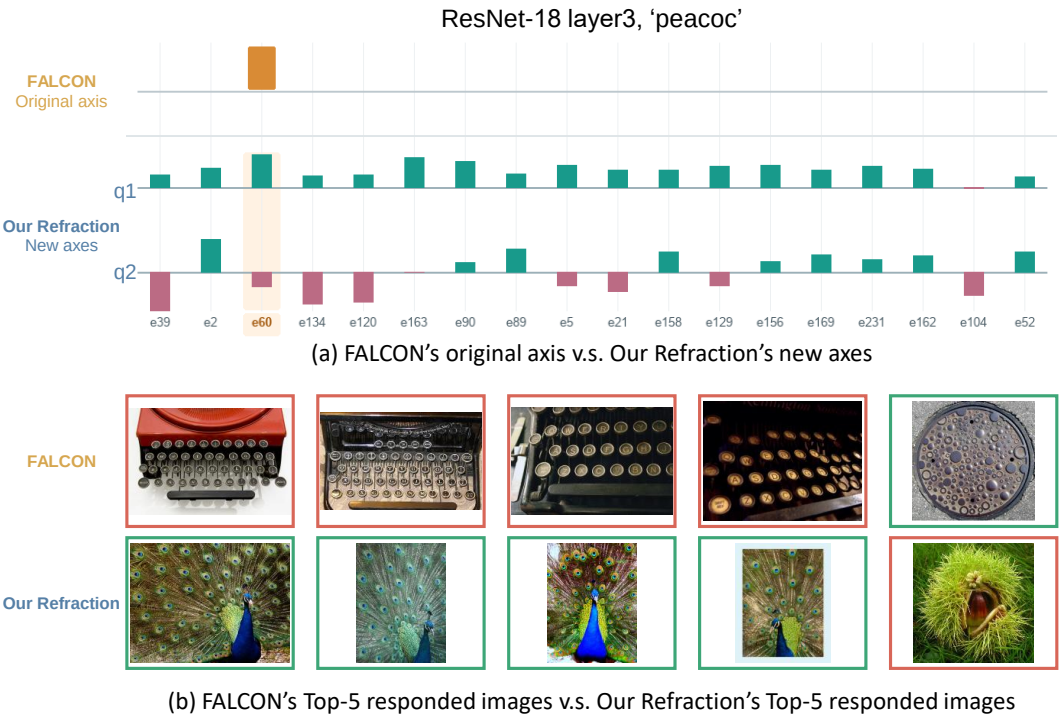


Figure 11: Comparison original axis of FALCON and new axes of our Refraction and the top-5 responded images on ResNet-18 layer 3, for concept ’peacoc’.

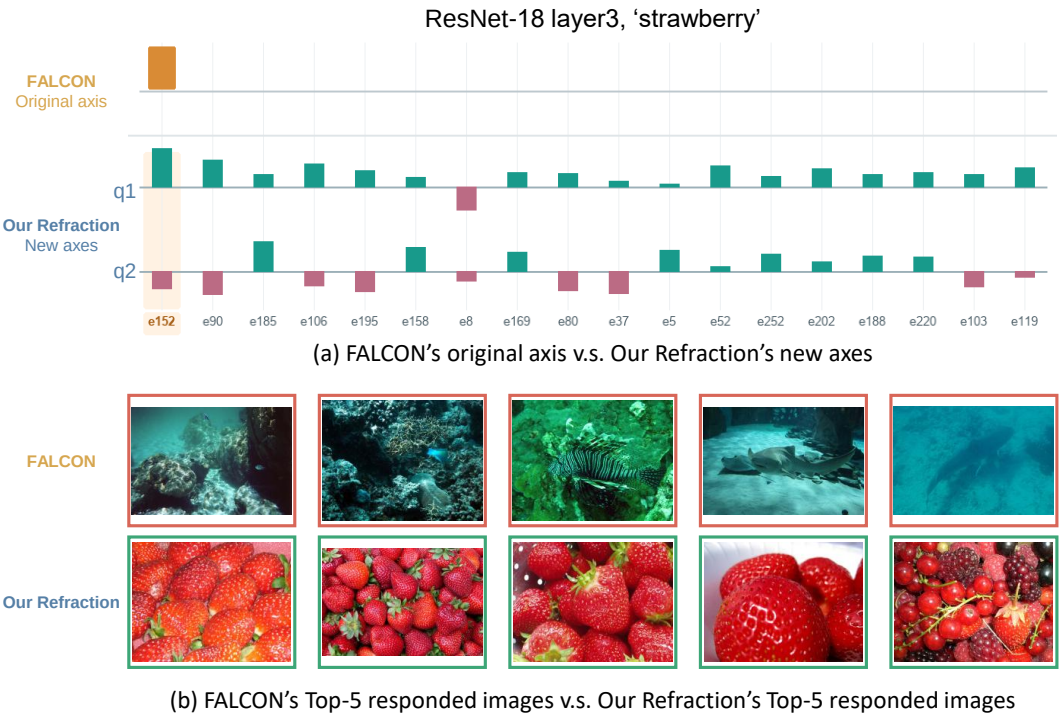


Figure 12: Comparison original axis of FALCON and new axes of our Refraction and the top-5 responded images on ResNet-18 layer 3, for concept ’strawberry’.

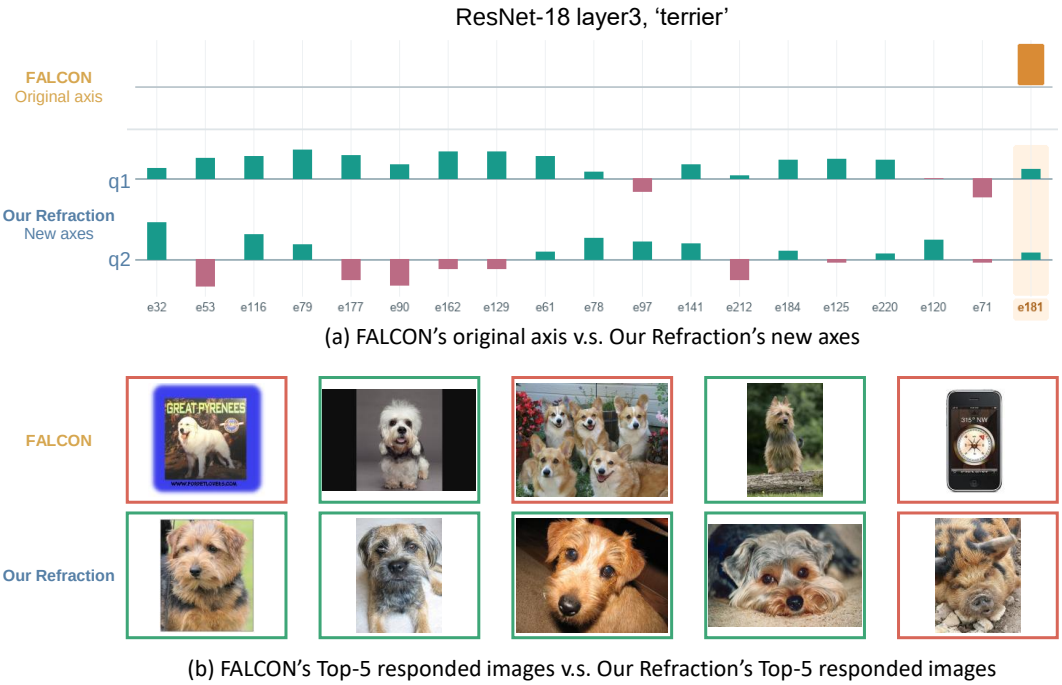


Figure 13: Comparison original axis of FALCON and new axes of our Refraction and the top-5 responded images on ResNet-18 layer 3, for concept 'terrier'.



Figure 14: Comparison original axis of FALCON and new axes of our Refraction and the top-5 responded images on ResNet-50 layer 3, for concept 'zebra'.

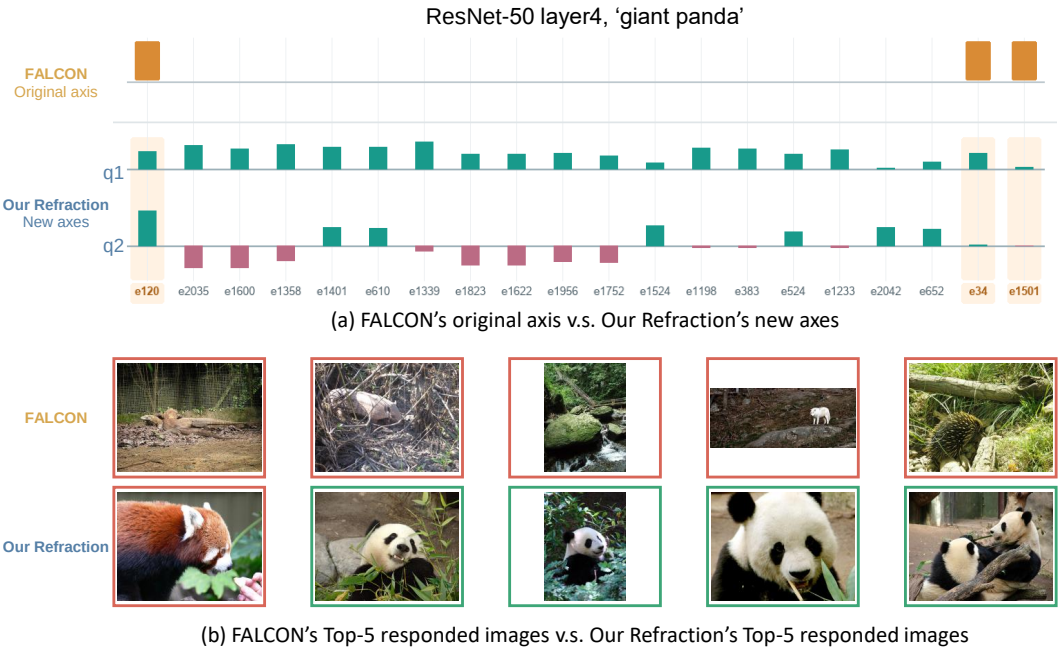


Figure 15: Comparison original axis of FALCON and new axes of our Refraction and the top-5 responded images on ResNet-18 layer 3, for concept 'giant panda'.

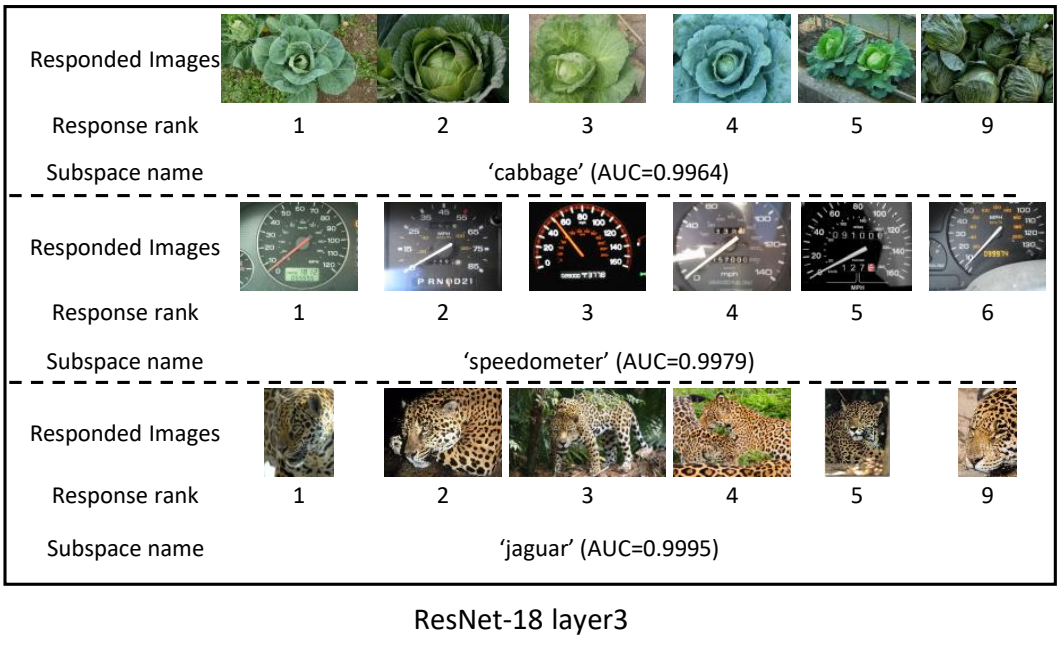
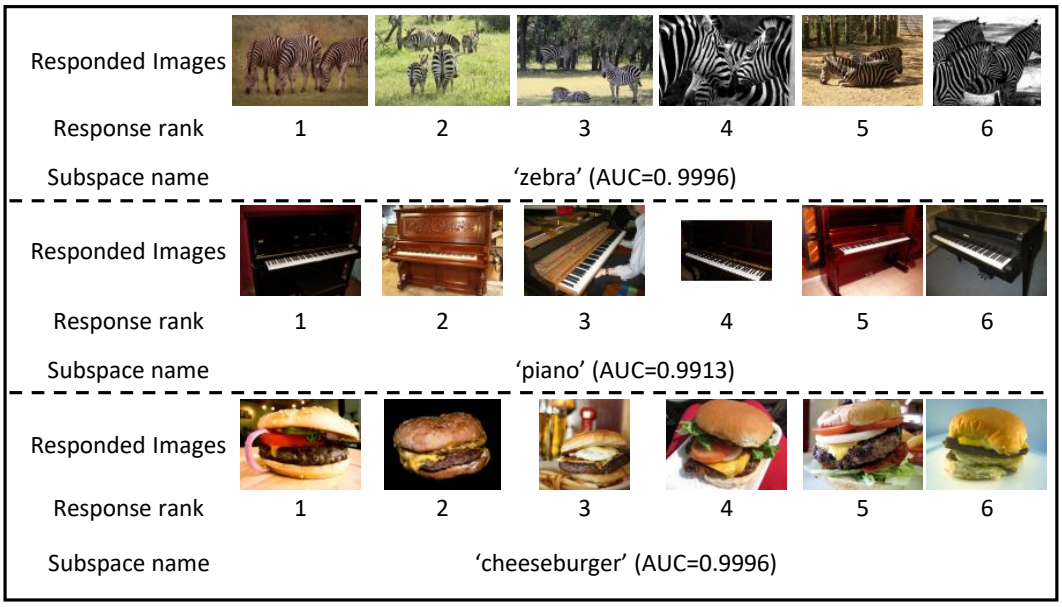
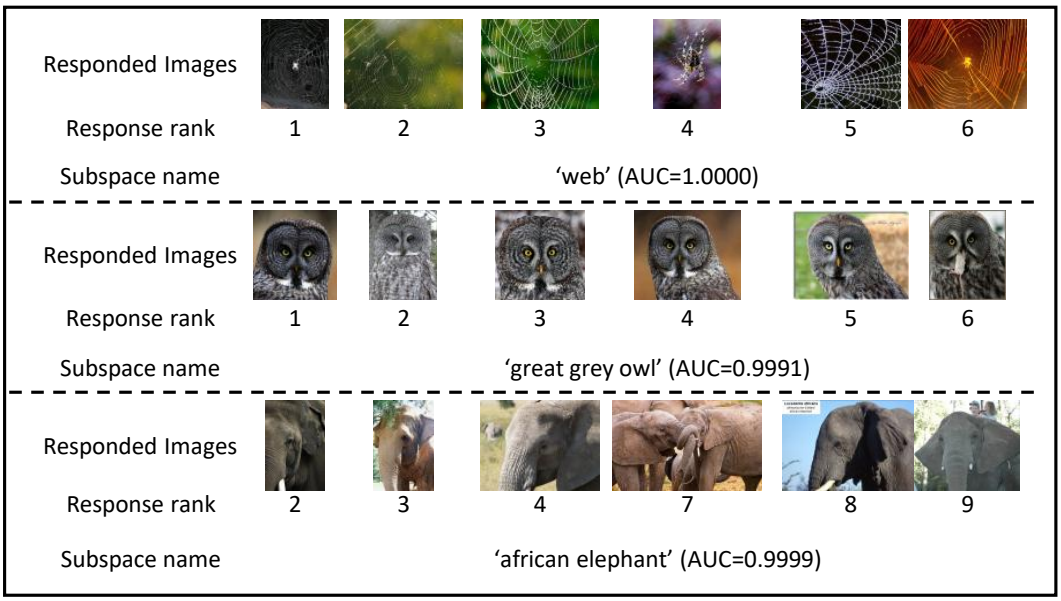


Figure 16: Interpretable subspaces discovered by Refraction but not covered by FALCON's reported results on ResNet-18 layer 3. Each row shows six high-response images, their response ranks, and the assigned textual concept with its MN-CoSy AUC.



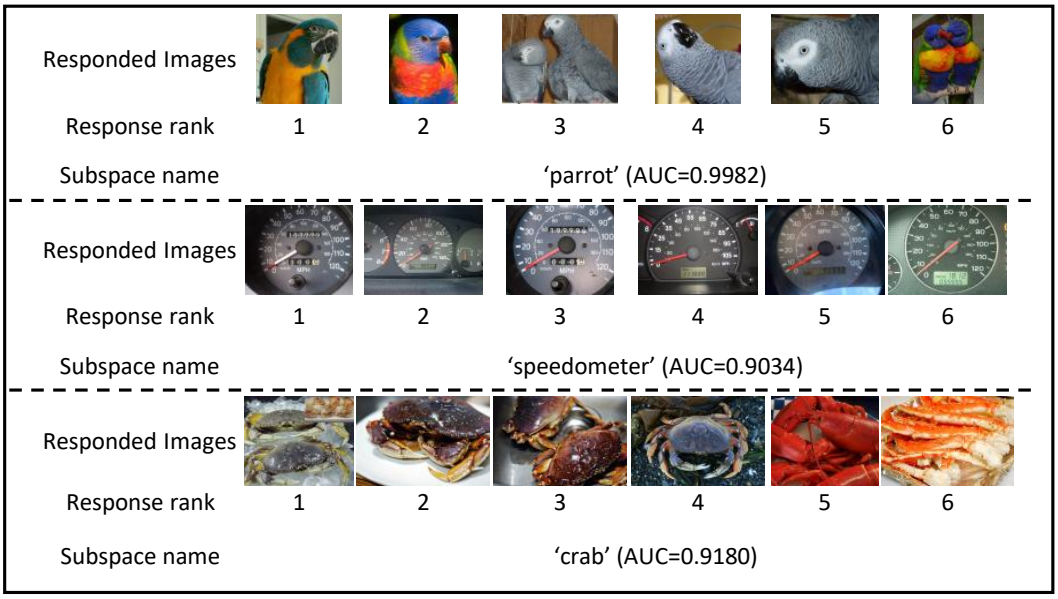
ResNet-18 layer4

Figure 17: Interpretable subspaces discovered by Refraction but not covered by FALCON's reported results on ResNet-18 layer 4. Each row shows six high-response images, their response ranks, and the assigned textual concept with its MN-CoSy AUC.



ViT-B/16 block8

Figure 18: Interpretable subspaces discovered by Refraction but not covered by FALCON's reported results on ViT-B/16 block 8. Each row shows six high-response images, their response ranks, and the assigned textual concept with its MN-CoSy AUC.



ViT-B/16 block12

Figure 19: Interpretable subspaces discovered by Refraction but not covered by FALCON's reported results on ViT-B/16 block 12. Each row shows six high-response images, their response ranks, and the assigned textual concept with its MN-CoSy AUC.