BOSTON UNIVERSITY

TEXAS The University of Texas at Austin

VU

UMEÅ UNIVERSITY

Vision-and-language training helps deploy taxonomic knowledge but does not fundamentally alter it

Yulu Qin,* Dheeraj Varghese,* Adam Dahlgren Lindström, Lucia Donatelli, Kanishka Misra,† Najoung Kim†

*,† indicate equal contribution **Contact:** yuluqin@bu.edu; d.varghese@uva.nl

tinlab

VIS LAB VIDEO & IMAGE SENSE LAB

1. Motivation

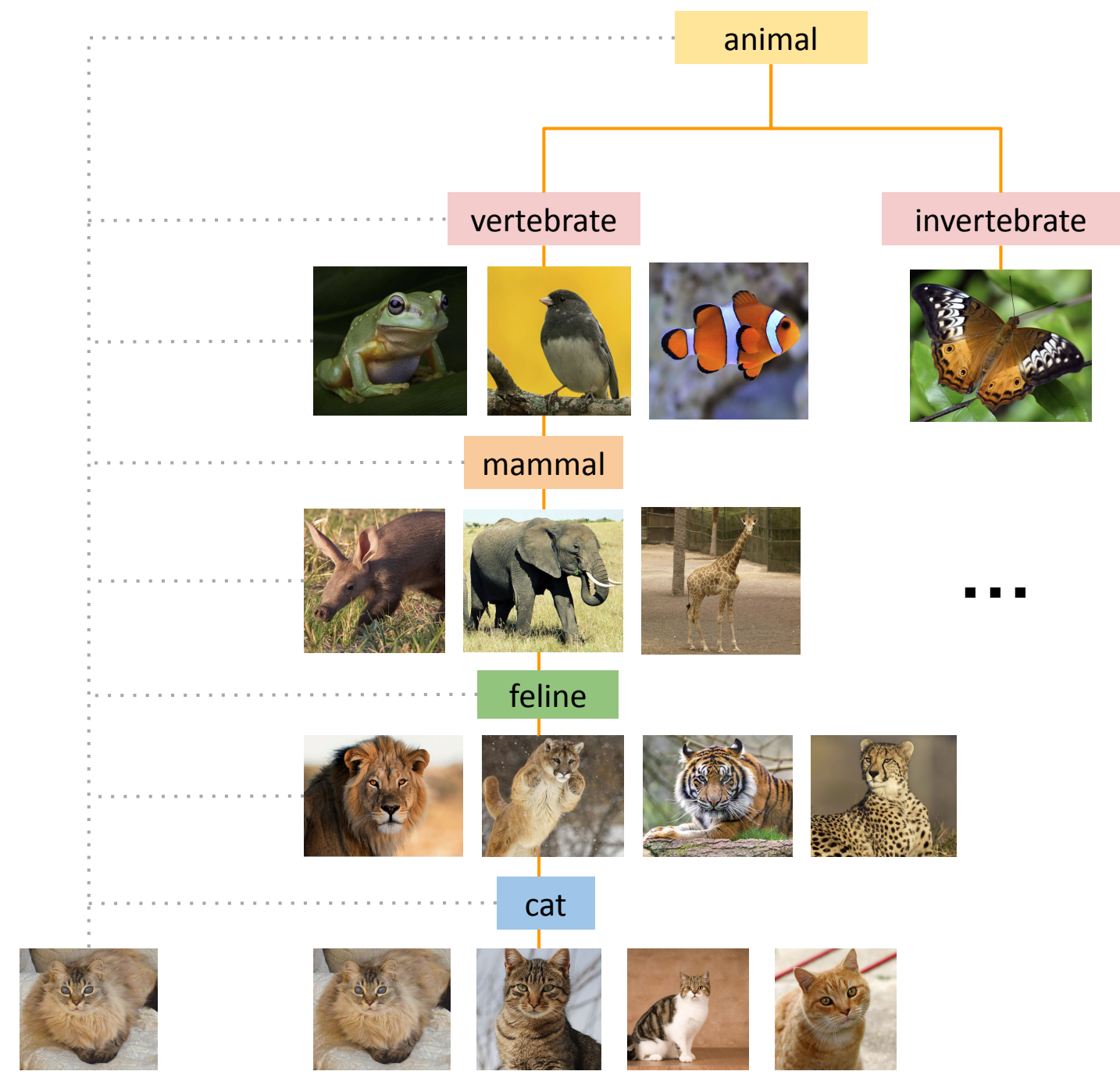
Question: Does **vision-and-language training** change how **language** is represented and used in **VLMs** compared to their **text-only counterparts**?

Gap: Most past work has found little to no positive impact of VL-training on language tasks
(e.g.,Yun et al., 2021; Amariucai & Warstadt, 2024; Wang et. al. 2023)

Domain: Taxonomy - (some) hierarchical concept understanding is naturally associated with visual understanding

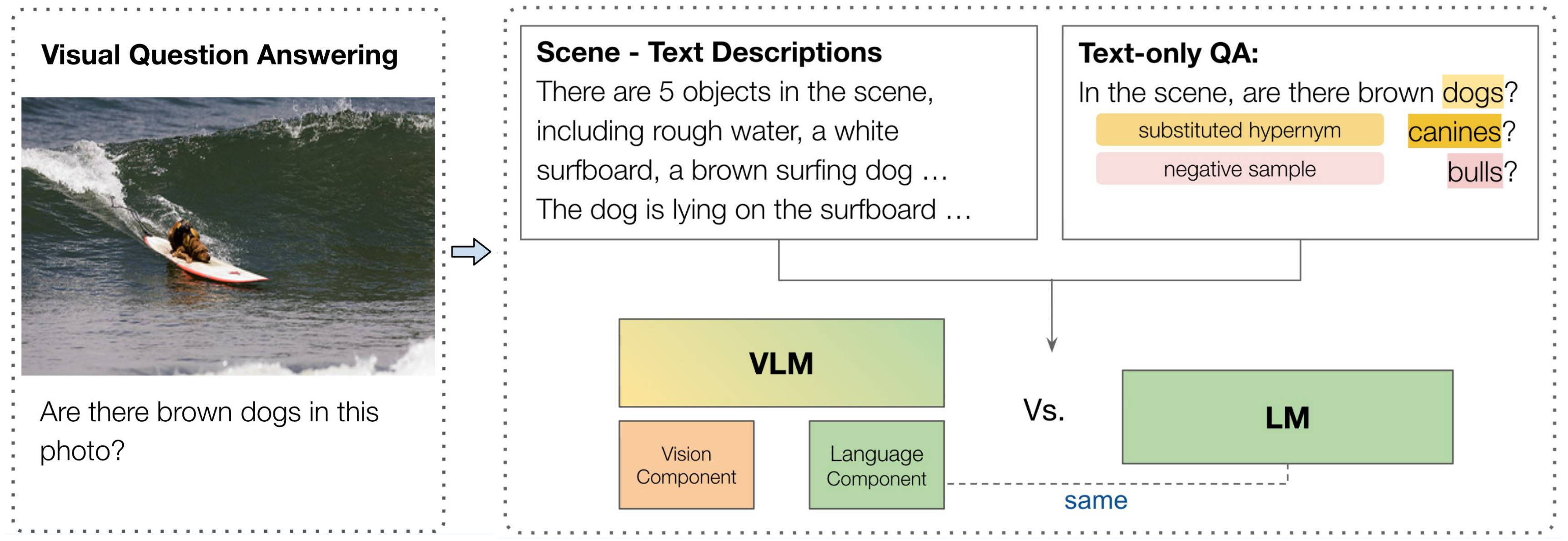
Intuition:

- The same visual referent may be labeled using concepts at different levels of the taxonomy.
- Same level (especially lower in the tree) share similar visual features



2. Task design: GQA → TaxonomiGQA

Task: We transformed the **GQA** task *(Hudson & Manning, 2019)* into a **text-only** QA task



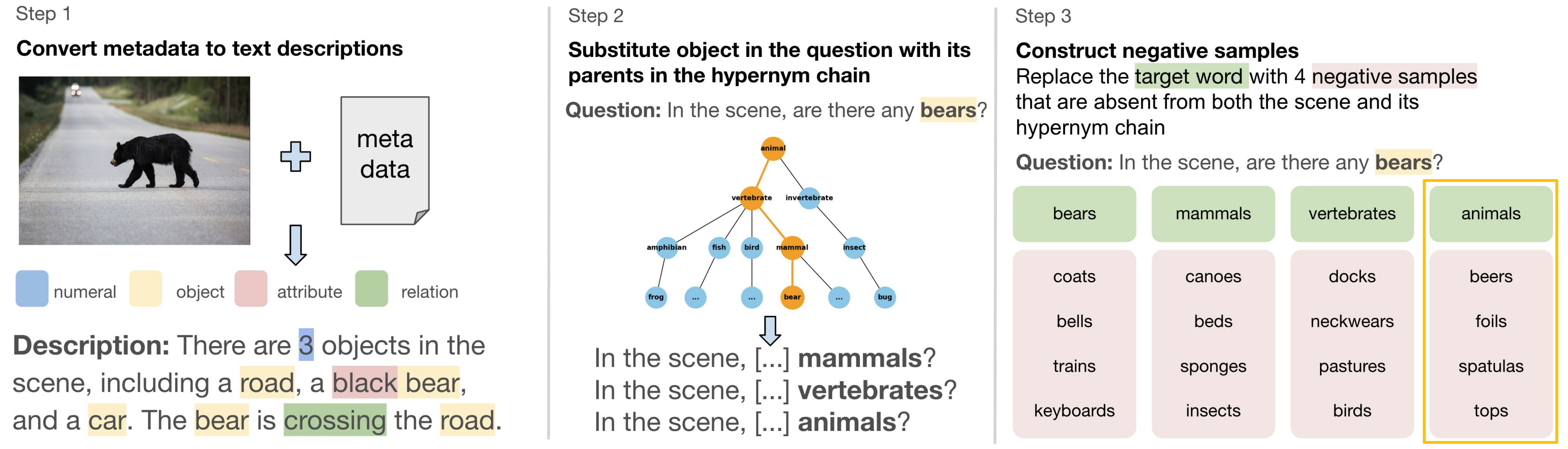
Visual Question Answering

Are there brown dogs in this photo?

Scene - Text Descriptions
There are 5 objects in the scene, including rough water, a white surfboard, a brown surfing dog …
The dog is lying on the surfboard …

Text-only QA:
In the scene, are there brown dogs?
substituted hypernym canines?
negative sample bulls?

Dataset: We applied a three-step transformation to GQA to create TaxonomiGQA



Step 1
Convert metadata to text descriptions

Description: There are 3 objects in the scene, including a road, a black bear, and a car. The bear is crossing the road.

Step 2
Substitute object in the question with its parents in the hypernym chain
Question: In the scene, are there any **bears**?

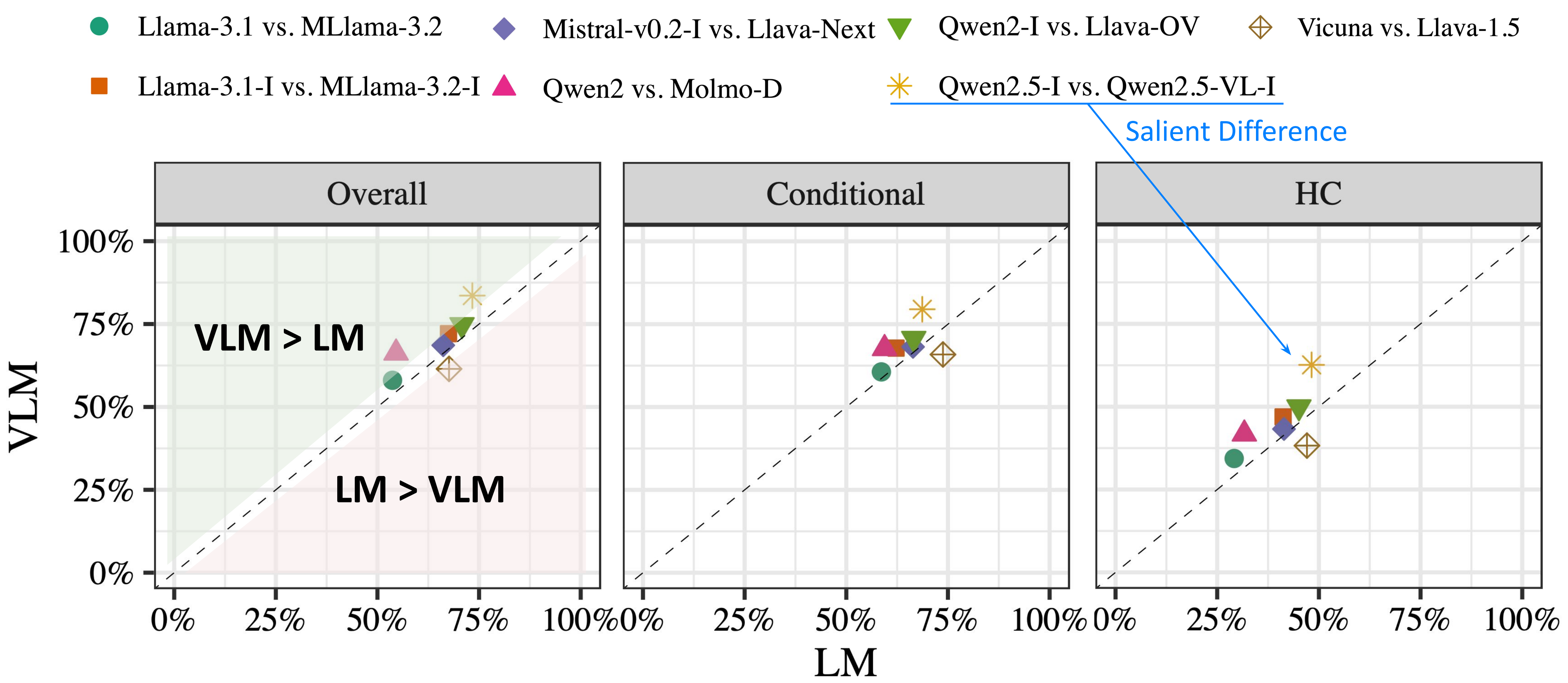
In the scene, [...] **mammals**?
In the scene, [...] **vertebrates**?
In the scene, [...] **animals**?

Step 3
Construct negative samples
Replace the target word with 4 negative samples that are absent from both the scene and its hypernym chain
Question: In the scene, are there any **bears**?

bears	mammals	vertebrates	animals
coats	canoes	docks	beers
bells	beds	neckwears	foils
trains	sponges	pastures	spatulas
keyboards	insects	birds	tops

3. Behavioral results: VLMs > LMs

Models: We tested on ***7 minimally different*** VLM-LM pairs



- **One data instance:** A question targeting taxonomic knowledge and 4 negative samples.
- **Conditional Accuracy:** accuracy of hypernym-substituted instances given the original instance (question about leaf-level hyponyms) was answered correctly.
- **Hierarchical Consistency (HC)**: answer original **and** all hypernym-substituted instances correctly (Wu et al., 2024)

Across 7 VLM-LM minimal pairs, most VLMs (6/7) *outperform their LM counterparts* on a text-only QA task that requires sensitivity to taxonomic knowledge.

References:

Yun, T., Sun, C., & Pavlick, E. (2021). Does vision-and-language pretraining improve lexical grounding? EMNLP.

Amariucai, T., & Warstadt, A. (2024). Acquiring linguistic knowledge from multimodal input. BabyLM.

Wang, W., Vong, W. K., Kim, N., & Lake, B. M. (2023). Finding structure in one child's linguistic experience. Cognitive science.

Hudson, D. A., & Manning, C. D. (2019). GQA: A new dataset for real-world visual reasoning and compositional question answering. CVPR.

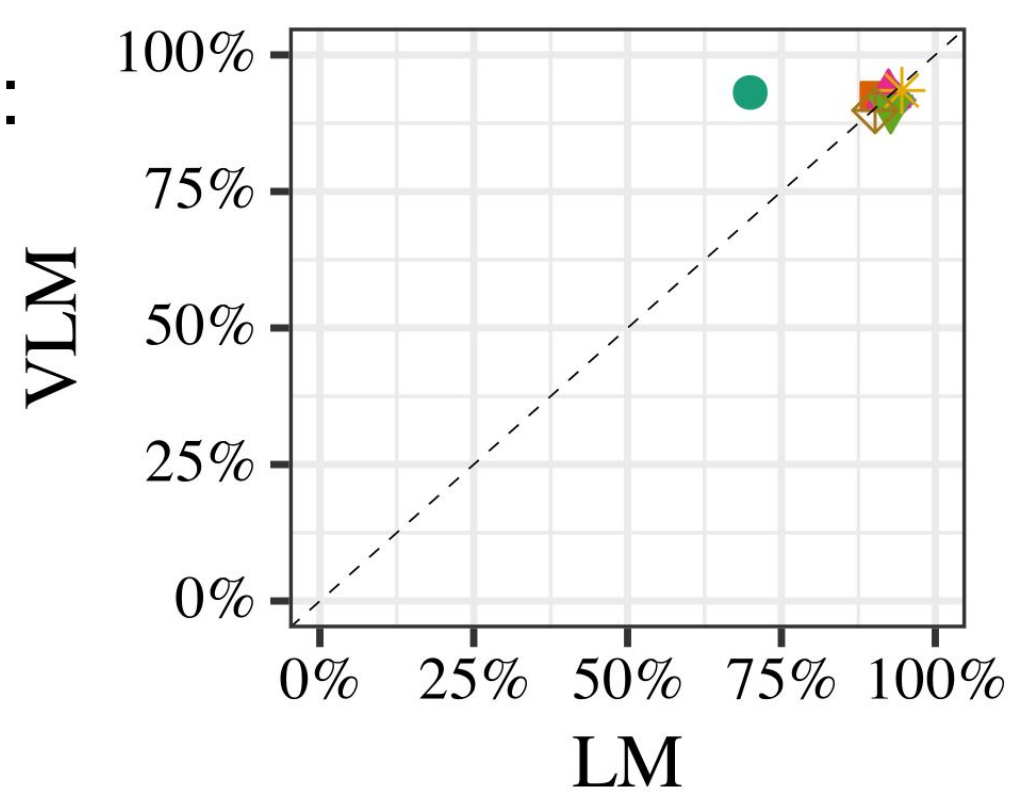
Wu, T. Y., Ho, C. H., & Vasconcelos, N. (2024). Protect: Prompt tuning for taxonomic open set classification. CVPR.

Park, K., Choe, Y. J., Jiang, Y., & Veitch, V. (2024) The Geometry of Categorical and Hierarchical Concepts in Large Language Models. ICLR.

4. Hypotheses

H1: VLMs' static taxonomic knowledge aligns better with the reference taxonomy

Behavioral eval on TAXOMPS: No significant difference between VLM v. LM in most cases.



Is it true that a cat is a feline?

Negative sample

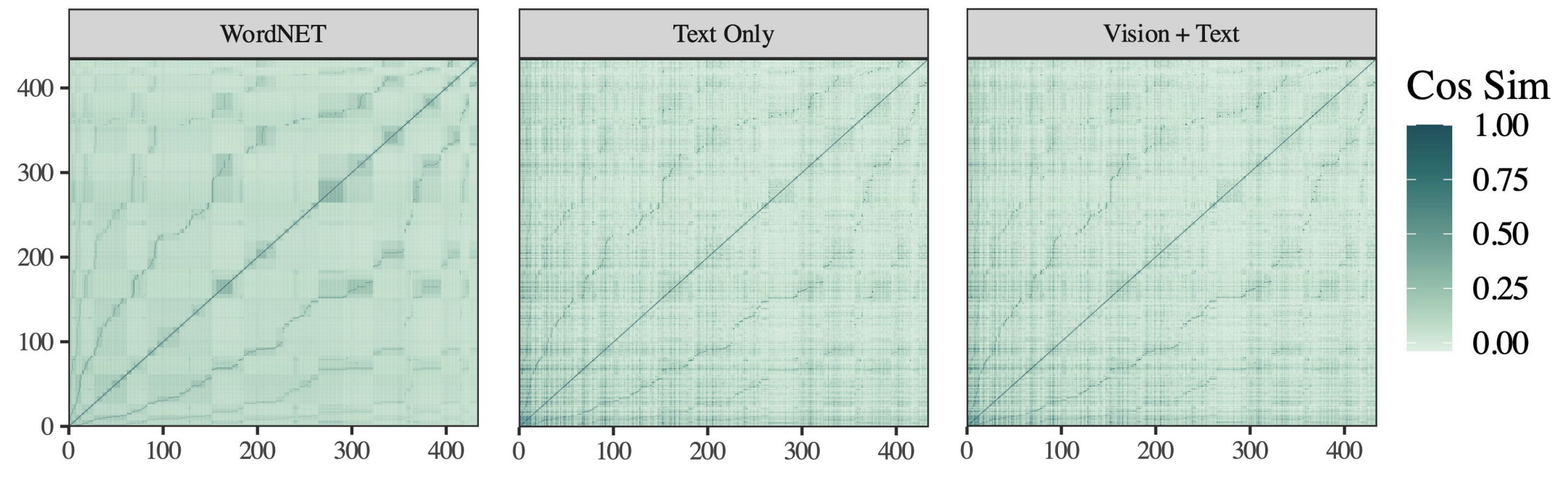
vehicle *fruit*
tool *vegetable*

Representational evals:

Pairwise Taxonomic Similarities in Transformed unembedding spaces *(Park et al, 2024)*

Δ_{model} = **sim**(hyponym, hypernym) - **sim**(hyponym, non-hypernym)

no salient difference



Qwen2.5-I and Qwen2.5-VL-I

Minimal Pairs	Raw Embeddings Δ_{VLM}	Raw Embeddings Δ_{LM}
Vicuna vs. Llava-1.5	0.02	0.02
Mistral-v0.2-I vs. Llava-Next	0.04	0.04
Qwen2.5-I vs. Qwen2.5-VL-I	**0.03**	**0.04**
Llama-3.1 vs. MLlama-3.2	0.04	0.04
Qwen2-I vs. Llava-OV	0.06	0.06
Qwen2 vs. Molmo-D	0.05	0.05
Llama-3.1-I vs. MLlama-3.2-I	0.04	0.04

H2: VL Training improves **deployment** of taxonomic knowledge – i.e., VL-trained models are better **when task contexts recruit taxonomic information**

Contextualized representational analysis:

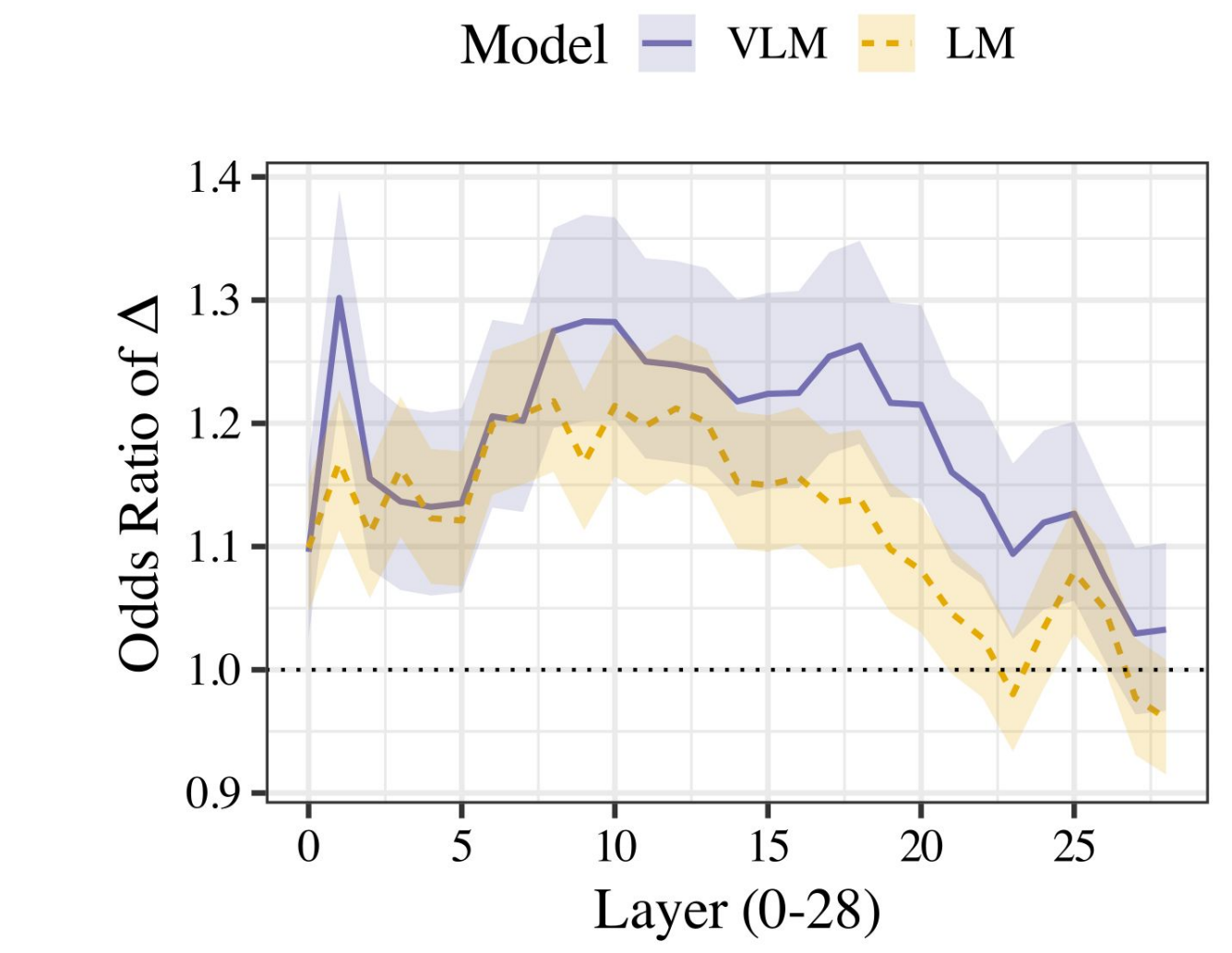
Full context (FC) = There is a **dog** (hypo) **…** In the scene, are there any **mammals**(hyper)?

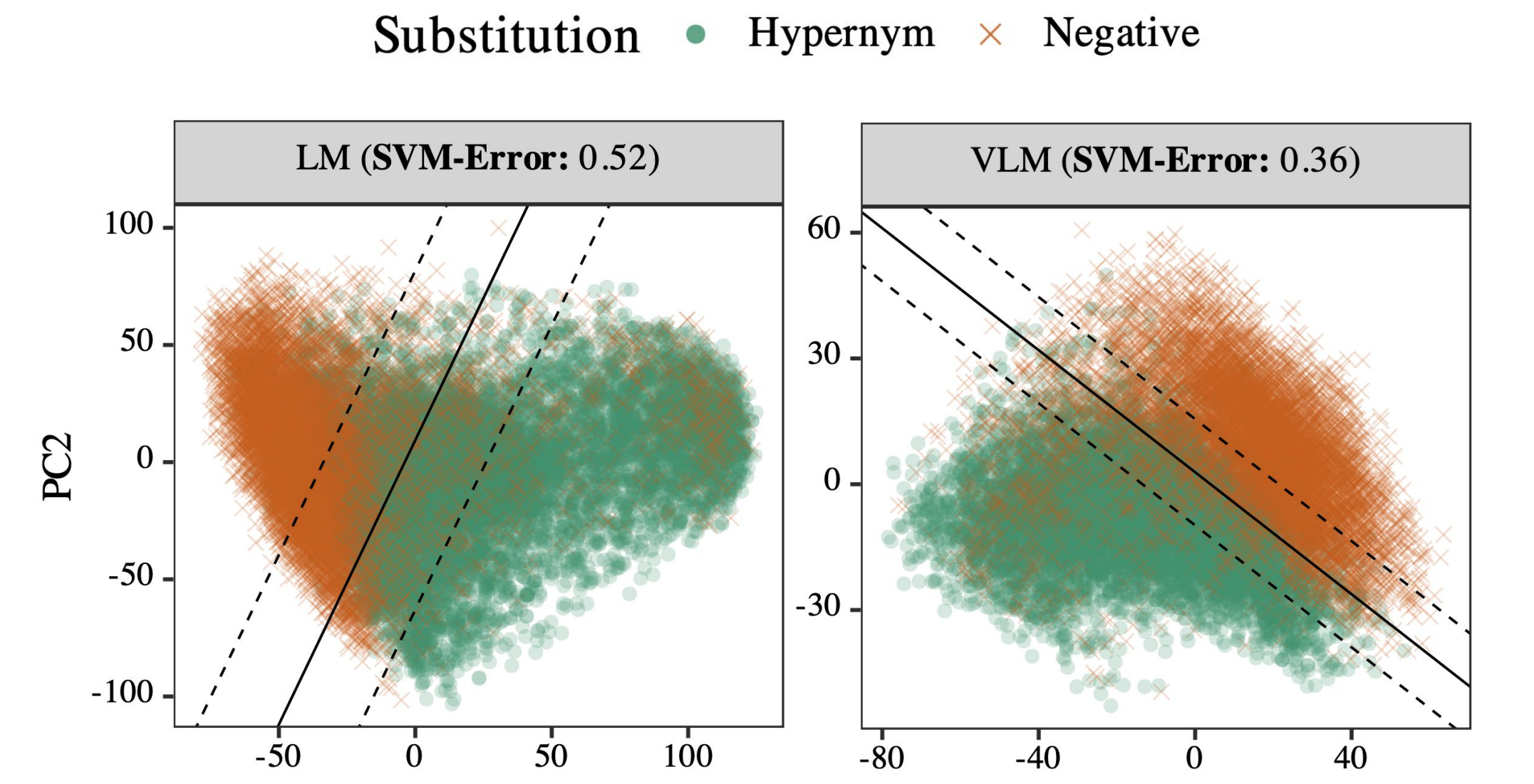
Δ_{model} logistic regression → model correctness (1/0)

PCA projections of the Last hidden state representations of FC

Contextualized embeddings **have a larger effect for predicting model correctness in the VLM** than in the LM.

The **linear separability** (hypernym vs. non-hypernym) of the task context is greater for VLM than for LM.





Conclusion: H2 > H1