SPACE: Shapley Projection Attribution for Concept Explanations

Anonymous authors
Paper under double-blind review

Abstract

Concept-based explanations connect a model's prediction to ideas people can name. Scored one at a time, they overcount: two concepts that point half the same way are each credited with the sensitivity they share, and the scores sum to no meaningful total. *SPACE* measures how much of the model's local gradient each group of concepts can express, and from that one game answers three questions. (1) How much of the model does the vocabulary reach? A coverage score, in closed form. (2) How does the captured sensitivity divide among the concepts? A Shapley allocation that is non-negative, sums exactly to the captured total, and credits shared sensitivity once. (3) Which few concepts should a reader keep? A subset selection on the same game, with a greedy guarantee, which the allocation does not solve. The game costs one gradient, one factorisation, and small per-coalition algebra, with no masked inputs, reference examples, or per-subset fits, at a per-coalition cost independent of the representation dimension. On ImageNet, a ResNet50 explained through 200,704 activation coordinates with fifteen texture and colour concepts, the shares add up exactly and the coverage is stated first. Keeping the three concepts ranked highest, the allocation's three explain 76% of what the vocabulary can, against 67% for one-at-a-time scores. The gap holds across five refits of every concept and is confirmed by moving the network itself. Selection on the same game comes within one point of the best possible three. On a rendered benchmark with planted factors, the allocation recovers the planted factor as often as any rule. When a factor has several near-synonymous descriptions, selection finds both planted factors on every image, while every per-concept ranking, the allocation included, fills the shortlist with synonyms. SPACE turns any set of concept directions, however obtained, into one accounted explanation.

1 Introduction

A concept-based explanation asks how much of a model's local sensitivity, the output change predicted for a very small representation change, can be expressed in terms people understand. Its standard instrument is the concept activation vector (CAV) (Kim et al., 2018): a unit direction $\mathbf{d}_k$ whose raw local score is the directional derivative $\langle \mathbf{g}, \mathbf{d}_k \rangle$. TCAV asks how often this derivative is positive across a class. ICS (Schrouff et al., 2021) explains one input by starting from a chosen reference and accumulating sensitivity to the same concept along a path to that input.

Concept directions estimated from data do not have to be perpendicular. Two may partly point the same way, and Nicolson et al. (2025) show that such overlap changes what a CAV explanation says. Each magnitude $|a_k|$ is a valid line score, but $\sum_k |a_k|$ decomposes nothing: overlapping directions credit shared components repeatedly, and even perpendicular ones are summed in ℓ_1 length while their joint effect has Euclidean length. The excess therefore mixes two effects, and our ImageNet study separates them rather than calling the sum a measure of overlap (Section 5).

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.

SHAP (Lundberg & Lee, 2017) uses Shapley averaging (Shapley, 1953) on a different total. It asks how far a feature pushed the prediction from what the model would have said with that feature hidden, substituting a replacement from a background distribution and re-running the model. SPACE asks how much of the sensitivity already present at this input each concept accounts for, so nothing is hidden, replaced, or re-run. ConceptSHAP allocates preserved predictive performance and CLIP-SHAP explains a concept-score bottleneck (Yeh et al., 2020; Kazmierczak et al., 2024). After one gradient, SPACE needs no masked inputs, replacement examples, subset-specific fits, or further model calls.

The value function. We keep the full gradient $\mathbf{g} = \nabla f(\mathbf{z}) \in \mathbb{R}^D$ rather than reducing it to one scalar per concept, and define the value of a coalition S as the norm of the gradient's orthogonal projection onto the span of its directions:

$$v(S) = \|\mathbf{P}_S\, \mathbf{g}\|, \qquad v(\emptyset) = 0. \tag{1}$$

Proposition 1 shows that $v(S)$ is the largest locally predicted change under any unit-length embedding perturbation confined to span(S). It needs no reference input because the gradient is evaluated at the instance itself.

Applying the Shapley value to (1) yields attributions ϕ_k that are nonnegative (Lemma 1), sum exactly to $\|\mathbf{P}_{[K]}\mathbf{g}\|$ by efficiency, and discount redundancy coalition by coalition: a direction that adds nothing to a coalition's span has zero marginal value there. A closed-form completeness score reports how much of the local gradient the supplied span covers. Given the directions, the attribution step uses no labels.

Cost. The expensive model computation is fixed at one backward pass. Because $v(S)^2 = (\mathbf{D}_S^\top \mathbf{g})^\top (\mathbf{D}_S^\top \mathbf{D}_S)^+ (\mathbf{D}_S^\top \mathbf{g})$, every coalition value is a function of the $K \times K$ Gram matrix and the K-vector $\mathbf{D}^\top \mathbf{g}$, so the attribution costs one backward pass plus $O(DK^2 + 2^K K^3)$ arithmetic with per-coalition cost *independent of* D; at $D = 200{,}704$ no $D \times D$ object is formed.

Contributions. One coalition game, computed from one gradient, serves three distinct operations, and the paper establishes which question each is licensed to answer. *Coverage*: the fraction of the local gradient the vocabulary reaches, in closed form, reported before any share is read. *Selection*: the m concepts whose span captures the most sensitivity, a subset problem on the same game with a greedy guarantee inherited from least-squares subset selection (Proposition 6); it is the operation that finds distinct factors when a vocabulary holds several descriptions of one, and it selects identically under v and v^2. *Allocation*: the Shapley value of the game or of its square, a family of two efficient nonnegative allocations that credit shared sensitivity once, differ in one stated property, and are reported side by side; the allocation is the accounting layer and not a shortlist. The primary study is ImageNet at $D = 200{,}704$, where all three operations are measured against the model itself and against five refitted vocabularies, with a rendered benchmark supplying planted ground truth. Concretely we contribute (i) the game and its operational reading, (ii) the three operations with their separate guarantees, (iii) computation from one backward pass and one factorisation, adapting the least-squares reduction of Bell et al. (2024) so that per-coalition cost is independent of the representation dimension, and (iv) evidence on where the operations agree and where they part: allocation changes what a reader keeps on ImageNet, only selection finds distinct factors under redundancy, and no allocation predicts a finite factor swap. Because the rule consumes only unit vectors it adds no supervision of its own, and the directions may come from any estimator, including a text alignment.

2 Related Work

Concept directions and per-concept scores. Kim et al. (2018) introduced the CAV and defined relevance as the fraction of class inputs with positive directional sensitivity, which is global and uses only the sign. ICS (Schrouff et al., 2021) combines the same CAV with Integrated Gradients (Sundararajan et al., 2017), accumulating the concept's effect along a path from a chosen reference, so its answer depends on that reference. SPACE gives

joint credit at the representation itself, with no baseline to choose. De Santis et al. (2026) extends concept attribution to saliency maps and Goldberg et al. (2025) locate reliable signal in the activation tail, both yielding directions SPACE can consume. Crabbé & van der Schaar (2022) instead replaces the direction with a kernel-defined region, which SPACE would need reduced to a local gradient first. Closest in motivation, Nicolson et al. (2025) show overlapping CAVs can mislead and Erogullari et al. (2025) refit directions toward perpendicularity. SPACE leaves the directions alone and changes their accounting.

Text-derived concepts. Label-free concept models define candidate concepts from language embeddings (Oikarinen et al., 2023), and cross-model alignment makes a backbone's representations comparable to text (Moayeri et al., 2023). Either route yields unit directions SPACE can consume; we evaluate the two standard example-based routes here and leave text-derived vocabularies to future work.

Shapley values over concepts. Concept-level Shapley methods differ in their players, value functions, and inputs. Jutte et al. (2025) allocate outcomes to temporal patterns, while Li et al. (2021) use pixel-masked image segments, requiring raw-input access and one model forward pass per coalition. Standard Shapley sampling and the implementations surveyed by Muschalik et al. (2024) can estimate the same set function when exact enumeration is too expensive. ConceptSHAP and CLIP-SHAP are the closest comparisons and are described at the end of this section.

Explained-variance games. For fixed $\mathbf{g}$, $v^2(S) = \mathbf{g}^\top \mathbf{P}_S \mathbf{g}$ is how much of the squared gradient length a least-squares fit on the directions in S reproduces, so its share of $\|\mathbf{g}\|^2$ is an *uncentred* R^2, measured against zero because the fit has no intercept. Splitting a regression's fit among its predictors by averaging over entry orders is relative-importance and dominance analysis (Lindeman et al., 1980; Budescu, 1993; Lipovetsky & Conklin, 2001; Israeli, 2007) and applied to variance in global sensitivity analysis (Owen, 2014). Bell et al. (2024) compute that allocation after one factorisation of the design matrix, at a per-coalition cost independent of the sample count; with the directions as design and the gradient as response, Section 4.5 adapts the same reduction. SPACE attributes the square root v instead, for the reason Proposition 1 gives.

Interaction indices. Grabisch & Roubens (1999) axiomatise the Shapley interaction index used in Section 4.6; Faith-Shap (Tsai et al., 2023) fits interaction terms instead and agrees with it for two or three concepts.

Impossibility results. Bilodeau et al. (2024) prove that complete-and-linear attribution can fail to predict behaviour on inputs it was not given, and You et al. (2025) lower-bound per-feature attribution in self-attributing networks; Section 7 separates those targets from SPACE's.

Closest comparisons and computation. ConceptSHAP (Yeh et al., 2020) asks how well each concept subset can reproduce the model's predictions over a dataset, relative to random accuracy. It fits a new mapping for every subset (Appendix C). SPACE asks a different, instance-level question and obtains its game from one gradient.

Two lines of work share SPACE's geometry. Concept Gradients (Bai et al., 2023) fits the same least-squares problem but returns the coefficient vector $\mathbf{D}^\dagger \mathbf{g}$, which says how to move, where SPACE keeps the *length* of the fitted part, the quantity that can be divided among concepts. Multi-dimensional concept discovery (Vielhaben et al., 2023) treats concepts as subspaces and, for a linear head, measures completeness as the fraction of the weight vector inside them, which is Equation (4) exactly. Our coverage carries that accepted quantity to a nonlinear head at one input, and the allocation on it is new.

CLIP-SHAP (Kazmierczak et al., 2024) explains a predictor built on CLIP concept scores with training examples filling in omitted concepts, and Explain Any Concept (Sun et al., 2023) scores image segments with masked model calls; SPACE applies to any differentiable head and needs neither examples nor further model calls.

3 Setup

Model. Let $\mathbf{z} \in \mathbb{R}^D$ be a representation from a frozen backbone and $f : \mathbb{R}^D \to \mathbb{R}$ a differentiable scalar head, such as a regression output or a class logit. SPACE needs nothing else from f. Write $\mathbf{g} = \nabla f(\mathbf{z})$, obtained by one backward pass.

SPACE does not assume that embeddings are normalised; a tangent-constrained variant for representations on the unit sphere replaces $\mathbf{g}$ by $(\mathbf{I} - \mathbf{z}\mathbf{z}^\top)\mathbf{g}$ and tangent-projects the directions, and is not needed for the unnormalised `layer3` representation studied here.

Concept directions. A concept k is a unit vector $\mathbf{d}_k \in \mathbb{R}^D$. We use two standard routes, neither a contribution of this paper:

Mean difference. $\mathbf{d}_k^A \propto \bar{\mathbf{z}}_k^+ - \bar{\mathbf{z}}_k^-$, the normalised difference of concept-positive and concept-negative mean embeddings, which equals Fisher's discriminant whenever the two groups share an isotropic covariance, and up to scale more generally whenever the mean difference is an eigenvector of the shared covariance. **Linear probe (CAV).** The normalised weight vector of a class-balanced linear classifier separating the same two sets (Kim et al., 2018). Both of these require labelled examples.

SPACE consumes unit vectors and is indifferent to their provenance: the estimation route affects direction quality, not the attribution rule, and a direction read off a text alignment enters the game the same way.

Independent directional scoring. The raw per-concept directional-derivative score is $a_k(\mathbf{z}) = \langle \mathbf{g}, \mathbf{d}_k \rangle$. It is the $|S| = 1$ special case of (1) up to sign, but $\sum_k |a_k|$ is not an additive decomposition of $v([K])$. We report $r_{\text{ind}} = \sum_k |a_k| / v([K])$ as a descriptive non-additivity ratio. It is not a measure of entanglement, because adding absolute scores differs from measuring a length even for perpendicular directions.

4 SPACE: Shapley Projection Attribution for Concept Explanations

4.1 The coalition value

Index the concepts by $[K] = \{1, \ldots, K\}$. For $S \subseteq [K]$ let $\mathbf{D}_S \in \mathbb{R}^{D \times |S|}$ hold the corresponding directions as columns. The orthogonal projector onto their span is

$$\mathbf{P}_S = \mathbf{D}_S \big(\mathbf{D}_S^\top \mathbf{D}_S\big)^+ \mathbf{D}_S^\top, \tag{2}$$

with $(\cdot)^+$ the Moore–Penrose pseudoinverse. This standard inverse-like operation remains defined when directions are duplicated or linearly dependent and makes the projector depend only on their true span. The coalition value is $v(S) = \|\mathbf{P}_S \mathbf{g}\|$ and $v(\emptyset) = 0$ as in (1).

We use this quantity because it answers a concrete question about the model:

Proposition 1 (Operational meaning)**.** *For any $S \subseteq [K]$,*

$$v(S) \;=\; \max_{\boldsymbol{\delta} \in \mathrm{span}(S),\ \|\boldsymbol{\delta}\| \le 1} \langle \mathbf{g}, \boldsymbol{\delta} \rangle,$$

the largest first-order change in the prediction achievable by a unit-norm perturbation of the embedding confined to the subspace spanned by the concepts in S. Here first order means the linear change predicted by the gradient for an infinitesimally small perturbation.

So $v(S)$ measures how much the concepts in S can jointly influence the prediction near this instance according to the gradient. Two concepts that point in the same direction have essentially the influence of one, which lets the Shapley step account for their redundancy without an added penalty.

Proposition 1 is exact about the gradient. Whether the first-order quantity tracks a finite perturbation is separate and empirical. Appendix E tests this on synthetic games and on the

ResNet50 evaluation below. These are local representation interventions, not input-space or causal fidelity.

The squared variant $v^2(S) = \|\mathbf{P}_S \mathbf{g}\|^2$ adds exactly across perpendicular directions where v is subadditive, so it serves the interaction index of Section 4.6; its singleton value is $\langle \mathbf{g}, \mathbf{d}_k \rangle^2$, the raw baseline squared.

4.2 The attribution

The SPACE attribution of concept k is the Shapley value of v:

$$\phi_k = \sum_{S \subseteq [K] \setminus \{k\}} \frac{|S|!\,(K - |S| - 1)!}{K!} \big[v(S \cup \{k\}) - v(S) \big]. \tag{3}$$

Proposition 2 (Axiomatic properties)**.** *Equation* (3) *satisfies the Shapley axioms with respect to v. In particular:* (i) Efficiency: $\sum_{k \in [K]} \phi_k = v([K]) = \|\mathbf{P}_{[K]} \mathbf{g}\|$. (ii) Symmetry: *concepts with identical marginal contributions to every coalition receive equal credit. In particular $\mathbf{d}_j = \mathbf{d}_k$ implies $\phi_j = \phi_k$.* (iii) Null player: *if $\mathbf{d}_k \perp \mathbf{g}$ and $\mathbf{d}_k \perp \mathbf{d}_j$ for every $j \neq k$, then $\phi_k = 0$.*

Neither hypothesis in (iii) can be dropped: a direction orthogonal to $\mathbf{g}$ but not to the others can still earn credit by completing a coalition's span (Appendix A), which Proposition 1 makes correct. Linearity adds no guarantee here because $\mathbf{g} \mapsto v_{\mathbf{g}}$ is not additive.

Lemma 1 (Monotonicity and nonnegativity)**.** *$S \subseteq T$ implies* $\operatorname{span}(S) \subseteq \operatorname{span}(T)$ *and hence $v(S) \leq v(T)$. Every marginal contribution in* (3) *is therefore nonnegative, and $\phi_k \geq 0$ for all k.*

One condition makes every $\mathbf{P}_S$ exact: $K \leq D$ and $\sigma_K(\mathbf{D}^\top) > \tau$ for the shared floor τ, since interlacing then lifts every coalition's smallest singular value above τ. It is sufficient, not necessary, and failure is inconclusive: a vocabulary whose discarded modes are exactly null, duplicates among them, is still allocated correctly, while one losing a small nonzero mode can lose monotonicity with it. Across 600 random vocabularies, about half near-duplicate by construction, every certified case is monotone (Appendix B).

Efficiency makes SPACE a decomposition and Lemma 1 makes it a nonnegative one, together giving the output the property independent directional scores lack: the numbers are shares of a common total.

One game, two allocations. The squared game v^2 admits the same treatment, and its Shapley value, rescaled by $v([K])/v^2([K])$, is also nonnegative, efficient for $v([K])$ and in output units. The two differ in one property. On mutually orthogonal directions the Shapley value of v^2 is $\langle \mathbf{g}, \mathbf{d}_k \rangle^2$ (Proposition 4), so adding an unrelated orthogonal concept leaves every existing raw credit unchanged, though the display rescaled to $v([K])$ moves with the total; the Shapley value of v averages achievable increments $v(S \cup \{k\}) - v(S)$ and redistributes even then, since with $\mathbf{g} = (3, 4)$ on two orthogonal directions it gives $(2, 3)$ where the rescaled squared allocation gives $(1.8, 3.2)$. Coverage and selection are identical under the two; the shares differ by a mean total-variation distance of 0.10 on ImageNet, and the paper reports both wherever they differ (Appendix D). We write ϕ_k for the allocation of v and call the pair the projection-game family.

The unexplained remainder. The projection splits the gradient into two perpendicular pieces. The first lies inside the concept span and the second, $(\mathbf{I} - \mathbf{P}_{[K]})\mathbf{g}$, lies outside it. Their squared lengths add by Pythagoras. The *completeness* score is

$$c(\mathbf{z}) = \|\mathbf{P}_{[K]} \mathbf{g}\|^2 / \|\mathbf{g}\|^2 \in [0, 1], \tag{4}$$

the fraction of squared gradient length inside the concept span. It is defined for $\mathbf{g} \neq \mathbf{0}$ and reported as zero otherwise, while shares and the ratio r_{ind} require $v([K]) > 0$ and are undefined without it. The unexplained fraction is $1 - c(\mathbf{z})$ and the unexplained length is $\|\mathbf{g}\| \sqrt{1 - c(\mathbf{z})}$. The two lengths add in square, not directly. This is a local analogue of ConceptSHAP's global completeness, available in closed form at every instance.

4.3 Selection is not allocation

The shares say how the captured sensitivity divides, not which m concepts to keep. That question is $\max_{|S|=m} v(S)$ on the same game, solved exactly from the enumerated values or by greedy forward selection, identically under v and v^2. The two operations can disagree: a concept whose span is largely shared earns a fair share of what it carries yet adds little to a shortlist that already holds one of its neighbours (Section 7).

4.4 Sign and units

v is monotone (Lemma 1), so $\phi_k \geq 0$ and the direction in which a concept pushes the prediction is not in ϕ_k. We restore it by

$$a_k = \langle \mathbf{g}, \mathbf{d}_k \rangle, \qquad \sigma_k = \begin{cases} +1, & a_k \geq 0, \\ -1, & a_k < 0, \end{cases} \qquad \tilde{\phi}_k = \sigma_k \phi_k. \tag{5}$$

This is a reporting step, not a new value function, and it does not preserve efficiency: $\sum_k \tilde{\phi}_k$ is in general not interpretable. A signed bar carries the magnitude ϕ_k and the sign of a_k and nothing more, and is read next to the unsigned total $v([K])$ and the completeness $c(\mathbf{z})$; when $a_k = 0$ the sign is a convention.

4.5 Computation

Equation (3) enumerates 2^K coalitions, but each costs far less than the projector formula (2) suggests, because $\mathbf{P}_S$ never needs to be formed. A projector is idempotent (projecting twice is projecting once) and self-adjoint ($\mathbf{P}_S^\top = \mathbf{P}_S$), so

$$v(S)^2 = \mathbf{g}^\top \mathbf{P}_S \mathbf{g} = \left(\mathbf{D}_S^\top \mathbf{g}\right)^\top \left(\mathbf{D}_S^\top \mathbf{D}_S\right)^+ \left(\mathbf{D}_S^\top \mathbf{g}\right), \tag{6}$$

where $\mathbf{a} = \mathbf{D}^\top \mathbf{g}$ collects the directional derivatives and $\mathbf{G} = \mathbf{D}^\top \mathbf{D}$ the pairwise dot products. So a coalition value is fixed by $\mathbf{a}_S$ and $\mathbf{G}_{S,S}$, with no object of size D in it.

We do not form $\mathbf{G}$, whose eigenvalues are squared singular values and so can make a well-resolved vocabulary look rank-deficient. Factoring $\mathbf{D}$ once and working in an orthonormal basis of its span gives the same values with every rank read off singular values that were never squared (Appendix B). A positive floor ε defines a thresholded game v_ε, so a run returns $\phi^\varepsilon = \text{Shapley}(v_\varepsilon)$ and $c_\varepsilon = v_\varepsilon([K])^2 / \|\mathbf{g}\|^2$. The certificate guarantees $\phi^\varepsilon = \phi$ and $c_\varepsilon = c(\mathbf{z})$, and equality may also hold when only exact null modes are removed. That factorisation costs $O(DK^2)$ and is cached for fixed directions, each coalition then costs $O(|S|\, r \min(|S|, r))$ for span rank $r \leq K$, independent of D, and the total is one backward pass plus $O(DK^2 + 2^K K^3)$ with no model evaluation afterwards, so exact enumeration suits modest vocabularies and permutation sampling estimates the same values beyond them. Every complete sampled permutation telescopes to $v_\varepsilon([K])$, so efficiency is retained and sampling error falls on the individual attributions (Appendix B).

4.6 Pairwise interactions

Efficiency says how credit was divided, not how two concepts relate. That separate question is measured by the pairwise Shapley interaction index (Grabisch & Roubens, 1999) applied to the squared game v^2, because squared projection lengths add exactly for perpendicular directions while unsquared lengths do not, so only the squared form gives a zero interaction to two concepts orthogonal to each other and to the rest of the vocabulary; pairwise orthogonality alone does not suffice. The index is a diagnostic, not part of the allocation, and its sign cannot be read from a pair's cosine alone (Appendix D). Near-synonymous concepts may instead be declared one grouped player spanning their union, which changes the granularity of the allocation (Appendix C).

5 ImageNet: one gradient, three answers

Setup. A torchvision ResNet50 is explained at `layer3`, whose activation tensor has 1024 channels over 14×14 positions. We flatten rather than average, so $D = 200{,}704$, where the dimension-independence of Section 4.5 matters in practice: no $D \times D$ object is ever formed, and one exact attribution allocates 46 MiB (Appendix B). The vocabulary is $K = 15$ TCAV-style textures and colours (`striped`, `dotted`, `zigzagged`, `bumpy`, `braided`, `honeycombed`, `spiralled`, `bubbly`, `cobwebbed`, `crystalline`, `red`, `blue`, `yellow`, `green`, `white`), each a linear probe of 120 concept images against 120 negatives. The explained scalar is the target-class logit on $n = 200$ ImageNet validation images, 50 each from `zebra`, `basketball`, `dalmatian` and `goldfish`; the target class is the top-1 prediction for 192 of them. This is TCAV's home ground: the same setting, vocabulary and CAV estimator.

Coverage, then accounting. Across all 200 instances the attributions sum to the projected gradient length to within 9.4×10^{-17}, floating-point noise, as (3) enumerated exactly must. The fitted vocabulary covers a mean 0.142% of the squared gradient length, a mean projected length fraction of 3.6%, and that number is reported before any share is read. It is small and not arbitrary: random 15-dimensional subspaces inside the top-200 principal subspace of the probe activations cover 0.063%, probes refit on shuffled labels 0.028%, and the fitted vocabulary exceeds the shuffled maximum on 96% of images (Appendix F). Independent magnitudes exceed the efficient total by 3.43× on average; most of that gap is the difference between adding magnitudes and measuring a length, and Appendix F separates the two. Table 1 gives the per-class picture and the agreement with TCAV, a rank correlation of 0.85 between the class-level TCAV score and the mean signed SPACE display over the same instances.

The gradient is what the model does. On 5,200 image–coalition pairs, stepping the `layer3` representation by $\varepsilon = 0.1$ along each coalition's best direction tracks the first-order prediction $\varepsilon v(S)$ with rank correlation 0.9998, a through-origin slope of 1.000 and a median normalised error of 0.4% (Appendix E), so the sampled values are changes the model realises.

Allocation changes what a reader keeps. A reader keeps the top few concepts. Table 2 scores each rule's top three by the fraction of the vocabulary's captured sensitivity their joint span commands, and by the logit change the model produces when stepped inside that span. The allocation's three reach 76% against 67% for normalised $|a_k|$, and the model moves 12% further along the allocation's choice; reaching 90% of the captured sensitivity takes seven concepts by allocation and ten by $|a_k|$. The single best concept is $|a_k|$'s by construction, and the gap opens from the second concept on, because $|a_k|$ keeps stacking near-parallel colour directions that the Shapley rule has already paid for. The gain is not a property of one fitted vocabulary: on five bootstrap refits of every probe it is 5 to 8 points with an image-level standard error below one point, and the allocation wins the paired comparison on 47 to 60% of images against 10 to 15% (Appendix F). It is not a property of the shared negative pool either: with negatives drawn per concept from disjoint pools, which lowers the mean inter-probe cosine from 0.19 to 0.10, the gap narrows to 2.5 points at five standard errors (Appendix F): the gain is the overlap being paid for, and it shrinks as the overlap does. ICS (Schrouff et al., 2021), the closest local competitor, ranks by the gradient integrated from a zero reference; its top three command 43%, no better than a random order, because a path average is not the sensitivity at the instance. Orthogonalising the directions before ranking selects as well as greedy but hands the reader mixtures of all fifteen concepts; grouping the colours narrows the gap to 3 points, and whitening the metric widens it to 11 (Appendix F).

Selection approaches the optimum. Greedy forward selection on v reaches 79% at three concepts against the exact best subset's 80%, read off the enumerated values, and Proposition 6 gives it a guarantee the allocation lacks. The Shapley value of v^2 selects marginally better than that of v, 77% against 76%, so the gain comes from paying for overlap coalition by coalition, not from the exponent. At $K = 52$, all DTD categories and the colours, sampled allocation beats $|a_k|$ by 6 points at three concepts and 15 at ten, and greedy beats both at a thirty-fifth of the cost (Appendix F).

Table 1: ImageNet $K = 15$ results over all 200 seed-0 instances. Spearman ρ compares the TCAV score against the mean signed SPACE display over the 15 concepts of each class. $r_{\text{ind}} = \sum_k |a_k| / v([K])$ is the non-additivity ratio, reported as a descriptive scale of the overcount and not as a measure of overlap, and $c(\mathbf{z})$ the completeness. Maximum efficiency error across all rows is 9.4×10^{-17}.

Class	n	Spearman ρ	mean r_{ind}	mean $c(\mathbf{z})$
basketball	50	0.950	3.23×	1.53×10^{-3}
dalmatian	50	0.769	3.91×	1.12×10^{-3}
goldfish	50	0.813	3.26×	1.23×10^{-3}
zebra	50	0.850	3.33×	1.82×10^{-3}
All	200	0.846	3.43×	1.42×10^{-3}

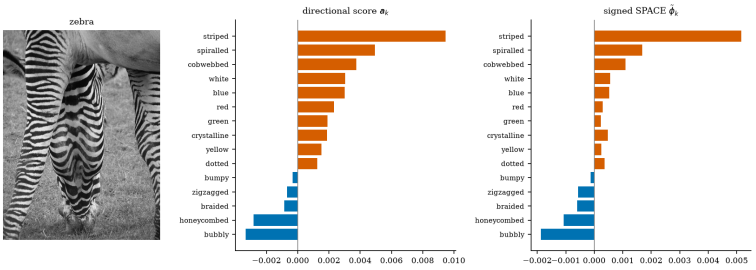


Figure 1: One ImageNet zebra instance, $K = 15$. *Left:* raw directional derivatives a_k, scored independently. *Right:* the signed display of the SPACE allocation, orange positive for the target logit, blue negative. The unsigned ϕ_k sum to $v([K]) = 0.01488$ at $c(\mathbf{z}) = 1.26 \times 10^{-3}$; the signed values do not sum (Section 4.4). `striped` takes 34.7% of the allocation against 23.1% of $\sum_k |a_k|$.

6 Controlled benchmark: planted ground truth

ImageNet has no correct concept explanation to test against, so we crossed pattern (striped/spotted) with colour (tan/gray) on one procedural quadruped under balanced nuisances (Figure 2, Appendix C): 90 rendered images per class, 54 to train a small head on frozen `layer3`, 18 to estimate pattern, colour and brightness-nuisance CAVs, 18 held out. Head and semantic CAVs are perfect on held-out data in every seed.

Recovery. Across five model and CAV seeds the planted factor takes the largest share in 18 of 20 seed–margin cases, and all 360 matched swaps move the margin in the expected direction. Normalised $|a_k|$ obtains the same 18/20 with the same two failures: with three near-independent directions there is no overlap to pay for, so the rules agree, as Remark 1 predicts (Appendix C).

Redundancy. The benchmark then does what ImageNet cannot. We refit each semantic probe on three disjoint thirds of the estimation cells, giving three independently estimated descriptions of colour and three of pattern at pairwise cosines near 0.9, and ask, on margins where both factors matter, whether a two-concept shortlist finds both. Selection on the coalition value finds both on every image of every seed and never spends a slot on the nuisance probe; every per-concept ranking, the allocation included, fills the budget with descriptions of the dominant factor (Table 2). Each description carries distinct span and is credited for it, which is exactly why credit is not a shortlist. No allocation predicts a finite factor swap, but a controlled shift of reliance shows what allocation is for: on a margin the model relies on colour for a measured 17%, three colour descriptions lead $|a_k|$ to credit colour 52%, the allocation of v 36% and of v^2 26%, stable across refits (Appendix C).

7 Scope

What the numbers answer, and over what. $v(S)$ is the largest first-order change a unit move inside the coalition's span can produce at one input (Proposition 1): a sensitivity

Table 2: SPACE allocation against SPACE selection, both on the same game. Left: sensitivity captured by each rule's top three ImageNet concepts, and the observed logit change inside that span relative to $|a_k|$. Right: controlled two-factor images whose two-concept shortlist holds both planted factors, with three colour and one or three pattern descriptions. Full tables in Appendices F and C.

	ImageNet, top three		Controlled, both factors found	
Rule	captured	logit gain	3c+1p+1n	3c+3p+1n
Exact best subset of v	80%	+19%	100%	100%
SPACE selection (greedy on v)	79%	+17%	100%	100%
SPACE allocation, v^2 (Shapley)	77%	+13%	13%	1%
SPACE allocation, v (Shapley)	76%	+12%	8%	0%
Normalised $\lvert a_k \rvert$, one at a time	67%	—	1%	0%
Random	44%	−38%	31%	49%

question answered exactly, not the counterfactual one of what changing the concept would do, which the impossibility results of Bilodeau et al. (2024) and You et al. (2025) bound while leaving this one untouched. The shares divide what the vocabulary captures, and SPACE reports how much that is, 0.142% on ImageNet (Section 5); one-at-a-time scores are equally exposed to small coverage while reporting none.

What the directions carry, the explanation carries. SPACE consumes unit vectors and adds no supervision, so concept semantics stay the estimator's job; estimation noise reaches every rule together, and none survives anisotropic rescaling (Proposition 3). Two hazards belong beside every bar chart. A concept nearly orthogonal to the gradient can still earn credit by completing a span (Proposition 2): on ImageNet, concepts with a derivative below 5% of their row's maximum hold 2.5% of the total on average and 31% on the worst instance (Appendix F). And the decomposition that adds up is unsigned; the signed display carries a direction and no efficiency guarantee (Section 4.4).

Allocation and selection, measured. Table 2 puts the two operations side by side. On ImageNet the allocation's top three capture 8 points more of the vocabulary's sensitivity than $|a_k|$'s, 5 to 8 across five refitted vocabularies, and the model confirms it. On the controlled benchmark, with three independently fitted descriptions of colour, every per-concept ranking, the allocation included, spends a two-concept budget on synonyms and finds both planted factors less often than a random pair, while selection on the coalition value finds them on every image. Shares are credit; the shortlist is a different operation on the same game.

8 Conclusion

Concept scores are usually read as though they were shares of something, and they are not. SPACE projects the model's local gradient onto the span of each group of concept directions, takes that length as the group's value, reports how much of the model the vocabulary covers, divides that total by the Shapley rule so that shared sensitivity is credited once, and selects a shortlist on the same game. The paper shows where the last two agree and where only selection finds distinct factors, and any estimator can supply the vocabulary.

AI Use Statement

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, and interpret results. We have not used generative AI tools to generate synthetic data sets: the controlled and synthetic data were produced by executing the released code. Assisting with translation and supporting qualitative and thematic data analysis are not ap-

plicable to this work. Additionally, we used generative AI tools to draft parts of a research paper, which here includes the first drafts of this paper's text, all of which the authors revised and approved, and to edit a research paper to improve readability, suggest a structure for a research paper, propose a title or keywords for a research paper, create or modify scientific figures or images, suggest experimental parameters, create or edit software code, create artifacts such as the anonymous code supplement, summarize or analyse existing literature, identify relevant literature, discover research topics or identify gaps, brainstorm, source or search for information, and format references. Every reported experimental value comes from executing the released code, not from language-model generation, and AI-assisted outputs were checked against executable tests, machine-readable records, and primary sources. The authors take full responsibility for the content of the paper and the artifact.

Reproducibility Statement

The anonymous supplement provides a single `reproduce.py` interface, fixed configurations, data-acquisition and controlled-data-generation instructions, deterministic rendering checks, the exact-replication, near-duplicate and distinct-factor tests, the top-m selection test with its bootstrap, negative-pool and coverage-null controls, the ICS baseline, the peak-allocation measurement, and the complete unit suite. The appendices provide proofs, protocols, and robustness analyses. Every reported number is produced by the supplement from ImageNet validation images, the DTD textures, the torchvision ResNet50 weights, and the released rendering protocol; nothing requires restricted inputs.

Ethics Statement

This work studies post-hoc explanations of existing image classifiers. It introduces no human-subject study, newly collected personal data, or deployed decision system. Datasets and pretrained checkpoints are used under their respective terms and are not redistributed. The main risk is over-interpreting an explanation. Sections 4.4 and 7 distinguish local projection accounting from causal or counterfactual claims and signed displays from efficient decompositions.

References

Andrew Bai, Chih-Kuan Yeh, Pradeep Ravikumar, Neil Y. C. Lin, and Cho-Jui Hsieh. Concept gradient: Concept-based interpretation without linear assumption. In *International Conference on Learning Representations*, 2023.

Logan Bell, Nikhil Devanathan, and Stephen Boyd. Efficient Shapley performance attribution for least-squares regression. *Statistics and Computing*, 34:149, 2024. arXiv:2310.19245.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024. doi: 10.1073/pnas.2304406120.

David V. Budescu. Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114(3):542–551, 1993. doi: 10.1037/0033-2909.114.3.542.

Jonathan Crabbé and Mihaela van der Schaar. Concept activation regions: A generalized framework for concept-based explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 2590–2607, 2022.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1057–1064, 2011.

Antonio De Santis, Riccardo Campi, Matteo Bianchi, and Marco Brambilla. Visual-TCAV: Concept-based attribution and saliency maps for post-hoc explainability in image classification. *Transactions on Machine Learning Research*, 2026. arXiv:2411.05698.

Eren Erogullari, Sebastian Lapuschkin, Wojciech Samek, and Frederik Pahde. Post-hoc concept disentanglement: From correlated to isolated concept representations. In *World Conference on Explainable Artificial Intelligence (xAI)*. Springer, 2025. arXiv:2503.05522.

Cassandra Goldberg, Chaehyeon Kim, Adam Stein, and Eric Wong. SuperActivators: Only the tail of the distribution contains reliable concept signals. *arXiv preprint arXiv:2512.05038*, 2025.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999.

Osnat Israeli. A Shapley-based decomposition of the R-square of a linear regression. *The Journal of Economic Inequality*, 5:199–212, 2007. doi: 10.1007/s10888-006-9036-6.

Annemarie Jutte, Faizan Ahmed, Jeroen Linssen, and Maurice van Keulen. C-SHAP for time series: An approach to high-level temporal explanations. *arXiv preprint arXiv:2504.11159*, 2025.

Rémi Kazmierczak, Eloïse Berthier, Goran Frehse, and Gianni Franchi. CLIP-QDA: An explainable concept bottleneck model. *Transactions on Machine Learning Research*, 2024. arXiv:2312.00110.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2668–2677, 2018.

Jiahui Li, Kun Kuang, Lin Li, Long Chen, Songyang Zhang, Jian Shao, and Jun Xiao. Instance-wise or class-wise? a tale of neighbor Shapley for concept-based explanation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM)*, pp. 3664–3672, 2021.

Richard H. Lindeman, Peter F. Merenda, and Ruth Z. Gold. *Introduction to Bivariate and Multivariate Analysis*. Scott, Foresman, Homewood, Illinois, 1980.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001. doi: 10.1002/asmb.446.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4765–4774, 2017.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: Shapley interactions for machine learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Angus Nicolson, Lisa Schut, J. Alison Noble, and Yarin Gal. Explaining explainability: Recommendations for effective use of concept activation vectors. *Transactions on Machine Learning Research*, 2025. arXiv:2404.03713.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023.

Art B. Owen. Sobol' indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014. doi: 10.1137/130936233.

Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. Best of both worlds: Local and global explanations with human-understandable concepts. *arXiv preprint arXiv:2106.08641*, 2021. arXiv:2106.08641.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games*, volume 2, pp. 307–317. Princeton University Press, 1953. Annals of Mathematics Studies 28.

Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. Explain any concept: Segment anything meets concept-based explanation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-Shap: The faithful Shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

Johanna Vielhaben, Stefan Blücher, and Nils Strodthoff. Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research*, 2023.

Chih-Kuan Yeh, Been Kim, Sercan Ö. Arık, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Weiqiu You, Helen Qu, Marco Gatti, Bhuvnesh Jain, and Eric Wong. Sum-of-parts: Self-attributing neural networks with end-to-end learning of feature groups. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. arXiv:2310.16316.

A Proofs

Proof of Proposition 1. Let $U = \operatorname{span}(S)$. For $\boldsymbol{\delta} \in U$, $\langle \mathbf{g}, \boldsymbol{\delta} \rangle = \langle \mathbf{P}_U \mathbf{g}, \boldsymbol{\delta} \rangle \leq \|\mathbf{P}_U \mathbf{g}\| \|\boldsymbol{\delta}\| \leq \|\mathbf{P}_U \mathbf{g}\|$, with equality at $\boldsymbol{\delta} = \mathbf{P}_U \mathbf{g} / \|\mathbf{P}_U \mathbf{g}\|$ when the projection is nonzero. The zero case is immediate. □

Proposition 3 (Equivariance and the choice of metric)**.** *Let $Q \in \mathbb{R}^{D \times D}$ be orthogonal and let the representation be reparameterised as $\mathbf{z}' = Q\mathbf{z}$ with $f'(\mathbf{z}') = f(Q^\top \mathbf{z}')$, so the model function is unchanged. Then $\mathbf{g}' = Q\mathbf{g}$, $\mathbf{d}'_k = Q\mathbf{d}_k$, and every $v(S)$, ϕ_k, and $c(\mathbf{z})$ is unchanged. Under a uniform rescaling $\mathbf{z}' = \alpha\mathbf{z}$, $\alpha > 0$, every ϕ_k scales by $1/\alpha$ and the shares $\phi_k / v([K])$ and $c(\mathbf{z})$ are unchanged. Neither statement extends to an anisotropic rescaling $\mathbf{z}' = \Lambda\mathbf{z}$ with Λ diagonal and non-scalar.*

Proof. For orthogonal Q, $\langle \mathbf{g}', \mathbf{d}'_k \rangle = \langle Q\mathbf{g}, Q\mathbf{d}_k \rangle = \langle \mathbf{g}, \mathbf{d}_k \rangle$ and $\mathbf{D}'^\top \mathbf{D}' = \mathbf{D}^\top \mathbf{D}$, so by (6) every coalition value is unchanged, hence so is every Shapley value. The equality $\|\mathbf{g}'\| = \|\mathbf{g}\|$ gives the completeness claim. For $\mathbf{z}' = \alpha\mathbf{z}$ the chain rule gives $\mathbf{g}' = \mathbf{g}/\alpha$ while unit directions are unchanged, so $v'(S) = v(S)/\alpha$ for every S. The Shapley value is positively homogeneous in the game, and the ratios are therefore invariant. For diagonal Λ, $\mathbf{g}' = \Lambda^{-1}\mathbf{g}$ while $\mathbf{d}'_k = \Lambda\mathbf{d}_k / \|\Lambda\mathbf{d}_k\|$, so both $\mathbf{D}'^\top \mathbf{g}'$ and $\mathbf{D}'^\top \mathbf{D}'$ change non-conformally. A counterexample suffices for the negative claim: with $\mathbf{d}_1 = e_1$, $\mathbf{d}_2 = e_2$ and $\mathbf{g} = (1, 1)$, symmetry gives $\phi_1 = \phi_2 = \sqrt{2}/2$. After $\Lambda = \operatorname{diag}(\alpha, 1)$ the same semantic axes give $\mathbf{g}' = (1/\alpha, 1)$ and $\phi'_1 \neq \phi'_2$ for $\alpha \neq 1$. □

The same counterexample changes $|a_k|$ from $(1, 1)$ to $(1/\alpha, 1)$, so the per-concept baseline is equally metric-dependent: fixing unit concept directions in an ambient inner product is a property of the CAV construction, not of the Shapley step. Section 7 states the assumption explicitly. A whitened or covariance-based metric is a coherent alternative, but its estimand would differ and would not be metric-free.

The proposition's scope should be read narrowly. It transports concept directions that have *already been supplied*, as geometric vectors, under the stated change of coordinates. It does not claim that a probe refitted in the rescaled representation would recover the transformed direction: a regularised logistic or ridge fit is not equivariant under anisotropic rescaling of its inputs, so the estimated direction and the transported one can differ. The statement is about the attribution rule given directions, not about the estimator that produces them.

Proposition 4 (Modularity of v^2 on orthogonal blocks)**.** *If $\mathbf{d}_j \perp \mathbf{d}_k$ for all $j \in S$, $k \in T$ with $S \cap T = \emptyset$, then $v^2(S \cup T) = v^2(S) + v^2(T)$.*

Proof. Orthogonal subspaces have projectors satisfying $\mathbf{P}_{S\cup T} = \mathbf{P}_S + \mathbf{P}_T$. Then $v^2(S\cup T) = \|(\mathbf{P}_S + \mathbf{P}_T)\mathbf{g}\|^2 = \|\mathbf{P}_S\mathbf{g}\|^2 + 2\langle \mathbf{P}_S\mathbf{g}, \mathbf{P}_T\mathbf{g}\rangle + \|\mathbf{P}_T\mathbf{g}\|^2$, and the cross term vanishes. □

Proposition 5 (Stability under direction estimation error)**.** *Let $\mathbf{D}, \mathbf{D}' \in \mathbb{R}^{D\times K}$ have unit columns, write $\mathbf{E} = \mathbf{D}' - \mathbf{D}$, and suppose $\|\mathbf{E}\|_{\mathrm{op}} < s := \sigma_K(\mathbf{D})$. Then for every coalition S and the same gradient $\mathbf{g}$,*

$$|v'(S) - v(S)| \le \frac{2\|\mathbf{g}\|\,\|\mathbf{E}\|_{\mathrm{op}}}{s - \|\mathbf{E}\|_{\mathrm{op}}}, \qquad |\phi'_k - \phi_k| \le \frac{4\|\mathbf{g}\|\,\|\mathbf{E}\|_{\mathrm{op}}}{s - \|\mathbf{E}\|_{\mathrm{op}}}.$$

Proof. Fix S and write $\mathbf{A} = \mathbf{D}_S$, $\mathbf{A}' = \mathbf{D}'_S$, $\mathbf{E}_S = \mathbf{A}' - \mathbf{A}$. Column subsets interlace, so $\sigma_{\min}(\mathbf{A}) \ge s$, and $\|\mathbf{E}_S\| \le \|\mathbf{E}\|$, so by Weyl's inequality $\sigma_{\min}(\mathbf{A}') \ge s - \|\mathbf{E}\| > 0$ and both projectors have full rank. Since $(\mathbf{I} - \mathbf{P})\mathbf{A}' = (\mathbf{I} - \mathbf{P})\mathbf{E}_S$ for $\mathbf{P} = \mathbf{P}_S$, we get $\|(\mathbf{I}-\mathbf{P})\mathbf{P}'\| = \|(\mathbf{I}-\mathbf{P})\mathbf{E}_S\mathbf{A}'^{+}\| \le \|\mathbf{E}\|/(s - \|\mathbf{E}\|)$, and symmetrically $\|\mathbf{P}(\mathbf{I}-\mathbf{P}')\| \le \|\mathbf{E}\|/s$. The identity $\mathbf{P}' - \mathbf{P} = (\mathbf{I} - \mathbf{P})\mathbf{P}' - \mathbf{P}(\mathbf{I} - \mathbf{P}')$ then gives $\|\mathbf{P}' - \mathbf{P}\| \le 2\|\mathbf{E}\|/(s - \|\mathbf{E}\|)$, and $|v'(S) - v(S)| \le \|(\mathbf{P}' - \mathbf{P})\mathbf{g}\|$ proves the first bound. Each ϕ_k is a convex combination of differences $v(S \cup \{k\}) - v(S)$, each of which moves by at most twice the first bound. □

The condition is the one the near-duplicate construction of Appendix C violates: for $\mathbf{d}_1 = e_1$ and $\mathbf{d}_2 \propto e_1 + \epsilon e_2$ the smallest singular value is of order ϵ, so the bound is void and the allocation can jump. The proposition gives a sufficient condition for stability, estimation error small relative to the vocabulary's smallest singular value, and says nothing about necessity. It is a stronger requirement than the numerical rank certificate of Section 4, which only guards the factorisation. On the ImageNet vocabulary the bound does not certify the result: the five bootstrap refits of Appendix F differ from the primary directions by an operator norm of 0.88 to 1.03, against a smallest singular value of 0.26, so the hypothesis fails by a factor of about four and the empirical sensitivity reported there is the honest statement of stability.

Proposition 6 (Greedy selection inherits a least-squares guarantee)**.** *$v^2(S) = \mathbf{g}^\top \mathbf{P}_S \mathbf{g}$ is the uncentred explained sum of squares of regressing $\mathbf{g}$ on the columns of $\mathbf{D}_S$, so the spectral analysis of greedy subset selection for R^2 by Das & Kempe (2011) applies without change. Let $\lambda_{\min}(\mathbf{G}, r)$ be the smallest eigenvalue over principal submatrices of the Gram matrix $\mathbf{G} = \mathbf{D}^\top\mathbf{D}$ of order at most r. Then the submodularity ratio of v^2 at budget m is at least $\gamma_m = \lambda_{\min}(\mathbf{G}, 2m)$, and greedy forward selection on v, which coincides with greedy on v^2, returns a set S_m with*

$$v^2(S_m) \;\ge\; \left(1 - e^{-\gamma_m}\right) \max_{|S|=m} v^2(S), \qquad v(S_m) \;\ge\; \sqrt{1 - e^{-\gamma_m}} \max_{|S|=m} v(S).$$

Proof. In Das & Kempe (2011) the R^2 of a subset S equals $\mathbf{b}_S^\top \mathbf{C}_S^{-1} \mathbf{b}_S$ for the predictor covariance matrix $\mathbf{C}$ and response covariances $\mathbf{b}$, with all predictors normalised to unit variance. Equation (6) is the same expression with $\mathbf{G}$ for $\mathbf{C}$ and $\mathbf{D}^\top\mathbf{g}$ for $\mathbf{b}$, unit directions playing the role of unit-variance predictors, and the certificate of Section 4 supplying the nonsingularity they assume. Their Lemma 2.4 bounds the submodularity ratio below

by $\lambda_{\min}(\mathbf{C}, 2k)$ and their Theorem 3.2 gives forward regression the $(1 - e^{-\gamma})$ guarantee. Monotonicity of the square root transfers the set inequality to v. □

The guarantee is honest about its constant. On the ImageNet vocabulary $\lambda_{\min}(\mathbf{G}, 6) = 0.069$, so at $m = 3$ greedy is guaranteed only 26% of the best three-concept value of v, while it achieves 99% (Table 8); the near-parallel colour directions make the worst case loose. The bound tightens as the directions approach orthogonality, where γ_m approaches one, and it is a guarantee the allocation does not have.

Remark 1 (Perpendicular vocabularies keep the singleton order)**.** *If the directions are mutually orthogonal then $v^2(S) = \sum_{k \in S} a_k^2$ by Proposition 4, so the gain from adding concept k to any coalition S is $\sqrt{v^2(S) + a_k^2} - \sqrt{v^2(S)}$, which increases with a_k^2. Comparing concepts j and k coalition by coalition with the same Shapley weights, $|a_j| > |a_k|$ implies $\phi_j > \phi_k$. On a perpendicular vocabulary SPACE and $|a_k|$ therefore rank the concepts identically, and every reranking reported in this paper is produced by overlap alone.*

Corollary 1 (The norm form is not usable as an overlap index)**.** *Under the hypotheses of Proposition 4, $v(S \cup T) = \sqrt{v^2(S) + v^2(T)} < v(S) + v(T)$ whenever both are nonzero. More generally, if $\mathbf{d}_j$ and $\mathbf{d}_k$ are orthogonal to each other and to every other direction, then the pairwise index* (7) *computed on v rather than v^2 is* strictly negative *whenever $\langle \mathbf{g}, \mathbf{d}_j \rangle$ and $\langle \mathbf{g}, \mathbf{d}_k \rangle$ are both nonzero.*

Proof. Write $s = v^2(S)$ for a conditioning coalition S and $\alpha = \langle \mathbf{g}, \mathbf{d}_j \rangle^2$, $\beta = \langle \mathbf{g}, \mathbf{d}_k \rangle^2$. Under the stated orthogonality, Proposition 4 gives the bracket of (7) on v as $h(s) = \sqrt{s + \alpha + \beta} - \sqrt{s + \alpha} - \sqrt{s + \beta} + \sqrt{s} = [\,\varphi(s + \alpha + \beta) - \varphi(s + \alpha)\,] - [\,\varphi(s + \beta) - \varphi(s)\,]$ with $\varphi(x) = \sqrt{x}$. As φ is strictly concave and $s + \alpha \geq s$, the increment over the same step β is strictly smaller at the larger base, so $h(s) < 0$ for every S. A positively weighted sum of strictly negative terms is strictly negative. □

This is why Section 4.6 uses v^2. The negative value is not an error about the game v. Subadditivity under a shared unit budget is real and makes the primary attribution well posed. However, the value is nonzero for concepts that are independent by construction, so it cannot separate overlap from concavity.

Proposition 7 (Zero pairwise interaction)**.** *$\phi_{jk} = 0$ if $K = 2$ and $\mathbf{d}_j \perp \mathbf{d}_k$. For general K, it is zero if $\mathbf{d}_j$ and $\mathbf{d}_k$ are each orthogonal to $\mathbf{d}_m$ for every $m \in [K] \setminus \{j, k\}$, in addition to $\mathbf{d}_j \perp \mathbf{d}_k$.*

Proof. The $K = 2$ case is (8) at $c = 0$. For general K under the stated condition, $\mathrm{span}\{\mathbf{d}_j, \mathbf{d}_k\} \perp \mathrm{span}(S)$ for every $S \subseteq [K] \setminus \{j, k\}$, so Proposition 4 applies to the bipartitions $\{S, \{j, k\}\}$, $\{S, \{j\}\}$ and $\{S, \{k\}\}$, giving $v^2(S \cup \{j, k\}) = v^2(S) + v^2(\{j, k\})$, $v^2(S \cup \{j\}) = v^2(S) + v^2(\{j\})$ and $v^2(S \cup \{k\}) = v^2(S) + v^2(\{k\})$. Since $\mathbf{d}_j \perp \mathbf{d}_k$, a further application within $\{j, k\}$ gives $v^2(\{j, k\}) = v^2(\{j\}) + v^2(\{k\})$. Substituting, every bracket in (7) is $\big(v^2(S) + v^2(\{j\}) + v^2(\{k\})\big) - \big(v^2(S) + v^2(\{j\})\big) - \big(v^2(S) + v^2(\{k\})\big) + v^2(S) = 0$. □

Remark 2 (Pairwise orthogonality is insufficient for $K > 2$)**.** *Let $\mathbf{d}_1 = (1, 0)$, $\mathbf{d}_2 = (0, 1)$, $\mathbf{d}_3 = (1/\sqrt{2}, 1/\sqrt{2})$ and $\mathbf{g} = (1, 0)$, so $K = 3$ and $\mathbf{d}_1 \perp \mathbf{d}_2$ while $\mathbf{d}_3$ is orthogonal to neither. Then $v^2(\{1\}) = 1$, $v^2(\{2\}) = 0$, $v^2(\{3\}) = 1/2$, $v^2(\{1, 2\}) = v^2(\{1, 3\}) = v^2(\{2, 3\}) = v^2(\{1, 2, 3\}) = 1$. The two conditioning coalitions for the pair $(1, 2)$ carry weight $\frac{1}{2}$ each. At $S = \emptyset$ the bracket is $1 - 1 - 0 + 0 = 0$, exactly as Proposition 4 predicts. At $S = \{3\}$ it is $1 - 1 - 1 + \frac{1}{2} = -\frac{1}{2}$. Hence $\phi_{12} = \frac{1}{2}(0) + \frac{1}{2}(-\frac{1}{2}) = -\frac{1}{4} \neq 0$ despite $\mathbf{d}_1 \perp \mathbf{d}_2$. The entire interaction comes from the coalition containing $\mathbf{d}_3$, whose correlation with both members breaks the modularity that held at $S = \emptyset$. This is why Proposition 7 requires orthogonality to every other direction and not merely to the partner.*

Proposition 8 (Closed form at $K = 2$)**.** *Let $\mathbf{d}_1, \mathbf{d}_2$ be unit directions with $c = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle$, $c \neq \pm 1$, and let $a = \langle \mathbf{g}, \mathbf{d}_1 \rangle$, $b = \langle \mathbf{g}, \mathbf{d}_2 \rangle$. Then $\phi_{12} = v^2(\{1, 2\}) - v^2(\{1\}) - v^2(\{2\}) = c\big[c(a^2 + b^2) - 2ab\big]/(1 - c^2)$.*

Proof. With $\mathbf{D} = [\mathbf{d}_1\, \mathbf{d}_2]$ the Gram matrix is $\mathbf{D}^\top\mathbf{D} = \left(\begin{smallmatrix} 1 & c \\ c & 1 \end{smallmatrix}\right)$ with inverse $\frac{1}{1-c^2}\left(\begin{smallmatrix} 1 & -c \\ -c & 1 \end{smallmatrix}\right)$. By (6), $v^2(\{1,2\}) = [a\ b](\mathbf{D}^\top\mathbf{D})^{-1}[a\ b]^\top = (a^2 - 2cab + b^2)/(1-c^2)$. Subtracting a^2 and b^2 and collecting terms gives the result. □

Remark 3 (The sign is not a function of c)**.** *Take $c = 0.5$, $a = 1$, $b = 0$: then Proposition 8 gives $\phi_{12} = 0.5(0.5)/0.75 = 1/3 > 0$, a* positive *interaction between non-orthogonal directions. Geometrically, $\mathbf{g} \perp \mathbf{d}_2$ so concept 2 contributes nothing alone, but the two directions jointly span a plane capturing more of $\mathbf{g}$ than the $\mathbf{d}_1$ line does, since $v^2(\{1,2\}) = a^2/(1-c^2) > a^2$. Correlation is therefore necessary but not sufficient for the negative interaction that signals redundancy. The full condition is* (9).

Remark 4 (Zero projections do not imply zero interaction for $K > 2$)**.** *Let $\mathbf{d}_1 = (1,0,0)$, $\mathbf{d}_2 = (0,1,0)$, $\mathbf{d}_3 = (1,1,1)/\sqrt{3}$ and $\mathbf{g} = (0,0,1)$. Then $a = b = 0$ exactly, while $\langle \mathbf{g}, \mathbf{d}_3 \rangle = 1/\sqrt{3}$. Direct computation gives $v^2(\{1\}) = v^2(\{2\}) = v^2(\{1,2\}) = 0$, $v^2(\{3\}) = 1/3$, $v^2(\{1,3\}) = v^2(\{2,3\}) = 1/2$ and $v^2(\{1,2,3\}) = 1$. The $S = \emptyset$ bracket is 0, the $S = \{3\}$ bracket is $1 - \frac{1}{2} - \frac{1}{2} + \frac{1}{3} = \frac{1}{3}$, so $\phi_{12} = \frac{1}{6} \neq 0$. The mechanism is a genuine three-way effect: $\mathbf{d}_3$ carries all the gradient mass and correlates with both members, and completing the basis with $\{\mathbf{d}_1, \mathbf{d}_2\}$ on top of $\{\mathbf{d}_3\}$ recovers the rest of $\mathbf{g}$. The homogeneity of* (8) *in (a,b) therefore does not survive $K > 2$: vanishing individual projections silence the $S = \emptyset$ bracket only.*

B Algorithm and complexity

Algorithm 1: SPACE (exact enumeration)

Input: embedding $\mathbf{z}$, directions $\{\mathbf{d}_k\}_{k=1}^K$, head f, rank floor ε. Steps 3–5 evaluate Equation (6) in an orthonormal basis of the directions' span, never forming the Gram matrix. When no direction is discarded at the floor, the computed game equals v on every coalition; that is the certificate reported with every run.

1. $\mathbf{g} \leftarrow \nabla f(\mathbf{z})$ (one backward pass)
2. $\mathbf{a} \leftarrow \mathbf{D}^\top\mathbf{g}$ ($O(DK)$, the display signs)
3. $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \leftarrow \mathbf{D}^\top$ (which is $K \times D$) by a thin SVD, keep the r singular values above $\varepsilon\sigma_1$, and set $\mathbf{M} \leftarrow \mathbf{U}\mathbf{\Sigma} \in \mathbb{R}^{K\times r}$, one row per concept ($O(DK^2)$, cacheable for fixed directions)
4. $\mathbf{h} \leftarrow \mathbf{V}^\top\mathbf{g} \in \mathbb{R}^r$ ($O(Dr)$, once per gradient)
5. for each $S \subseteq [K]$: let $\mathbf{R}_S \in \mathbb{R}^{r\times q}$ hold as *columns* the right singular vectors of $\mathbf{M}_S$ whose singular values clear the same floor, and $v_\varepsilon(S) \leftarrow \|\mathbf{R}_S^\top\mathbf{h}\|$ ($O(|S|r\min(|S|,r))$, independent of D)
6. $\phi_k^\varepsilon \leftarrow \sum_{S\subseteq[K]\setminus\{k\}} \frac{|S|!(K-|S|-1)!}{K!}[v_\varepsilon(S\cup\{k\}) - v_\varepsilon(S)]$
7. $\tilde{\phi}_k^\varepsilon \leftarrow \phi_k^\varepsilon$ if $\mathbf{a}_k \geq 0$, and $\tilde{\phi}_k^\varepsilon \leftarrow -\phi_k^\varepsilon$ otherwise (reporting only, see Section 4.4)

Output: unsigned $\{\phi_k^\varepsilon\}$, signed display $\{\tilde{\phi}_k^\varepsilon\}$, $\sum_k \phi_k^\varepsilon = v_\varepsilon([K])$, and c_ε. Efficiency is exact against $v_\varepsilon([K])$, and against $\|\mathbf{P}_{[K]}\mathbf{g}\|$ when the run certifies its rank.

Step 5 is Equation (6) in the factorised basis: with $\mathbf{D}_S^\top = \mathbf{M}_S\mathbf{V}^\top$, $v^2(S) = \mathbf{h}^\top\mathbf{M}_S^\top(\mathbf{M}_S\mathbf{M}_S^\top)^+\mathbf{M}_S\mathbf{h} = \|\mathbf{R}_S^\top\mathbf{h}\|^2$. The total cost is one backward pass plus $O(DK^2 + 2^K K^3)$ arithmetic, with no further model evaluation. For vocabularies too large to enumerate, the same game is estimated by sampling random orderings of the concepts and averaging each concept's marginal gain; every complete ordering still sums exactly to the full-vocabulary value.

Runtime, memory and numerical checks. On a synthetic $D = 64$ instance, exact enumeration takes about a hundredth of a second at $K = 8$, three seconds at $K = 15$ and a hundred seconds at $K = 20$ on one CPU; the exponential trend is the point, not the machine-specific values. At $K = 15$, 2,000 sampled permutations reproduce the exact

ranking with a mean relative L^1 error of 1.3% in under half the exact time. One exact attribution at the ImageNet geometry ($D = 200{,}704$, $K = 15$) allocates 46 MiB, about $2K$ copies of the gradient, where a dense $D \times D$ array would be 300 GiB. The released test suite checks the factorised coalition values against directly built projectors on random, collinear, duplicate and rank-deficient direction sets, and reproduces the counterexamples of Remarks 2 and 4. The rank floor is 10^{-8} on the singular values of the direction matrix; on the ImageNet vocabulary every coalition's smallest singular value is at least 0.26, and sweeping the floor from 10^{-12} to 10^{-6} leaves every share bit-identical, so no thresholded value enters the reported numbers.

C Controlled benchmark

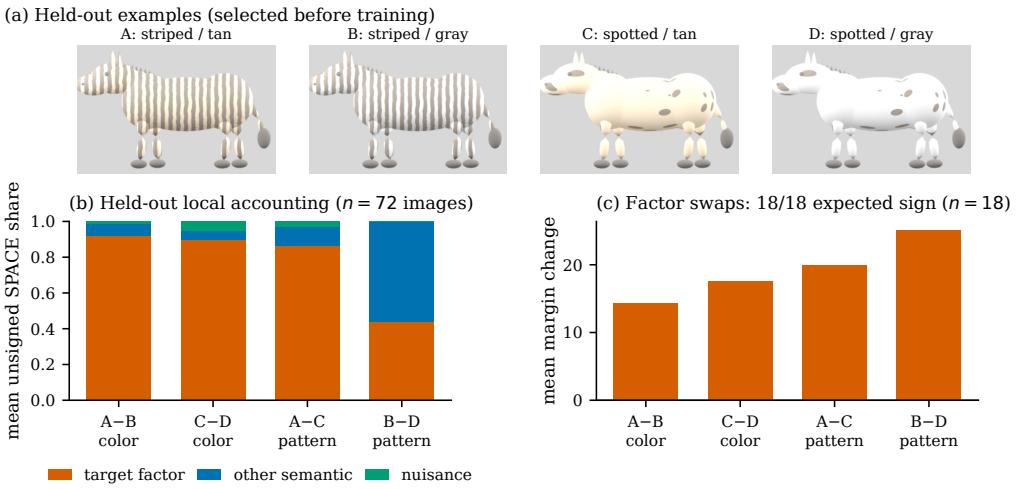


Figure 2: Controlled benchmark. *(a)* One held-out set: same animal, pose, camera, lighting and background, differing only in pattern and colour. *(b)* Mean SPACE shares for the four decisions. *(c)* Change in each decision when only the named factor is altered in the picture; all 18 pairs move the expected way.

The benchmark was repeated for five model and CAV seeds with images, splits, frozen activations, head architecture and training settings fixed. The classifier and both semantic CAVs are perfect on held-out data in every seed. Table 3 reports the mean share of each direction with its spread across seeds, and every matched swap in every seed has the expected sign. ConceptSHAP (Yeh et al., 2020), run on the same three directions with one fitted mapping per coalition, allocates a dataset-level completeness gain rather than a local total, so its panel is not comparable number for number; it recovers the planted factors and takes eight fitted mappings per seed against one backward pass per instance.

An exact copy of a concept. The three benchmark directions barely overlap, so the benchmark cannot show the overlap problem on its own. We therefore add an exact copy of each decision's planted direction to the vocabulary, giving four concepts, and recompute; nothing is refitted. Table 4 shows the effect. One-at-a-time scoring keeps the planted factor first in 18 of 20 cases but now credits the copied factor more, in every case. The SPACE shares also credit the copied factor more, and the planted factor is first in only 16 of 20 cases when each copy competes alone, because the credit is split between the two copies. Told which two entries are copies, SPACE treats them as one player and everything is exactly as before the copy was added; one-at-a-time scoring could use the same knowledge to merge them. So the copy test shows that the grouped game is coherent, not that SPACE wins.

Near-duplicate stress test. Exact replication is a limit. Appending instead a direction at a declared cosine $\rho < 1$ to the planted factor, built from a refit of the same probe on a

Table 3: Controlled benchmark over five seeds. Top: the mean SPACE share of the planted factor, the other factor and the nuisance for each decision, with the spread across seeds in brackets, and how often the planted factor is first. Bottom: ConceptSHAP's allocation of its own dataset-level total, for reference.

SPACE: local unsigned shares, mean (sample SD) over five seeds						
Margin	planted	target	other	nuisance	$c(\mathbf{z})$	target first
A–B	color	.938 (.017)	.044 (.019)	.019 (.004)	.0432 (.0047)	5/5
C–D	color	.919 (.035)	.048 (.035)	.033 (.011)	.0312 (.0109)	5/5
A–C	pattern	.670 (.173)	.319 (.180)	.011 (.011)	.0358 (.0135)	4/5
B–D	pattern	.737 (.220)	.258 (.220)	.006 (.002)	.0327 (.0117)	4/5

ConceptSHAP: global completeness allocation over five seeds		
Concept	mean (sample SD)	range
pattern	.4667 (.0086)	[.4599, .4815]
color	.5074 (.0060)	[.5000, .5154]
nuisance	.0259 (.0047)	[.0185, .0309]

Table 4: Adding an exact copy of the planted direction. Shares are the copied factor's, averaged over images; "increased" counts the 20 seed–decision cases in which the share rose.

Quantity	Before	After	Mean shift	Increased
$v([K])$ grand value	within declared tolerance		—	0/20
Normalised $\lvert a_k \rvert$ share	0.7665	0.8574	+0.0910	**20/20**
SPACE summed members	0.8159	0.8495	+0.0336	20/20
SPACE grouped player	0.8159	0.8159	within tolerance	**0/20**
Target first, normalised $\lvert a_k \rvert$	18/20	18/20	—	—
Target first, SPACE members	18/20	**16/20**	—	—
Target first, SPACE grouped	18/20	18/20	—	—

disjoint half of the estimation split, from a bootstrap refit, or from a random orthogonal axis, gives Table 5. At the concept level both rules put the planted factor first in 20/20 cells at every $\rho \geq 0.9$. On the stricter single-column criterion $|a_k|$ holds 18/20 throughout, ungrouped SPACE stays at the 16/20 of exact replication and recovers only once $\rho \leq 0.8$, and grouping restores 18/20. Because $v(S)$ depends on a coalition's span, any $\rho < 1$ adds the same plane and the grand value takes one value across the whole open interval, at most 1.3% above the $K = 3$ value, returning to it only at $\rho = 1$. On a synthetic $K = 8$ game, appending an exact duplicate leaves coverage unchanged but moves the twins' combined credit by a mean 0.015 away from the original player's; treating the twins as one grouped player recovers it to rounding error.

Table 5: Adding a near-copy at cosine ρ to the planted factor, built three ways. $\Delta v([K])$ is the change in the vocabulary's total. "Largest single share" asks whether one column of the planted factor beats every rival; "largest summed share" first adds the factor's two columns.

			Largest single share			Largest summed share	
Tilt axis	ρ	$\Delta v([K])$	$\lvert a_k \rvert$	SPACE	Grouped	$\lvert a_k \rvert$	SPACE
Disjoint half	1.00	0	18/20	16/20	18/20	20/20	20/20
	0.99	+0.14%	18/20	16/20	18/20	20/20	20/20
	0.90	+0.14%	18/20	16/20	18/20	20/20	20/20
	0.50	+0.14%	18/20	18/20	18/20	20/20	19/20
Bootstrap	1.00	0	18/20	16/20	18/20	20/20	20/20
	0.99	+1.34%	18/20	16/20	19/20	20/20	20/20
	0.90	+1.34%	18/20	16/20	19/20	20/20	20/20
	0.50	+1.34%	18/20	17/20	19/20	19/20	19/20
Random axis	1.00	0	18/20	16/20	18/20	20/20	20/20
	0.99	+0.01%	18/20	16/20	18/20	20/20	20/20
	0.90	+0.01%	18/20	16/20	18/20	20/20	20/20
	0.50	+0.01%	18/20	18/20	18/20	20/20	19/20

Finding both factors with two concepts. Refitting the colour probe on three separate sets of pictures gives three descriptions of colour that agree to a cosine of about 0.9, and likewise for pattern. With vocabularies holding one to three descriptions of each factor plus the nuisance, on the two decisions that depend on both factors, we keep two concepts and ask whether both factors are among them. Ground truth is rendered: swapping one factor in the matched picture and reading the decision gives each factor's measured effect. Table 6 gives the results over five seeds and 144 images each. When one factor has several descriptions, every per-concept ranking keeps two of them: one at a time finds both factors on at most 1% of images, the SPACE shares on 8% with three colour descriptions and under 1% with three of each, below a random pair. The descriptions are not copies, so each really carries a little of its own and the shares credit it correctly; credit is just not a shortlist. Selecting by group value finds both factors on every image of every seed, never picks the nuisance, and needs no knowledge of which entries are synonyms. The same holds with three concepts kept.

The swaps also test the shares directly, and the answer is no for every rule. On these decisions colour accounts for 43% of the measured change on average; the rules' colour shares are 69% (one at a time), 75% (SPACE) and 82% (squared), and across images the shares and the measured fractions correlate at about 0.2. A share describes the response to a tiny move at one point; a swap is a large change through a nonlinear network; they are different questions, as Section 7 says.

A controlled shift in reliance. For the same image, the decision "A or B" depends on colour and the decision "A or C" on pattern, so the model's reliance shifts from one factor to the other between them. We ask how much colour each rule reports on each decision, over five seeds and three independent refits of every probe. Every rule sees the shift on every image, and every detection agrees across refits, so detection does not separate the rules. The size does. On the pattern decision the model's measured reliance on colour, by swapping colour in the picture, is 17%. One-at-a-time scoring reports 31% with one description of colour and 52% with three, since each description is credited in full. The SPACE shares report 28% and 36%, the grouped game 28% and 29%, and the squared shares 24% and 26%. Everyone overstates, but with synonyms in the vocabulary the one-at-a-time score says the model relies on colour for over half of a decision it relies on colour for a sixth of, and the share rules stay within a factor of two, because a second description of the same factor adds little and earns little. This is what the allocation is for, measured against something the projection does not define.

Table 6: Keeping two concepts on decisions that depend on both factors, five seeds and 144 images each. Vocabularies are written as colour descriptions + pattern descriptions + nuisance. Each cell: fraction of images whose two concepts cover both factors, then the fraction of the measured effect those factors carry.

Rule (budget $m = 2$)	3c+1p+1n	1c+3p+1n	2c+2p+1n	3c+3p+1n	1c+1p+1n
Best subset (exact)	100% / 100%	100% / 100%	100% / 100%	100% / 100%	100% / 100%
Greedy forward	100% / 100%	100% / 100%	100% / 100%	100% / 100%	100% / 100%
Shapley on v^2	13% / 50%	100% / 100%	1% / 43%	1% / 43%	100% / 100%
SPACE	8% / 47%	100% / 100%	1% / 43%	0% / 43%	100% / 100%
Normalised $\lvert a_k \rvert$	1% / 43%	99% / 99%	0% / 43%	0% / 43%	100% / 100%
Random order	31% / 62%	28% / 68%	37% / 69%	49% / 74%	31% / 66%

D The pairwise interaction index

The pairwise Shapley interaction index of Grabisch & Roubens (1999), applied to the squared game, is

$$\phi_{jk} = \sum_{S \subseteq [K] \setminus \{j,k\}} \frac{|S|!\,(K - |S| - 2)!}{(K-1)!} \big[v^2(S \cup \{j,k\}) - v^2(S \cup \{j\}) - v^2(S \cup \{k\}) + v^2(S) \big]. \quad (7)$$

For two concepts with cosine c and derivatives a and b it reduces to

$$\phi_{jk} = \frac{c\left[c(a^2+b^2) - 2ab\right]}{1-c^2}, \tag{8}$$

which is negative, signalling redundancy, exactly when the cosine and the product of the derivatives have the same sign and the cosine is small enough,

$$c\,ab > 0 \quad \text{and} \quad |c| < \frac{2|ab|}{a^2+b^2}; \tag{9}$$

at $c = 0$ the interaction is zero. So two concepts can be strongly correlated and same-signed and still interact positively: with $a = 2$ and $b = 1$ the threshold is 0.8, and a cosine of 0.7 gives a negative interaction while 0.9 gives a positive one. The index is a diagnostic of pairwise structure; it does not add up to anything and is no part of the allocation.

Where the two allocations differ. On the 200 ImageNet images the allocations of v and of v^2 agree on the top concept on 96% of images and rank the concepts alike (correlation 0.94), but their shares differ by a mean total-variation distance of 0.10, and the squared allocation selects marginally better (Appendix F). The squared allocation is not just the squared derivatives: it names a different top concept from them on 30% of images.

E Does the group value match the network?

For each of the 200 images we took 26 groups of concepts, all fifteen single concepts, the full vocabulary, and two fixed groups at each of the sizes 2, 4, 8, 12 and 14, moved the representation by a small step in the group's best direction, ran the rest of the network, and compared the change in the target logit with the change the group value predicts. The network after this layer is piecewise linear, so the two agree exactly unless the step crosses a switching point; the errors in Table 7 measure how often that happens, not anything about what the concepts mean. At the step size used in the paper the typical error is 0.4%, larger for single concepts (2.7% on average) than for large groups (0.2%), and a few pairs miss completely. This is a check that the group value is a real change in the network, and no more.

Table 7: Agreement between the predicted and observed logit change over 5,200 image–group pairs at each step size. "nonnegative" is the fraction of pairs whose observed change has the predicted sign.

ε	med. $\varepsilon/\|\mathbf{z}\|$	Spearman ρ	med. error	mean error	95th error	nonnegative
0.10	2.04×10^{-4}	.999817	.00364	.01696	.05501	.99827
0.25	5.11×10^{-4}	.999453	.00634	.02740	.09889	.99519
0.50	1.02×10^{-3}	.998887	.00975	.03854	.14457	.99250
1.00	2.04×10^{-3}	.997814	.01426	.05226	.21199	.98885
2.00	4.09×10^{-3}	.995889	.02087	.06809	.30316	.98423

F ImageNet: evidence behind the main text

What to compare the coverage with. A vocabulary of 15 random directions in the 200,704-dimensional activation space would cover 15/200,704 of the gradient, about 0.0075%; the fitted vocabulary covers 0.142%, nineteen times more. That comparison flatters, because real activations occupy a small part of that space. Two fairer references: random 15-dimensional subspaces drawn inside the principal subspace of the probe activations cover 0.063%, and probes refit on the same pictures with shuffled labels cover 0.028%. The fitted vocabulary covers 2.2 and 5.0 times those, and beats the shuffled-label maximum on 96% of images. The honest statement is that the vocabulary captures a few times what an equally sized random piece of the activation space does, and that this is still a small part of the model.

How the concepts were fitted. Each texture concept is a linear probe of 120 DTD pictures of that texture against one shared pool of 120 other pictures; each colour concept uses 120 generated colour patches against the same pool. None of these pictures is among the 200 target images. Sharing one negative pool makes the probes more alike, which is part of the overlap the allocation corrects; the negative-pool control below tests what happens without it.

Refitting the concepts. Fitting each concept again on a resampled set of its 120 pictures, five times, gives directions that agree with each other only to a cosine of 0.85 on average, and 0.46 for the worst concept. That noise reaches the shares: rankings from two refits agree on the top concept on 67% of images, and only 45% of images keep one top concept across all five. So a single ranked explanation should be read with its spread across refits beside it; the released outputs carry that spread for every image. Coverage barely moves across refits (0.118% to 0.136%), so coverage is a stable conclusion while individual top concepts are not.

The one-at-a-time score is just as unstable on the same refits: it keeps one top concept across all five on 39% of images against SPACE's 45%, and its rankings correlate across refits at 0.78 against 0.76. Neither rule is steadier; the instability comes from the fitted directions, not from the rule.

Credit without a derivative of its own. A concept whose own directional derivative is near zero can still earn credit by completing a span with others (Proposition 2). Concepts with a derivative below 5% of the largest in their image hold 2.5% of the total on average and 31% on the worst image. At the top of the ranking, where a reader looks, this is rare: on 3 of 200 images the concept with the largest share has a derivative below a quarter of the largest. In the middle of the ranking it is common.

Keeping the top three. A ranking is used by keeping its top few entries. For each image and each rule we take the three concepts ranked first, and score them by how much of the vocabulary's captured sensitivity those three explain together, as a fraction of what all fifteen explain. The best possible three, found by checking every triple, give the ceiling; a random three give the floor. Table 8 averages over the 200 images. With one concept kept, the one-at-a-time score is best by definition, since a single concept's value is its own derivative. From two concepts on the allocation overtakes it, and at five the allocation's pick explains 85% of what the vocabulary can, against 75%, within five points of the ceiling. The squared allocation picks slightly better still, so the gain comes from charging for overlap, not from the square root. Moving the network's representation inside each pick and reading the logit confirms the criterion: the logit moves 12% further inside the allocation's three than inside the one-at-a-time three, further on 57% of images and less far on 10%. Refitting every concept five times keeps the gap between 5 and 8 points with an image-level standard error below one point each time, and the allocation wins the paired comparison on 47 to 60% of images against 10 to 15%. Per class the gap runs from 5 points (basketball) to 11 (goldfish).

Table 8: Keeping the top m concepts by each rule on the 200 ImageNet images: fraction of the vocabulary's captured sensitivity the kept concepts explain together, the mean number of concepts needed to reach 90%, and the observed logit change inside the kept span relative to one-at-a-time scoring.

	Captured fraction $v(S_m)/v([K])$, mean over 200 images					Observed logit gain	
Selection rule	$m=1$	$m=2$	$m=3$	$m=5$	m_{90}	$m=3$	$m=5$
Best subset (exact)	54%	72%	80%	90%	5.2	+19%	+21%
Greedy forward	54%	71%	79%	89%	5.6	+17%	+19%
Shapley on v^2	49%	66%	77%	87%	6.0	+13%	+17%
SPACE	50%	66%	76%	85%	6.8	+12%	+14%
Normalised $\lvert a_k \rvert$	54%	63%	67%	75%	10.3	+0%	+0%
Random order	21%	34%	44%	58%	11.6	−38%	−25%

Controls and other rankings. Six checks on the top-three gap, all on the same images. *Separate negative pools.* Refitting all fifteen concepts with negatives taken from the 37 DTD textures that are not concepts, once as one shared pool and once as fifteen disjoint pools, makes the probes less alike (mean cosine 0.19 shared, 0.10 disjoint). The allocation's gap over one-at-a-time scoring is 3.3 points with the shared texture pool and 2.5 with disjoint pools, against 8.3 on the main vocabulary, each several standard errors from zero. The gap persists without a shared pool and shrinks as the overlap it charges for shrinks. *ICS.* The closest local competitor (Schrouff et al., 2021) integrates the gradient along a path from a blank reference and ranks by the result; its top three explain 43%, the same as a random three. The path average is not the sensitivity at the image. This is a comparison on our criterion, not a verdict on what ICS is for. *Least-squares coefficients.* Ranking by the size of each concept's regression coefficient explains 71% at three concepts and reaches 90% in 6.2 concepts, against 6.8 for the allocation. *Orthogonalising first.* Replacing the fifteen directions by the closest mutually perpendicular set with the same span, then scoring one at a time, explains 79%, the same as greedy selection. The price is that each new direction is a blend of all fifteen concepts, so the shortlist no longer names anything a user asked about; greedy on the original directions reaches the same value with the names intact. *Grouping the colours.* Treating the five colours as one concept and keeping three concepts, the allocation explains 85% against 82% for one-at-a-time scoring: grouping narrows the gap from 8 points to 3 and does not close it. *A different metric.* Rescaling every coordinate by its spread across the probe pictures and recomputing everything (Proposition 3) gives 70% for the allocation against 59% for one-at-a-time scoring and 76% for greedy, so the gap widens rather than closes.

A larger vocabulary. Every DTD texture plus the five colours gives 52 concepts on the same 200 images, fitted against the same negative pool. Checking every group is out of reach at this size, so the allocation is estimated from 500 random orderings; two independent estimates differ by 8% and agree on the top three concepts on 69% of images, and one estimate takes 12 seconds per image against 0.35 seconds for greedy selection. Coverage rises to 0.33%, still small. The gap between the allocation and one-at-a-time scoring is 6 points at three concepts and 15 at ten, and reaching 90% takes 22 concepts against 37, while greedy needs 16. The rank certificate held on every image. At this size the sampled allocation is both noisier and slower than selection, which is the regime the paper's guidance is for: select by the group value, allocate when credit is the question.

One image end to end. Figure 1 shows one zebra image. Scored one at a time, the fifteen concepts' absolute scores add to 0.0411; the shares add to 0.0149, the most the vocabulary can explain, which is 0.13% of the gradient's squared length. `striped` is first under both rules, with 34.7% of the shares against 23.1% of the one-at-a-time sum. Kept three at a time, the one-at-a-time pick (striped, spiralled, cobwebbed) explains 64% of what the vocabulary can, the allocation's pick (striped, bubbly, spiralled) 77%, and greedy's (striped, bubbly, zigzagged) 82%; the second and third one-at-a-time picks point where striped already points and add nothing.