Multi-dimensional concept discovery in multimodal encoders

Michał Puścian

2026

Warsaw University of Technology
Faculty of Mathematics and Information Science

Outline

Motivation

What are concepts in c-XAI?

MCD: Multi-dimensional Concept Discovery

My method

Experimental setup

Results

Examples

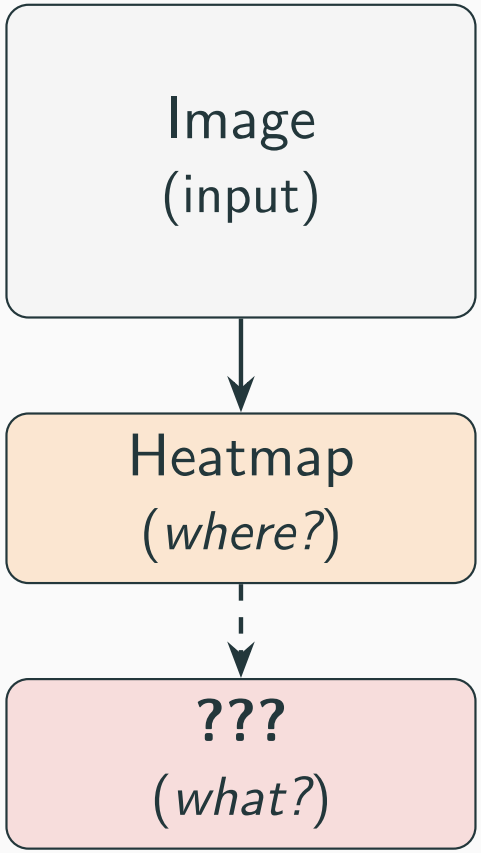
Current challenges

Motivation

The problem with feature attribution

Feature attribution methods (LIME, SHAP, GradCAM, . . .) answer **where** the model looks – but not **what** it sees.

- A heatmap highlighting a mouth does not distinguish between "shape of the smile" and "whiteness of teeth"
- User must construct a *subjective narrative* to bridge pixels → semantics
- Prone to confirmation bias



⇒ **Concept-based explanations** address this semantic gap.

My contribution

Concept-based XAI aligns model representations with human-understandable concepts.

Gaps in the literature

1. Most c-XAI methods model concepts as single directions.
   Some concepts are inherently multi-dimensional (e.g. "car" $\supset$ SUV, sports car, ambulance).
2. Concept discovery has been studied mostly for unimodal CNNs.
   Multimodal vision-language encoders (CLIP, SigLIP) are increasingly deployed but lack concept-based interpretability tools.

My goal: Extend unsupervised multi-dimensional concept discovery to dual-encoder multimodal architectures.

What are concepts in c-XAI?

Unimodal concept-based methods

Method	Discovery	Approach	Dim
TCAV (Kim, 2018)	Supervised	Linear classifier	1D
ACE (Ghorbani, 2019)	Unsupervised	Segment + K-Means	1D
ICE (Zhang, 2021)	Unsupervised	NMF	1D
CRAFT (Fel, 2023)	Unsupervised	Recursive NMF	1D
MCD (Vielhaben, 2023)	Unsupervised	SSC + PCA	***n*D**

Fel et al. (2023) show that ACE, ICE, CRAFT are all special cases of **dictionary learning** – each concept is a single direction.

MCD breaks this pattern: concepts are **linear subspaces of arbitrary dimensionality**.

Limitation: All of the above are designed for unimodal CNNs.

TCAV: supervised concept discovery (Kim et al. 2018)

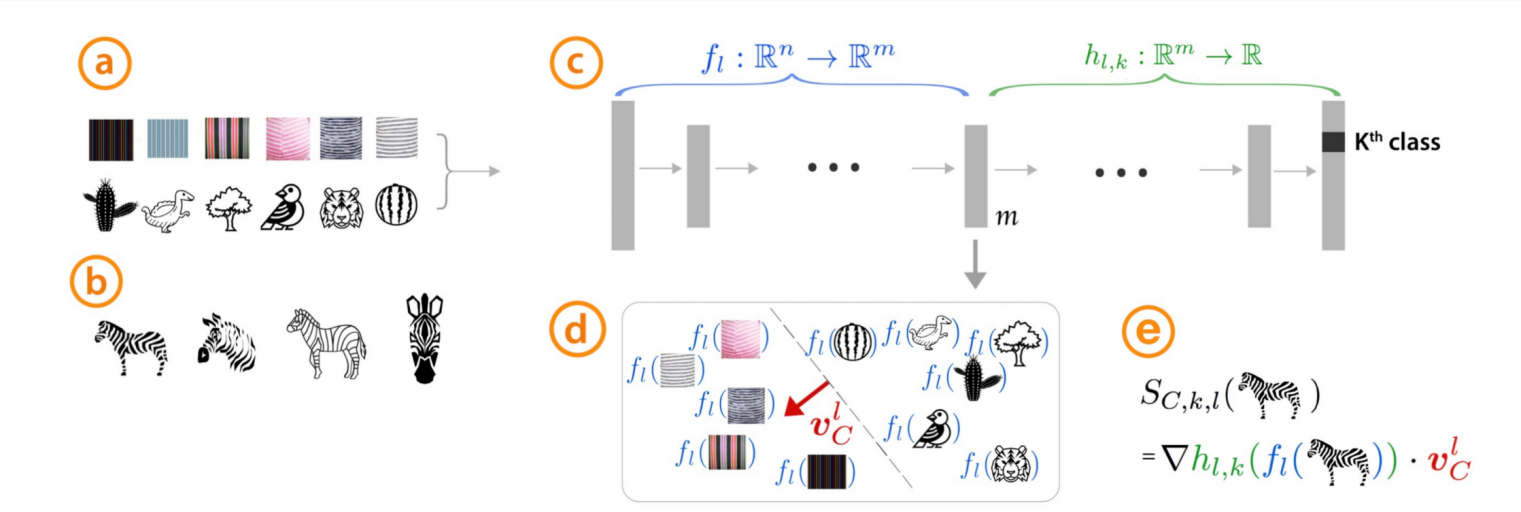


Figure 1: TCAV overview. User-defined concept examples (a) and class examples (b) are passed through a network (c). A linear classifier finds the concept direction v_C^l (d), and conceptual sensitivity is measured via directional derivatives (e). Source: Kim et al. (2018)

Unsupervised concept discovery via clustering (Fel et al. 2023)

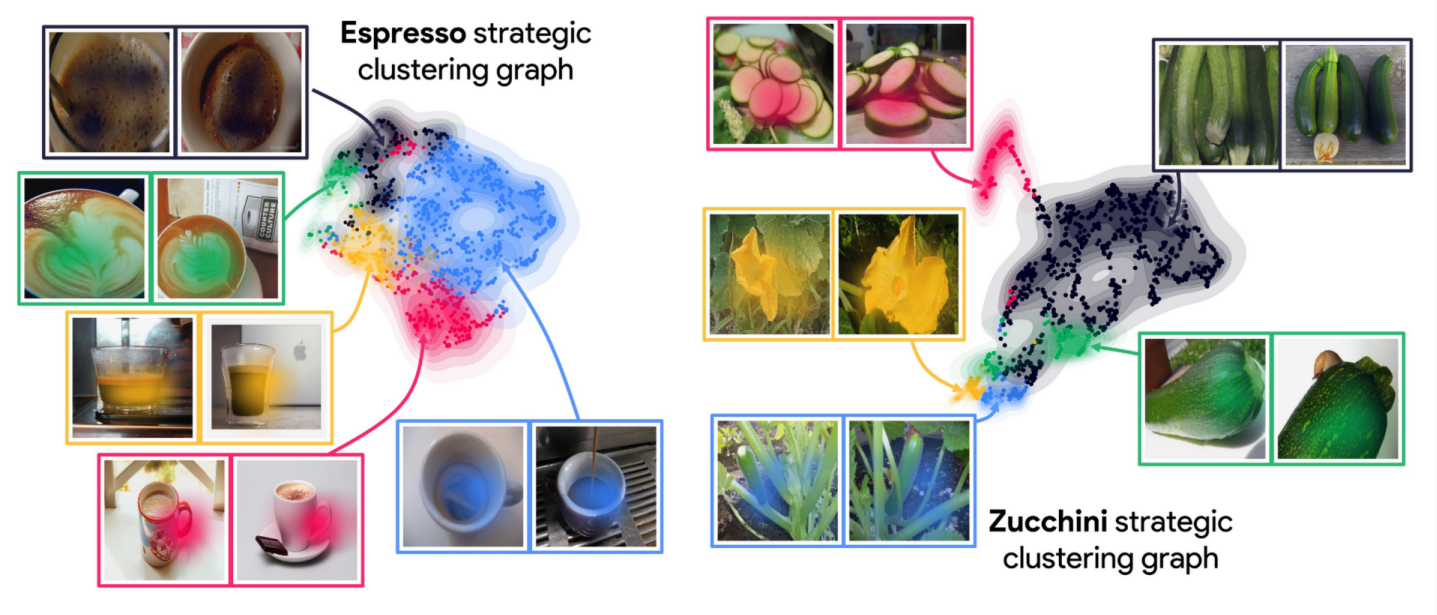


Figure 2: Concepts discovered by clustering activation vectors. Each color represents a concept cluster – e.g. for "espresso": latte art, black coffee, brewing equipment, cups. Source: Fel et al. (2023)

Decomposition types in the literature

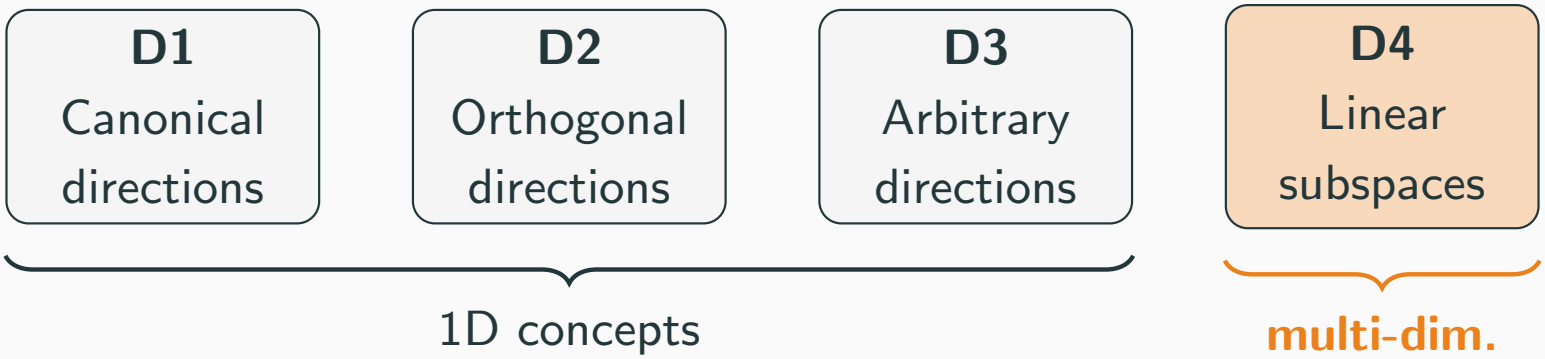


Figure 3: Decomposition categories (Vielhaben et al. 2023)

Most existing methods fall into D1–D3. MCD operates in **D4**: concepts as **linear subspaces of arbitrary dimensionality**.

MCD: Multi-dimensional Concept Discovery

Visualization of MCD

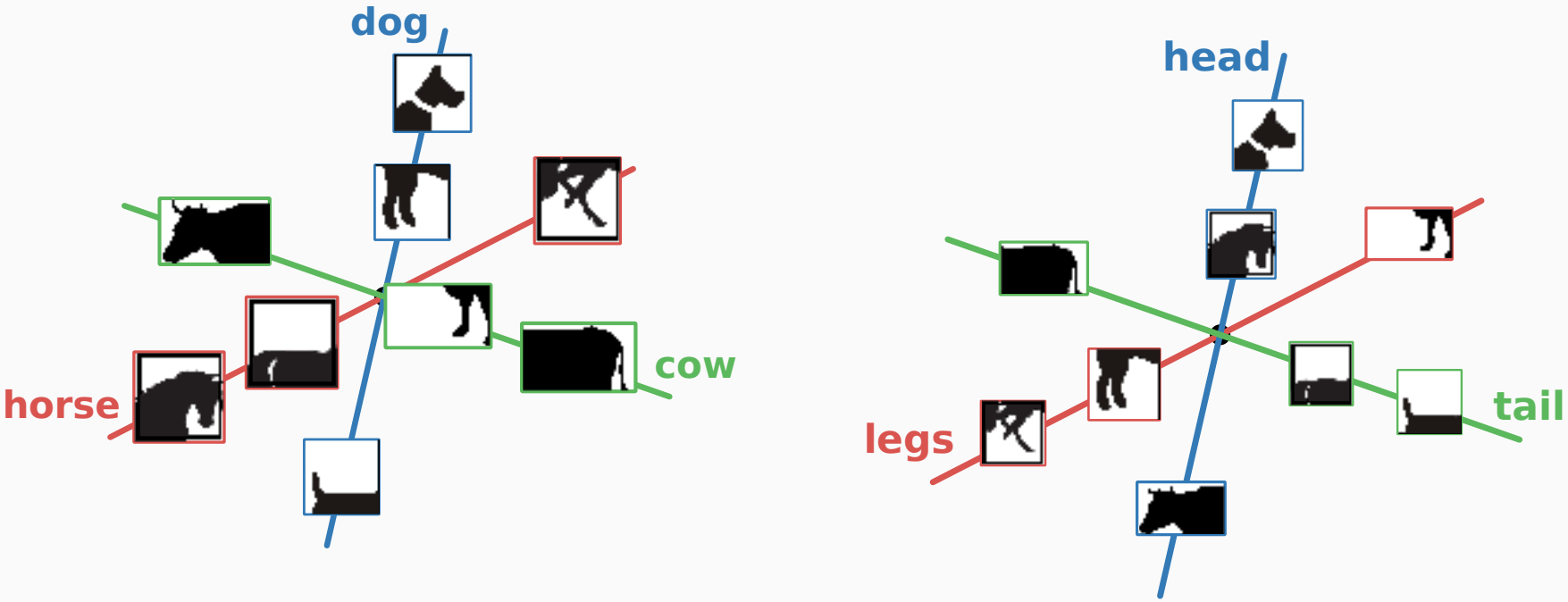
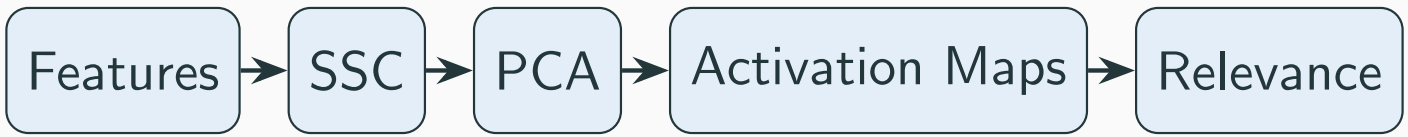


Figure 4: Two paradigms for visual concept discovery.

MCD method overview (Vielhaben et al. 2023)

1. **Feature extraction** – Obtain activation maps from a chosen layer
2. **Sparse Subspace Clustering (SSC)** – Cluster feature vectors into groups that lie on common subspaces
3. **PCA per cluster** – Extract orthonormal basis vectors for each concept subspace
4. **Explanation** – Decompose activations and classifier weights into:
   - Concept Activation Maps (where does concept appear?)
   - Concept Relevance Heatmaps (how relevant is each concept?)
   - Global Relevance Scores (which concepts matter most?)



Sparse Subspace Clustering

Key idea: Data points lying on a union of subspaces can be clustered by finding a sparse self-representation.

Given data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, solve:

$$\min_{\mathbf{C}} \|\mathbf{C}\|_1 \quad \text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{C}, \quad \operatorname{diag}(\mathbf{C}) = 0$$

- Each column $\mathbf{c}_i$ encodes how $\mathbf{x}_i$ can be expressed as a sparse combination of other points
- Points from the *same subspace* receive non-zero coefficients
- Affinity matrix $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}|^\top$ fed to spectral clustering

After clustering: **PCA within each cluster** $\rightarrow$ orthonormal basis $\{\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_{d_k}^{(k)}\}$ for concept k.

Concept Activation Maps

Given concept subspace bases $\{\mathbf{U}^{(k)}\}_{k=1}^{K}$ and a patch embedding $\mathbf{h}_i^v \in \mathbb{R}^d$:

Concept Activation Score for patch i and concept k:

$$a_k(\mathbf{h}_i^v) = \|\mathbf{U}^{(k)\top}\mathbf{h}_i^v\|_2^2 = \|\phi^k\|_2^2$$

where $\phi^k = \mathbf{U}^{(k)}\mathbf{U}^{(k)\top}\mathbf{h}_i^v$ is the projection of $\mathbf{h}_i^v$ onto subspace k.

- MCD assumes subspaces are **mutually orthogonal** $\Rightarrow$ unique decomposition $\mathbf{h}_i^v = \sum_{k=1}^{K} \phi^k + \phi^{\perp}$
- Apply to all 196 patches $\rightarrow$ spatial map over the image

Note

If subspaces overlap, then decomposition is no longer unique.

Concept Relevance

Setup: Train a linear probe with weights $\mathbf{w}$ on frozen SigLIP embeddings for a downstream classification task.

Decompose the prediction into per-concept contributions:

$$f(\mathbf{h}) = \mathbf{w}^\top \mathbf{h} = \sum_{k=1}^{K} \underbrace{\mathbf{w}^\top \mathbf{U}^{(k)} \mathbf{U}^{(k)\top} \mathbf{h}}_{r_k(\mathbf{h})}$$

- $r_k(\mathbf{h})$ is the **relevance of concept** k to the prediction for input $\mathbf{h}$
- Decomposition is exact when concept subspaces are orthogonal and span $\mathbb{R}^d$
- Apply per-patch $\rightarrow$ **concept relevance heatmap** (where does concept k contribute to the decision?)

Global Relevance Scores

Aggregate per-patch relevance scores across an image or dataset to rank concepts globally:

Per-image global relevance of concept k:

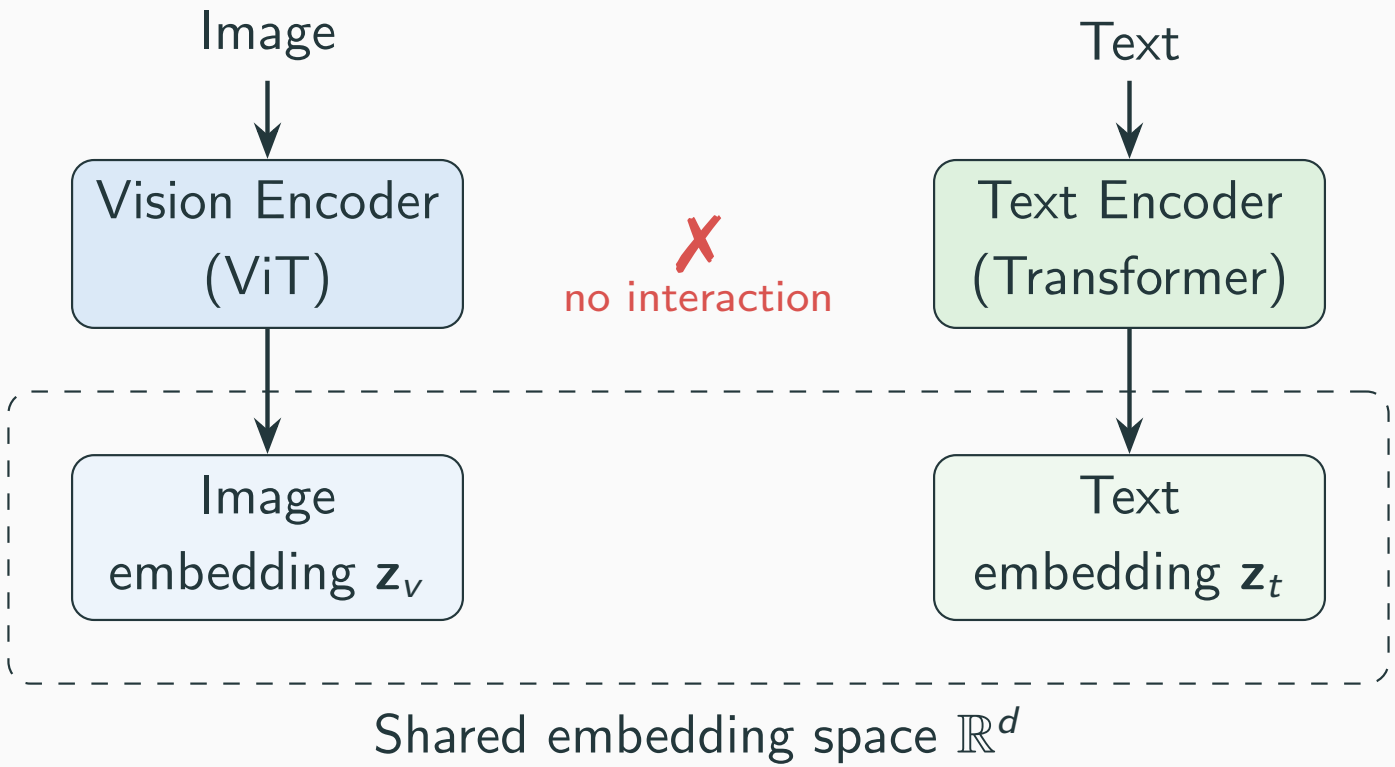
$$R_k = \sum_{i=1}^{196} r_k(\mathbf{h}_i^v)$$

- Rank concepts by $|R_k|$ to find which concepts drive the prediction most
- Average over a class $\rightarrow$ **class-level concept importance**
- Enables compact explanations: "this image is classified as *bird* mainly because of concepts #3 (wing shape) and #7 (beak texture)"

My method

SigLIP architecture

We use SigLIP model:



Note: Modalities do not interact during encoding so we perform concept discovery separately per encoder and study concepts post-hoc.

Feature extraction

Vision encoder (SigLIP ViT):

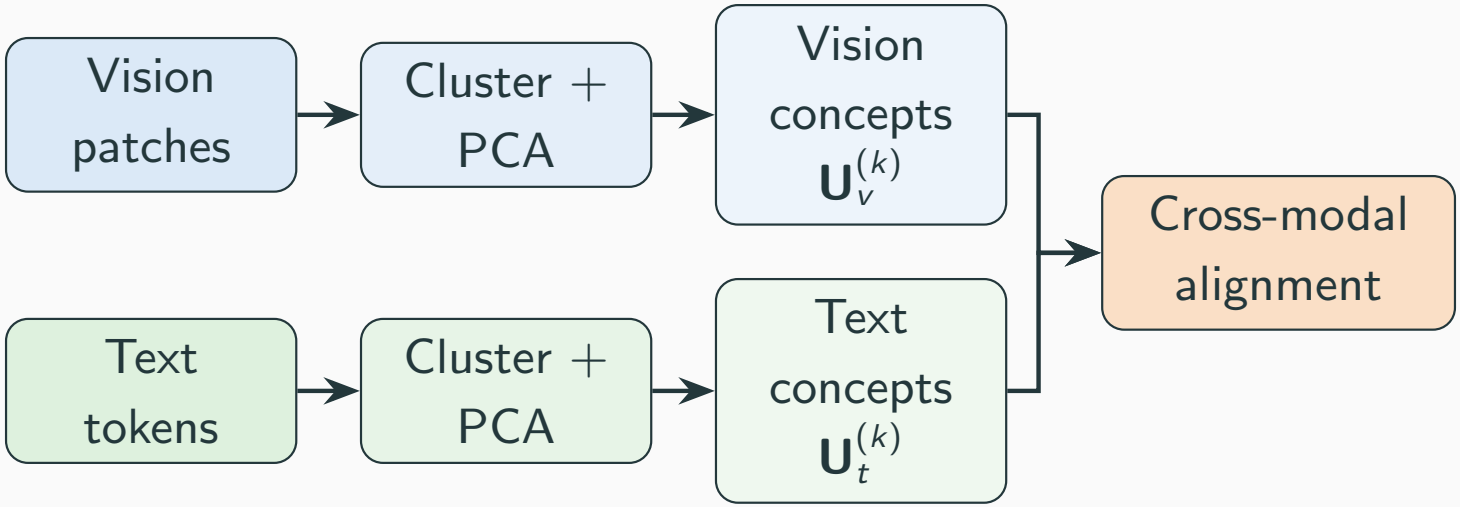
- Image $\rightarrow$ sequence of **196 patch tokens**
- Final image embedding = **mean pool** over all patch tokens
- We use the **individual patch embeddings** $\mathbf{h}_i^v \in \mathbb{R}^d$ for concept discovery

Text encoder (SigLIP):

- Caption $\rightarrow$ sequence of token embeddings
- Final text embedding = **EOS token** embedding $\mathbf{h}_{\text{EOS}}^t \in \mathbb{R}^d$
- We extract **token-level embeddings** for concept discovery

Pooling over dataset
Collect samples of patch (or token) embeddings across images (or captions) in a dataset and perform clustering.

Adapting MCD to dual encoders



Pipeline applied **independently** to each modality, then concept subspaces compared in the shared embedding space.

The modality gap

CLIP/SigLIP embeddings of images and text occupy **separate regions** of $\mathbb{R}^d$.

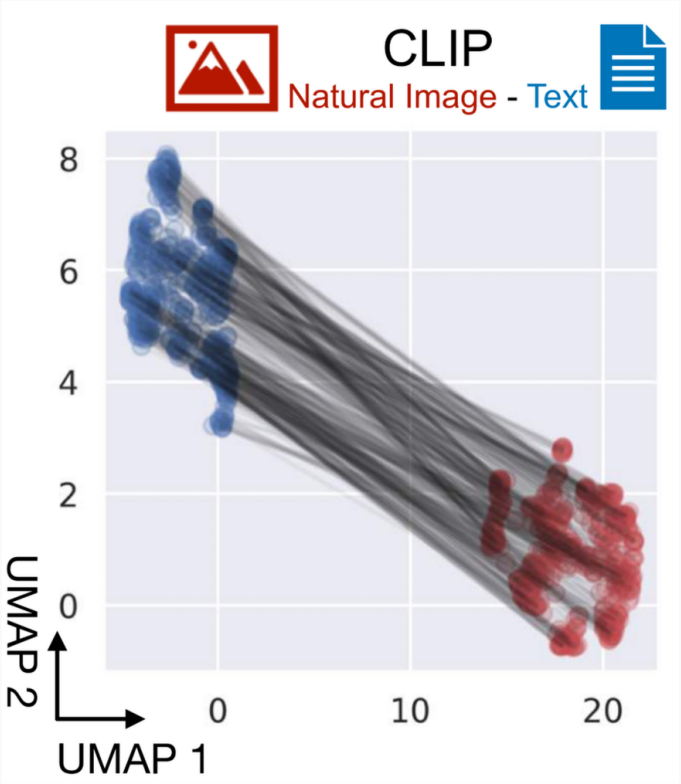


Figure 5: UMAP of CLIP embeddings – image (red) and text (blue) clusters are clearly separated. Source: Liang et al. (2022)

Embedding geometry before and after attention-sink removal

Metric	Before	After
Image patches		
Mean pairwise cosine similarity	0.274	0.007
Half-angle (°)	74.1	89.6
Effective rank	348.6	371.9
Uniformity	−2.78	−3.89
Text tokens		
Mean pairwise cosine similarity	0.625	0.000
Half-angle (°)	51.3	90.0
Effective rank	196.2	226.7
Uniformity	−1.47	−3.91
Cross-modal		
Mean cross-modal cosine similarity	0.028	0.000
Gap distance ($\|\mu_{\text{img}} - \mu_{\text{txt}}\|_2$)	0.918	0.085

Table 1: All embeddings are L2-normalised. Top-2 principal directions removed.

Procrustes superimposition

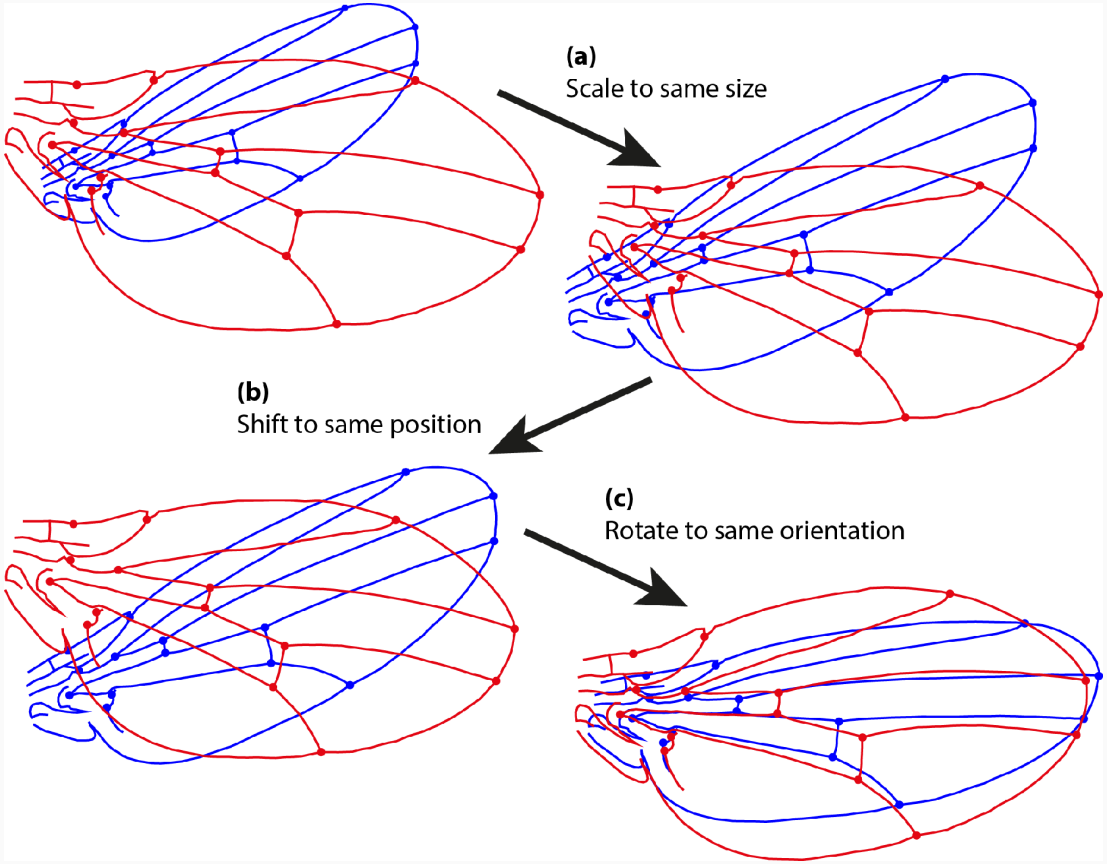


Figure 6: Three transformation steps of an ordinary Procrustes fit: (a) scaling both configurations to the same size; (b) translation to the same centre of gravity; (c) rotation to minimise the sum of squared distances between corresponding landmarks. Source: Klingenberg (2015)

Embedding geometry before and after attention-sink removal

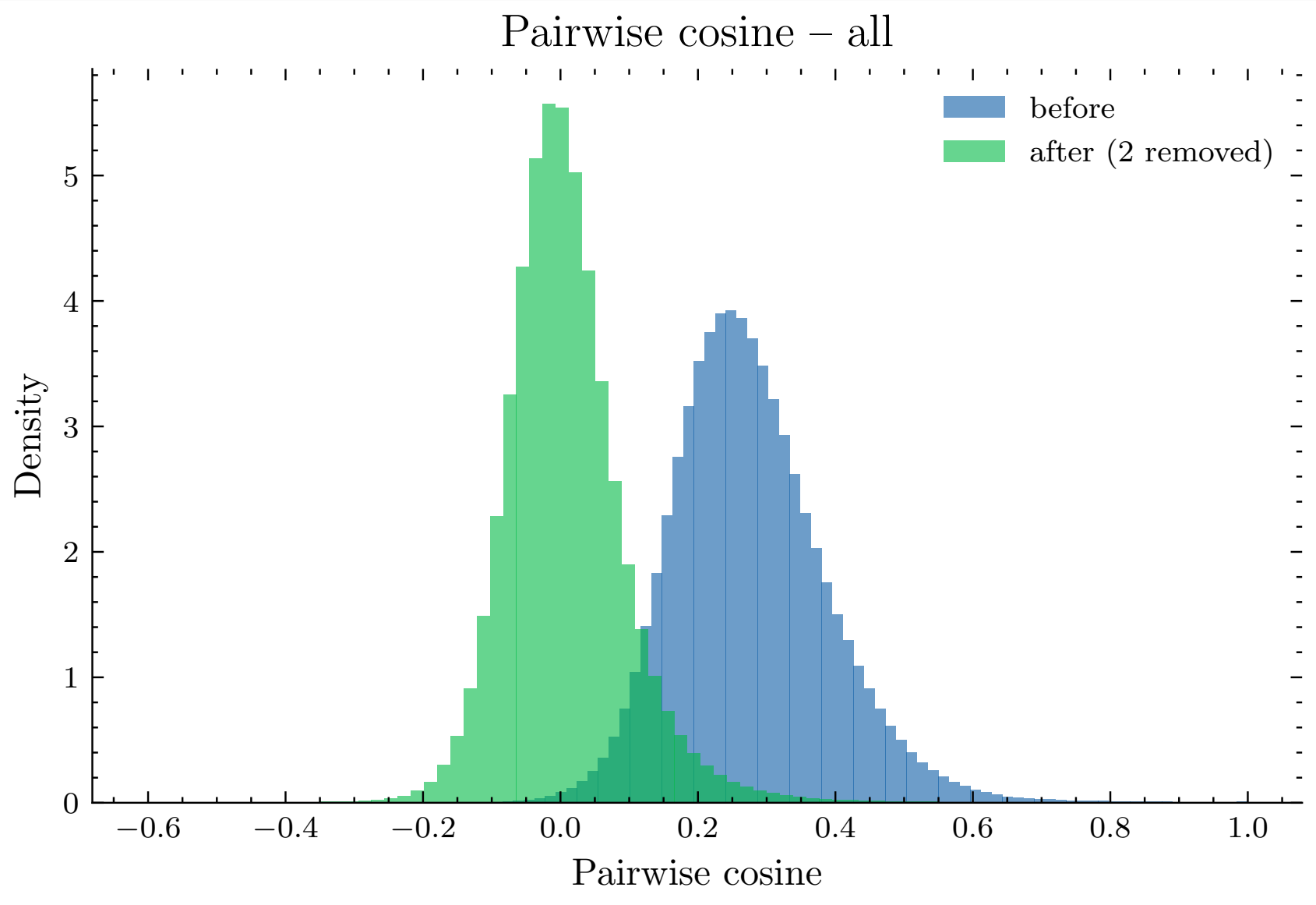


Figure 7: Histogram of cosine similarities between modalities.

Procrustes superimposition

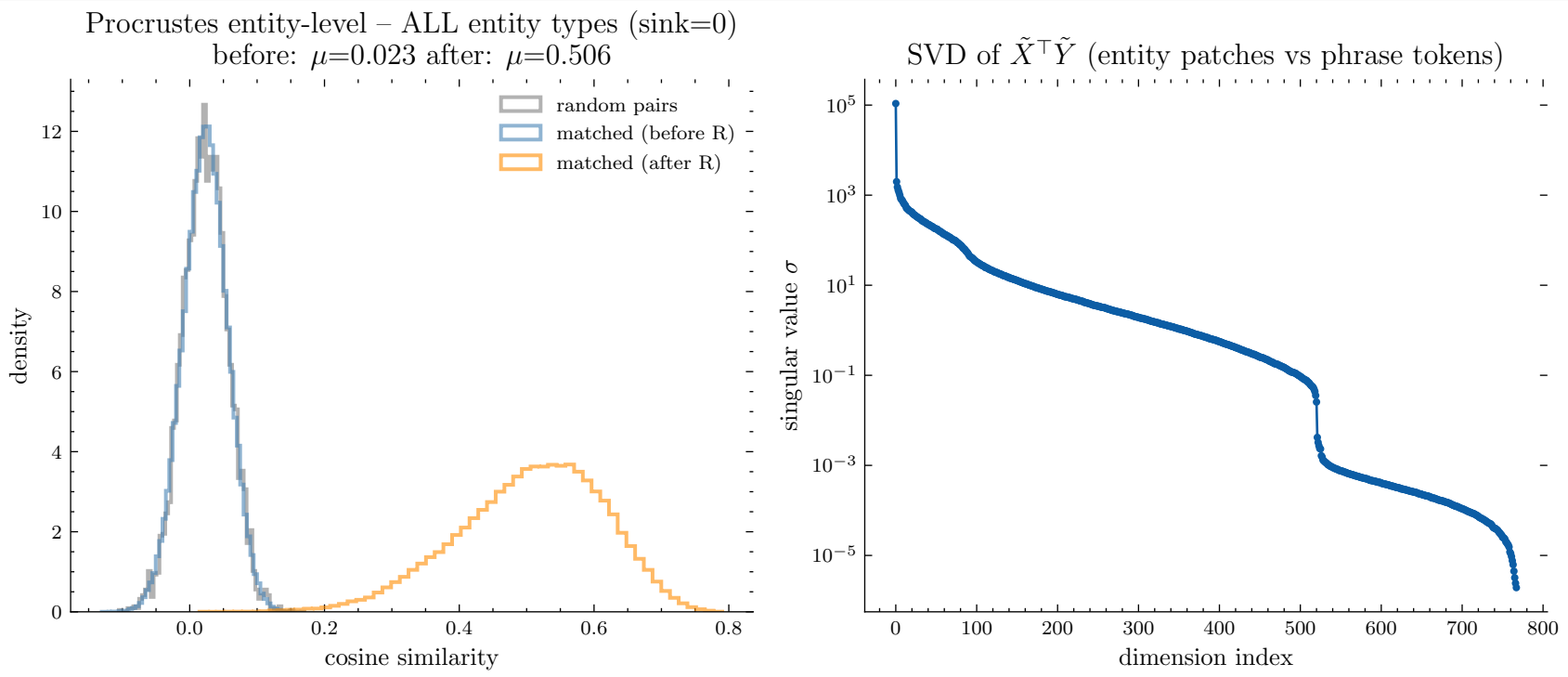


Figure 8: One singular value strongly dominates — the modality gap collapses the shared space onto a single direction.

Procrustes superimposition

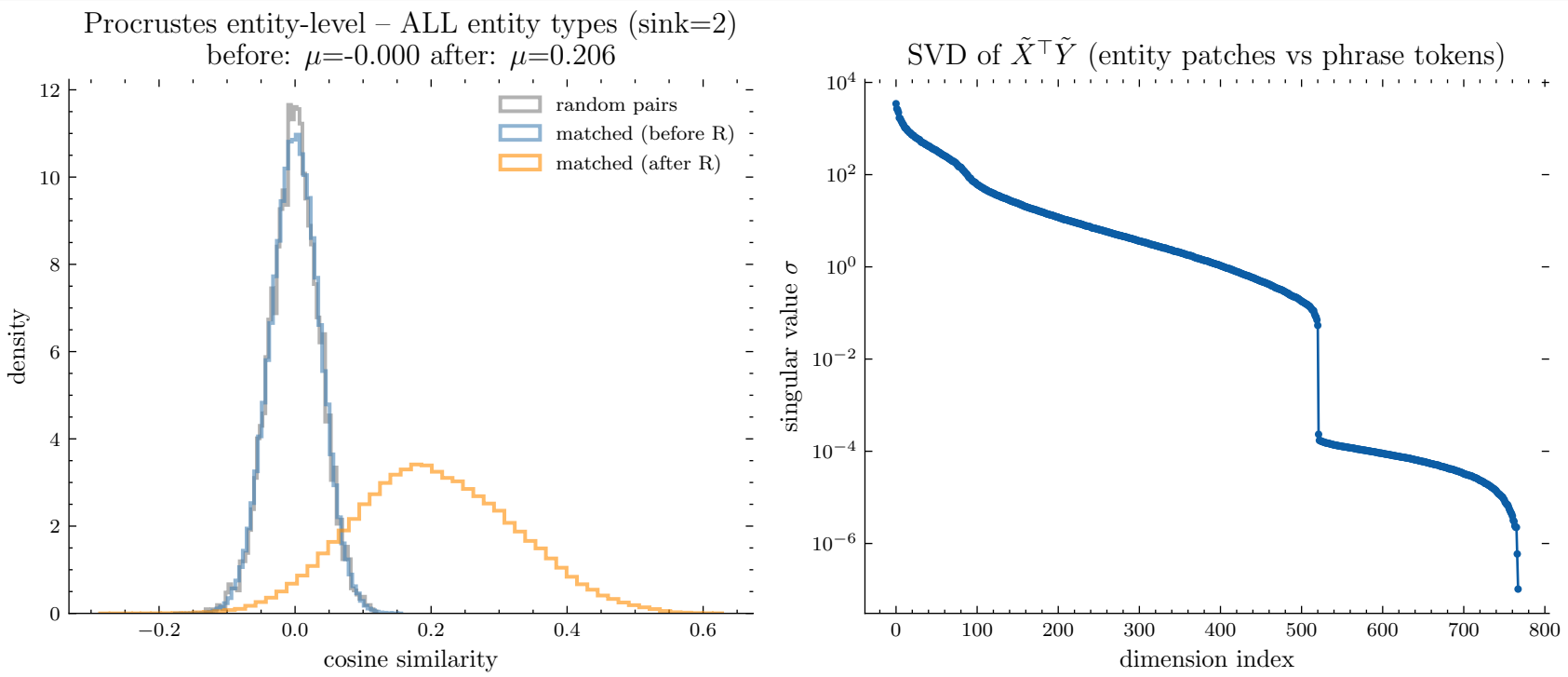


Figure 9: After removing 2 directions.

Experimental setup

Experimental setup

Dataset: Flickr30k Entities
7 entity types: people, animals, clothing,
vehicles, bodyparts, scene, other

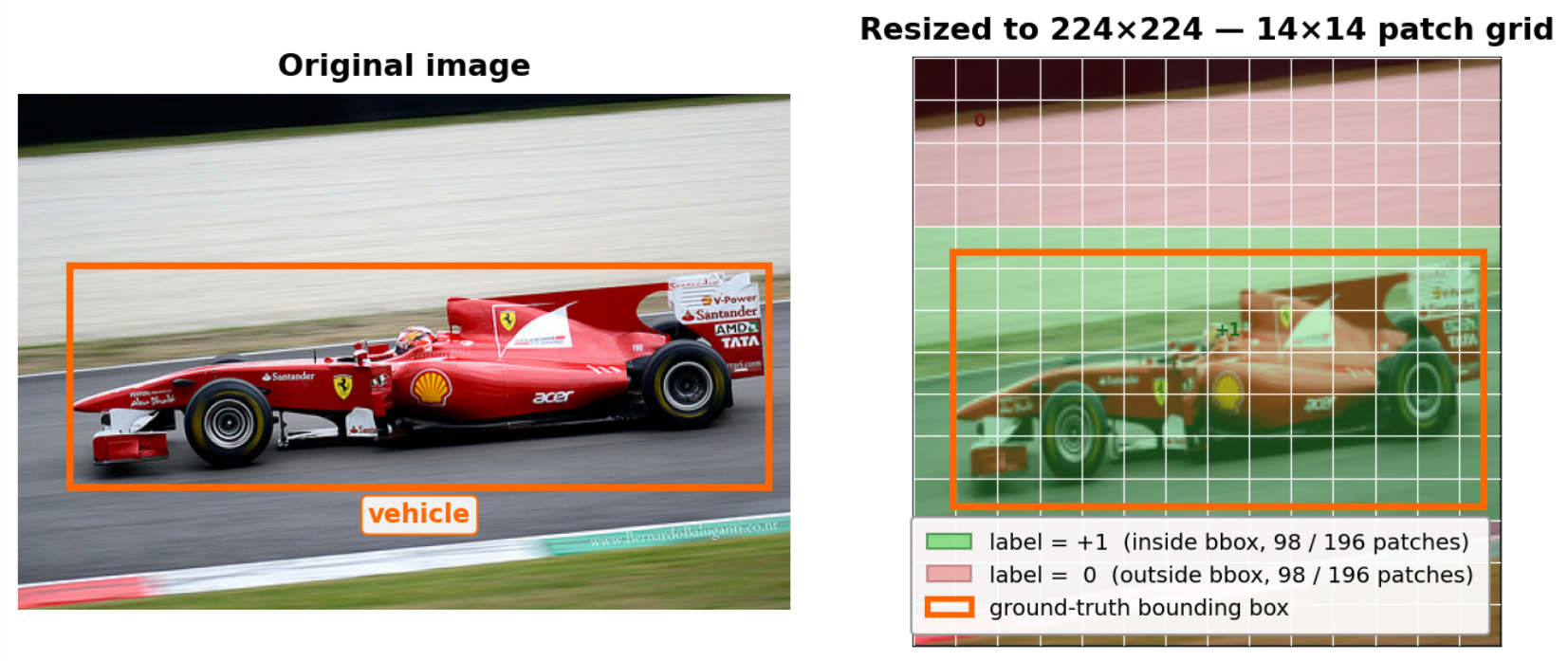
Model: SigLIP ViT-B/16
196 patches $\times$ 768 dims per image

Concept discovery (K=6):
SSC on 5 000 subsampled patches
PCA per cluster + complement basis

Probe: Logistic regression
Labels: per-patch bounding-box

Metrics:
F1, AUROC — grounding accuracy
Completeness — bases span the space?
η — fraction of logit from concepts
Principal angles — cross-modal alignment
Co-activation — image–text correlation

Entity grounding: patch-level labels from bounding boxes



We train a logistic regression probe to predict which patches belong to the entity. Patches overlapping the ground-truth bounding box receive label $+1$, the rest label 0.

Results

Grounding performance across entity types (K=6)

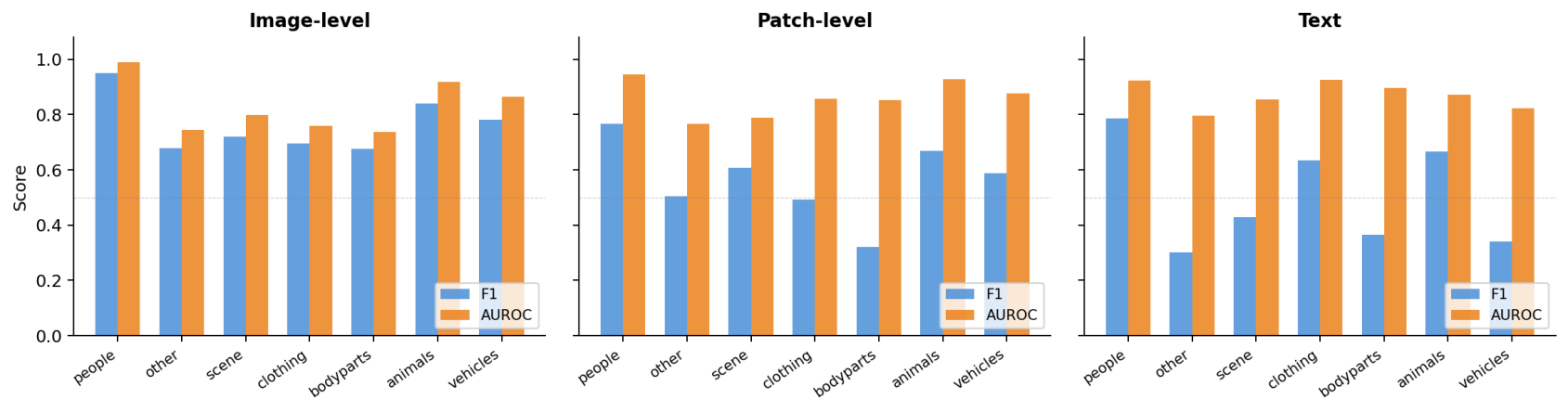


Image-level probes achieve the highest F1; patch-level and text probes are harder due to finer-grained labels and the modality gap.

Co-activation by entity type

Co-activation: Pearson r between image-concept activation and text-concept activation on matched (image, caption) pairs.

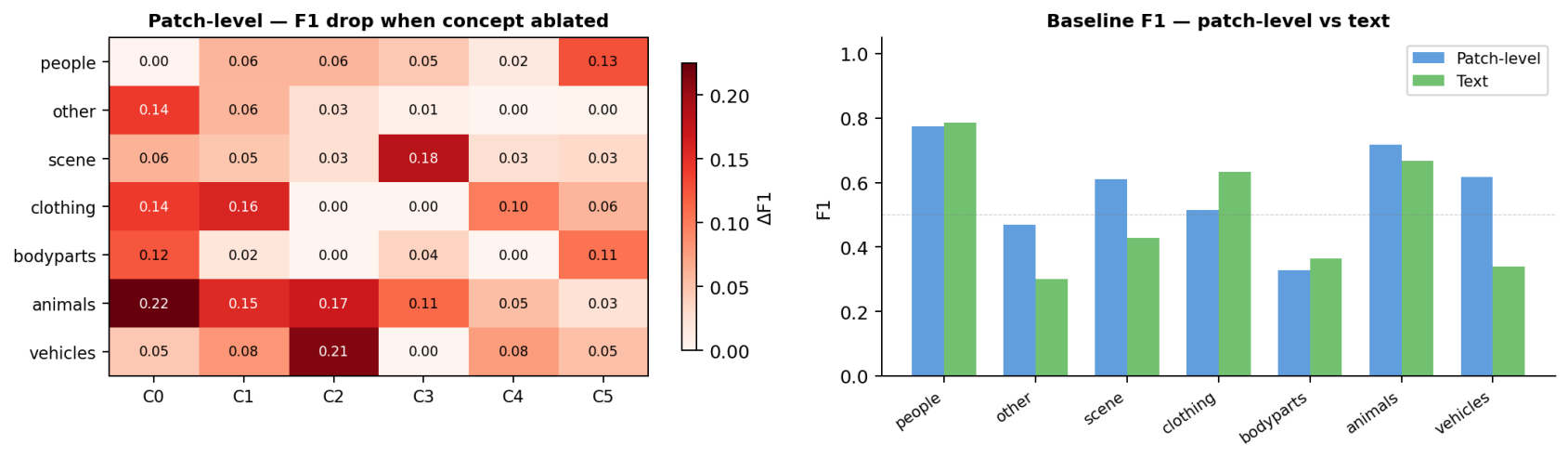
Aligned (positive):

- **Clothing +0.86** — "red jacket" maps cleanly across modalities
- People +0.36 — moderate alignment
- Other +0.27 — weak positive

Anti-correlated (negative):

- **Vehicles −0.47** — image: shape/texture; text: semantics
- Scene −0.35 — depth/texture vs. "outdoor"
- Animals −0.24 — high intra-class variation

Faithfulness ablation across entity types (K=6)



Each cell = F1 drop when that concept is removed. Concepts are not equally important — one dominant concept per entity.

Examples

How to read relevance CAMs

Each panel = one concept subspace $\mathbf{U}^{(k)}$ (or complement). **Title** = pooled logit contribution $\sum_i r_k(\mathbf{h}_i)$.

Image panels:

Red patch $\Rightarrow r_k > 0$ (pushes *toward* entity)

Blue patch $\Rightarrow r_k < 0$ (pushes *away*)

Text panels:

Same colour logic per token.

Comp = complement basis (null space of all K concept subspaces).

Large positive **Comp** $\Rightarrow$ probe fires on directions *outside* the K concepts.

All panels sum exactly to the probe logit (minus bias).

Vehicles (img 07, rel)

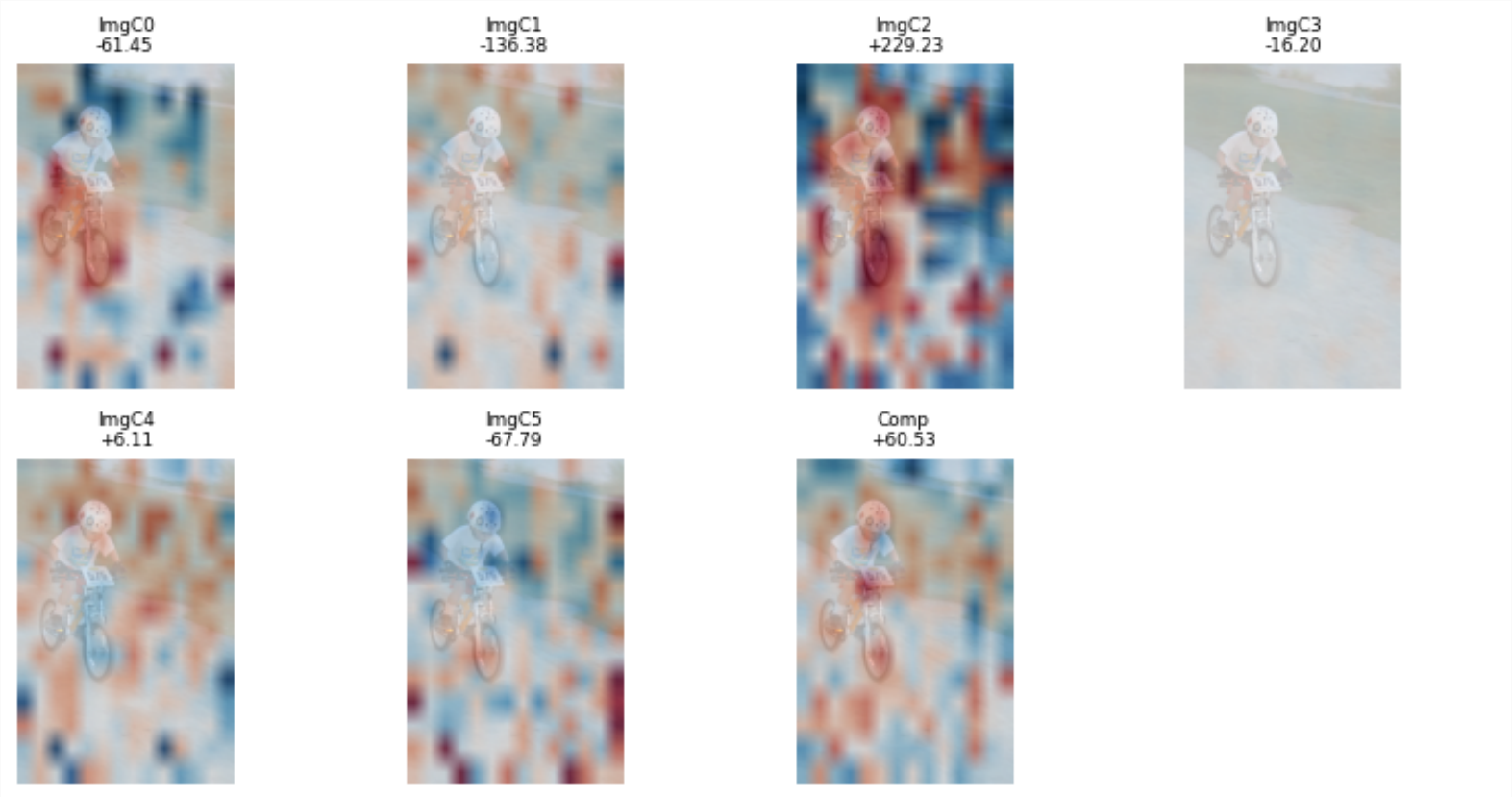


Figure 10: C2 (+229) dominates — patches covering the *rider's body* are red, not the bike frame. C3 is nearly blank (low intrinsic dimension, almost no signal). The probe learns "person doing a vehicle activity", not "vehicle shape".

Vehicles (img 06, raw)

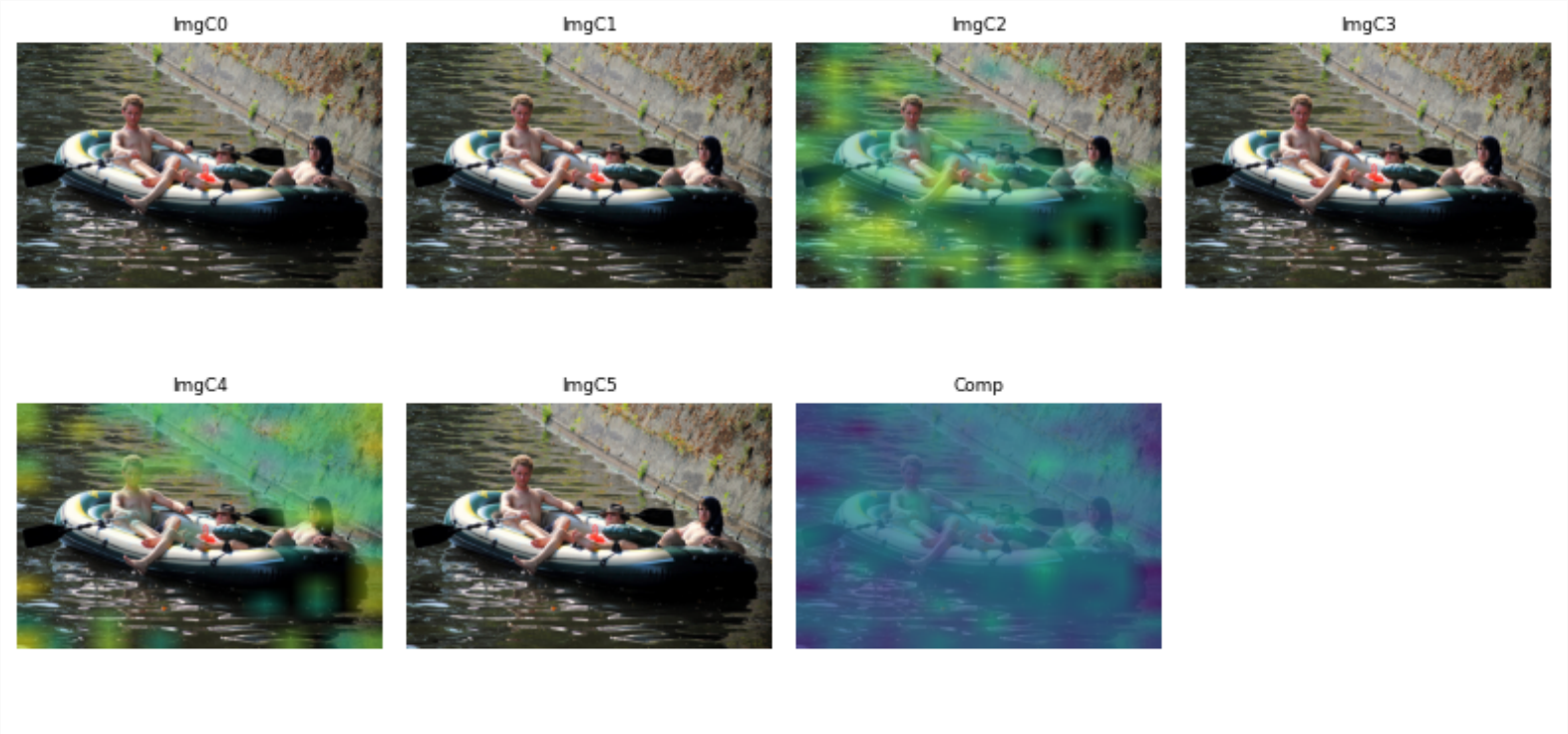


Figure 11: Two distinct concepts: C2 covering the boat, C4 covering the wall.

Vehicles (img 04, rel)

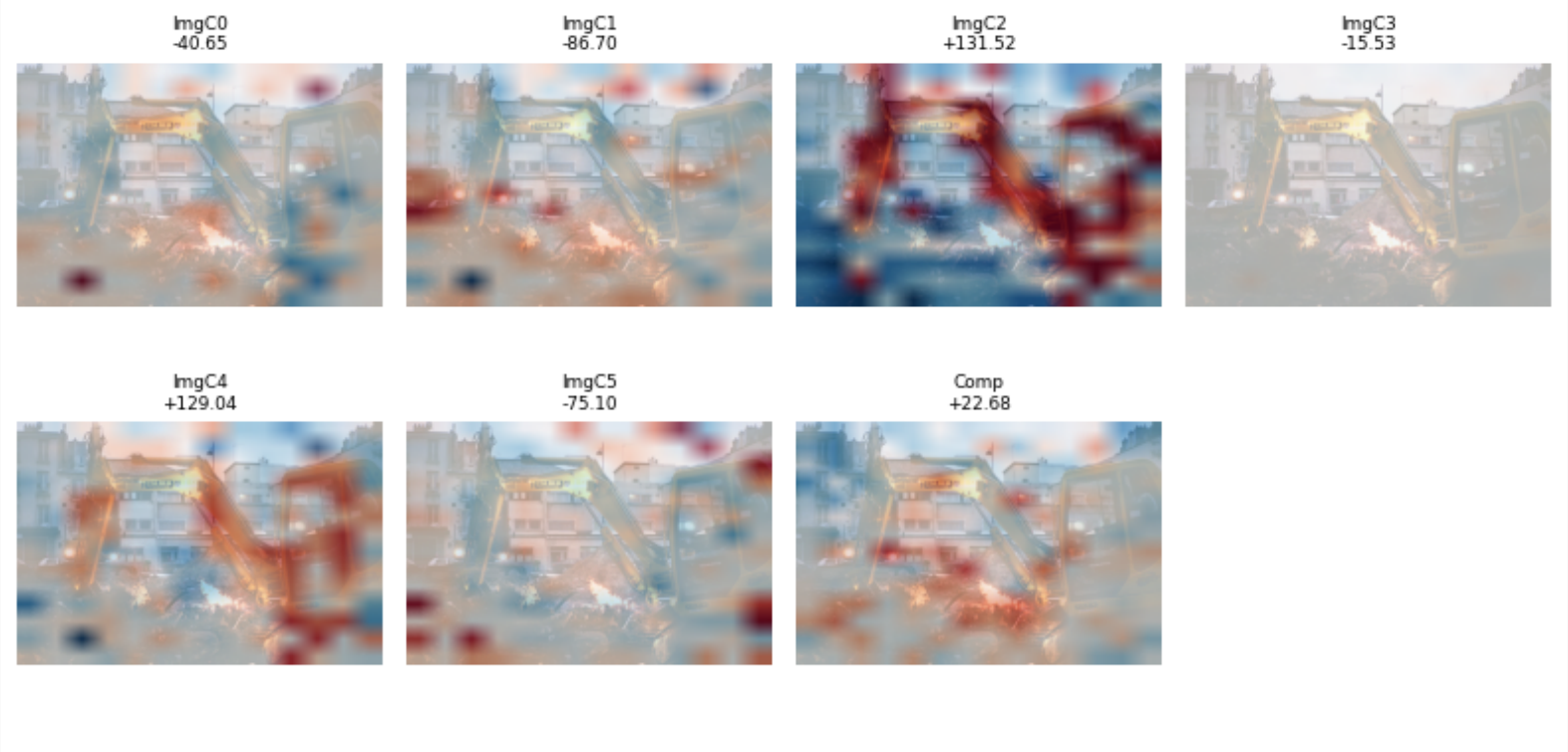


Figure 12: Excavator. **C2 (+132)** and **C4 (+129)** are the dominant positives; **C1 (−89)** and **C5 (−75)** suppress. Large opposite-sign contributions cancel without producing a coherent vehicle hotspot.

Vehicles (img 00, raw)

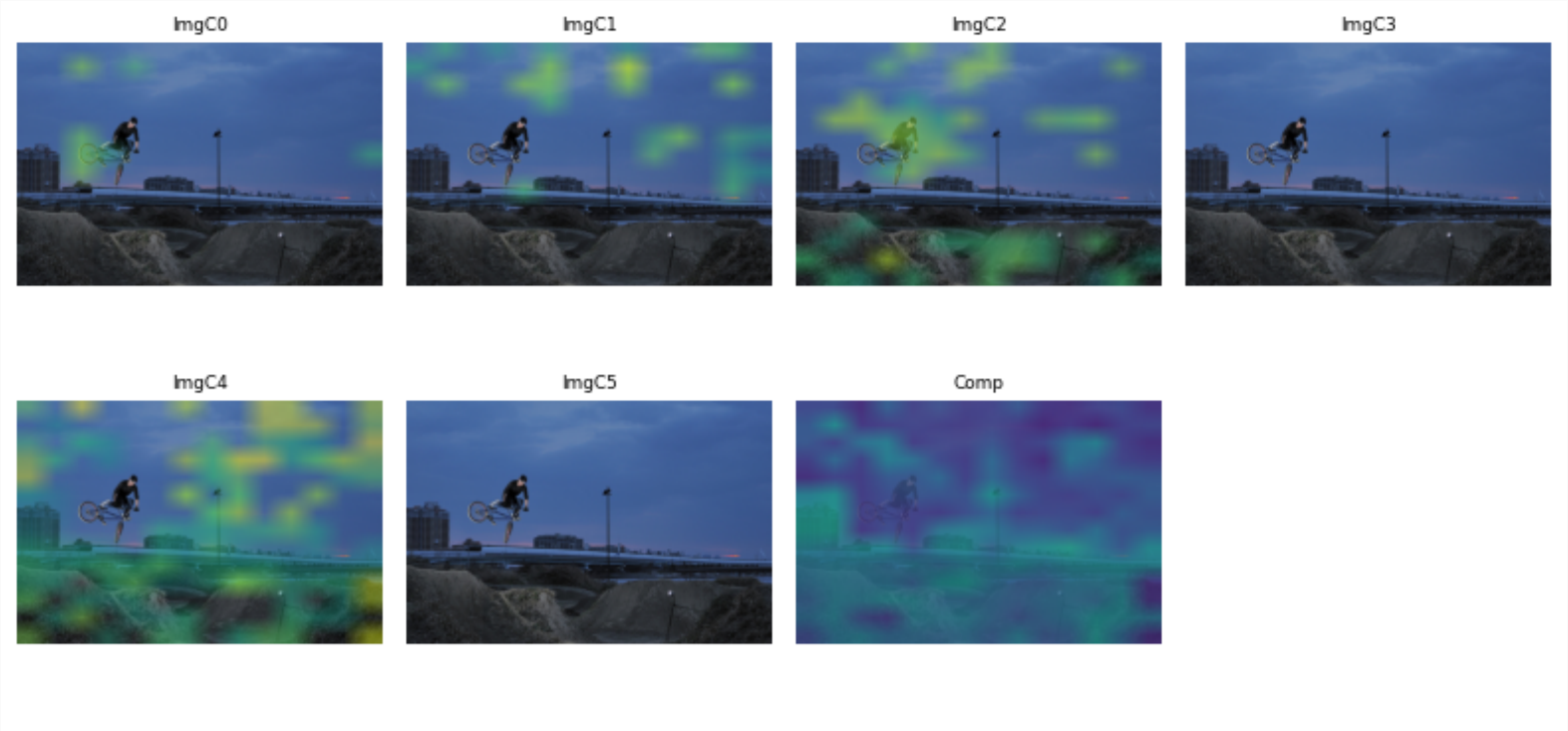


Figure 13: C0 covers the bicycle, C2 the rider and ground, C4 the surrounding area.

Vehicles (img 00, rel)

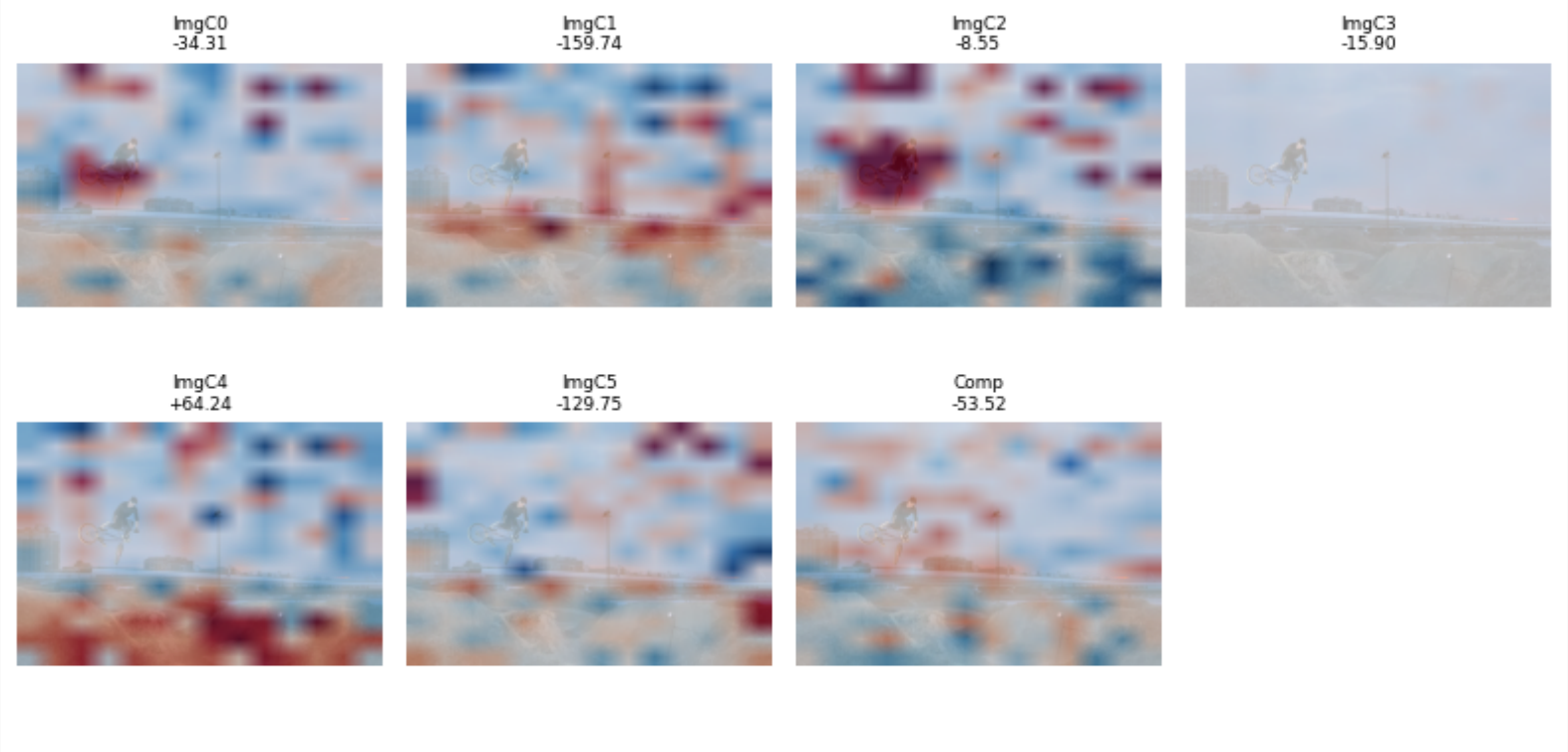


Figure 14: Same image. **C2 (+64)** fires on the rider and ground; **C1 (−140)** and **C5 (−130)** broadly suppress. Probe fires on the rider area (C2), not the bicycle (C0).

Vehicles (img 00, text)

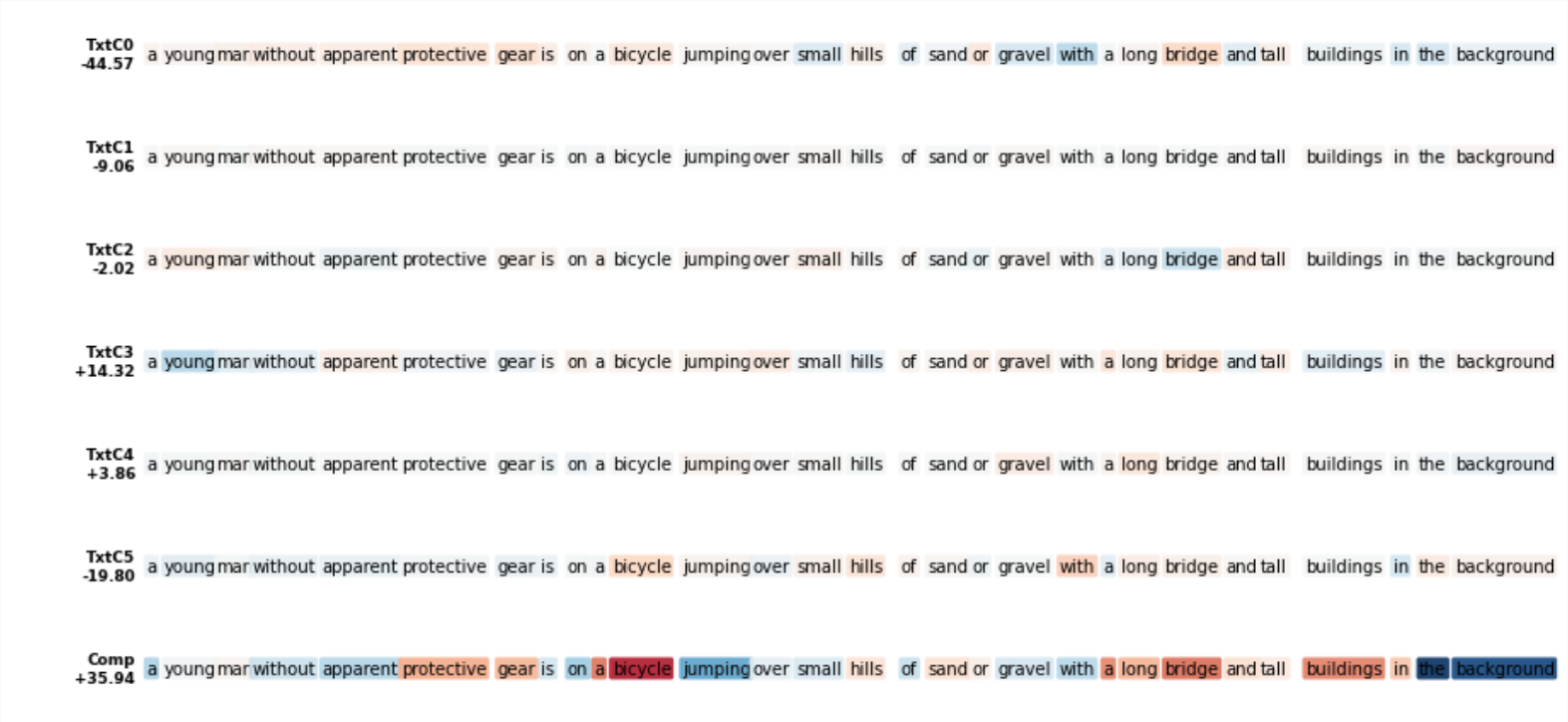


Figure 15: No visible pattern.

Animals (img 08, raw)

Figure 16: C0 surprisingly accurately covers the horse face; C2 covers the person.

Animals (img 08, rel)

Animals (img 07, raw)

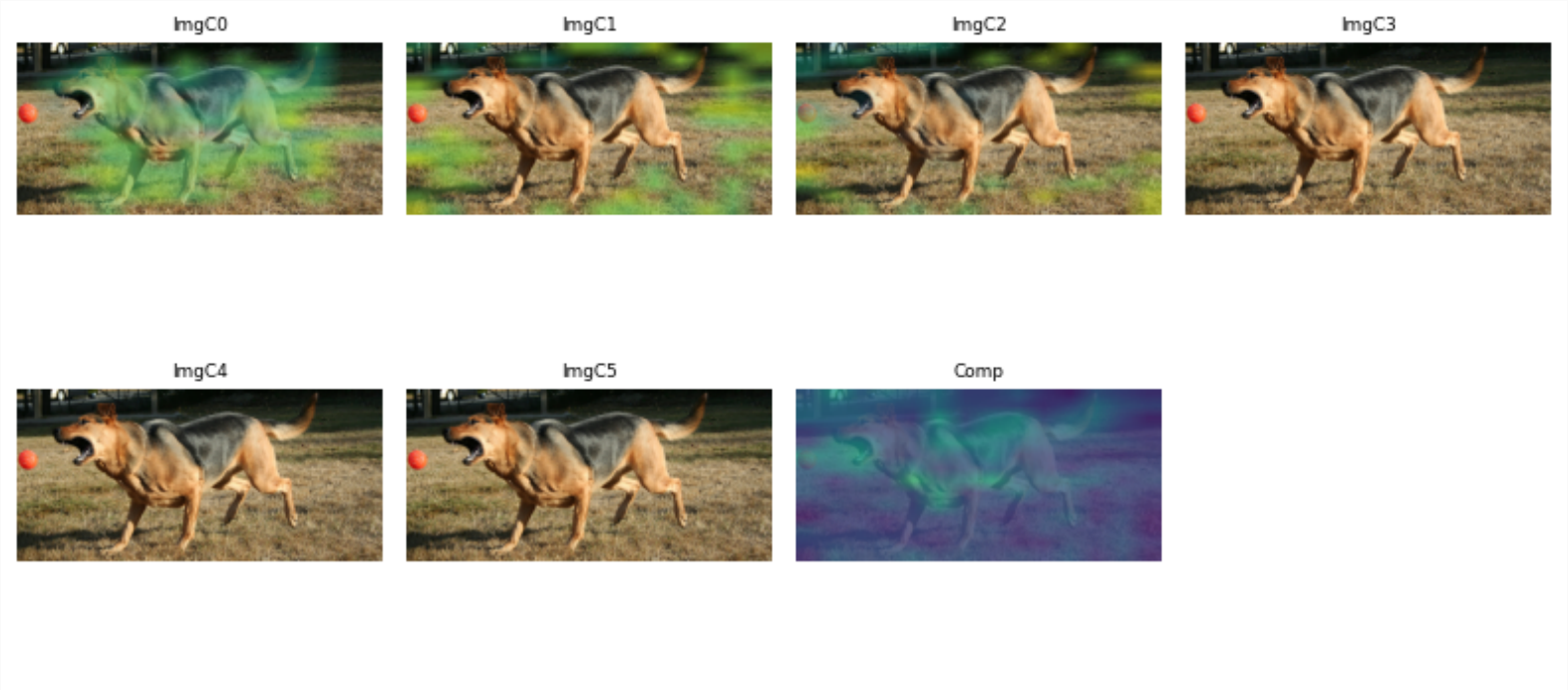


Figure 17: C0 — dog, C1 — ground, C2 — ball.

Animals (img 07, rel)

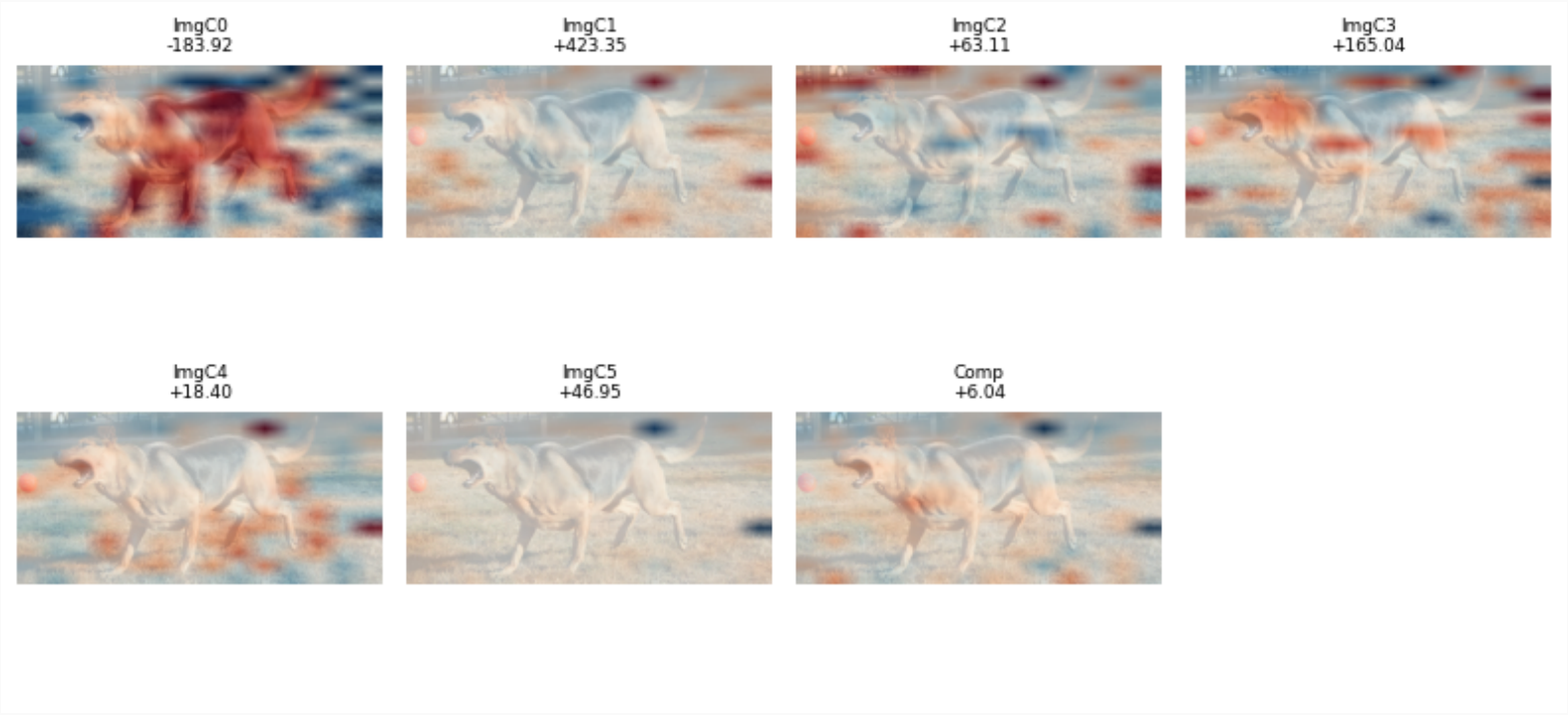
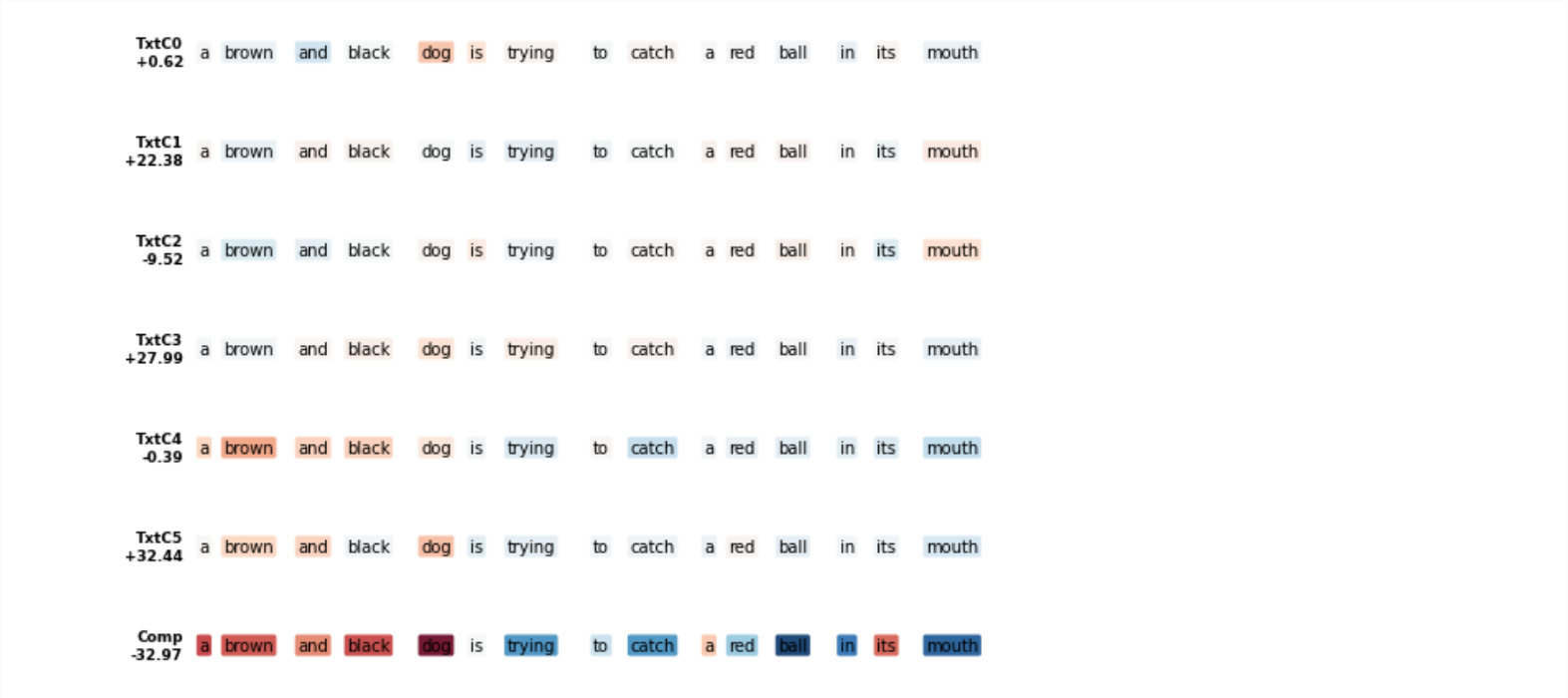


Figure 18: Same image. **C1 (+423)** is the dominant positive — the dog silhouette is clearly outlined in red. **C0 (−184)** suppresses the background. Best animal localisation observed across all examples.

Animals (img 07, text)



Animals (img 05, rel)

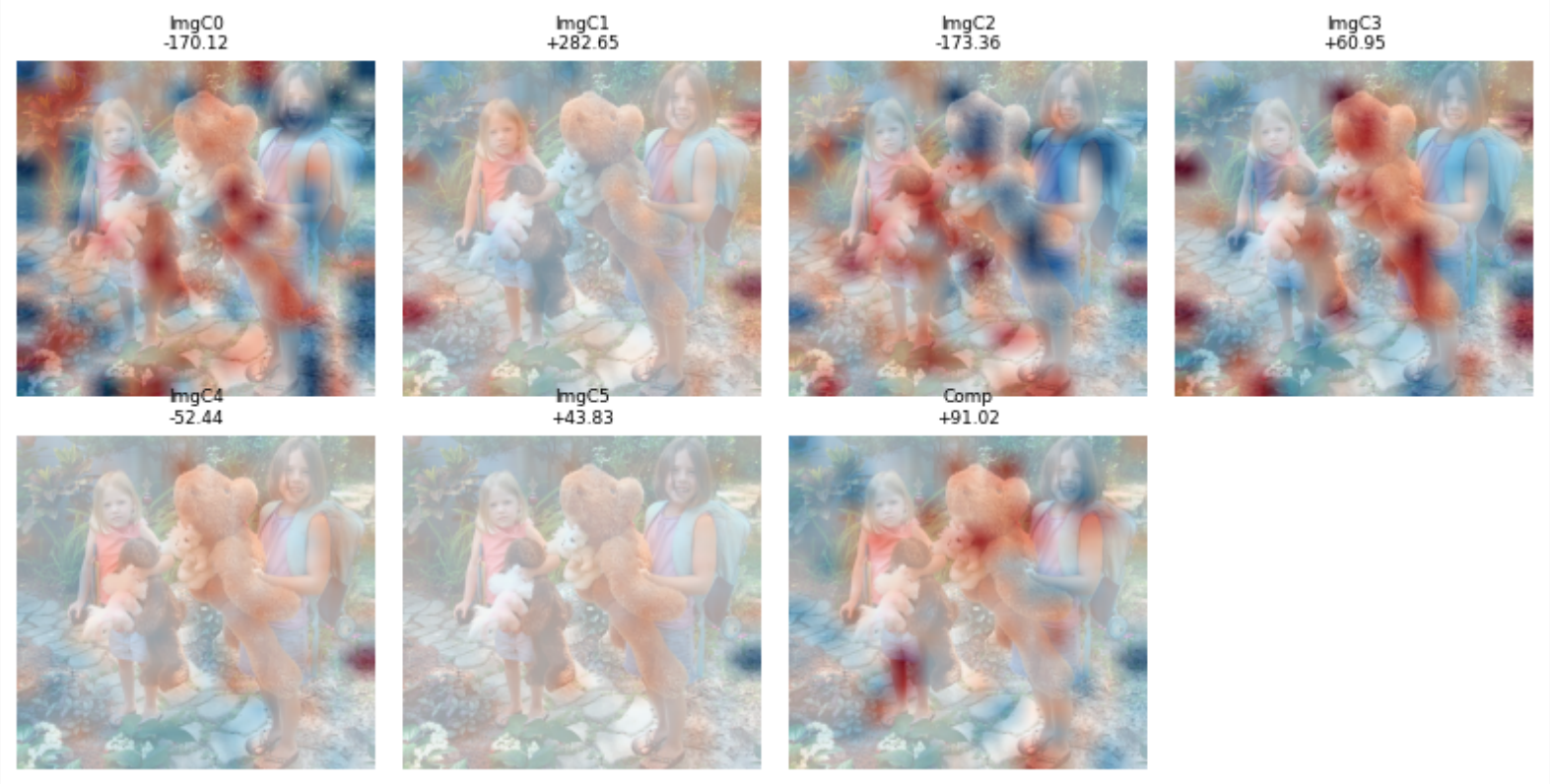


Figure 19: Mixed concepts found — no single concept specialises on any one animal.

Animals (img 04, raw)

Figure 20: Interestingly, C2 precisely covers the leash and C0 covers the dog.

Animals (img 04, rel)

Animals (img 02, raw)

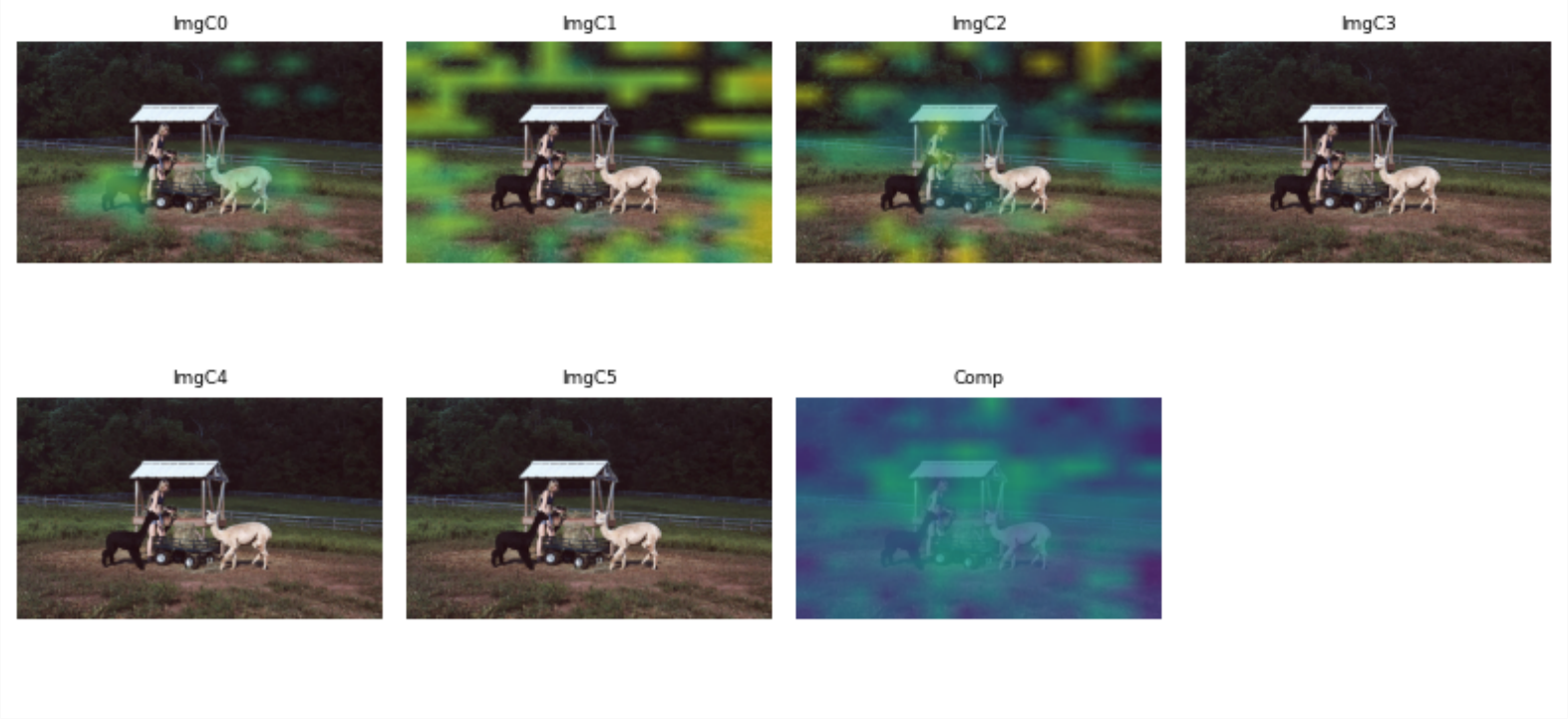


Figure 21: C0 — animals.

Animals (img 01, raw)

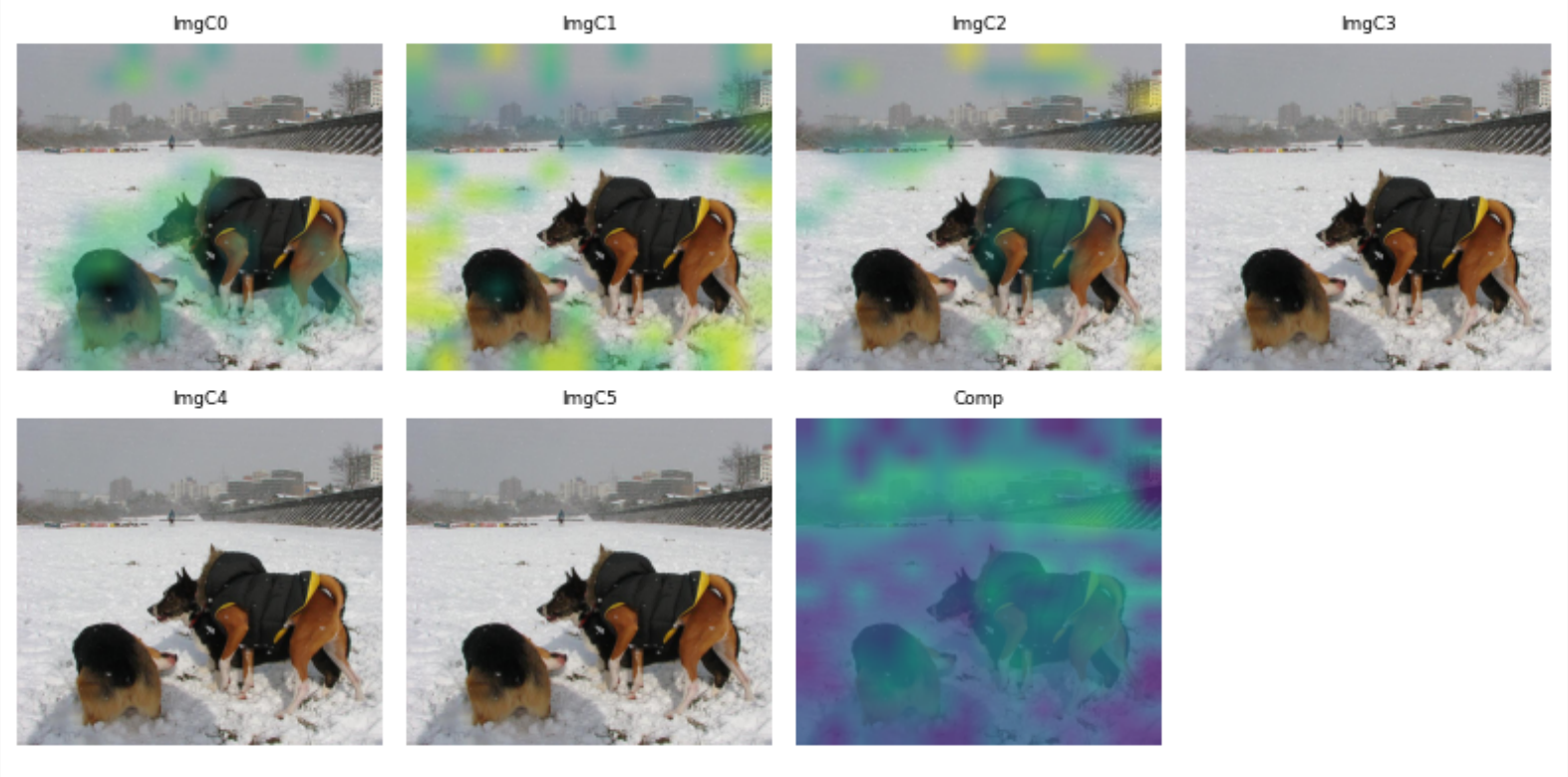
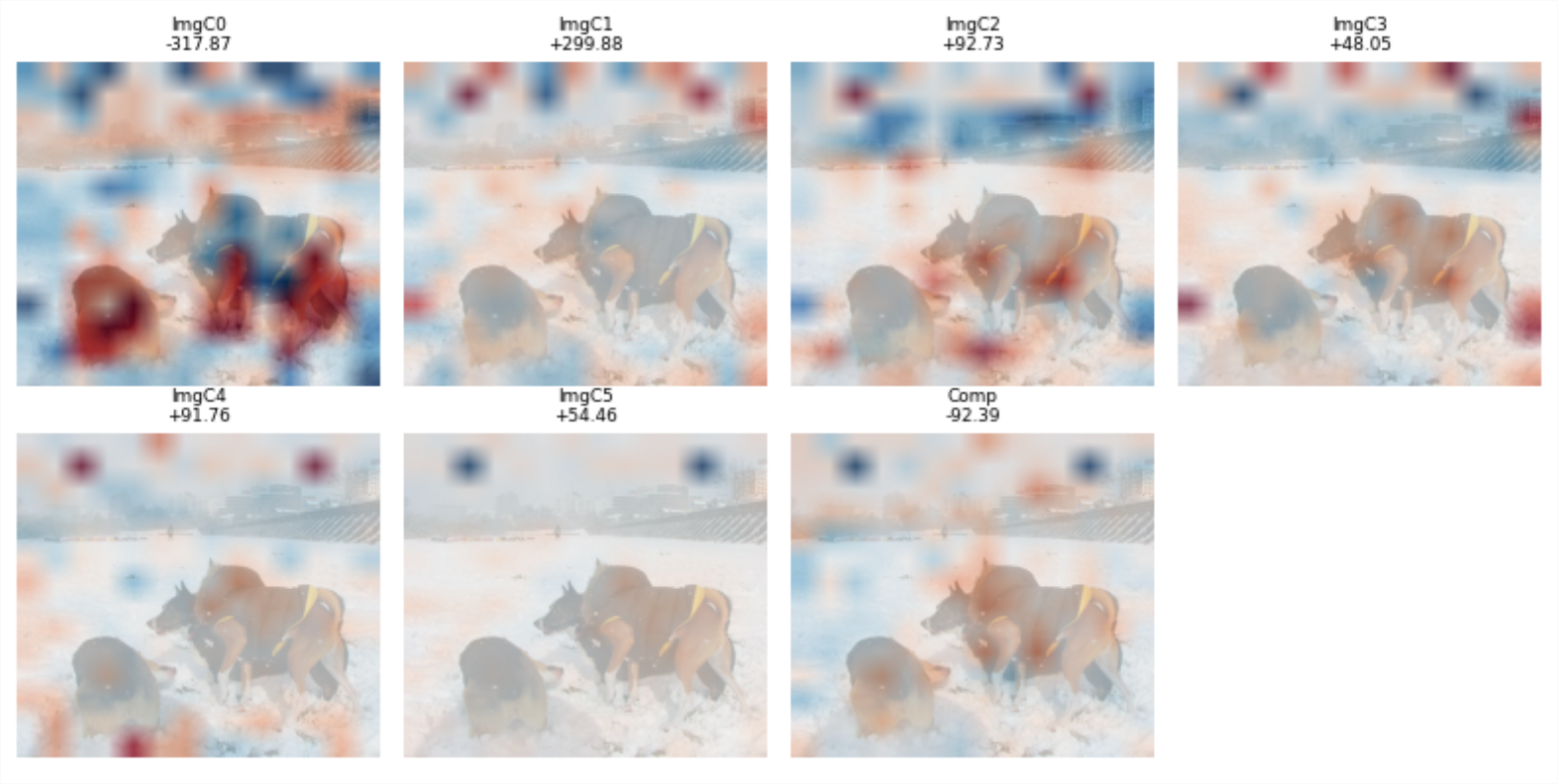


Figure 22: C0 surprisingly precisely outlines the dog bodies, excluding the coat. C2 focuses on the coat itself.

Animals (img 01, rel)



Current challenges

Current challenges

- Must concepts always be linear subspaces?
- Concept discovery is sensitive to noise in patch embeddings
- Training a logistic probe for relevance CAMs introduces a classification bias
- SigLIP encodes scene-level semantics rather than fine-grained part-level details
- Entangled concepts are difficult to interpret individually
- Discovered concepts have no semantic labels without external supervision
- Is the method truly unsupervised if class labels are provided?

References

References i

1. Vielhaben, Bluecher, Strodthoff. *Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees.* TMLR, 2023.
2. Kim, Wattenberg, Gilmer et al. *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV).* ICML, 2018.
3. Ghorbani, Wexler, Zou, Kim. *Towards automatic concept-based explanations (ACE).* NeurIPS, 2019.
4. Zhang, Bargal, Plummer, Sclaroff. *Invertible concept-based explanations for CNN models with non-negative concept activation vectors (ICE).* AAAI, 2021.
5. Fel, Boutin, Cadène et al. *CRAFT: Concept recursive activation factorization for explainability.* CVPR, 2023.
6. Liang, Zhang, Kwon, Yeung, Zou. *Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning.* NeurIPS, 2022.
7. Klingenberg. *Analyzing fluctuating asymmetry with geometric morphometrics: Concepts, methods, and applications.* Symmetry, 2015.