Concept Learning for Interpreting Large Vision–Language Models

Which visual concepts does a vision–language model really use when it answers?

Md Abdul Kadir[1 2] · Omair Shahzad Bhatti[1] · Daniel Sonntag[1 2]

[1] German Research Center for AI (DFKI), Saarbrücken [2] University of Oldenburg

NeurIPS 2026

Prompt: "What are the main objects in each grid?"



Gemma-3n-E4B: hotdog, bear, rabbit, tiger

Grounded in the image, or in language priors?

The problem: concepts inside LVLMs are entangled

Opaque answers

Grounded in the image, or driven by language priors? Hard to tell, and a source of hallucinations.

Built for classifiers

Concept methods expect one activation per input. LVLMs give many token activations that mix image and text.

Prior LVLM method

Needs object-name tokens and full-image activations, so concepts stay entangled.

Prior work (CoX-LMM): one concept vector, many meanings

Decoded words and top image patches of a single concept vector:

pepper · canine · chili · sandwich · sausage



wild · resting · striped · lion · tiger



Gemma-3n-E4B on ImageNet (paper Fig. 4)

Key insight: contrast image patches, not tokens

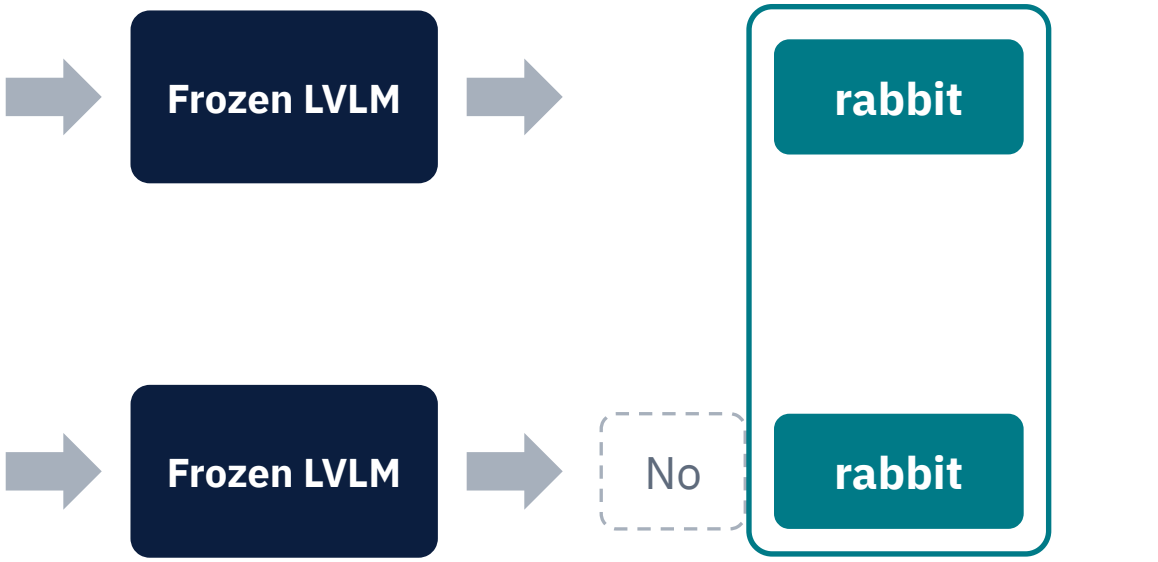
Same question for every patch: *"Does the image contain rabbit? If yes, output rabbit; otherwise, output No rabbit."*



concept present



concept absent

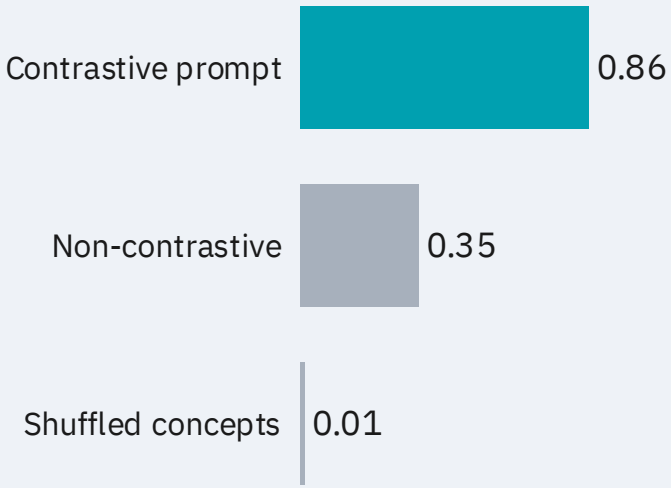


Pool over the shared word only

Same words in both rows → the contrast comes from the image, not the text.

The contrast is what works

Concept F1 ↑ on a fixed 10-category COCO vocabulary



Paper Table 4. Shuffled concept names are a random control.

TGCL in four steps, on a frozen LVLM

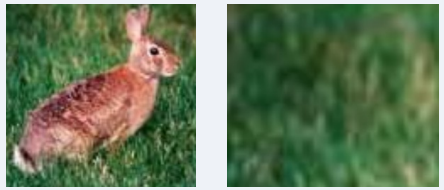
1 Discover

The LVLM lists objects, parts and textures in unlabeled images → concept vocabulary.

dog | fur | stripe | rabbit

2 Build bags

For each concept, crop its images (random or SAM). Crops with and without the concept, unlabeled.



3 Probe & factorize

Yes/no prompt, pool residuals over the concept words, Semi-NMF with 2 bases: concept vs. no concept.

concept | no concept

4 Select & ground

Keep the basis whose top crops the LVLM labels as the concept; decode it via the output head.



→ **“stripe”**

Explain any answer: cosine similarity between the model's activations and each concept vector

Linear · stays in the model's activation space · no proxy model · no token supervision · ~10× faster than CoX-LMM

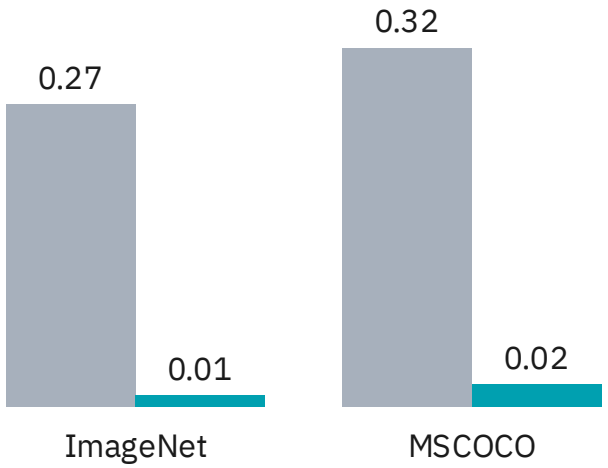
Cleaner, better-aligned, more faithful concepts

Gemma-3n-E4B · TGCL vs. the strongest prior method, CoX-LMM

CoX-LMM TGCL (ours)

Distinct concepts

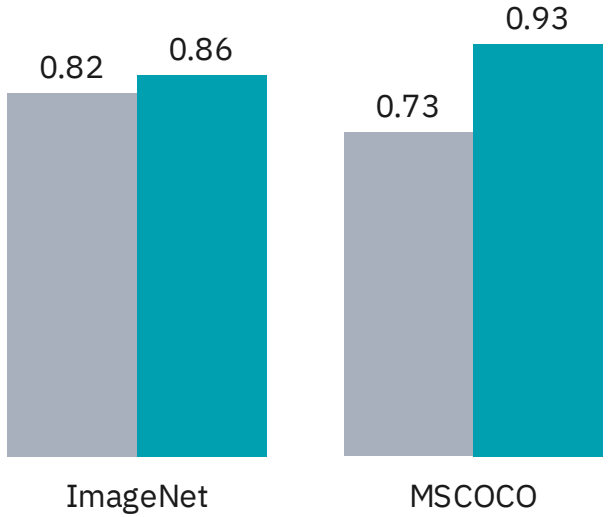
Concept overlap ↓ (pairwise cosine)



0.27–0.32 → 0.01–0.02: one vector, one concept

Grounded explanations

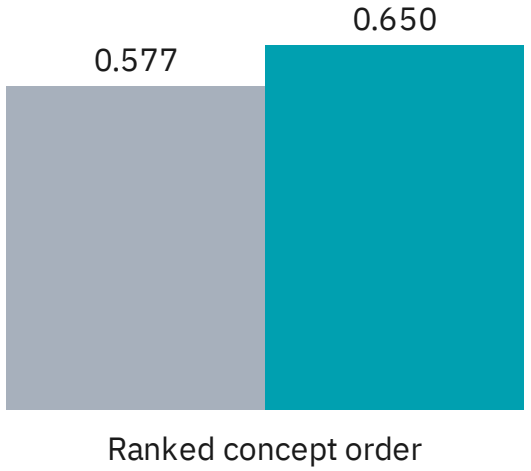
BERTScore ↑, 2×2 image-grid setting



+20 points on MSCOCO

Faithful ranking

Insertion – deletion gap ↑ (MSCOCO-10)



Rank spread 0.016 → 0.108: top concept matters most

Also evaluated on CIFAR-100 and DTD, and with Qwen2-VL-7B and Qwen2.5-VL-7B (see paper)

Takeaway

Let the image, not the tokens, make the difference.

- Token-free, weakly supervised concept discovery in autoregressive LVLMs
- Contrastive, patch-level probing yields sparse, stable, monosemantic concepts
- Better-aligned and more faithful explanations than prior concept methods

Scan for project page & code

mak-ta-reque.github.io/TGCL-LVLM

Concepts TGCL learned in Gemma-3n-E4B (ImageNet)

hot dog



fur



stripe

