Vision-and-Language Training Helps Deploy Taxonomic Knowledge but Does Not Fundamentally Alter It

Yulu Qin,[1,*] **Dheeraj Varghese,**[2,*] **Adam Dahlgren Lindström,**[3] **Lucia Donatelli,**[4]
Kanishka Misra,[5,6,†] and **Najoung Kim**[1,†]
[1]Boston University, [2]University of Amsterdam, [3]Umeå University,
[4]Vrije Universiteit Amsterdam, [5]TTIC, [6]The University of Texas at Austin

https://taxonomigqa.github.io

Abstract

Does vision-and-language (VL) training change the linguistic representations of language models in meaningful ways? Most results in the literature have shown inconsistent or marginal differences, both behaviorally and representationally. In this work, we start from the hypothesis that the domain in which VL training could have a significant effect is lexical-conceptual knowledge, in particular its taxonomic organization. Through comparing *minimal pairs* of text-only LMs and their VL-trained counterparts, we first show that the VL models often outperform their text-only counterparts on a text-only question-answering task that requires taxonomic understanding of concepts mentioned in the questions. Using an array of targeted behavioral and representational analyses, we show that the LMs and VLMs do not differ significantly in terms of their taxonomic knowledge itself, but they differ in how they represent questions that contain concepts in a taxonomic relation vs. a non-taxonomic relation. This implies that the taxonomic knowledge itself does not change substantially through additional VL training, but VL training does improve the *deployment* of this knowledge in the context of a specific task, even when the presentation of the task is purely linguistic.

1 Introduction

Humans readily integrate perceptual and linguistic signals to form generalizable mappings from semantic information to language, allowing them to reason about concepts beyond their immediate environment [57, 18]. Approaches to concept grounding in AI, which traditionally relied on annotated datasets to specify how language links to people, objects, and events [72, 26], have rapidly shifted in light of the impressive capabilities of vision-language models (VLMs).

Many standard VLMs [33, 30, i.a.] often build on top of a pretrained language model (LM) by adding visual conditioning to its next token prediction task, often also updating the parameters of the language model. Analyses of VLM capabilities often focus on the multimodal tasks this additional modality enables. But (how) does this vision-and-language (VL) training change the linguistic capacity of the model? Answering this question requires comparing VL-tuned LMs to their original LM counterparts. Empirical evidence in this literature is rather sparse, often comparing such "VLM-LM minimal pairs" on general benchmarks such as MMLU [17] and GLUE [65]. In this paper, we consider a more targeted investigation (like [73]) of VLM-LM pairs in a particular domain:

[*,†]: Equal contribution; Code can be found at https://github.com/tinlaboratory/taxonomigqa

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

lexical-conceptual knowledge, specifically its taxonomic organization (e.g., *a **cat** is an **animal***). Evaluation of taxonomic knowledge has been of continued interest within the Natural Language Processing [14, 32, 43, 45] and Computer Vision communities [2, 62, 48]—however, to the best of our knowledge no work so far has compared *minimally* differing VLM-LM pairs in terms of how well they can reason taxonomically.

To this end, we develop **`TaxonomiGQA`**, a synthetically augmented *text-only* version of the popular visual-question answering (VQA) dataset GQA [19], where a subset of WordNet [40] hierarchy is used to create questions that require taxonomic knowledge. On comparing 7 widely used VLM-LM minimal pairs, we find most VLMs to consistently outperform their LM counterparts, despite the fact that the QA task is text only. We put forth two hypotheses to explain these results. **H1:** VL training fundamentally alters the (task-agnostic) taxonomic knowledge in LMs; and **H2:** VL training improves the ability of the LM to *deploy* its (largely unchanged) taxonomic knowledge in tasks that require its usage. Through a series of controlled behavioral and representational analyses, we find evidence that supports H2 relative to H1. Finally, we conduct a preliminary investigation where we relate the successes of VLMs over LMs to the visual similarities between the hyponym-hypernym categories we have tested in our work. Here we find initial evidence that suggests that VLMs especially perform well at answering questions about hyponym-hypernym pairs that are visually similar, leaving open areas of interesting future research for a more precise characterization of the role of visual input.

2 Related Work

Influence of vision on language in VLMs There are two main strands of empirical work measuring the influence of the additional visual modality on models' *linguistic* behavior and representations. The first line of work compares VLM and LM performance on downstream text-only benchmarks. The results are mixed: for instance, FLAVA [60] noted around 8% point gains over the base masked language model on GLUE-style NLP tasks (although the evaluation setting involved finetuning). On the other hand, Molmo has been reported to be outperformed by its base LM, Qwen, on text-only benchmarks like MMLU [8]. Generally, more evidence exists in favor of multimodal training hurting text-only task performance [21, 37] and this observation has been used to argue for freezing the language part of the model during multimodal training [12]. The second line of work conducts more targeted comparisons of VLMs and LMs, examining whether additional vision training leads to differences in representations of syntactic categories [66] and performance on tasks that require more "grounding" [73], but the findings overall have indicated no substantial differences. We contribute to this line of work by showcasing a context where there is a non-trivial difference brought about as a result of VL training. In particular, we show that while VL training does not fundamentally alter task-agnostic representations of taxonomic knowledge in LMs (in line with prior work), it does improve the *deployment* of this knowledge in the context of a question-answering task.

Taxonomic knowledge and its deployment Taxonomic knowledge has long been a topic of interest in cognitive psychology [35, 46], and has also often been used to analyze conceptual organization in LMs [14, 32, 45, 44]. Work that tests its functional consequences, such as property inheritance [43, 58, 56] and inductive generalization [42, 13], found strong evidence that while LMs do learn explicit taxonomic knowledge, they struggle to deploy it in taxonomically sensitive tasks [43]. Taxonomic knowledge has also been evaluated and analyzed in multimodal models. For instance, Pach et al. [48] show that the internal structure of neurons in models such as CLIP are often in alignment with existing taxonomies. Our work contributes to this line of work by proposing a level-ground comparison between minimally differing LMs and VLMs, narrowing in on the precise ways in which additional VL training may or may not alter the nature of this knowledge.

3 Behavioral testing of minimal pair VLMs and LMs with `TaxonomiGQA`

The question we are interested in answering concerns the *change* that VL training introduces to the lexical-conceptual knowledge of a model. This requires a shared evaluation that can be applied to both VLMs and LMs. We discuss below how we designed this evaluation as well as our findings about a range of VLM-LM pairs from this evaluation.

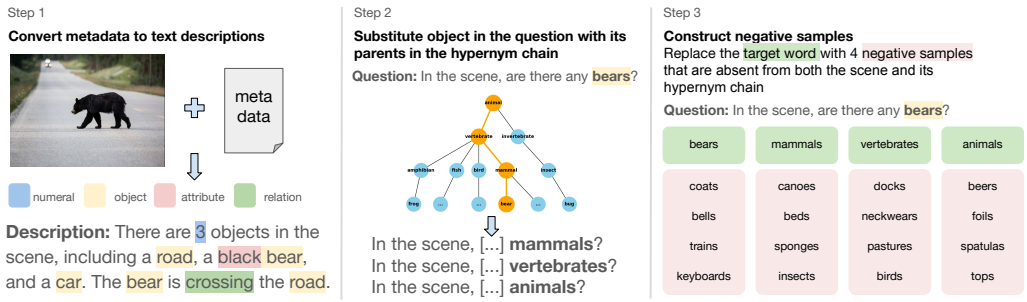


Figure 1: The three-step pipeline to create **TaxonomiGQA**.

3.1 Dataset design

We created a QA dataset that requires taxonomic understanding based on GQA [19], named **TaxonomiGQA**. A datapoint in GQA consists of an image of a scene, a question about this scene, and metadata that includes a scene graph of the objects, their attributes, and relations between the objects. We applied a three-step modification (each step illustrated in Figure 1) to create a *text-only* dataset, since our goal is to systematically compare the LM and VLMs' taxonomic competence. (1) Convert the scene graph into a purely textual description of the scene programmatically using hand-crafted templates; (2) For each question that contains a word that corresponds to a node in our reference taxonomy, substitute the word to its hypernym; (3) For each substitution, create four negative samples following Misra et al. [43], by substituting the target word with a word that is not in its hypernymy chain. Below we describe the resources, filtering, and sampling details used to create **TaxonomiGQA**.

Taxonomy To construct the reference taxonomy, we first extracted all unique noun lemmas ($N = 1216$) that appeared in the GQA questions, and annotated their senses in the WordNet taxonomy, these serve as the leaf nodes of our taxonomy. Next, for each noun, we extracted its hypernym chain (e.g., dog < canine < mammal < vertebrate < animal) from WordNet, rejecting hypernyms that were too abstract (determined manually), e.g., *entity, material, conveyance*. 315 concepts were removed as a result, many of which often had abstract entities in their hypernym chains or had non-ideal WordNet categorization (e.g., *bubble* as a member of *ball*), leaving us with 901 unique chains. (See Appendix C for more details.)

Dataset construction We applied a multi-stage filtering process to the validation split of GQA (10,696 images/scenes and 488,293 questions) to obtain our base questions. We first applied scene-level filtering by excluding scenes containing more than 20 annotated objects or any repeated object labels to avoid ambiguity in referring expressions in text. For each remaining scene, we applied question-level filtering to retain questions that refer to a single object (excluding any that mention multiple objects) and whose hypernyms do not overlap with those of any other object in the scene. Next, we balanced the dataset by randomly sampling 40 questions per scene in proportion to each scene's question type ratios. We further filtered the questions by answer type and restricted the dataset to yes/no questions to facilitate the substitution step. This reduced our taxonomy to 314 unique chains. In the base questions remaining after filtering, we substituted each target concept with each of its hypernyms in its hypernym chain to obtain the substituted questions. Then, we created negative sample questions by substituting the target concepts with concepts that are not in their hypernym chain, discarding question types where this substitution was not possible due to the introduction of presupposition failure (e.g., questions such as *Is the color of the dog brown?* when there is no dog in the image). This ended up eliminating more hypernym chains (which were only present in the discarded question types), leaving us with 126 final chains. More details about this negative sampling pipeline is given in Appendix D.

Dataset statistics The final dataset contains 1,342 unique images/scenes, 29,604 positive sample instances (9,334 targeting leaf node concepts, 20,270 targeting hypernym-substitutions), and 4 negative samples for each positive sample, amounting to 148,020 total instances. There are 276 hyponym-hypernym pairs, 126 unique hypernym chains, 88 unique hypernyms, and 24 top-level categories (e.g., *animal, vehicle,* etc.).

3.2 Metrics

We propose metrics designed to be sensitive to taxonomic structure (cf. [62]). The design principles are: (1) be sensitive to hierarchical relationships between two concepts; (2) anchor expectations on taxonomic knowledge conditioned on the model's success at foundational or prerequisite tasks; and (3) provide insight into robustness, including contrasting the performance on both positive and negative samples. By grounding our metrics in these properties, we move beyond correctness and toward a more systematic assessment of model performance.

As preliminaries, each instance, $X_i = (q, q^n_{1\ldots4})$ in **TaxonomiGQA** consists of a positive sample question, q, about some leaf-level category (target concept), coupled with a set of 4 negative sample questions, $\{q^n_{1\ldots4}\}$, where the target concept in the original question is now replaced by a negative-sample concept, as described in Section 3. Next, for each instance, we have a set of k_i hypernym-substituted instances, $\{X^{s,i}_1, \ldots, X^s_{k_i}\}$, where each item X^s_j is an instance but with the original category *substituted* with a category in its hypernym chain, along with their own 4 negative samples. Finally, we use a function, `correct(.)` $\rightarrow \{0, 1\}$, which accepts an instance X, and returns 1 iff. the model correctly answers the positive sample question *and* all four negative sample questions, and 0 otherwise. Using these preliminaries, we propose the following metrics:

Overall Accuracy measures the proportion of time the model correctly answers all original, unsubstituted, and hypernym-substituted instances, treating each instance as separate item.

$$\text{Overall} = \frac{1}{N + \sum_{i=1}^{N} k_i} \left[\sum_{i=1}^{N} \left(\texttt{correct}(X_i) + \sum_{j=1}^{k_i} \texttt{correct}(X^s_j) \right) \right] \tag{1}$$

Conditional Accuracy measures the proportion of time the model correctly answers hypernym-substituted instances, conditioned on the fact that the model correctly answered the original, unsubstituted instance correctly. That is, if there are N_{sel} original instances that the model answered correctly, the metric is calculated by:

$$\text{Conditional} = \frac{1}{\sum_{i=1}^{N_{sel}} k_i} \sum_{i=1}^{N_{sel}} \sum_{j=1}^{k_i} \texttt{correct}(X^s_j) \tag{2}$$

Hierarchical Consistency proposed by Wu et al. [68, originally named "Hierarchical Consistence Accuracy"] measures a stricter form of accuracy relative to the previous ones, as the proportion of time the model correctly answers the original unsubstituted instance *and* all of its corresponding hypernym-substituted instances. Using our notation, this is measured as:

$$\text{HC} = \frac{1}{N} \sum_{i=1}^{N} \texttt{correct}(X_i) \prod_{j=1}^{k_i} \texttt{correct}(X^s_j) \tag{3}$$

All of the metrics incorporate robustness to negative samples (using the `correct()` function). Conditional Accuracy is stricter than Overall, ruling out cases where the model succeeds at higher level categories without correctly answering questions about the target object. HC requires the model to answer all questions about a hypernym chain correctly, being the strictest measure. That is, if the model fails to answer questions about *canines* then all *dog/wolf/fox* questions will be penalized. This is the most faithful to the chain in the reference taxonomy but may be considered overly strict.

3.3 Models

We selected seven LM-VLM model pairs, where the LM has been reported to be the base model that the VLM has been trained on top of, following the approach of [24]. The selected pairs are: (1) **Llama-3.1-8B** vs. **MLlama-3.2-11B** [12]; (2) their **instruct versions**; (3) **Vicuna** vs. **Llava-1.5-7B** [33]; (4) **Mistral-v0.2-Instruct** [22] vs. **Llava-Next** [34]; (5) **Qwen2-7B** [70] and **Molmo-7B-D** [8]; (6) **Qwen2-7B-Instruct** vs. **Llava-OneVision** [29]; and (7) **Qwen2.5-7B-Instruct** [71] vs. **Qwen2.5-7B-VL-Instruct** [4]. See Appendix A for more details. Since **TaxonomiGQA** consists of Yes/No questions, we sampled from a constrained probability distribution of `Yes` and `No` tokens from the models' output vocabulary, allowing for surface form variation such as casing and space-prefixing.

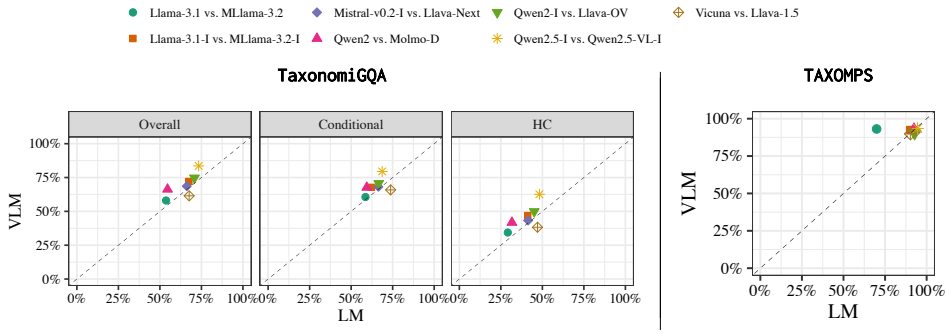


Figure 2: Performance of VLM-LM model pairs on **TaxonomiGQA** (Section 3) and **TAXOMPS** (Section 4.1). Points above the line indicate that VLM outperforms LM.

3.4 Results

The results are shown in Figure 2: points above the diagonal denote model pairs where the VLM outperforms the LM counterpart, and points below denote model pairs where the LM outperforms the VLM. We observe a consistent trend (with a single exception of Vicuna vs. Llava-1.5) where the VLMs outperform their LM counterpart, even though the presentation of the task was purely linguistic.[1] We rule out the possibility that the VLMs are performing better due to having been trained on GQA directly by running a control experiment where we give only the question (without the scene description) to the VLMs—if VLMs have encountered the original GQA instances during training, they may have learned question-label associations used in our dataset. Our results (Appendix E) show that most VLMs do not perform substantially above chance. Since the text descriptions are newly introduced in **TaxonomiGQA**, we can safely rule out the hypothesis that VLMs' improvements are due to having been trained on GQA. Accepting the trend of VLMs outperforming LMs on **TaxonomiGQA** as a genuine improvement, we conduct analyses that aim to explain this result more in the subsequent sections. When we are not analyzing all model pairs, we focus on Qwen 2.5-Instruct vs. Qwen 2.5-VL-Instruct, since the performance gap between VLM and LM was the most salient with this pair, especially on stricter metrics (conditional accuracy and HC).

Generalizability of our finding We conducted a supplementary experiment using the dataset from Rodriguez et al. [56]—another case where task contexts require the deployment of taxonomic knowledge—to verify the generalizability of our finding outside of **TaxonomiGQA**. Here, LMs were evaluated on their projection of novel properties (e.g., *is daxable, has feps*, etc.) from hypernyms to their hyponyms. Results on this dataset (shown in Table 5 in Appendix F) are qualitatively in line with those on **TaxonomiGQA**: VLMs were substantially better than their LM counterparts in 5 out of 7 VLM-LM pairs. Furthermore, concurrent work by Tan et al. [63] that also investigates taxonomic understanding in minimally differing LMs and VLMs corroborates our finding. In their results, VLMs (with the exception of LlaVa-OneVision, and to a milder extent, InternVL-8B) showed improved performance on 4 out of 5 taxonomic understanding benchmarks.

4 H1: VLMs' taxonomic knowledge aligns better with reference taxonomy

One possible reason that VLMs are performing better on **TaxonomiGQA** could be due to the difference in their underlying (task-agnostic) taxonomic knowledge, and in particular, in a way that better aligns with the reference taxonomy used to create the hypernym-substituted questions. We test this hypothesis about taxonomic knowledge difference in three different ways: (1) through a QA task that directly elicits taxonomic judgments; (2) through an analysis of the hierarchical organization of concepts in the models' representation space; and (3) through similarity analysis on the embeddings.

[1] For the curious: see Appendix E for how the VL models perform on our text-only QA vs. VQA.

4.1 Directly eliciting taxonomic judgments through Taxonomic Minimal Pairs (TAXOMPS)

Since **TaxonomiGQA** *presupposes* taxonomic knowledge rather than eliciting it directly, we first checked whether VLMs and LMs differed in their ability to directly answer questions about taxonomic relations. To this end, we introduce **TAXOMPS** (Taxonomic Minimal Pairs), a dataset which consists of questions of the form "*Is it true that a* C_1 *is a* C_2*?*" where C_1 (*cat*) and C_2 (*feline*) are concepts that are in a hypernymy relation, and negative samples where C_2 is replaced by a concept that is not the hypernym of C_1 (*vehicle*), following Misra et al. [43]. We constructed **TAXOMPS** directly from the final taxonomy used in our **TaxonomiGQA** analysis—i.e., 276 total hyponym-hypernym pairs, each coupled with 4 negative samples (same as in **TaxonomiGQA**), yielding 1380 questions. We use Overall Accuracy as our performance measure (since there is no conditional analog), following Section 3. That is, an instance is considered correct iff. the model answers questions with the hyponym-hypernym pairs (*Is it true that a* ***cat*** *is an* ***animal****?*) with a `Yes` while answering `No` to the negative sample questions (*Is it true that a* ***cat*** *is a* ***vehicle/fruit/tool/vegetable****?*).

Figure 2 shows our results. With the exception of Llama-3.1 vs. MLlama 3.2, most VLM-LM pairs perform quite similarly (and well) on **TAXOMPS**. This suggests that additional VL training does not in general alter the basic taxonomic membership judgments of a language model.

4.2 Lexical representations of taxonomic knowledge

Can the alignment with reference taxonomy be observed representationally, although not by direct elicitation? We tested whether the lexical representations in VLMs align better with the reference taxonomy than LMs via their hierarchical organization and hypernym-hyponym embedding similarity.

4.2.1 Hierarchical taxonomic structure

Park et al. [52] propose a method to analyze the latent hierarchical taxonomic structure of an LM, based on ideas including the linear representational hypothesis [39, 53] and causal separability of concepts [67], finding that taxonomic hierarchies (*dog* < *canine* < *mammal*...) are encoded as orthogonalities in LMs' transformed unembeddings. Therefore, one way we may observe the effect of VL training on the taxonomic knowledge of the LM is via differences in this hierarchical structure.

We applied Park et al. [52]'s method to transform the unembedding space in our models to a space where the inner product between two concepts' vectors is sensitive to the hierarchical relation between them. Then, we compared VLM-LM pairs in terms of their pairwise cosine similarities between concepts in their unembeddings' causal inner product space (as established in [53]). In addition, we used the large WordNet hierarchy (a superset of our taxonomy) originally used by [52] to compare the pairwise similarities of concepts in VLM and LM to that of the pairwise path-similarities between concepts in WordNet. We conducted these comparisons using Representational Similarity Analysis [25], which computes the Spearman's correlation between two matrices' (flattened) upper triangular matrices, treating it as the representational similarity between the two spaces. We conducted RSA between three representational spaces: VLM, LM, WordNet. Greater RSA value between two spaces indicates greater similarity between. To account for potential variance, we sampled 100 subsets (of size 100×100 each) from the full pairwise matrices and report the mean and standard deviation of the RSA correlations across all subsets.

This analysis (Table 1, left) shows that the hierarchical organization of concepts (as defined by [52]) is mostly shared between the VLM and LM, indicated by the consistently similar RSA scores when comparing VLMs and LMs to WordNet, as well as the high similarity between the VLM and LM when directly compared (all RSA scores ≥ 0.95). Interestingly, the Qwen 2.5 and Molmo pairs, the two model pairs that showed the most salient advantage of VLMs in Figure 2 had the lowest VLM-LM RSA scores: 0.95 and 0.96, respectively. However these values are still very high in terms of raw correlation, suggesting that they are still fundamentally similar. The pairwise similarities for VLMs, LMs, and WordNet can be visually inspected in Figure 5 in Appendix G.

4.2.2 Embedding similarities of taxonomic relations

Another way in which taxonomic relations can be investigated is via vector similarity—we tested whether the lexical embeddings (i.e., uncontextualized representations) corresponding to concepts in our reference taxonomy are more similar to embeddings of their hyponyms, relative to embeddings

Table 1: **Left:** RSA comparisons of hierarchically sensitive pairwise similarities [52] in the unembedding spaces of VLM-LM pairs, and pairwise path-similarities from the WordNet (WN) Noun Hierarchy. Subscripts show standard deviation (hidden if under 0.01). **Right:** Differences (Δ) in cosine similarities between positive concept pairs (i.e., in a hypernymy relationship) and negative samples from the taxonomy in **TaxonomiGQA**, computed using VLM and LM static-embedding layers.

Minimal Pairs	**RSA using** Park et al. [52]			**Raw Embeddings**			
	(VLM, WN)	(LM, WN)	(VLM, LM)	Δ_{VLM}	Δ_{LM}	t	p
Vicuna vs. Llava-1.5	$0.43_{\pm 0.04}$	$0.43_{\pm 0.04}$	0.99	0.02	0.02	1.09	0.27
Mistral-v0.2-I vs. Llava-Next	$0.42_{\pm 0.04}$	$0.42_{\pm 0.03}$	0.99	0.04	0.04	1.19	0.23
Qwen2.5-I vs. Qwen2.5-VL-I	$0.38_{\pm 0.05}$	$0.39_{\pm 0.04}$	**0.95**	**0.03**	**0.04**	-7.51	$<.001$
Llama-3.1 vs. MLlama-3.2	$0.40_{\pm 0.04}$	$0.41_{\pm 0.04}$	0.99	0.04	0.04	1.34	0.18
Qwen2-I vs. Llava-OV	$0.40_{\pm 0.04}$	$0.40_{\pm 0.05}$	0.99	0.06	0.06	0.82	0.41
Qwen2 vs. Molmo-D	$0.38_{\pm 0.04}$	$0.39_{\pm 0.04}$	**0.96**	0.05	0.05	-	-
Llama-3.1-I vs. MLlama-3.2-I	$0.40_{\pm 0.04}$	$0.40_{\pm 0.04}$	0.99	0.04	0.04	-0.09	0.92

of non-hyponyms in our taxonomy. We computed the similarity between each target concept and its hyponym, as well as between the target concept and four randomly sampled non-hyponym concepts (same as in Section 3.1). Then, we computed the difference between target-hyponym similarity and the average similarity between the target and the negative samples. We tested whether this difference is greater in VLMs than LMs, which would mean that VLM embeddings encode hypernym-hyponym relations more similarly than non-hypernym-hyponym relations. This hypothesis is *not* borne out: Table 1 (right) shows that this holds for no VLM-LM pairs (the significant effect in Qwen2.5-I vs. Qwen2.5-VL-I is in the opposite direction).

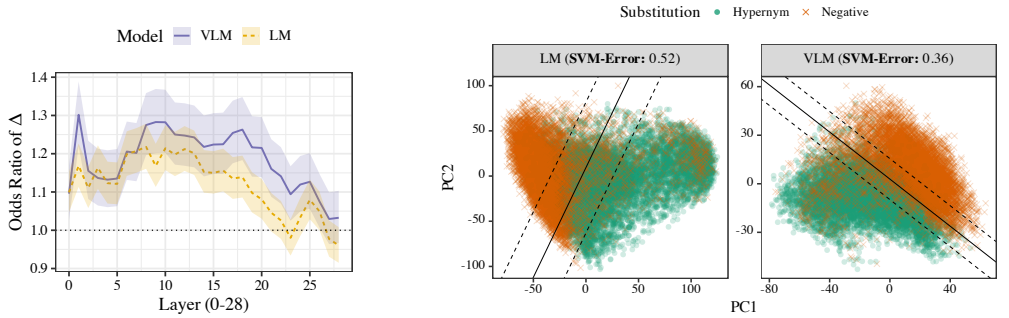
5 H2: VLMs are better at deploying taxonomic knowledge

As mentioned in Section 4.1, solving a downstream task presupposes the domain knowledge recruited, and requires this knowledge to be correctly deployed in the context of the specific task. Hence, solving **TaxonomiGQA** requires (1) taxonomic knowledge and (2) its deployment specifically for scene description-based QA. Our analyses in the previous section did not show convincing evidence in support of the hypothesis that the underlying taxonomic knowledge differs substantially in our VLM-LM pairs. In light of this mostly negative result, we turn to our second hypothesis: VLMs are better at *deployment* of taxonomic knowledge. To test whether there is a difference when taxonomic knowledge is incorporated into the specific task context, we used *contextual* similarity of lexical representations and a Principal Component Analysis (PCA) of representations of questions. These analyses let us examine both the contextualized lexical representations as well as the holistic representation of the full question context. We used the Qwen2.5 pair in both analyses.

Data To control for the confounding effect of the target label (`Yes`/`No`) when analyzing contextualized representations, we used a subset of **TaxonomiGQA** that has the same ground truth label for both positive and negative samples. In our dataset, this only includes cases where the ground truth answer is `No`. We further filtered this dataset to instances where the models got the original, unsubstituted question right, and used the models' Conditional Accuracy on substituted questions as the target of study. This gave us us 37,790 and 40,145 samples for Qwen2.5-I and Qwen2.5-VL-I, respectively.

5.1 Contextualized representation similarity

Our first analysis aims to relate the behavioral outcome of a model for each question to the representational structure of the concepts in context. To this end, we investigated the contextualized representations of a target concept in the scene description in terms of their similarity to representations of its hypernym in the question (e.g., *There is a* ***dog***$_{hypo}$ *on a yellow surfing board [...]. In the scene, are there any* ***mammals***$_{hyper}$*?*). The quantitative hypothesis is that greater contextualized hypernym-hyponym similarity (e.g., sim(**dog**, **mammal**) compared to hyponym similarity with negative samples (e.g., sim(**dog**, **fruit**) from *There is a* ***dog***$_{hypo}$ *[...]. In the scene, are there any* ***fruits***$_{neg}$?) would predict how well the model can answer the **TaxonomiGQA** questions. We used the 4 negative samples from **TaxonomiGQA**, and then fit a logistic regression model to predict model



(a) Odds ratio of the difference between cosine similarity of hyponym and hypernym, and max cosine similarity of hyponym and negative samples (Δ) in estimating model correctness, across layers.

(b) PCA projections of the last hidden state representations of questions containing hypernym (●) vs. negative sample substitutions (×), extracted from Qwen2.5-I (LM) and Qwen2.5-VL-I (VLM), along with separation lines fitted using an SVM classifier. SVM Error denotes the margin error *and* the classification error.

Figure 3: Contextualized representational analysis on Qwen2.5-I and Qwen2.5-VL-I.

correctness (measured using the `correct()` function from Section 3.2) using the difference in cosine similarity between hypernym-hyponym, and maximum[2] cosine similarity of 4 hyponym-negative sample pairs. Here, an odds ratio (of the difference term) being greater than 1 indicates that the similarity of the hyponym to hypernym (relative to negative samples) is more strongly associated with the model correctly answering questions, while the opposite is true if the odds ratio is less than 1. We performed this analysis using representations from every layer in the Qwen2.5 model pair, and took the maximum similarity in cases where the hyponym is mentioned in the scene more than once.

Figure 3a shows the layerwise odds ratios of the difference in similarities between concept pairs (sim-Δ; discussed above), in predicting model correctness, for both models. For most layers, we see odds ratios greater than 1.0, indicating a positive association between sim-Δ and model correctness for both model classes. At the same time, the VLM odds ratios are often greater than those of the LM, with the LM odds ratios sometimes even veering off below the 1.0 level (which would suggest an association of sim-Δ with *wrongness* as opposed to correctness. Overall, this suggests that VL training helps establish stronger connections between model representations and behavior in task contexts requiring deployment of taxonomic knowledge.

5.2 Principal Component Analysis (PCA) of question representations

Like in the previous analysis, we focused on the distinguishability of hypernym-hyponym relations from non-hypernym-hyponym relations, but considered whether this is captured in the representation of the question context from data used in the previous section. Following Alhamoud et al. [1] (who tested negation sensitivity in VLMs), we took the last hidden state of the final layer of the text decoder to be the summary representation of the full context. Then we asked whether representations of questions that contain a hypernym-hyponym relation (e.g., *dog-canine*) are separated from representations of questions that contain a non-hypernym-hyponym relation (e.g., *dog-bird*) via PCA.

Figure 3b shows the first two principal components (PCs) of the question representations from the VLM & LM, with hypernym (green) vs. negative sample substitutions (orange) color coded. We see that the two types are largely visually distinct in both models, suggesting that their question representations do encode differences in terms of the taxonomic relations tested. To quantify (linear) separability, we fit a soft-margin support vector machine (SVM) classifier [7] on the first two principal components of the representations extracted from each model separately, and measured its error on the PC-representations—greater the error, the poorer the separability. We found that the SVM error of the PCs of VLM representations is substantially lower than that of the LM, demonstrating that taxonomic distinctions are more linearly separable in VLM question representations. This complements the results from the previous analysis of contextualized embeddings, and suggests genuine differences in the representational states of the VLM and LM when the task contexts require taxonomic reasoning.

[2]We note that taking the average instead of maximum results in substantially weaker trends.

5.3 On the distinction between knowledge possession and deployment

Collectively, our results highlight that VL training does not change the underlying taxonomic knowledge within LMs, but rather affects its *deployment* in task contexts that require sensitivity to taxonomic knowledge. We see two specific reasons why this distinction could be important.

First, storing or representing knowledge differs from learning its "functional consequences" [46]. A model may robustly encode category information (e.g., that *robins are birds*), yet fail to recruit this knowledge when the context demands it. **TaxonomiGQA** is designed precisely to probe this aspect of deployment: in order to be successful, a model must not only *store* taxonomic knowledge, but *use* it appropriately when answering questions. Practically, teasing apart knowledge possession and deployment can inform decisions about (post-)training data selection: if a model's limitations stem from representational gaps, additional encyclopedic knowledge may help; if the issue lies in deployment, more diverse contextual supervision, involving contexts in which such knowledge is recruited, could be more effective. This distinction can also motivate solutions; for instance transferring task vectors from models that are better at deployment [16].

Second, the distinction has implications for cognitive and philosophical interpretations of multimodal learning—in particular, for drawing appropriate conclusions about the roles of linguistic versus (added) extralinguistic exposure. For instance, the platonic representation hypothesis [20] suggests that models trained on sufficiently large amounts of data converge toward similar internal representations, irrespective of modality. Our findings provide a complementary perspective to this hypothesis. While independent unimodal models might converge to similar taxonomic representations, combining information from multiple modalities can result in non-trivial changes that go beyond representational convergence (in our case, in terms of how knowledge is accessed and deployed).

6 *Why* might vision training help?

Our analyses so far have pinpointed *where* the meaningful behavioral and representational differences lie in the context of a taxonomic task when comparing a VLM-LM pair. However, we have not discussed *why* vision training would be beneficial. We present a preliminary investigation here, hypothesizing that visual similarity between members of concepts in a hypernym-hyponym relation is helpful information that VLMs can leverage. Some examples would be the visual similarity of members of *equine* and *horse* or *root vegetable* and *radish*. Of course, visual similarity will not be informative cues for *all* such relations, e.g., it would not be very helpful in better understanding the relation between *vertebrates* and its hyponyms, since there are few salient visual features shared by members of *vertebrate* (e.g., *fish, mammal, amphibian...*). This motivates a hypothesis that links visual information to model performance: high visual similarity between members of a hypernym and its hyponym would have a positive effect on model performance on questions probing that relation, but the effect would substantially vary depending on the target concepts.

Method To test this hypothesis, we first estimated hypernym-hyponym visual similarities by computing the cosine similarity between the image representation of a leaf node object and the image representations of other objects within its parent node (i.e., its hypernym) for concepts in our taxonomy. The image representations are extracted from the target VLM's (Qwen2.5-VL-I) vision encoder. Importantly, the images themselves are sourced from an independent dataset (THINGS [15]) so that our conclusion is not tied to specific images in GQA. Rather, they are intended as estimates of visual similarity more broadly. More details about the image similarity computation is in Appendix H. Then, we tested the extent to which this similarity predicts Conditional Accuracy of the VLM on hypernym questions where it outperforms its LM counterpart, using a linear mixed-effects model. Specifically, we predicted Conditional Accuracy of the VLM between each hyponym-hypernym pair using the pair's visual similarity as a fixed effect, and included random slopes and intercepts for each hypernym (model formula: $\texttt{cond\_acc} \sim \texttt{viz\_sim} + (1 + \texttt{viz\_sim} \mid \texttt{hypernym})$).

Results We found a significant global effect of visual similarity in predicting Conditional Accuracy ($b = 0.52, \text{SE} = 0.19, p < .01$). These results are much weaker when using the text-only LM's Conditional Accuracy as the dependent variable ($b = 0.23, \text{SE} = 0.17, p = 0.18$), suggesting that image similarity captures interesting properties related to the success of the VLM and uniquely so for VLMs. We also found interesting hypernym-specific random effects, where the effect of similarity

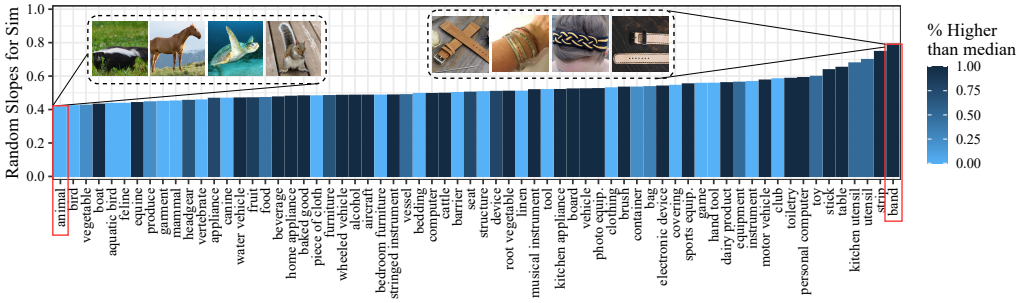


Figure 4: Hypernym-specific random effects of image similarity in predicting VLM accuracy on **TaxonomiGQA**. Greater values indicate closer association of visual similarity to model accuracy. Bar colors indicate percentage of hypernym-hyponym pairs that have above-median similarity.

varies greatly depending on the hypernym. Figure 4 shows the random slopes for each hypernym. This substantial individual variation aligns with our initial intuition that visual similarity would help some relations more than others. To quantify this intuition concretely, we annotated the higher level concepts in our taxonomy in terms of the % of the time the visual similarity to their hyponyms were above the median. This is meant to capture the difference between *equine* and *animal* we discussed earlier: members of *equine* are more similar to each other, more so than members of *animal* are (i.e., visually cohesive). The colors of the bars in Figure 4 are mapped to the degree of visual cohesion, where darker bars mean more cohesive. We see that the degree of cohesion generally lines up with effect sizes of similarity on predicting VLM performance, with mostly lighter bars on the left edge and darker bars on the right edge. The figure zooms into the concepts on either edge (left: *animal*, right: *band*), showing a sample of images corresponding to those concepts to illustrate the low visual cohesion of *animal* and high visual cohesion of *band*. Overall, the results present a promising lead into elucidating the source of improvement in VLMs, establishing a potential link between visual similarity, visual cohesion, and behavioral QA performance.

7 Conclusion

By building **TaxonomiGQA**, a text-only QA dataset that requires taxonomic understanding, we identified an interesting performance gap between VLM and their LM counterparts. That is, most VLMs consistently outperformed LMs under all metrics we adopted, despite this task being purely text based. We set out to pinpoint the source of this gain in VLMs. The first set of findings show that both behaviorally and representationally, there was no substantial difference between VLMs and LMs in their taxonomic knowledge, corroborating the general implications of [73, 66] that additional vision training does not fundamentally restructure the underlying knowledge. However, our second set of analyses show: (1) VLMs' contextual representation similarity of concepts in taxonomic relation in higher layers better predict success on **TaxonomiGQA**, and (2) VLM representations of questions containing taxonomic relations and questions that do not are better linearly separable, suggesting that VLMs have an advantage over LMs in adequately *deploying* taxonomic knowledge. We furthermore conducted a preliminary investigation on *why* vision training helps, testing the hypothesis that visual similarity of members in the extension set of hypernym/hyponyms help VLMs learn more useful representations of these words for taxonomic tasks. The results showed that VLMs' behavioral success on **TaxonomiGQA** can be predicted by visual similarity between members of concepts in a taxonomic relation, and the prediction strength is modulated by the visually cohesion of the hypernym.

Limitations and future work Our analyses do not provide causal evidence for the relation between behavior on **TaxonomiGQA** and the analyzed representations. Gaining causal evidence would require analyses more closely tied to the training data and objective, which is challenging due to the scale of the models as well as the scarcity of open data models. Additionally, a caveat to our results is that VL-tuned models do encounter more text-data in addition to visual supervision. This confound can be teased apart in future work by training LMs on the text-only portion of the VL training data. Furthermore, our SVM-based separability analysis is only applicable to taxonomic distinctions that are linearly encoded, leaving room for future work to extend this to non-linear separability.

Acknowledgments

We thank Yukyung Lee for her advice on creating better visualizations and Mahir Patel for earlier discussions on constructing the dataset, as well as for their general support. We also thank the anonymous NeurIPS reviewers and the Area Chair for helpful feedback. Pilot experiments for this work were conducted with the support of the PaliGemma Academic Program GCP Credit Award from Google awarded to the team. An in-person workshop partially dedicated to this work was funded by the Royal Netherlands Academy of Arts and Sciences (KNAW) Early Career Partnership. We acknowledge that the computational work reported on in this paper was performed on the Shared Computing Cluster which is administered by Boston University's Research Computing Services, as well as on the `burrata` machine at TTIC. Kanishka Misra is supported by the Donald D. Harrington Faculty Fellowship at UT Austin.

References

[1] Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29612–29622, 2025.

[2] Morris Alper and Hadar Averbuch-Elor. Emergent visual-semantic hierarchies in image-text representations. In *European Conference on Computer Vision*, pages 220–238. Springer, 2024.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv:2502.13923*, 2025.

[5] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.703. URL https://aclanthology.org/2020.emnlp-main.703/.

[6] Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross A Knepper, and Yoav Artzi. Learning to map natural language instructions to physical quadcopter control using simulated flight. *3rd Conference on Robot Learning (CoRL)*, 2019.

[7] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine learning*, 20(3): 273–297, 1995.

[8] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025.

[9] Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. Modality-aware integration with large language models for knowledge-based visual question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2417–2429, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.132. URL https://aclanthology.org/2024.acl-long.132/.

[10] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Can we talk models into seeing the world differently? In *Thirteenth International Conference on Learning Representations*, 2025.

[11] Deepanway Ghosal, Navonil Majumder, Roy Lee, Rada Mihalcea, and Soujanya Poria. Language Guided Visual Question Answering: Elevate Your Multimodal Language Model Using Knowledge-Enriched Prompts. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12096–12102, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.809. URL https://aclanthology.org/2023.findings-emnlp.809/.

[12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv:2407.21783*, 2024.

[13] Simon Jerome Han, Keith J Ransom, Andrew Perfors, and Charles Kemp. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155, 2024.

[14] Michael Hanna and David Mareček. Analyzing BERT's knowledge of hypernymy via prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 275–282, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.blackboxnlp-1.20. URL https://aclanthology.org/2021.blackboxnlp-1.20.

[15] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

[16] Roee Hendel, Mor Geva, and Amir Globerson. In-Context Learning Creates Task Vectors. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

[17] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[18] Douglas R. Hofstadter and Emmanuel Sander. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic books, 2013.

[19] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[20] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic Representation Hypothesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BH8TYy0r6u.

[21] Taichi Iki and Akiko Aizawa. Effect of Visual Extensions on Natural Language Understanding in Vision-and-Language Models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2189–2196, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.167. URL https://aclanthology.org/2021.emnlp-main.167/.

[22] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv:2310.06825*, 2023.

[23] Chenchen Jing, Yunde Jia, Yuwei Wu, Xinyu Liu, and Qi Wu. Maintaining Reasoning Consistency in Compositional Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2022.

[24] Najoung Kim, Sebastian Schuster, and Shubham Toshniwal. Code Pretraining Improves Entity Tracking Abilities of Language Models. *arXiv:2405.21068*, 2024.

[25] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:249, 2008.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[27] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

[28] Brenden M. Lake, Wojciech Zaremba, Rob Fergus, and Todd M. Gureckis. Deep neural networks predict category typicality ratings for images. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 37, 2015.

[29] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[31] Zhenyang Li, Yangyang Guo, Kejie Wang, Yinwei Wei, Liqiang Nie, and Mohan Kankanhalli. Joint answering and explanation for visual commonsense reasoning. *IEEE Transactions on Image Processing*, 32:3836–3846, 2023.

[32] Ruixi Lin and Hwee Tou Ng. Does BERT know that the IS-a relation is transitive? In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 94–99, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.11. URL https://aclanthology.org/2022.acl-short.11/.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

[34] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[35] Joan Lucariello, Amy Kyratzis, and Katherine Nelson. Taxonomic knowledge: What kind and when? *Child development*, 63(4):978–998, 1992.

[36] Caroline Lyon, Chrystopher L. Nehaniv, Joe Saunders, Tony Belpaeme, Ambra Bisio, Kerstin Fischer, Frank Förster, Hagen Lehmann, Giorgio Metta, Vishwanathan Mohan, et al. Embodied language learning and cognitive bootstrapping: methods and design principles. *International Journal of Advanced Robotic Systems*, 13(3):105, 2016.

[37] Avinash Madasu and Vasudev Lal. Is multimodal vision supervision beneficial to language? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2642, 2023.

[38] Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. End-to-end image captioning exploits distributional similarity in multimodal space. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 381–383, Brussels, Belgium, November

2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5455. URL https://aclanthology.org/W18-5455/.

[39] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

[40] George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38 (11):39–41, 1995.

[41] Kanishka Misra. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv:2203.13112*, 2022.

[42] Kanishka Misra, Allyson Ettinger, and Julia Rayz. Do language models learn typicality judgments from text? In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*, 2021.

[43] Kanishka Misra, Julia Rayz, and Allyson Ettinger. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.213. URL https://aclanthology.org/2023.eacl-main.213/.

[44] Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2331–2350, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.127. URL https://aclanthology.org/2024.acl-long.127/.

[45] Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. Are large language models good at lexical semantics? A case of taxonomy learning. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1498–1510, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.133/.

[46] Gregory L. Murphy. *The Big Book of Concepts*. MIT press, 2002.

[47] Jerry Ngo and Yoon Kim. What do language models hear? Probing for auditory representations in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5435–5448, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.297. URL https://aclanthology.org/2024.acl-long.297/.

[48] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv:2504.02821*, 2025.

[49] Letitia Parcalabescu and Anette Frank. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.223. URL https://aclanthology.org/2023.acl-long.223/.

[50] Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.329. URL https://aclanthology.org/2024.acl-long.329/.

[51] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.567. URL https://aclanthology.org/2022.acl-long.567/.

[52] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv:2406.01506*, 2024.

[53] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=UGpGkLzwpP.

[54] Karalyn Patterson and Matthew A. Lambon Ralph. The hub-and-spoke hypothesis of semantic memory. In *Neurobiology of language*, pages 765–775. Elsevier, 2016.

[55] Karalyn Patterson, Peter J. Nestor, and Timothy T. Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, 8 (12):976–987, 2007.

[56] Juan Diego Rodriguez, Aaron Mueller, and Kanishka Misra. Characterizing the role of similarity in the property inferences of language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11515–11533, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.574/.

[57] Timothy T. Rogers and James L. McClelland. *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT press, 2004.

[58] Chen Shani, Jilles Vreeken, and Dafna Shahaf. Towards concept-aware large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13158–13170, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.877. URL https://aclanthology.org/2023.findings-emnlp.877/.

[59] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of caption. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7454–7464, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.664. URL https://aclanthology.org/2020.acl-main.664/.

[60] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, June 2022.

[61] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005.

[62] Vésteinn Snæbjarnarson, Kevin Du, Niklas Stoehr, Serge Belongie, Ryan Cotterell, Nico Lang, and Stella Frank. Taxonomy-aware evaluation of vision-language models. *arXiv:2504.05457*, 2025.

[63] Yuwen Tan, Yuan Qing, and Boqing Gong. Vision LLMs are bad at hierarchical visual understanding, and LLMs are the bottleneck. *arXiv:2505.24840*, 2025.

[64] Gabriella Vigliocco, Pamela Perniss, and David Vinson. Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130292, 2014.

[65] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv:1804.07461*, 2018.

[66] Wentao Wang, Wai Keen Vong, Najoung Kim, and Brenden M. Lake. Finding structure in one child's linguistic experience. *Cognitive science*, 47(6):e13305, 2023.

[67] Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=SGlrCuwdsB.

[68] Tz-Ying Wu, Chih-Hui Ho, and Nuno Vasconcelos. Protect: Prompt tuning for taxonomic open set classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16531–16540, 2024.

[69] Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FrFQpAgnGE.

[70] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report, 2024. *arXiv:2407.10671*, 2024.

[71] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2024.

[72] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542, 2016.

[73] Tian Yun, Chen Sun, and Ellie Pavlick. Does vision-and-language pretraining improve lexical grounding? In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.370. URL https://aclanthology.org/2021.findings-emnlp.370/.

[74] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

Table 2: Overview of model configurations used in our experiments, including size, modality, and model training details. The "Training Details" columns indicate whether the model was pretrained on GQA, trained with video data, or instruction-tuned. A checkmark (✓) denotes the presence of the corresponding training signal, (✗) indicates its absence, a bold question mark (**?**) represents unknown or unclear training status, and a blank cell indicates that the category is not applicable. Hugging Face identifiers are provided for reproducibility.

Model	**Size**	**Modality**	**Training Details**			**HF Identifier**
			GQA Pretrained	Video Involved	Instruction Tuned	
Qwen2	7B	L			✗	`Qwen/Qwen2-7B`
Molmo-D	7B	VL	✗	✗	✓	`allenai/Molmo-7B-D-0924`
Llama-3.1	8B	L			✗	`meta-llama/Llama-3.1-8B`
MLlama-3.2	11B	VL	**?**	**?**	✗	`meta-llama/Llama-3.2-11B-Vision`
Vicuna	7B	L			✓	`lmsys/vicuna-7b-v1.5`
LLaVA-1.5	7B	VL	✓	✗	✓	`llava-hf/llava-1.5-7b-hf`
Qwen2-I	7B	L			✓	`Qwen/Qwen2-7B-Instruct`
LLaVA-OV	7B	VL	✓	✓	✓	`llava-hf/llava-onevision-qwen2-7b-ov-hf`
Mistral-v0.2-I	7B	L			✓	`mistralai/Mistral-7B-Instruct-v0.2`
LLaVA-Next	7B	VL	✓	✗	✓	`llava-hf/llava-v1.6-mistral-7b-hf`
Llama-3.1-I	8B	L			✓	`meta-llama/Llama-3.1-8B-Instruct`
MLlama-3.2-I	11B	VL	**?**	**?**	✓	`meta-llama/Llama-3.2-11B-Vision-Instruct`
Qwen2.5-I	7B	L			✓	`Qwen/Qwen2.5-7B-Instruct`
Qwen2.5-VL-I	7B	VL	✗	✓	✓	`Qwen/Qwen2.5-VL-7B-Instruct`

A Selected Model Pairs

Table 2 shows a list of model pairs used in this work, along with their metadata – parameters, modality, huggingface identifier, etc.

B Extended Related Work

Multimodal Semantic Representations in Humans and Language Models. A central question in cognitive science and linguistics is how humans integrate perceptual and linguistic signals to form generalizable mappings from semantic and conceptual knowledge to language. Research exploring the cognitive and neural underpinnings of such knowledge supports the idea that language learning and processing is inherently multimodal, grounded in visual, motor, and affective experience [61, 64]. At the neural level, conceptual knowledge is proposed to be coordinated by a transmodal "semantic hub," allowing humans to flexibly attend to the modality that provides the most informative cue in context and to abstract over modality-specific input [55, 54]. Several NLP tasks now commonly employ multimodal representations [5], notably image captioning [59, 38] and visual commonsense reasoning [74, 31]. In embodied agents, linking physical actions to explicit linguistic representations has been shown to facilitate more effective concept learning [36, 6].

Computational representations can be optimized by identifying and exploiting semantic structure shared across modalities. Models trained on different modalities and objectives may converge on similar representations as they grow in size, forming a "platonic" structure shaped by statistical correlations across input that is modality agnostic [20]. Unified representations and architectures have been argued to better support multimodal processing and reasoning by reshaping how models reference and access perceptual and linguistic features, both reflecting the "semantic hub" structure found in humans and partially mitigating common biases found in unimodal models [60, 10, 69]. These methods can enable the implicit grounding of language in perceptual features such as spatial awareness and sound, even in text-only models [47].

For vision and language modalities, the Visual Question Answering (VQA) task [3] has inspired work on joint language and image understanding using on compositionality, consistency metrics, and knowledge-enriched prompting [23, 11, 9]. Focused benchmarks like VALSE [51], which tests

models' ability to ground linguistic phenomena in the visual modality, and interpretability methods such as MM-SHAP [49] and CC-SHAP [50], measure how VLMs integrate and rely on visual versus textual information. Findings show that VLMs often underuse visual input for reasoning, yet rely on it more heavily for generating explanations, especially in chain-of-thought (CoT) settings. These findings highlight that contributions of each modality in VLMs are uneven and task-dependent, challenging assumptions of uniform multimodal integration. An open question thus remains as to whether multimodal training indeed changes conceptual content, or instead how that content is accessed and applied. Our research explores this in a unique setting where VLM/LM minimal pairs share the textual component.

C Taxonomy Filtering and Annotation

Before manual annotation and removal of specific chains, we first identified 52 highly abstract concepts[3] (e.g. *entity, conveyance, act*, etc.) to be removed from all chains. After generating the initial hypernym chains, we conducted a second round of manual inspection to identify and address cases of "non-ideal" categorizations, defined as instances where either (a) the assigned hypernym was not the canonical category of the object (e.g., *bubble* categorized as *circle*), or (b) the chain consisted solely of abstract elements that were missed during the first filtering step. Through this process, we identified 611 problematic cases. Of these, 296 were corrected by querying WordNet for alternative hypernym chains, leaving 315 unfixable cases; these were subsequently removed (as reported in the main text).

D Negative Sampling Details

Table 3: Nine question types in **TaxonomiGQA**. Each question type is illustrated with an example and the total number of instances of that category. Question types ending in "C" have "no" as the correct answer ("C" stands for counterfactual); all others have "yes" as the correct answer, consistent with the design of GQA.

Question Type	Sample Question	Counts
exist	Are there any dogs?	29030
existAttr	Are there any boats that are white?	16405
existAttrNot	Are there dogs in this scene that are not white?	15300
existAttrC	Do you see dogs that are white?	16010
existAttrNotC	Do you see a fork that is not silver?	16440
existThat	Are there any tables in the picture that are wooden?	20435
existThatNot	Is there a television in the image that is not off?	4120
existThatC	Is there a boat that is green?	19985
existThatNotC	Is there a watch in the image that is not on?	3670
existMaterial	Do you see a fence that is made of wood?	1750
existMaterialNot	Is there a bench that is not made of wood?	1465
existMaterialC	Are there any lace tablecloths?	1650
existMaterialNotC	Are there forks that are not made of metal?	1760

After dataset filtering, we identified 32 types of questions. 19 of them were excluded because substituting the object in the question with one not present in the scene could result in presupposition failures. For the remaining types, we sampled four negative objects for each question based on the following three criteria: the sampled argument is (1) not present in the scene, (2) not in the original argument's hypernym chain, and (3) associated with the same set of attributes as the original arguments defined in GQA metadata. Due to inconsistencies and errors in the GQA metadata, we manually[4] verified the attribute matches to ensure the naturalness and validity of each substitution.

[3]These concepts are listed in the `GQA Hierarchies First Pass.ipynb` notebook located under `notebooks/` in our repository.

[4]We remove attributes that introduce non-standard property attribution, such as "happy trees", "swimming flowers", "fluffy apple".

This process resulted in a final dataset consisting of 13 question types and reduced our taxonomy to 126 unique chains. Details of question types, examples, and statistics can be found in Table 3.

E VQA vs. Text and a Question-only Control

Table 4 shows results from evaluating VLMs on the original GQA questions across different formats: (1) the original VQA setup, conditioned on an image; (2) the Text setup, where they are conditioned on the scene description; and (3) a Question-only control where we condition them only on the question, without any context.

While it is difficult to compare between the VQA and the text setup, we see stark differences in the absolute values of the accuracies. The VLMs seem to answer the (positive sample, unsubstituted) questions with very high accuracy (sometimes near-perfect) relative to their performance on the subset of the VQA task we have used in this work. Next, the VLMs are substantially worse at the question-only baseline than they are in the text version, often times being closer to chance (50%). This question-only control is especially relevant for any potential concern readers might have about VQA data being present in the models' training set. Since models are largely worse off at these relative to the text version of the dataset, the potential presence of VQA in the model's training set is of little concern. One interesting observation here is that MLlama-3.2 (non-instruct tuned) performs similarly at the Question-only task and at the VQA task. This could suggest that it might not have been trained on VQA after all.

Table 4: Accuracies of VLMs on GQA questions when evaluated using standard VQA-based setup (i.e., with images) vs. Text (i.e., with scene descriptions), as well as a Question-only control (No image and no text). Evaluation data consists only the positive sample version of the dataset with unsubstituted questions taken verbatim from GQA [19]. Chance performance is 50%.

Model	**VQA**	**Text**	**Question-only**
Molmo-D	0.79	0.89	0.52
MLlama-3.2-I	0.79	0.92	0.58
MLlama-3.2	0.63	0.91	0.60
Llava-1.5	0.78	0.95	0.53
Qwen2.5-VL-I	0.81	0.98	0.49
Llava-Next	0.84	0.98	0.52
Llava-OV	0.87	0.99	0.60

F Supplementary Experiment on the Rodriguez et al. Dataset

Property inheritance testing datasets such as those introduced by Rodriguez et al. [56] involve attributing a novel, nonsense property (e.g., *is daxable*) to concepts given that their parents are attributed with it—e.g., *Given that birds are daxable, is it true that robins are daxable? Answer with Yes or No*. This particular dataset includes a robustness check in the form of a single negative sample per concept and includes 2016 positive samples, spanning 44 superordinate categories and 1281 subordinate categories.[5] We evaluated our model pairs on this dataset with minimal prompt tuning,[6] and found that, as shown in Table 5, all VLMs (except MLlama-3.2 and Llava-OV) outperformed their LM counterparts, reinforcing the robustness of the observed VLM $>$ LM trend on models' behavioral performances on taxonomic understanding within a QA setting.

[5]This dataset has a larger set of categories than TaxonomiGQA, as its taxonomy is not constrained by visual scenes.

[6]We added the quantifier "all" to the premise - e.g., *Given that **all** birds are daxable, is it true that robins are daxable? Answer with Yes or No*, to make the logical inference more coherent. This modification improved the QA accuracy significantly, particularly for Qwen2.5-VL-I and MLlama-3.2.

Table 5: Performance comparison between LMs and VLMs on the Rodriguez et al. dataset. Δ(VLM–LM) denotes the difference in accuracy between the VLM and its corresponding LM.

Pair	LM	VLM	Δ(VLM–LM)
Llama-3.1 vs. MLlama-3.2	0.50	0.68	**0.18**
Llama-3.1-I vs. MLlama-3.2-I	0.51	0.50	-0.01
Mistral-v0.2-I vs. Llava-Next	0.69	0.78	**0.09**
Qwen2 vs. Molmo-D	0.60	0.81	**0.21**
Qwen2-I vs. Llava-OV	0.84	0.81	-0.02
Qwen2.5-I vs. Qwen2.5-VL-I	0.78	0.83	**0.05**
Vicuna vs. Llava-1.5	0.50	0.74	**0.24**

G RSA Heatmaps

We depict heatmaps showing the pairwise cosine similarities computed for the transformed Unembedding vectors of the Qwen2.5 model pair, as well as pairwise path-similarity from WordNet, in Figure 5. The LM and VLM matrices look quite similar, while the WordNet matrices are more sparse, showing clearer depiction of hierarhical structure. We computed similar plots for all other models but left them out due to large file sizes. Full plots can be viewed on https://github.com/tinlaboratory/taxonomigqa.

H More Details about Visual Similarity Analysis

To compute visual cosine similarity between two nodes–a leaf node object (e.g., *dog*) and one of its hypernyms (e.g., *vertebrate*)– we first needed a sufficient number of images for both. We used images from THINGS [15], a dataset with 26,107 high-quality, manually curated object-centric images of 1,854 diverse object concepts. Since the taxonomy in THINGS is more coarse-grained than ours, we aligned the taxonomies through the following steps: (1) Identify intermediate nodes that are missing in THINGS (e.g., *vertebrate*); (2) Collect leaf node objects present in THINGS and prompt a large language model (GPT-4o and Gemini 2.5 Pro) to identify which of them can be classified under the given intermediate node (e.g., *vertebrate*); 3) Manually verify the correctness of the selected objects. After aligning the taxonomies, we obtained visual representations for each node in our taxonomy from the Qwen 2.5VL-7B Instruct model. To do so, we modified the model's forward pass to extract hidden states immediately after the `merger.ln_q` RMSNorm layer within the `Qwen2_5_VLPatchMerger` module, and before the `merger.mlp` layer. These intermediate hidden states served as patch-level embeddings, which we mean-pooled to produce a 1280-dimensional representation for each image. We then computed cosine similarities between the visual representation of the leaf node (e.g., *dog*) and each of its hypernyms (e.g., *vertebrate*) by taking the mean of all image embeddings for the intermediate category node—similar to the prototype approach in [28]—and compute pairwise cosine similarities with each image from the leaf node.

I Experimental Resources

Dataset filtering for **TaxonomiGQA** was performed using multithreaded processing across 8 CPU cores and completed in approximately 3 hours. Negative sampling was carried out on a single CPU core and took approximately 5 minutes.

Model Inference was conducted using vLLM[27]. Vision tasks were processed on a single NVIDIA A40 GPU (48GB) over 3 hours, while text-only tasks were run on two NVIDIA L40 GPUs (48GB each) for approximately 1.5 hours. Image representation extraction for Qwen2.5VL was also performed on a single A40 GPU and completed in roughly 2.5 hours. Static embeddings were computed in under 10 minutes on an L40 GPU.

TAXOMPS, RSA on unembedding layer vectors, contextualized representational similarity analysis, and PCA analysis were conducted on a single NVIDIA RTX6000 Ada (48GB) GPU, and took a total of 1 hour, 1.5 hours, 4 hours, and 1 hour, respectively. Representation extraction and **TAXOMPS** behavioral analyses were performed using the `minicons` library [41]. All plots were produced using the `ggplot2` library in the R programming language.

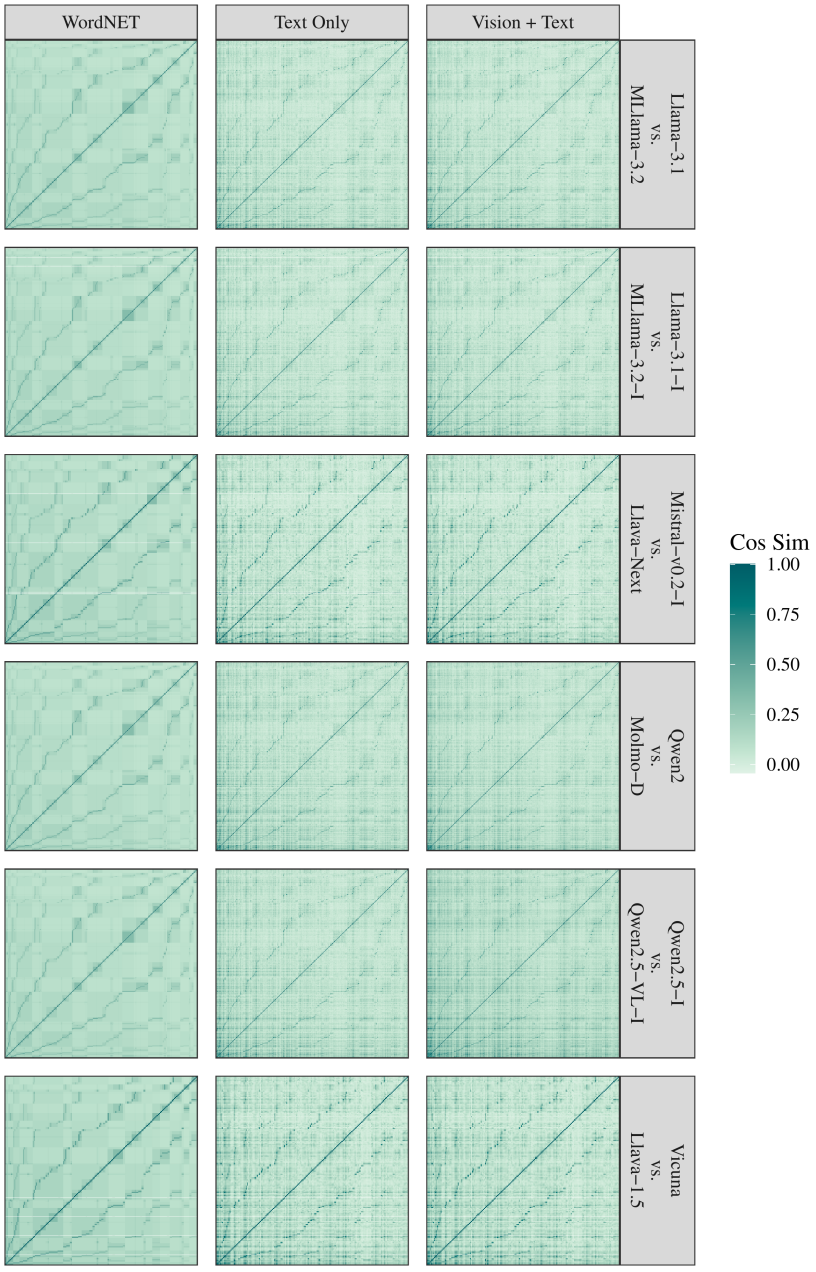


Figure 5: Pairwise similarities between concepts in WordNet, and the transformed unembedding spaces in Qwen2.5-I (LM) vs. Qwen2.5-VL-I (VLM) (computed using Park et al. [52]'s method), across all pairs.

We estimate that the total compute cost, including preliminary and unsuccessful experiments, is approximately **3x** the sum of the runtimes reported above.

J License Information

The original GQA dataset was released under CC BY 4.0 and we downloaded the dataset from https://cs.stanford.edu/people/dorarad/gqa/download.html. We follow this and release **TaxonomiGQA** under the same license, CC BY 4.0.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, they accurately reflect the paper's contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we have a separate Limitations and Future Work paragraph in the Conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the process of dataset creation and how the additional analyses were run in detail. We provide additional descriptions about the models used (e.g., Huggingface identifiers) and computing environments in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We submit our data and code as part of Supplemental Material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide experimental details necessary to understand the results in the main text and Appendix. Code is also submitted as part of Supplemental Material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide statistical analysis wherever appropriate.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we provide this detail in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics and confirm that there is no violation.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work primarily concerns evaluation and foundational understanding of existing models.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we credit the original creator of the asset we used to create our dataset and release ours under the same license as the original.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, the new assets will be released as a Github repository which would contain a full README for the dataset. The repository is submitted as part of Supplemental Material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subject experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM classification was used in data preparation for one of the analyses (combined with post-hoc human verification) and we discuss this in the Appendix.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.