VISaGE: Understanding Visual Generics and Exceptions

Anonymous ACL submission

Abstract

While Vision Language Models (VLMs) are trained to learn conceptual representations (generalized knowledge across many instances), 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 VLMs are typically dominated by the semantic prior, which arises from the language modality, when answering queries about instances. In contrast, conceptual understanding degrades when the assumption of congruency underlying the pragmatic prior is violated with incongruent images.

1 Introduction

Vision-language models (VLMs) are typically used to analyze *instances*: what is going on in a particular image? However, during training they learn a set of *conceptual* representations, generalized knowledge that holds over many instances. 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 their conceptual representations, based on exposure to typical instances, have long been analyzed (e.g., Bruni et al., 2014; Silberer et al., 2013; Collell and Moens, 2016), the potential tension between instance and concept representations, as arises in atypical instances, is currently under-explored.

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].

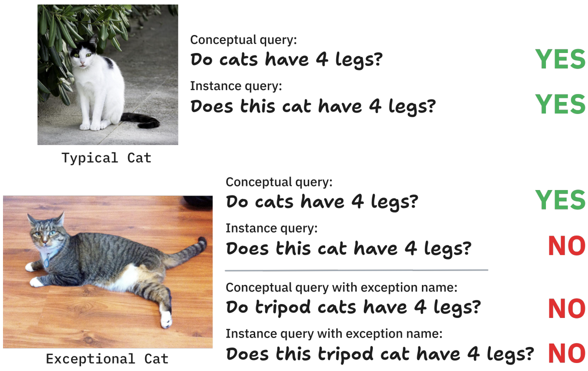


Figure 1: VISaGE contains both typical category instances and exceptional instances for which generics do not hold. We probe VLMs for conceptual and instance-level understanding, which is congruent in the typical case (top pair) but conflicts in the case of exceptional instances of a category `cat` (middle pair). However, the same exceptional instance can also be a typical member of the exception category (bottom pair).

Unlike language, which can 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 attributeshas 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).

In this vein, we introduce a new evaluation dataset, **VisaGE:** Visual Generics and Exceptions, consisting of conceptual categories with images of

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

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 conceptual information impact VLMs' ability to recognize *instance attributes*?
2. **(RQ2)** How does visual grounding to (potentially atypical) instances impact a model's ability to access *conceptual information*?

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: In RQ1, given an atypical instance, the pragmatic prior to focus on the current context must overrule the semantic prior of typicality, while for RQ2, the atypical image must be ignored, and the semantic prior should be followed.

We test a set of contemporary VLMs and find evidence that their conceptual representations do not recognize possible variation in attributes. Specifically, we find evidence that models rely on explicit textual cues to recognize instantiations of exceptional attributes in images, suggesting a strong semantic prior for the generic. Additionally, we observe that the models' pragmatic prior often interferes with conceptual understanding (and the semantic prior) when visual grounding is incongruent with the text. This suggests that VLMs' visually grounded conceptual representations only include typical or generic instances of the category.

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 do not sufficiently recognize within-category variation (exceptions).

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

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 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. Additionally, our experiments, in which we manipulate the 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; Parcalabescu and Frank, 2024).

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 1698 exceptional image examples for 441 exception subcategories, derived from 972 category-attribute relations (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-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. Finally, for each generic (category-attribute statement) we generate a set of exceptions

[3]Dataset & code will be released on publication (CC-BY).

$e_{c,a}$ 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. These images have been specifically collected to be typical object instances; we further 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 $n_{c,a}$; category-attribute relations that are not visually salient (*birds can sing*) or are not exhibited across images are discarded. 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 Fig. 2 for an overview. Specifically, we vary 1. the type of knowledge being queried (conceptual versus instance); 2. the type of image input (typical versus exceptional images); and, 3. the noun-phrase used to refer to the concept (category-name versus 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 1.

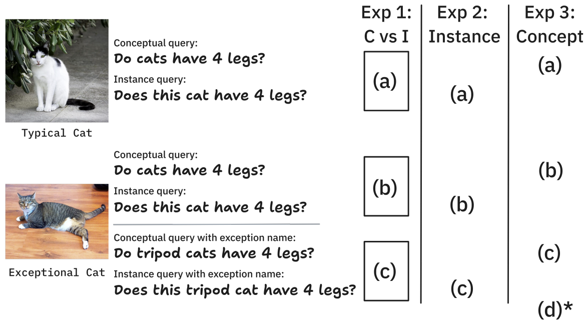


Figure 2: Summary of experiments and conditions: Exp. 1 measures the difference in accuracy between probing at the conceptual level vs. instance level. Exp. 2 tests models' ability to reason about instances, while Exp 3 tests models' conceptual understanding. Exp. 3 also includes condition (d), not shown, involving typical images with queries about exception categories.

4.1 Conceptual vs. Instance Queries

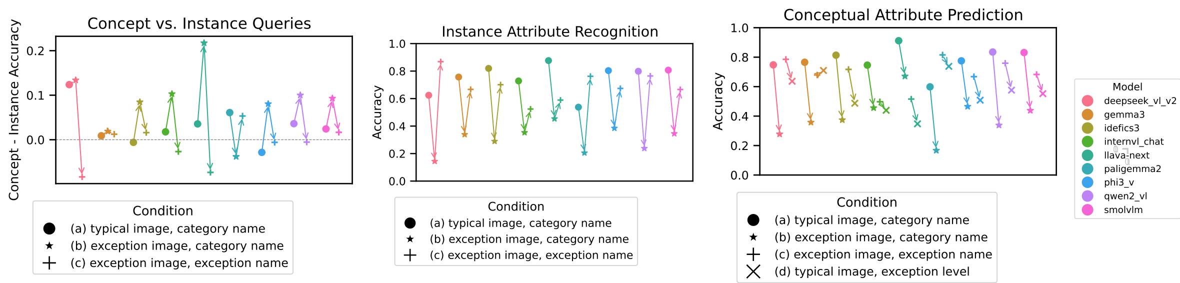
In an initial analysis, we test the ability of VLMs to distinguish between conceptual and instance queries. Specifically, we measure the difference in accuracy between the conceptual and instance queries in three conditions: (1a) typical images with category names, (1b) exceptional images with category names, and (1c) exceptional images with exception names. Condition (1b) is the critical condition, in which the correct prediction is different for conceptual and instance queries (see Fig.1; numerical results are in App. Table F).

We observe (Figure 3a) that instance queries are in fact harder for VLMs than conceptual queries. When visual input and category name are congruent ((1a) and (1c)), we observe minimal differences (near zero) between the conceptual and instance queries. In contrast, when models are required to consider specific visual features of the input, rather than the semantic information from the category name, as for instance queries in (1b), we observe that most models fail to do this (conceptual accuracy is higher than instance accuracy). The accuracy difference that is visible *only* with incongruent inputs emphasizes the importance of considering how image instances interact with conceptual representations.

4.2 Instance Attribute Recognition

Having shown that VLMs struggle with instance queries requiring visual grounding (§4.1), our

[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}?`



(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

Figure 3: Results: See Fig. 2 for setup. Exp 1: The difference between conceptual and instance accuracy is highest for incongruent pairs (b). Exp 2: Instance attribute recognition declines for exceptional images (b), unless they are named as such (c). Exp 3: Conceptual attribute prediction accuracy decreases for incongruent inputs (b and d).

second set of analyses investigates the role of language-based conceptual activation in misleading models. We compare category-name instance queries ("Does this cat have four legs?") in two conditions: with (2a) typical and (2b) exceptional images. The third condition (2c), *exception-name* instance queries with exceptional images, provides an explicit language cue to the model about which conceptual representation it should use (the exception rather than the category).

Our results (Figure 3b) show that, despite instance queries directing the model to consider the image, all models still appear to ignore the visual features, relying instead on language-based conceptual cues. In particular, we again see that models have higher accuracy in the conditions ((2a) and (2c)) where the text and image are congruent. When the text and image are *in*congruent (condition (2b)), the models appear to rely on language-based conceptual information. Since the image in (2b) is exceptional for the category, conceptual information activated by the category name does not apply to the image, resulting in a substantial drop in accuracy. This leads to the v-shaped pattern in accuracy. Note that if the models instead prioritized using the visual features of the input, their performance would be relatively stable across conditions.

4.3 Conceptual Attribute Prediction

Since we have shown that visual grounding is often ignored by VLMs when answering instance queries, our final set of experiments studies how it impacts conceptual queries. Specifically, we use two pairs of conditions to investigate the impact of text-image congruency. The first pair of conditions uses the category name in conceptual queries with: (3a) typical (congruent), and (3b) exceptional (incongruent) images. The second pair of conditions similarly queries conceptual information about the *exception* subcategory with: (3c) exceptional (congruent), and (3d) typical (incongruent) images.

Our results (Figure 3c) 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 in both pairs of conditions when comparing the congruent condition to the incongruent condition ((3a) vs. (3b) and (3c) vs. (3d)). This suggests that the pragmatic prior (considering the image relevant) interferes with the conceptual representation; that is, the image distracts the model from what is actually being asked in the query.

We also observe that incongruency in the input has less impact on accuracy when the queries are about the exception subcategory. One reason for this may be that the generic category is necessarily a well-established concept, since it is derived from a conceptual norm, while the exception subcategory may not be. Models may therefore lack a well-developed multimodal conceptual representation for the exception, resulting in them treating the conceptual query as an instance query. Our results that VLMs rely primarily on language cues for instance queries (§4.2), support this hypothesis.

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*).

Compute limitations restricted the testing of very large VLMs (`llama4, pixtral`).

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.

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.

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.

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*.

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. 2024. 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.

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.

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 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 1689 tuples).

Short Name	HF Model Name
`deepseek_vl_v2`	`deepseek/deepseek-vl2-tiny`
`gemma3`	`google/gemma-3-4b-it`
`idefics3`	`HuggingFaceM4/Idefics3-8B-Llama3`
`internvl_chat`	`OpenGVLab/InternVL2-2B`
`llava-next`	`llava-hf/llava-v1.6-mistral-7b-hf`
`paligemma2`	`google/paligemma2-3b-ft-docci-448`
`phi3_v`	`microsoft/Phi-3.5-vision-instruct`
`qwen2_vl`	`Qwen/Qwen-VL`
`smolvlm`	`HuggingFaceTB/SmolVLM2-2.2B-Instruct`

Table 1: Models used in experiments.

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. We only evaluated the first output token.

E AI Agent Use

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

F Full Results

Numerical results for all experiments and conditions are in Table 4.

G Annotation Tool

See Figure 5.

Conds	prompt	image type	name	`deepseek`	`idefics3`	`qwen2_vl`	`phi3_v`	`paligemma2`	`gemma3`	`llava-next`	`internvl_chat`	`smolvlm`
(1a, 3a)	concept	generic	category	0.75	0.81	0.83	0.78	0.60	0.77	0.91	0.75	0.83
(1a, 2a)	instance	generic	category	0.62	0.82	0.80	0.80	0.54	0.76	0.88	0.73	0.81
(1b, 3b)	concept	exception	category	0.28	0.37	0.34	0.47	0.17	0.36	0.67	0.46	0.44
(1b, 2b)	instance	exception	category	0.14	0.29	0.24	0.38	0.20	0.34	0.45	0.35	0.35
(1c, 3c)	concept	exception	exception	0.79	0.72	0.76	0.67	0.82	0.68	0.52	0.50	0.68
(1c, 2c)	instance	exception	exception	0.87	0.70	0.76	0.67	0.76	0.67	0.59	0.52	0.67
(3d)	concept	generic	exception	0.64	0.49	0.58	0.51	0.74	0.71	0.35	0.44	0.55
-	instance	generic	exception	0.34	0.61	0.53	0.60	0.36	0.37	0.68	0.54	0.53

Figure 4: Accuracy results for all experiment conditions.

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

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

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

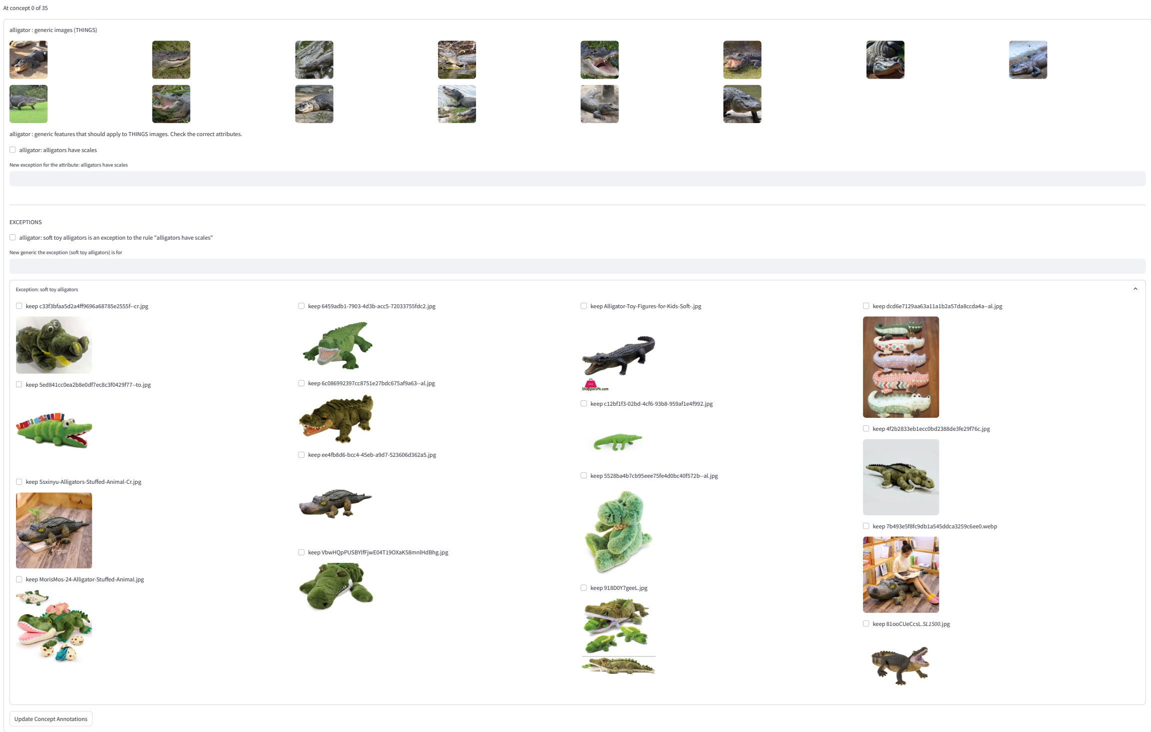


Figure 5: Annotation interface for dataset validation and expansion