VISaGE: Understanding Visual Generics and Exceptions

Stella Frank
University of Copenhagen, DK
stfr@di.ku.dk

Emily Allaway
University of Edinburgh, UK
emily.allaway@ed.ac.uk

Abstract

While Vision Language Models (VLMs) learn conceptual representations, in the form of generalized knowledge, during training, they are typically used to analyze individual instances. When evaluation instances are atypical, this paradigm results in tension between two priors in the model. The first is a *pragmatic prior* that the textual and visual input are both relevant, arising from VLM finetuning on congruent inputs; the second is a *semantic prior* that the conceptual representation is generally true for instances of the category. In order to understand how VLMs trade off these priors, we introduce a new evaluation dataset, VISaGE, consisting of both typical and *exceptional* images. In carefully balanced experiments, we show that conceptual understanding degrades when the assumption of congruency underlying the pragmatic prior is violated with incongruent images. This effect is stronger than the effect of the semantic prior when querying about individual instances.

1 Introduction

Vision-language models (VLMs) are typically used to analyze *instances*: what is going on in a particular image? Moreover, during training they learn a set of *conceptual* representations: generalized knowledge that holds over many instances. However, *exceptions* to generalizations, in the form of atypical instances, disrupt the alignment of in-context instance understanding and in-weights conceptual knowledge. While VLMs have been thoroughly tested on their ability to discern minimal differences between image instances (e.g., Johnson et al., 2017; Thrush et al., 2022; Tong et al., 2024), and likewise their conceptual representations (based on exposure to typical instances) have also been analyzed (e.g., Bruni et al., 2014; Silberer et al., 2013; Collell and Moens, 2016; Oneata et al., 2025), the potential tension between instance and concept representations, as arises in atypical instances, is currently under-explored.

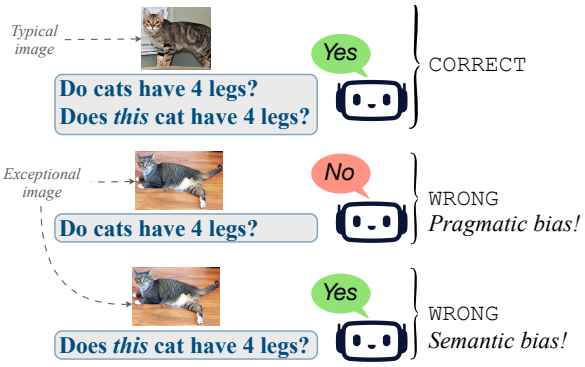

Figure 1: VISaGE contains both typical and exceptional instances of a category, with respect to some conceptual norm. We probe VLMs for conceptual and instance-level understanding, with congruent (top) and incongruent (middle, bottom) text-image pairs. Model failures indicate effects of both pragmatic and semantic biases.

In language, the attributes associated with a conceptual category are often expressed through *generics*: generalizations without quantifiers (e.g., *cats have four legs*). This lack of quantification means that generics remain true regardless of exceptions (tripod cats—cats missing one leg—do not impact the truth of *cats have four legs*). In other words, the attribute is associated as characteristic of the category regardless of how frequent it actually is.[1]

Unlike language, which can both denote on this generic or conceptual level, as well as refer to a particular instance, VLMs are always grounded in a particular visual instance. Work that has probed for conceptual attributes has used typical instances to stand in for the concept. This conflates instance and conceptual representations. In order to separate the two, visual *exceptions* are required: instances of a category that violate the generic (see Figure 1).

To this end, we introduce a new evaluation

[1]This is a substantial simplification of the semantics of generics (cf. Krifka, 1987).

dataset, **VisaGE:** Visual Generics and Exceptions, consisting of conceptual categories with images of both typical and exceptional instances. Specifically, exceptions are always with regard to a particular generic norm, i.e., a typical attribute: a `tripod cat` is an exception for *cats have four legs*, but is typical for *cats have a long tail*. The category-attribute pairs in VISaGE, along with their exceptions, are extracted from textual generics and carefully manually validated, together with the image instances.

Using VISaGE, we investigate two questions:

1. **(RQ1)** How does visual grounding to (potentially atypical) instances impact a model's ability to access *conceptual information*?
2. **(RQ2)** How does conceptual information, as recruited by text labels, impact VLMs' ability to recognize *instance attributes*?

These research questions examine the effects of two priors in VLMs. The first is a *pragmatic prior*, arising from VLM finetuning, that the textual and visual input are congruent and both relevant; the second is a *semantic prior* that the category-attribute generic is generally true.[2] In the exceptional image settings we explore with VISaGE, these two priors can conflict: when asking about conceptual knowledge (RQ1), the atypical image must be ignored and the semantic prior followed, while for instance queries (RQ2), given an atypical instance, the pragmatic prior to focus on the current context must overrule the semantic prior of typicality.

We test a set of contemporary open-weight VLMs and find evidence that their conceptual representations do not recognize possible variation in attributes. We observe that the models' pragmatic prior interferes with conceptual understanding (and the semantic prior) when visual grounding is incongruent with the text. This suggests that VLMs do not correctly differentiate between generic concepts and specific instances. Results are more mixed when models are tasked to recognize instantiations of exceptional attributes in images: while most models do show evidence of a semantic bias, this effect is less strong.

Our contributions are: (1) a new dataset, VISaGE, consisting of concept-attribute pairs with images of both typical (generic) and exceptional instances; (2) experimental evidence that VLM conceptual representations are visually grounded only in typical or generic instances and are not sufficiently robust to within-category variation.

2 Background

Previous work has investigated the semantics of generics with LMs (Ralethe and Buys, 2022; Collacciani et al., 2024; Cilleruelo et al., 2025). These studies show LMs often struggle to account for and reason about exceptions in both probing (Allaway et al., 2024) and reasoning (Allaway and McKeown, 2025) tasks. However, they have not considered generics in VLMs, particularly how visual grounding interacts with generic's semantics.

For evaluating VLMs, most visual benchmarks test situational and configurational instance understanding (Thrush et al., 2022; Li et al., 2024), sometimes with (synthetic) atypical examples (Bitton-Guetta et al., 2023). Although Saleh et al. (2013) create a small dataset of exceptional object images, these are not annotated with semantic attributes, unlike VISaGE. More recently, Luo et al. (2025); Vo et al. (2025) create datasets using synthetic image generation to manipulate object attributes.

Additionally, our experiments, in which we manipulate image–text congruency, contribute to a line of work investigating the relative importance of different modalities in VLMs (Gat et al., 2021; Frank et al., 2021; Fu et al., 2024; Tong et al., 2024; Li et al., 2024; Parcalabescu and Frank, 2025).

3 Dataset

Our dataset VISaGE is constructed by first collecting text pairs ($n_{c,a}$, $e_{c,a}$) where $n_{c,a}$ is a conceptual norm for category c with attribute a and $e_{c,a}$ is an exception to that norm (i.e., a subcategory of c that does not have the attribute a). Then for each pair, we retrieve two sets of images corresponding to cases where the norm applies (generic images V_c) and where it does not (exception images V_e). The resulting dataset then consists of tuples ($n_{c,a}$, $e_{c,a}$, V_c, V_e). Finally, we manually validate and expand the dataset (details in Appendix A).

VISaGE contains 1601 exceptional image examples for 437 exception subcategories, derived from 296 category-attribute relations (generics/conceptual norms) for 171 categories, balanced with the same number of typical images.[3]

Norm-Exception Text Pairs For our initial set of concept-attribute norms, we intersect the category-

[2] This is analogous to the Gricean maxims of relevance and quality (truthfulness) (Grice, 1975).

[3] Dataset and code at `github.com/scfrank/visage1601`

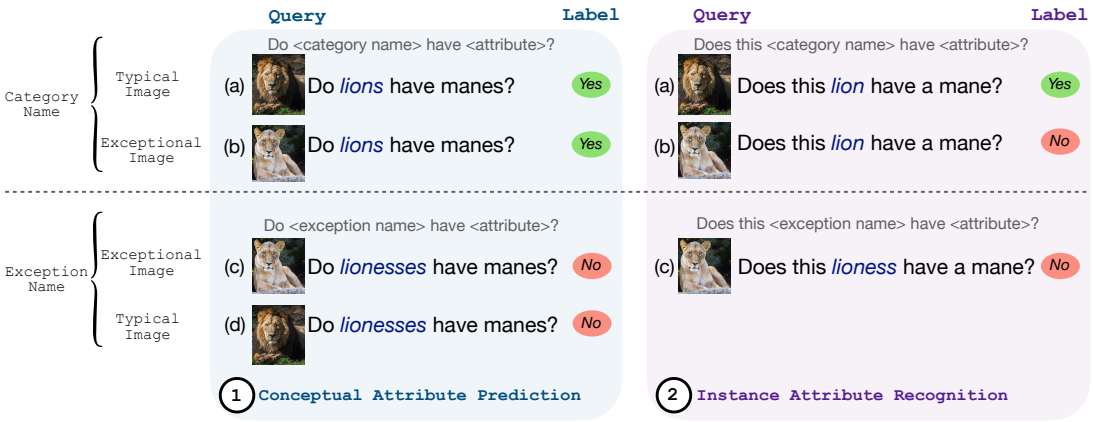

Figure 2: Summary of experiments and conditions: Exp. ① tests models' conceptual understanding of generics, while Exp. ② tests models' ability to reason about instances.

attribute lists of XCSLB (Devereux et al., 2014; Misra et al., 2022) and the McRae norms (McRae et al., 2005), with the categories in the THINGS object image dataset (Hebart et al., 2019). This results in a robust set of conceptual norms expressed as generics. Then, for each generic (category-attribute statement) we generate a set of exceptions $e_{c,a}$ using the LM prompting framework proposed by Allaway et al. (2024). We retain the short exceptions, ideally corresponding to subcategories.

Images We retrieve a large set of images for each exception subcategory using Bing Image Search by querying for the exception name $e_{c,a}$. Subsequent human validation (see below) selects the best images, resulting in a mode of 4 images per exception. A matched number of generic images for each category are taken from the THINGS dataset. THINGS images were specifically collected to be typical object instances; we further manually validate the applicability of the generic conceptual norms.

Validation We collect three types of validation annotations for each tuple. First, we validate that the images V_c retrieved from THINGS exhibit the conceptual norms $n_{c,a}$. Second, we validate that each $e_{c,a}$ is actually an exception to the norm $n_{c,a}$. With this we filter out exception subcategories that are hallucinated (e.g., `strawberry blonde cheetah`) or incorrectly related (e.g., not exceptional or not actually subcategories). Finally, we validate that the retrieved images V_e for each exception are correct. We exclude images that are the wrong category (e.g., images of Ryan Gosling retrieved for the category `gosling`) or that are the wrong style (e.g., not object-centered photographs).

4 Experiments

Using VISaGE, our experiments query VLMs about conceptual and instance attributes across a number of conditions: see Figure 2 for an overview. Specifically, we vary (1) the type of knowledge being queried (conceptual vs. instance); (2) the type of image input (typical vs exceptional images); and (3) the noun-phrase used to refer to the concept (category-name vs exception-name reference).

Models We test a suite of open-weights VLMs: these are listed in Appendix B. We use the `vllm` library to wrap our prompts[4] in the correct model-specific formats.

Evaluation We report the percentage of correct (yes/no) responses for each model, using the first token of the model output. Note that the correct response depends on the condition: see Figure 2.

4.1 Conceptual Attribute Prediction

Our first experiment ① targets RQ1 and examines how visual groundings impacts model responses to queries about conceptual information. Specifically, we use two pairs of conditions to investigate the impact of text-image congruency (see Fig. 2 ①). The first pair of conditions uses the category name in conceptual queries ("Do lions have manes?"): (1a) typical (congruent), and (1b) exceptional (incongruent) images. The second pair of conditions similarly queries conceptual information but about the *exception* subcategory ("Do lionesses have manes?"): (1c) exceptional (congruent), and (1d) typical (incongruent) images.

[4]Conceptual prompt template example: `Answer yes or no. Do {concept-pl} have {attribute}?`
Instance prompt template example: `Answer yes or no. Does this {concept-sg} have {attribute}?`

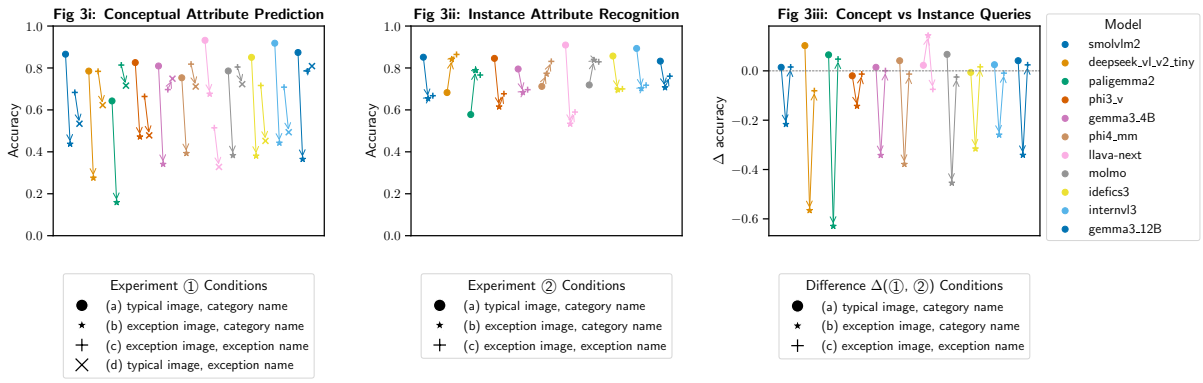

Figure 3: Results: See Fig. 2 for setup. Exp 1: Conceptual attribute prediction accuracy decreases for incongruent inputs (b and d). Exp 2: Across most models, instance attribute recognition declines for exceptional images (b), unless they are named as such (c). Fig 3iii shows data from Exps 1&2: The difference between conceptual and instance accuracy is highest for incongruent pairs (b). Most models have higher instance accuracy in condition (b). Models are ordered by size. Numbers in Appendix G.

Our results (Fig. 3i) show that VLMs' ability to answer conceptual questions degrades when the visual grounding is incongruent with the text input. That is, we observe a drop in accuracy from congruent to incongruent inputs in both pairs of conditions ((1a) → (1b) and (1c) → (1d)). This suggests that the *pragmatic prior*, i.e., assuming the image is relevant, is overriding the correct retrieval of conceptual knowledge.

Intriguingly, we also observe that incongruency in the input has less impact (interferes less) on accuracy when the queries are about the exception subcategory, and the incongruent image is of a typical instance, rather than vice-versa. This could be due to models' processing atypical images more attentively than typical images; we return to this in Section 4.3.

4.2 Instance Attribute Recognition

Our second experiment ② aims to answer RQ2 by examining how conceptual information impacts VLMs' ability to answer instance-specific queries. We specifically focus on the role of language-based conceptual activation; that is, can VLMs override conceptual generalizations to recognize specific attributes of individual instances. We compare category-name instance queries ("Does this lion have a mane?") in two conditions: with (2a) typical and (2b) exceptional images (see Fig. 2 ②). The third condition (2c) uses the *exception-name* in instance queries with exceptional images, providing an explicit language cue to the model to consider the exception rather than the category conceptual representation.

Our results (Figure 3ii) indicate that, despite instance queries directing the model to consider the image, many models still appear to ignore the visual features, relying instead on language-based conceptual cues. Specifically, we observe a drop in accuracy from condition (2a) to (2b). When the text and image are *in*congruent (in (2b)), the conceptual information activated by the category name (semantic prior) does not apply to the image, since the image is exceptional. Inasmuch as a model is relying on its semantic prior, this will result in a substantial drop in accuracy. Note that if the models instead prioritized using the visual features of the input, performance would be relatively stable across conditions.

Congruence alone does not explain the results: Despite the image and text being congruent in (2c), the accuracy is comparable to (2b), indicating that presenting the correct semantic information is of limited use for improving instance attribute recognition in atypical instances. One reason for this could be that the exception categories are lower frequency than the general categories, so exception attribute knowledge may be less well developed (see also (1a) vs. (1c)). In addition, many of the exception names include the category name (e.g., *lion*ess, tripod *cat*), which could lead to semantic priming of the general category and its attributes.

Not all models show semantic bias effects: some are more accurate when queried for exception images, rather than typical images. We speculate that these VLMs are sensitive to the (a)typicality of images and are also considering the pragmatics of the query. That is, since the query is asking about

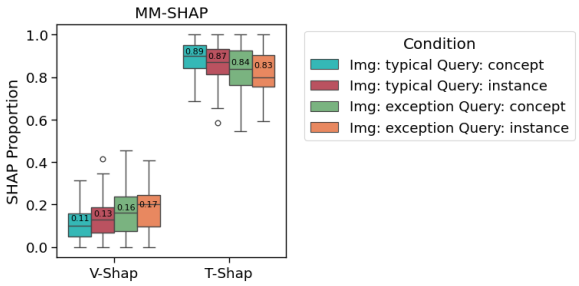

Figure 4: MM-SHAP calculated with `smolVLM2`. V-SHAP measures the proportion of Shapley values coming from the image, while T-SHAP is the proportion from text tokens.

the application of a conceptual norm to a specific instance, there must be something special about that instance: it might be expected to be a violation of the norm. This behavior would lead to lower accuracy in (2a), where the image isn't exceptional, and higher accuracy in both (2b) and (2c), which we observe with four of the VLMs we evaluated.

Finally, we observe (Figure 3iii) that the effect of pragmatic prior violation is stronger than violations of the semantic prior. When visual input and category name are congruent ((a) and (c)), models perform similarly for both conceptual and instance queries (difference is near-zero). In contrast, when models are required to disregard the image (conceptual queries), performance suffers significantly, compared to instance queries in which the semantic prior is incongruent with the exception image (condition (b): difference is negative). The accuracy difference that is visible *only* with incongruent inputs emphasizes the importance of considering exceptional examples to test how image instances interact with conceptual representations.

4.3 Feature Attribution with Shapley Values

We perform a Shapely value analysis (Shapley, 1952; Lundberg and Lee, 2017) in order to understand how different parts of the input contribute to model predictions. Specifically, does the image contribute more to the model prediction in the instance query condition, compared to the concept query condition?

In our setting, the Shapley value represents the contribution of a single input feature (e.g., a text token or the image, which we process as a single feature) to the model's prediction using the full input (i.e., the generation of a "yes" or "no" token in response to the query). Following MM-SHAP (Parcalabescu and Frank, 2023, 2025), we calculate the proportional contribution of the image input (V-SHAP) and the text tokens (T-SHAP) aggregated over the dataset. See Appendix E for details.

We calculate MM-SHAP values for a small model (`smolvlm2`) that behaved similarly to other models in our experiments, comparing instance and conceptual queries with typical and exception images. The results (Figure 4) do not show a qualitative difference in use of the image between concept and instance conditions, confirming a general pragmatic bias of attending to the image in all contexts. However, quantitatively, there is an increase in V-SHAP for each instance query condition, compared to the corresponding concept query condition (all differences are significant at $p < 0.01$ with a paired t-test; effect size, as measured with Cohen's d, between the two typical image conditions is $d = 0.3$, while for the exception image conditions $d = 0.19$, both indicating a small effect). Furthermore, conditions with exception images have higher V-SHAP compared to typical images ($p < 0.01$; $d = 0.62$ for concept queries and $d = 0.56$ for instance queries, both indicating a medium effect), suggesting that the models are recruiting visual information to supplement and possibly counteract the conceptual information from the text, in the cases where these are incongruent or atypical.

5 Conclusion

VLMs must balance learned priors with the requirements of the current context. With the use of a new dataset of visual exceptions, VISaGE, we have shown that VLMs have not yet solved this task: models neither reliably attend to the exception instance, ignoring the conceptual semantic prior, nor can they reliably ignore distractor images to answer generic conceptual queries.

Limitations

Our categories and attributes are limited to conceptual norms in American English. This is because the typical images we use for visual grounding (derived from THINGS) are based on American English definitions of categories. Conceptual spaces are language-dependent and different languages will make different conceptual distinctions, attending to different attributes. However, we believe the general patterns of results would hold across languages and models, since the distinction between instance-level and conceptual-level reasoning is common across languages.

The data collection process focused on quality rather than recall; we may have inadvertently omitted particular important exception types. In particular, exceptions that are rare, hard to see, or unlikely to be photographed, are missing (e.g., `insomniac owl` as an exception for *owls sleep in the day*, `cheetah with a broken leg` as an exception for *cheetahs are fast*).

Risks The concepts in our dataset correspond to concrete object categories. However, the difficulty of appropriately distinguishing (exceptional) instances vs. conceptual generalizations can also apply to categories that group people, where overgeneralization can lead to stereotyping. Understanding VLM capabilities and limitations is a step towards mitigating these risks.

Acknowledgments

This work was supported in part by the Pioneer Centre for AI, DNRF grant number P1 and an Edinburgh–Copenhagen Strategic Partnership Award.

References

Emily Allaway, Chandra Bhagavatula, Jena D. Hwang, Kathleen McKeown, and Sarah-Jane Leslie. 2024. Exceptions, Instantiations, and Overgeneralization: Insights into How Language Models Process Generics. *Computational Linguistics*, pages 1291–1355.

Emily Allaway and Kathleen McKeown. 2025. Evaluating defeasible reasoning in LLMs with DEFREASING. In *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 10540–10558, Albuquerque, New Mexico. Association for Computational Linguistics.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2616–2627, Paris, France. IEEE.

E. Bruni, N. K. Tran, and M. Baroni. 2014. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Gustavo Cilleruelo, Emily Allaway, Barry Haddow, and Alexandra Birch. 2025. Generics are puzzling. can language models find the missing piece? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6571–6588, Abu Dhabi, UAE. Association for Computational Linguistics.

Claudia Collacciani, Giulia Rambelli, and Marianna Bolognesi. 2024. Quantifying generalizations: Exploring the divide between human and LLMs' sensitivity to quantification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11811–11822, Bangkok, Thailand. Association for Computational Linguistics.

Guillem Collell and Marie-Francine Moens. 2016. Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences between Visual and Linguistic Representations. In *Coling*.

Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, 46(4):1119.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. 2024. IsoBench: Benchmarking Multimodal Foundation Models on Isomorphic Representations. In *Proceedings of the Conference on Language Modeling (COLM)*.

Itai Gat, Idan Schwartz, and Alex Schwing. 2021. Perceptual Score: What Data Modalities Does Your Model Perceive? In *Advances in Neural Information Processing Systems*, volume 34, pages 21630–21643. Curran Associates, Inc.

Herbert P. Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

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

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, Honolulu, HI. IEEE.

Manfred Krifka. 1987. An outline of genericity. In *Seminar für natürlich-sprachliche Systeme der Universität Tübingen*.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples. *Advances in Neural Information Processing Systems*, 37:17044–17068.

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2025. Probing Visual Language Priors in VLMs. In *Forty-Second International Conference on Machine Learning*.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Kanishka Misra, Julia Taylor Rayz, and Allyson Ettinger. 2022. A property induction framework for neural language models. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

Dan Oneata, Desmond Elliott, and Stella Frank. 2025. Seeing What Tastes Good: Revisiting Multimodal Distributional Semantics in the Billion Parameter Era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24174–24191, Vienna, Austria. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2025. Do Vision & Language Decoders use Images and Text equally? How Self-consistent are their Explanations? In *The Thirteenth International Conference on Learning Representations*.

Sello Ralethe and Jan Buys. 2022. Generic Overgeneralization in Pre-trained Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3187–3196, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Babak Saleh, Ali Farhadi, and Ahmed Elgammal. 2013. Object-Centric Anomaly Detection by Attribute-Based Reasoning. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–794, Portland, OR, USA. IEEE.

Lloyd S. Shapley. 1952. A Value for N-Person Games. Technical report.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of Semantic Representation with Visual Attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238, New Orleans, LA, USA. IEEE.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578, Seattle, WA, USA. IEEE.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2025. Vision Language Models are Biased. *Preprint*, arXiv:2505.23941.

A Dataset Construction

The McRae norms are conceptual norms elicited from humans (McRae et al., 2005). Devereux et al. (2014) builds on these in the XCSLB dataset and then (Misra et al., 2022) further revise them. Each norm can be expressed as a generic.

To generate exceptions to the conceptual norms, we use the framework proposed by Allaway et al. (2024). This framework proposes specific prompt templates for generating exceptions from LLMs, along with a filtering process to ensure the generated exceptions are true and salient. We use these templates with GPT-3.5 (Ouyang et al., 2022)[5] to generate candidate exceptions and remove false

[5] `gpt-3.5-turbo-0613`

ones. We keep the top 5 candidates ranked by perplexity to use in our dataset.

VISaGE includes substantial human validation, including an iterative process of adding new attribute norms and exceptions. During validation, annotators can revise and expand the dataset by adding additional exceptions and category-attribute relations. Specifically, for valid category-attribute relations annotators, can provide an additional exceptional subcategory $\hat{e}_{c,a}$. Additionally, for each exception, annotators can provide a new category-attribute relation $n_{c,\hat{a}}$ that the exception corresponds to. This allows us to capture subcategories that are exceptional for the category but not for the original attribute a. For example, pixie-bob cats are an exception to *cats have long tails* but not to the original norm *cats have tails*. The tuples with the new category-attribute norms ($n_{c,\hat{a}}$, $e_{c,\hat{a}}$, V_c, V_e)[6] are added directly into the dataset while for the new exceptions $\hat{e}_{c,a}$, new images $V_{\hat{e}}$ are first retrieved and validated before being added to the dataset as ($n_{c,a}$, $\hat{e}_{c,a}$, V_c, $V_{\hat{e}}$).

The annotations were conducted by the authors of this paper. Through the revision and expansion process, we added 121 new tuples of conceptual-norm-and-exception (along with their corresponding images). Combined with the added conceptual norms, we nearly doubled the size of our dataset (an increase from 872 tuples to the final 1601 tuples).

B Models

See Table 1 for the details of the models used. Models are downloaded from HuggingFace; model details can be found at `https://huggingface.co/MODEL_NAME`.

C Compute

Experiments were performed using either Nvidia A100 or A4500 GPUs. On average, each evaluation (single model, condition) took approximately 15m, including model loading.

D Experimental Details

We used the `vllm`[7] package, version `0.8.5.post1` with `transformers v4.52.0.dev0` and `torch v2.6.0`. Models were evaluated with default settings, apart from limiting the model's output size in order to deal with memory limitations. Generation temperature was set to 0.05. We only evaluated the first output token.

E Shapley Experiments

We use the ExactExplainer from the `shap` library (version `0.48.0`): treating the image as a single part, together with the short texts, makes this feasible. We mask the image by removing it entirely (rather than replacing it with uniform pixels); language tokens are masked with `_`.

F AI Agent Use

We used coding agents (`copilot`) to assist with code development. We did not use any AI agents for writing.

G Full Results

Numerical results for all experiments and conditions are in Table 2.

H Annotation Tool

See Figure 5.

[6] Note that $e_{c,\hat{a}} = e_{c,a}$; the changed index is for clarity.

[7] `https://docs.vllm.ai`

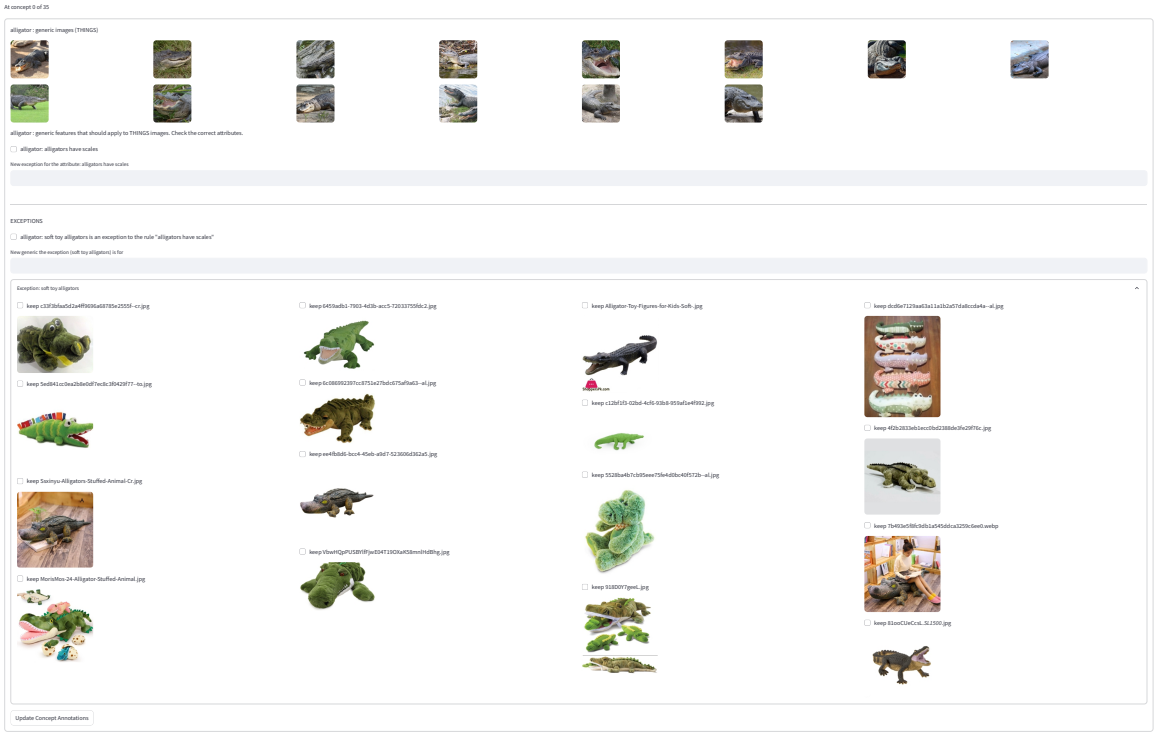

Figure 5: Annotation interface for dataset validation and expansion

Short Name	HF Model Name	Size
smolvlm2	HuggingFaceTB/SmolVLM2-2.2B-Instruct	2.2B
deepseek-vl-v2	deepseek/deepseek-vl2-tiny	3.37B
paligemma2	google/paligemma2-3b-ft-docci-448	3.03B
phi3-v	microsoft/Phi-3.5-vision-instruct	4.15B
gemma3-4B	google/gemma-3-4b-it	4.3B
phi4_mm	microsoft/Phi-4-multimodal-instruct	5.57B
llava-next	llava-hf/llava-v1.6-mistral-7b-hf	7.57B
qwen2-vl	Qwen/Qwen2-VL-7B-Instruct	8.29B
qwen2.5-vl	Qwen/Qwen2.5-VL-7B-Instruct	8.29B
molmo	allenai/Molmo-7B-O-0924	7.67B
idefics3	HuggingFaceM4/Idefics3-8B-Llama3	8.46B
internvl3	OpenGVLab/InternVL3-8B	7.94B
gemma3-12B	google/gemma-3-12b-it	12.2B

Table 1: Models used in experiments.

prompt	image	name	smolvlm2	deepseek-vl-v2-tiny	paligemma2	phi3-v	gemma3-4B	phi3-v
concept	generic	base	0.8657	0.7851	0.6427	0.8257	0.8095	0.7533
concept	exception	base	0.4372	0.2755	0.1587	0.4716	0.3410	0.3929
concept	exception	exception	0.6833	0.7839	0.8139	0.6640	0.6971	0.8182
concept	generic	exception	0.5340	0.6227	0.7158	0.4785	0.7489	0.7114
instance	generic	base	0.8513	0.6827	0.5778	0.8457	0.7951	0.7121
instance	exception	base	0.6540	0.8413	0.7883	0.6146	0.6833	0.7714
instance	exception	exception	0.6677	0.8645	0.7664	0.6771	0.6964	0.8314

prompt	image	name	llava-next	molmo	idefics3	internvl3	gemma3-12B
concept	generic	base	0.9319	0.7858	0.8507	0.9182	0.8738
concept	exception	base	0.6758	0.3829	0.3804	0.4422	0.3648
concept	exception	exception	0.5147	0.8045	0.7164	0.7083	0.7851
concept	generic	exception	0.3279	0.7227	0.4522	0.4934	0.8089
instance	generic	base	0.9094	0.7189	0.8570	0.8932	0.8326
instance	exception	base	0.5328	0.8376	0.6964	0.7021	0.7064
instance	exception	exception	0.5896	0.8295	0.7002	0.7177	0.7608

Table 2: Accuracy results for all experiment conditions.