Mechanistic Interpretability Workshop

Spectral stratification of semantic abstraction in vision-language models

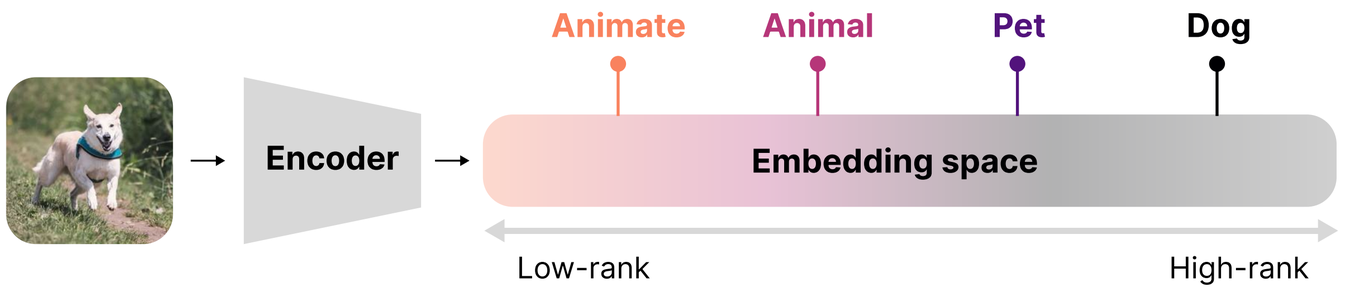
Jeonghwan Cheon[1], Marin Vogelsang[2], Lukas Vogelsang[2], Pawan Sinha[2]

[1] Independent Researcher, Seoul, Republic of Korea,
[2] Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA

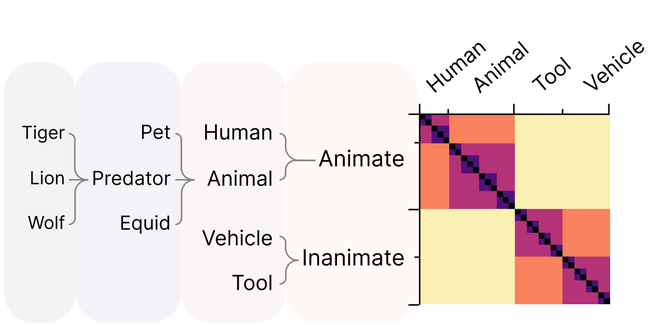
Me cheon.io
OpenReview 2ucMWYxTKU

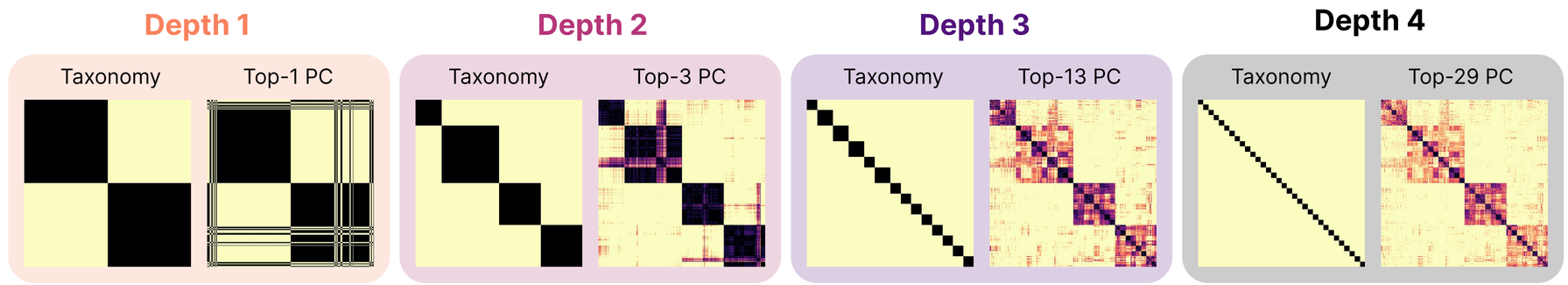
Takeaway

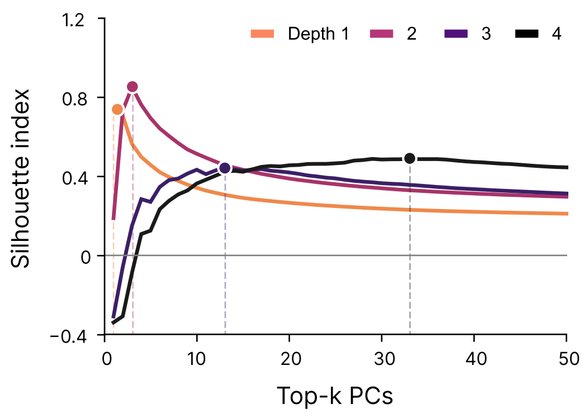
Semantic abstraction in VLMs is hierarchically organized along the principal component spectrum, with compact spectral subspaces improving human alignment.

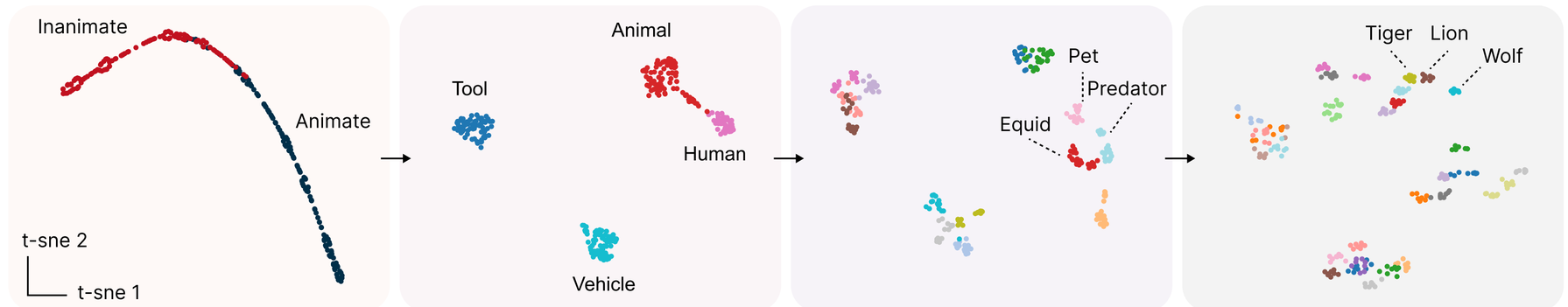


Hierarchical abstraction across principal component spectrums





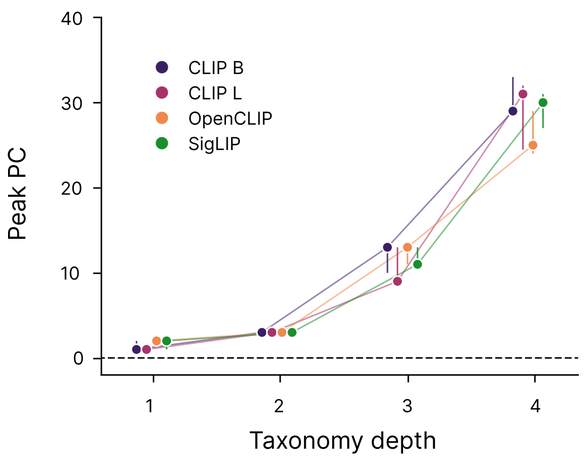




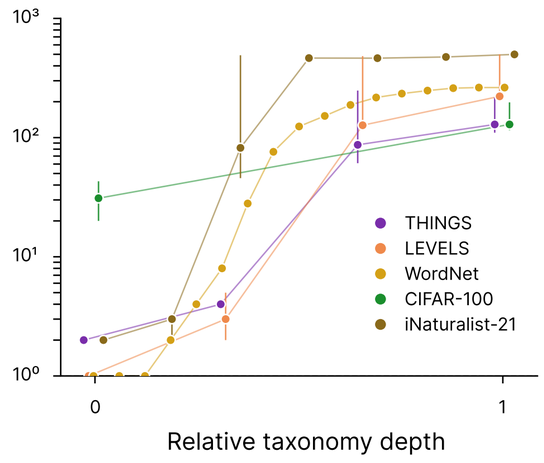
Low-rank PC subspaces encode broad conceptual structure, while higher-rank carry finer-grained distinctions.

Generalization across models and taxonomies

Various VLMs

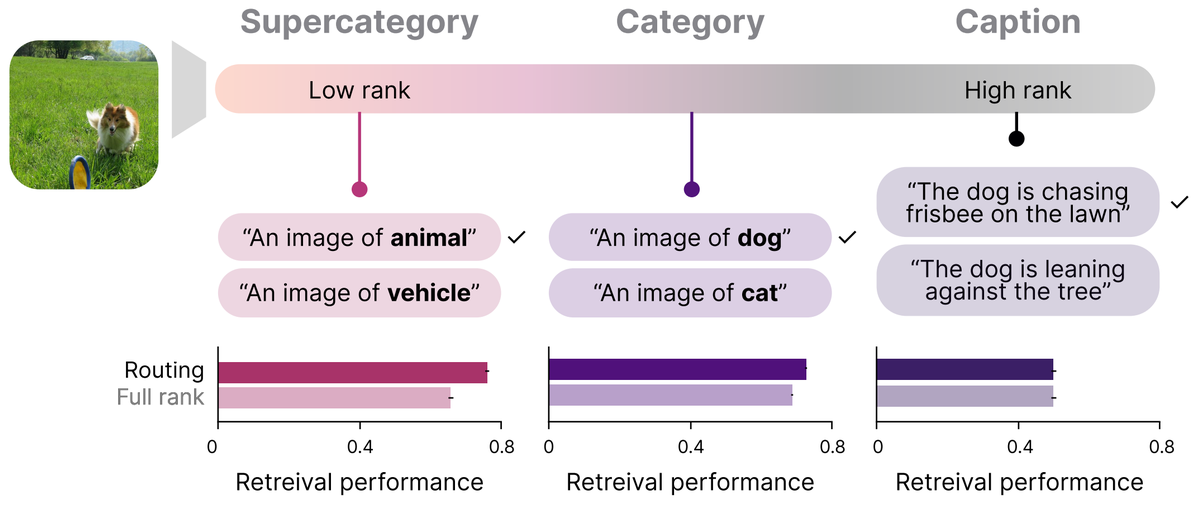


Various Taxonomies



Abstraction levels consistently separate along the PC spectrum.

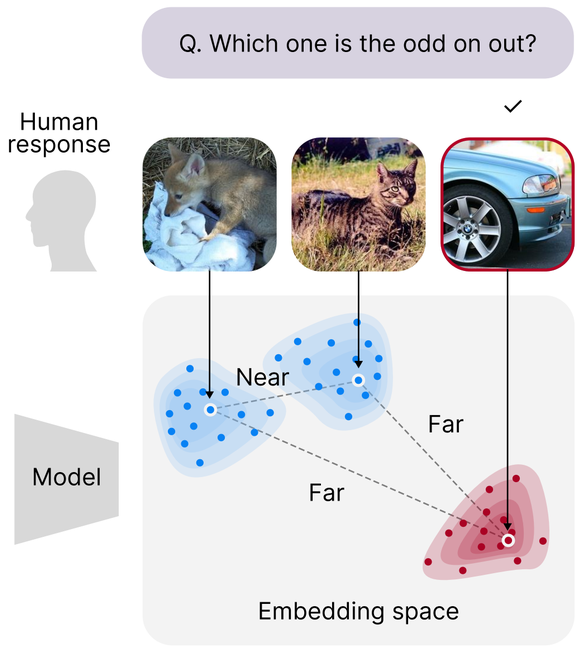
Rank-aware retrieval by abstraction level



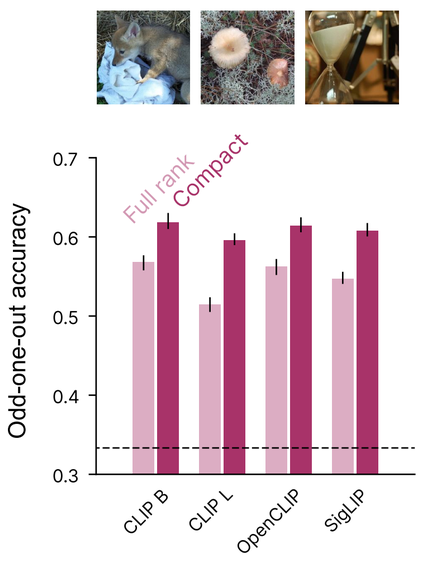
Matching PC ranks to query abstraction improves retrieval.

Compact spectral subspace achieves superior human alignment without supervision

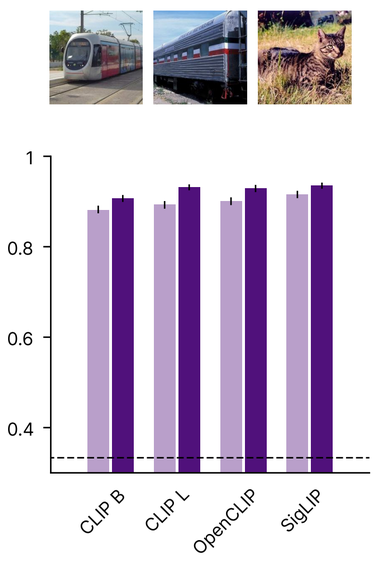
Odd-one-out task



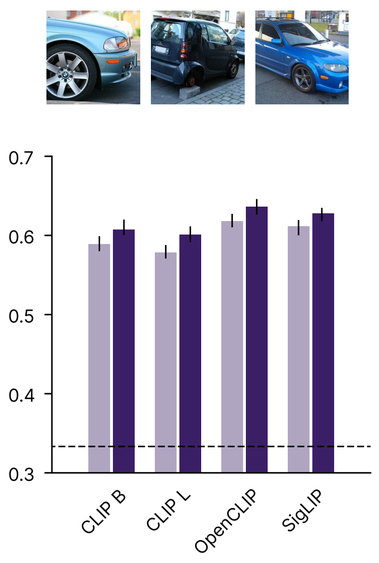
Coarse-grained



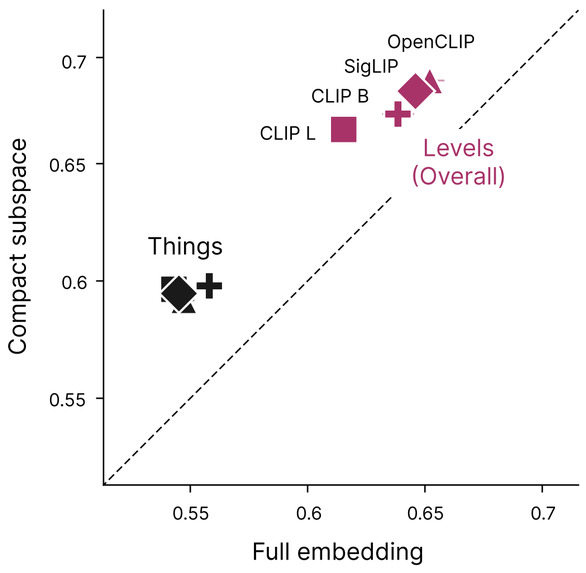
Class boundary



Fine-grained



Odd-one-out accuracy



Compact spectral subspaces exhibit superior alignment with human abstraction behavior than the full-rank model.