UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Naming, Describing, and Quantifying Visual Objects in Humans and LLMs

Testoni, A.; Sprott, J.; Pezzelle, S.

DOI
10.18653/v1/2024.acl-short.50

Publication date
2024

Document Version
Final published version

Published in
The 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024) : proceedings of the conference

License
CC BY

Link to publication

Citation for published version (APA):
Testoni, A., Sprott, J., & Pezzelle, S. (2024). Naming, Describing, and Quantifying Visual Objects in Humans and LLMs. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024) : proceedings of the conference: ACL 2024 : August 11-16, 2024* (Vol. 2, pp. 547-557). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-short.50

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Oct 2026

Naming, Describing, and Quantifying Visual Objects in Humans and LLMs

Alberto Testoni
ILLC, University of Amsterdam
a.testoni@uva.nl

Juell Sprott
University of Amsterdam
juell.sprott@student.uva.nl

Sandro Pezzelle
ILLC, University of Amsterdam
s.pezzelle@uva.nl

Abstract

While human speakers use a variety of different expressions when describing the same object in an image, giving rise to a distribution of plausible labels driven by pragmatic constraints, the extent to which current Vision & Language Large Language Models (VLLMs) can mimic this crucial feature of language use is an open question. This applies to common, everyday objects, but it is particularly interesting for uncommon or novel objects for which a category label may be lacking or fuzzy. Furthermore, similar patterns of variation are observed among human speakers for highly context-sensitive expressions, such as the quantifiers 'few' or 'most'. In our work, we evaluate VLLMs (FROMAGe, BLIP-2, LLaVA) on three categories (nouns, attributes, and quantifiers) where humans show great subjective variability concerning the distribution over plausible labels, using datasets and resources mostly under-explored in previous work. Our results reveal mixed evidence on the ability of VLLMs to capture human naming preferences at generation time: while some models are good at mimicking human distributions for nouns and attributes, all of them fail to assign quantifiers, a task that requires more accurate, high-level reasoning.

1 Introduction

Recent years have witnessed increasing popularity in the development of Large Language Models (LLMs) given their notable performance in following instructions, answering questions, and in many reasoning tasks, serving as general-purpose assistants (Huang and Chang, 2023; Zhao et al., 2023). In parallel, a new generation of powerful Vision and Language LLMs (VLLMs) with excellent visual understanding and generation capabilities have emerged (Gan et al., 2022; Li et al., 2023a). Rapidly, these models have outperformed previous approaches in many downstream tasks. In our work, we focus on the Natural Language Generation skills of powerful VLLMs by analyzing an important but under-explored problem, namely, their ability to capture human production variability (in terms of distribution over plausible labels/descriptions) in naming tasks.

Previous work highlighted that speakers display a wide range of variability when asked to utter sentences, resulting in inter-speaker variability but also variability over time for the same speaker (Levelt, 1993; Fan et al., 2018; Alva-Manchego et al., 2021; Takmaz et al., 2024). In particular, in object naming, speakers may refer to objects appearing in a visual scene in many different ways (Graf et al., 2016). Objects generally belong to multiple categories/super-categories, and all the lexicalized labels of such categories are valid (Brown, 1958). However, although multiple labels are valid, humans pragmatically adapt their naming preferences depending on the context (Olson, 1970; Rohde et al., 2012), resulting in some labels being more frequently uttered than others. For instance, 'mammal' is a correct label to describe a Gold Retriever, but pragmatically less likely than 'dog'. Similarly, speakers tend to prefer sub-ordinate words like 'car' instead of the potentially ambiguous super-ordinate word 'vehicle' in case multiple vehicles appear in the image. In our work, we are interested in capturing both these two features: while many labels are equally valid and acceptable when naming or describing entities, these labels distribute according to a certain likelihood distribution.

In our work, we investigate this issue, which has recently entered the NLP research community (Plank, 2022), in three different production conditions. First of all, we consider the ManyNames dataset (Silberer et al., 2020a,b), where annotators assign labels to describe common objects in images in a referential expression generation setting (Yu et al., 2016; Kazemzadeh et al., 2014). We also explore two additional resources that have not received much attention within the NLP community

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 547–557
August 11-16, 2024 ©2024 Association for Computational Linguistics

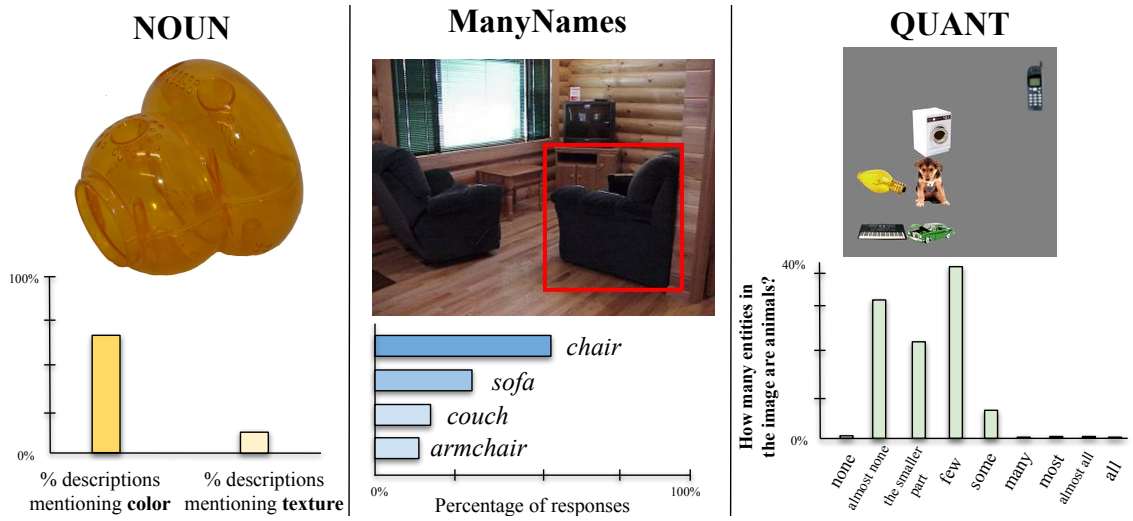


Figure 1: Datasets used in our experiments and distribution of human answers/labels. In NOUN (left), we focus on the frequency of color and texture attributes in the generated descriptions. In ManyNames (middle), each object is associated with the frequency of the nouns used to describe it. In QUANT (right), each image is associated with a probability distribution over a list of quantifiers that humans selected when answering the question 'How many of the objects are animals?'.

and that allow us to broaden the horizons of this phenomenon. First, we analyze the NOUN dataset (Horst and Hout, 2016), where speakers describe uncommon and novel objects: we focus on both the choice of the adjectives and how they distribute in the across-subject distribution. Finally, we investigate human production variability arising from the context-sensitive nature of non-numerical quantifiers using the data collected by Pezzelle et al. (2018).

We evaluate three VLLMs (FROMAGe, BLIP-2, LLaVA) on the above-mentioned tasks in a zero-shot setting. We sample multiple times from the model using nucleus sampling, mimicking various human speakers, and compare the generated samples against human production patterns using different metrics (Jensen–Shannon divergence and Pearson's correlation, depending on the task at hand). Our results show that models weakly to moderately mimic human distributions in naming common and uncommon objects. Instead, all of them fail to mimic human distributions when selecting quantifiers, as highlighted by our in-depth analyses.

2 Tasks and Datasets

We use the images and corresponding human labels or descriptions from three datasets in English, that we briefly describe below.

NOUN The Novel Object and Unusual Name (NOUN) dataset (Horst and Hout, 2016) contains 64 images of multipart, multicolored, and three-dimensional uncommon and novel objects. The dataset was originally created for behavioral studies on word learning and, to the best of our knowledge, it has not been used for NLP research. We focus on the *naming task*, where participants were asked to answer the question "What would you call this object?". The answers are sentences like: 'a plastic object with red stuff on top'. For each object, the proportion of colors (e.g., 'red', 'bronze') and textures (e.g., 'soft', 'rough') was calculated as the number of attributes given the number of responses. An example from the dataset is reported in Figure 1 (left), together with the ratio of colors and textures in human responses. In NOUN, we examine human production preferences on a high level, by looking at the frequency according to which certain adjectives (related to color and texture attributes) are used.

ManyNames In ManyNames (Silberer et al., 2020a,b), the authors collected names for 25K objects appearing in real-world images from VisualGenome (Krishna et al., 2017) by asking human annotators to generate a name for them. Each object (highlighted by a red box in the image) is associated with an average number of 35.3 annotations. More than 90% of the objects are associated with more than one unique label (5.7 average name types per object). An example is shown in Figure 1 (middle). When describing the object in the

NOUN

Q: What do you see in the image?

LLaVA 🌋**:**

1) The image features a green toy with an orange nose and two orange wheels.

2) The image features a green and orange toy shaped like a face, with a big nose.

ManyNames

Q: Please name the object in the red box with the first name that comes to mind.

LLaVA 🌋**:**

1) jet 2) plane

3) plane 4) jet

5) jet 6) jet

QUANT

Q: How many animals are there in the image? Pick the best among the following options: A: the smaller part, B: many, C: almost none, D: some, E: none, F: all, G: most, H: few, I: almost all

LLaVA 🌋**:**

1) D *(some)* 2) I *(almost all)*

3) D *(some)* 4) H *(few)*

Figure 2: Examples of the output generated by LLaVA (multiple samples with nucleus sampling decoding) for the three tasks analyzed in our work. For each task, a sample of the answers provided by the model is displayed. For space constraints, we only report a few random samples for each task.

red box, most annotators referred to it as 'chair', while around 30% said 'sofa', and the remaining ones used 'couch' and 'armchair'. The images in ManyNames are classified into 7 domains (e.g., vehicles, people, animals, etc.): for computational constraints, we evaluated 300 randomly sampled objects from each domain. Different from NOUN, we examine production preferences on a more fine-grained level using the actual distribution over multiple labels.

QUANT To study how quantifiers are used when referring to quantities grounded in images, Pezzelle et al. (2018) introduced a dataset of visual abstract scenes containing a variable number of animals and artifacts and asked human participants to answer the question "How many of the objects are animals?". Participants could select the answer from a list of nine pre-selected quantifiers: 'none', 'almost none', 'the smaller part', 'few', 'some', 'many', 'most', 'almost all', and 'all'. The authors used images with 17 different proportions of animals and artifacts (ranging from 0% to 100%). In our work, we tested 50 images for each of the 17 proportions in the dataset, resulting in a total number of 850 images.[1]

[1]The actual images used in our experiment come from Testoni et al. (2019), which built a large-scale dataset using the stimuli and pipeline by Pezzelle et al. (2018).

3 Experiments

3.1 Generation

In our work, we test the performance of three models in a zero-shot setting: BLIP-2 (Li et al., 2023b), FROMAGe (Koh et al., 2023), and LLaVA 1.5 (Liu et al., 2023b,a). All three models can be prompted for zero-shot generation. Additional details are discussed in Appendix A.4. For each of the three tasks described in Section 2, we used prompts that resembled the instructions provided to human annotators during the dataset collection. ManyNames: *Q: Please name the object in the red box with the first name that comes to mind. A:*. NOUN: *Q: What do you see in the image? A:*. QUANT: *Question: How many animals are there in the image? Pick the best among the following options:* , followed by the list of the nine quantifiers, each associated with a letter (from A to I). The ordering of the quantifiers is randomized at each inference step. Although investigating several variations of the above-mentioned prompts is beyond the scope of the paper, we discuss some insights on this aspect in Appendix A.5. We sample multiple times from each model using nucleus sampling decoding (Holtzman et al., 2019), with $p = 0.9$, $t = 0.5$ (different hyperparameter configurations did not significantly affect the overall results, as discussed in Appendix A.5). For each task, we sample the model 20 times and filter out ill-formed answers, such as empty strings or question repeti-

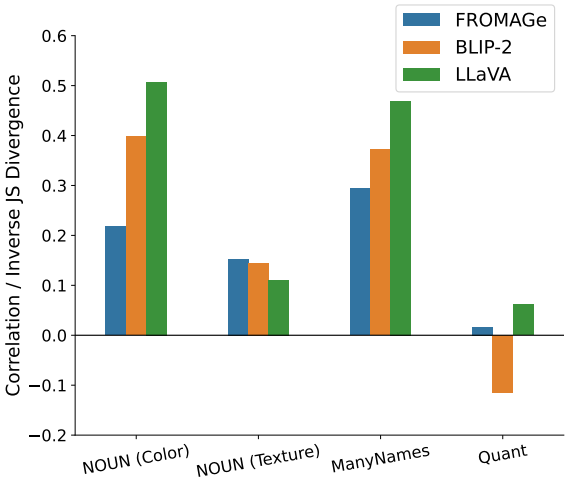


Figure 3: For NOUN and QUANT, the plot shows the correlation between human responses and model samples. For ManyNames, it shows the inverse JS divergence between the frequency of the nouns chosen by annotators and the ones generated by the model.

tions. After filtering, we randomly take 10 generations per image for ManyNames and NOUN, and 15 for QUANT. In this way, we have the same number of generations for each image/object. Some examples of the output generated by LLaVa for the three tasks analyzed are reported in Figure 2. We release our code at: https://github.com/albertotestoni/ndq_visual_objects.

3.2 Evaluation

Each object in **NOUN** is associated with color and texture saliency, i.e., how often speakers described the object using these attributes. We use a string-match approach (see Appendix A.2) to analyze the model output and compute color and texture saliency. We then compute the Pearson's *r* correlation between human and model saliency, considering all objects.

Each object in **ManyNames** is associated with H unique nouns assigned by human annotators and M unique nouns sampled from the model output, together with their frequency. Given $A = H \cup M$, we construct two term-frequency vectors for human and model output, h and m, respectively, with $|h| = |m| = |A|$. Each noun in A is mapped to a unique position in h and m and each vector is filled with its normalized frequency. We evaluate the models by computing the inverse Jensen–Shannon (JS, bounded between 0 and 1) divergence (Lin, 1991) between h and m. See Figure 6 in the Appendix for an example.

Each image in **QUANT** is associated with a probability distribution over 9 quantifiers, depending on the proportion of animals and artifacts. From the model outputs, we extract the relative frequency of each quantifier and compute Pearson's *r* correlation with the human distribution. We then average the correlation results over all images. Correlation is bounded between -1 and 1. Higher is better for all the metrics.

3.3 Results

As we can observe from Figure 3, the results for ManyNames and NOUN (color saliency) show a clear trend: all the models correlate, to some extent, with human production, with LLAVA obtaining the highest correlations for both tasks (around 0.5) and significantly outperforming (t-test, $p < 0.01$) both BLIP2 and FROMAGe.[2] These findings align with previous work showing the primacy of LLaVA over other models (Liu et al., 2023b,a). However, the remaining tasks show critical weaknesses for all models. First, none of the models achieve a statistically significant correlation for texture saliency (all have $p > 0.05$). We conjecture that texture attributes are less common for the models compared to colors, and thus they may be less accurate when generating them: we leave an in-depth analysis of this issue for future work. Despite the correlation results being similar across models, our manual inspection reveals interesting differences: while the low performance of FROMAGe is due to an under-generation of texture attributes, the opposite is true for LLaVA, with BLIP-2 being more flexible in terms of texture attribute generation but not aligned with human variability (see Figure 7 for an example). Finally, all models show almost no correlation in assigning quantifiers to visual scenes, highlighting a severe limitation of all models on this task. We scrutinize this issue in the following Section.

4 *The Curious Case* of Quantifiers

We run some analyses to investigate the poor performance of all models in the QUANT task. First of all, we acknowledge that the multiple-choice prompting used in QUANT is different and more complex than the prompts used for the other datasets. Still, it is unlikely that this is the main reason behind the poor performance of all models.

[2]Appendix A.1 shows per-domain results for ManyNames.

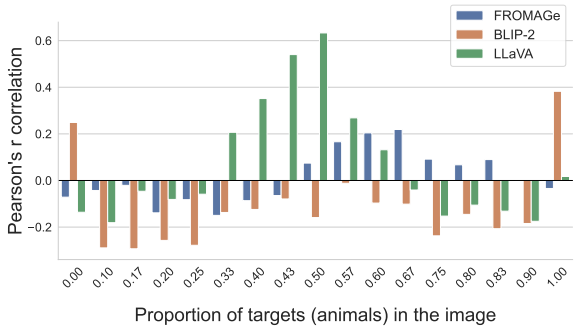


Figure 4: Pearson correlation results (y-axis) broken down by proportion of targets (animals) in the image (x-axis) in the QUANT task and dataset.

Figure 4 shows the correlation results broken down by the proportion of animals in the image. We observe that even though the overall correlation results are similar across models (Figure 3), they perform quite differently depending on the proportion of animals in the scene. While BLIP-2 performs relatively well on the 'extreme' proportions (no animals or all animals in the image, when speakers generally choose the quantifiers 'none' and 'all', respectively), LLaVA excels at intermediate proportions, and FROMAGe performs better on proportions above 50%. Can we conclude that models properly handle the task of assigning the most likely quantifiers for some proportions? These results evoke two hypotheses: (a) The models are capable of selecting plausible quantifiers only for some proportions or, vice versa, they understand only some of the quantifiers analyzed; (b) The models have a bias towards some specific quantifiers, regardless of the proportion of targets in the scene, leading to a decent perform on some proportions as a side effect. Our additional analyses, reported in Figures 8 and 9 in the Appendix, support hypothesis (b): FROMAGe has a strong bias towards selecting the quantifier 'many'; BLIP-2 frequently selects the *extreme* quantifiers 'none' and 'all', and its selection is not influenced by the proportion of targets; LLaVA has a bias towards selecting the quantifier 'some', regardless of the proportion of targets.

To further shed light on this result, we qualitatively assess the 'counting' skills of the models, a crucial skill to succeed in assigning quantifiers. As the examples in Figure 11 in the Appendix illustrate, all models struggle to successfully count how many animals appear in the image. We hypothesize that the reason for the poor performance in assigning quantifiers lies in the quantity estimation and comparison skills of the models. This observation is in line with recent research investigating the poor 'counting' skills of current models (Paiss et al., 2023).

5 Conclusion

While human speakers exhibit a wide range of human production variability in naming tasks, mirroring pragmatic constraints and subjective preferences, it is not clear to what extent VLLMs can mimic this peculiar trait of language use. In our work, we investigate this issue in three tasks: naming common objects, naming novel objects, and assigning quantifiers. Our results reveal that best-performing models achieve a moderate correlation with human patterns in some tasks (object names and color terms). However, all models dramatically fail when assigning quantifiers, the only production setup that requires some form of reasoning, i.e., the ability to reason over sets of objects and process quantities. Based on our analyses, we hypothesize that the reason behind this failure stems from the poor "counting" skills of the models.

Limitations

In the following, we discuss some limitations of our study that may inspire follow-up work in this direction. The poor performance on the quantification tasks may stem from the higher complexity of the prompt used (multiple choice prompting). Even though in our paper we discuss how analyzing the output variability allows us to gain valuable insights even when the model is not accurate, we can not rule out the possibility that a simpler prompt may lead to more accurate results. As an initial step, we used a prompt that corresponds to the instruction provided to the participants of the original experiment in Pezzelle et al. (2018). In Appendix A.5 we discuss the effect of re-phrasing the original prompt instructions.

Moreover, it is worth noting that the human production variability analyzed in our experiments is obtained by aggregating data coming from multiple speakers. Even though we do aim at this, we acknowledge that it is unlikely that one single model can mimic such a rich variability. Our study is more focused on understanding *to what extent* current Vision & Language LLMs can mimic this feature, showing the suitability of some tasks and datasets not explored in previous work.

Finally, we computed the color and saliency feature for NOUN using a string-matching approach based on a manually defined list of keywords (as described in Appendix A.2. We acknowledge that this approach may underestimate the color and texture saliency in the model output. Although in this case, the small size of the dataset allowed us to verify that this is not the case, we believe that it is important to take this point into account when running experiments on a larger scale. Moreover, as a limitation of the NOUN dataset (and not of our experimental setup), we do not have access to the actual color and texture labels used by human participants during the dataset collection. For this reason, in NOUN we do not consider the actual distribution of the attributes used by human speakers but just their overall frequency.

Acknowledgments

We would like to thank the Dialogue Modelling Group (DMG) at the University of Amsterdam for their feedback and support at the different stages of this work. We thank Brent Brakenhoff for his input on some preliminary experiments with the ManyNames dataset. Alberto Testoni is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455, PI R. Fernández).

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Roger Brown. 1958. How shall a thing be called? *Psychological review*, 65(1):14.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352.

Caroline Graf, Judith Degen, Robert XD Hawkins, and Noah D Goodman. 2016. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *CogSci*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Jessica S Horst and Michael C Hout. 2016. The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior research methods*, 48:1393–1409.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pages 17283–17300. PMLR.

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

Willem JM Levelt. 1993. *Speaking: From intention to articulation*. MIT press.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023a. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2):2.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

David R Olson. 1970. Language and thought: aspects of a cognitive theory of semantics. *Psychological review*, 77(4):257.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180.

Sandro Pezzelle, Raffaella Bernardi, and Manuela Piazza. 2018. Probing the mental representation of quantifiers. *Cognition*, 181:117–126.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Rohde, Scott Seyfarth, Brady Clark, Gerhard Jäger, and Stefan Kaufmann. 2012. Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In *The 16th workshop on the semantics and pragmatics of dialogue, paris, september*.

Carina Silberer, Sina Zarrieß, and Gemma Boleda. 2020a. Object naming in language and vision: A survey and a new dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5792–5801, Marseille, France. European Language Resources Association.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020b. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2024. Describing images fast and slow: Quantifying and predicting the variation in human signals during visuo-linguistic processes. *arXiv preprint arXiv:2402.01352*.

Alberto Testoni, Sandro Pezzelle, and Raffaella Bernardi. 2019. Quantifiers in a multimodal world: Hallucinating vision with language and sound. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 105–116, Minneapolis, Minnesota. Association for Computational Linguistics.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Appendix

A.1 ManyNames Appendix

Figure 9 shows the results on the ManyNames dataset broken down by the image domain. LLaVA outperforms other models in most of the domains, but for *clothing* and *people* it is comparable to BLIP-2. Note that all model reach have the poorest performance on these two domains. As highlighted by Silberer et al. (2020a,b) and confirmed by our manual inspection, models confuse people and clothing objects much more frequently than humans do. ManyNames is licensed under Creative Commons Attribution 4.0 International.

A.2 NOUN Appendix

We define the following list of color and texture attributes to analyze the samples generated by the model with a string-matching approach.

Colors = ["Red", "Orange", "Yellow", "Green", "Blue", "Purple", "Pink", "Brown", "Gray", "Black", "White", "Beige", "Turquoise", "Teal", "Magenta", "Lavender", "Indigo", "Maroon", "Gold", "Silver", "Bronze", "Copper", "Olive", "Navy", "Sky blue", "Cream", "Peach", "Rose", "Fuchsia", "Coral", "Mint", "Chartreuse", "Salmon", "Sienna", "Slate", "Tan", "Crimson", "Ivory", "Khaki", "Lilac", "Mauve", "Mustard", "Rust", "Scarlet", "Tangerine", "Vermilion", "Violet", "Wheat", "Brick red", "Caramel"]

Textures = ["Smooth", "Rough", "Fuzzy", "Soft", "Hard", "Bumpy", "Slick", "Sticky", "Grainy", "Sandy", "Slippery", "Jagged", "Sharp", "Coarse", "Silky", "Velvety", "Wet", "Dry", "Glossy", "Matte", "Sparkly", "Metallic", "Wooden", "Leathery", "Plastic", "Rubber", "Furry", "Woolly", "Feathery", "Smooth", "Satin", "Lace", "Crochet", "Knitted", "Embroidered", "Linen", "Silk", "Velvet", "Suede", "Corduroy", "Denim", "Felt", "Tweed", "Mesh", "Hairy", "Crisp", "Crumbly", "Flaky", "Puffy", "Spongy", "Crunchy", "Chewy", "Gummy", "Slimy", "Starchy", "Syrupy", "Icy", "Rocky", "Stony", "Sandy", "Peppery", "Salty", "Sour", "Sweet", "Tangy", "Tart", "Spicy", "Herbaceous", "Earthy", "Mossy", "Woody", "Smoky", "Smokey", "Rusty", "Corroded", "Weathered", "Rugged", "Smooth", "Polished", "Shiny", "Gleaming", "Dull", "Muddy", "Cloudy", "Milky", "Transparent", "Translucent", "Opaque"]

Figure 7 shows an example of the output of different models, together with their color and texture

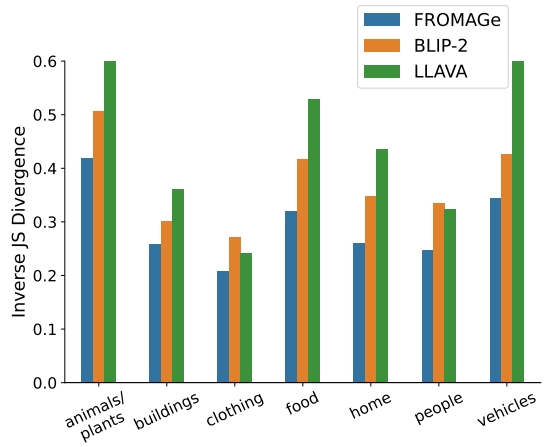


Figure 5: Inverse Jensen–Shannon divergence broken down by the image domain in ManyNames.

saliency as well as human saliency values. NOUN is released without a specific license.

A.3 QUANT Appendix

Figures 8 and 9 show additional analyses on the QUANT dataset. They are discussed in Section 4. Figure 11 shows some qualitative examples of the models' output when asked to answer the question 'How many animals are there in the image?'. The images are randomly selected from QUANT. QUANT is released without a specific license.

A.4 Models Appendix

While FROMAGe is trained with a contrastive learning objective for image captioning and it is shown to perform particularly well with longer textual contexts, BLIP-2 jointly optimizes three pre-training objectives that share the same input format and model parameters: image-text contrastive learning, image-grounded text generation, and image-text matching. The main innovation of LLaVA is the use GPT-4 generated visual instruction tuning data. Moreover, LLaVA has a simpler scheme to connect image and language representations compared to BLIP-2 and FROMAGe. We used `blip2-opt-2.7b` and `llava-v1.5-7b`, while for FROMAGe we used the model made available by Koh et al. (2023). FROMAGe and LLaVA are released with an Apache-2.0 license. BLIP-2 is distributed with BSD 3-Clause License. We run our experiments under the model license.

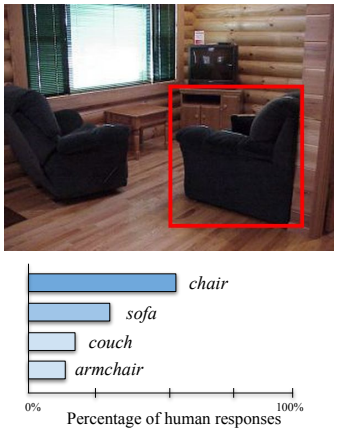
A.5 Generation Details

The Effect of Different *top_p* Values We experimented with various *top_p* values for nucleus sampling decoding. As illustrated in Figure 10 (showing the results for LLaVA, with similar results for the other models), we observe that this variable does not play a significant role in our experimental setup for all the tasks analyzed.

Different prompts In our experiments, we prompted the models with the same instructions provided to human annotators during the collection of the different datasets analyzed. We also experimented with small variations of the above-mentioned prompts, such as '*What is the object in the image?*' for the NOUN dataset, '*Name the object in the red box with the most appropriate single name*' for ManyNames, and a more detailed instruction for QUANT, such as '*Carefully examine the image. Can you determine the proportion of animals present, compared to objects? Please select the most accurate answer from the options below*'. While we do not observe any significant difference between NOUN and ManyNames, the revised prompt for QUANT leads to a slight improvement in the model performance, with LLaVA reaching a correlation of 0.29. Still, the low absolute correlation coefficient highlights that computational models struggle to accurately assign quantifiers to visual scenes. This result demonstrates that the prompt may influence the performance of the models on this task. Although exploring which prompts work best was beyond the scope of this paper, we leave a systematic exploration of this aspect to future research.

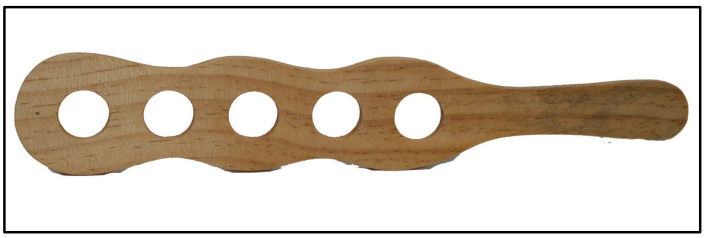
A.6 Additional Details

The data used in our work do not contain any information that names or uniquely identifies individual people or offensive content. FROMAGe has 5M trainable parameters and a total number of around 7.2B parameters. BLIP-2 has 188M trainable parameters and 2.7B total parameters. LLaVA has 7B parameters. All the models are evaluated on a single GPU (NVIDIA RTX A5000). We experimented with a few configurations of hyperparameters for nucleus sampling generation (described in Section 3). We did not find significant differences across different hyperparameters. We used the `SciPy` library (`https://scipy.org/`) to compute the correlation/divergence results. We used the NLTK library (`https://www.nltk.org/`) to extract nouns from the model output for ManyNames.



	chair	sofa	couch	armchair	window	Inverse Jensen–Shannon divergence vs. humans
Humans	0.51	0.24	0.14	0.11	0	**1**
Model 1	0.09	0.71	0.02	0.18	0	**0.59**
Model 2	0.15	0	0.29	0	0.56	**0.38**

Figure 6: Toy example to show how the effect of over-generating a noun that was not often assigned by humans (model 1) and generating a noun that was not selected by humans (model 2) on the inverse Jensen–Shannon divergence metric.



Human color saliency: 3%
Human texture saliency: 55%

FROMAGe

- A hand with a knife.
- A person holding a knife with a long handle.
- The image is a visual representation of the tool.
- The image shows a **wooden** board with a nail on it.
- …

Color saliency: 0%
Texture saliency: 20%

BLIP-2

- A **wooden** stick with eight holes on it
- A **wooden** toothbrush holder with five holes in it
- A **wooden** spoon holder with five holes
- A **wooden** spatula is sitting on top of a **white** surface
- …

Color saliency: 10%
Texture saliency: 80%

LLaVA

- The image features a **wooden** object with four holes in it, resembling a wooden paddle
- The image features a **wooden** pizza cutter with a handle, which is designed to cut through
- The image features a **wooden** spoon with a handle
- …

Color saliency: 0%
Texture saliency: 100%

Figure 7: Example of the model output in the NOUN dataset.

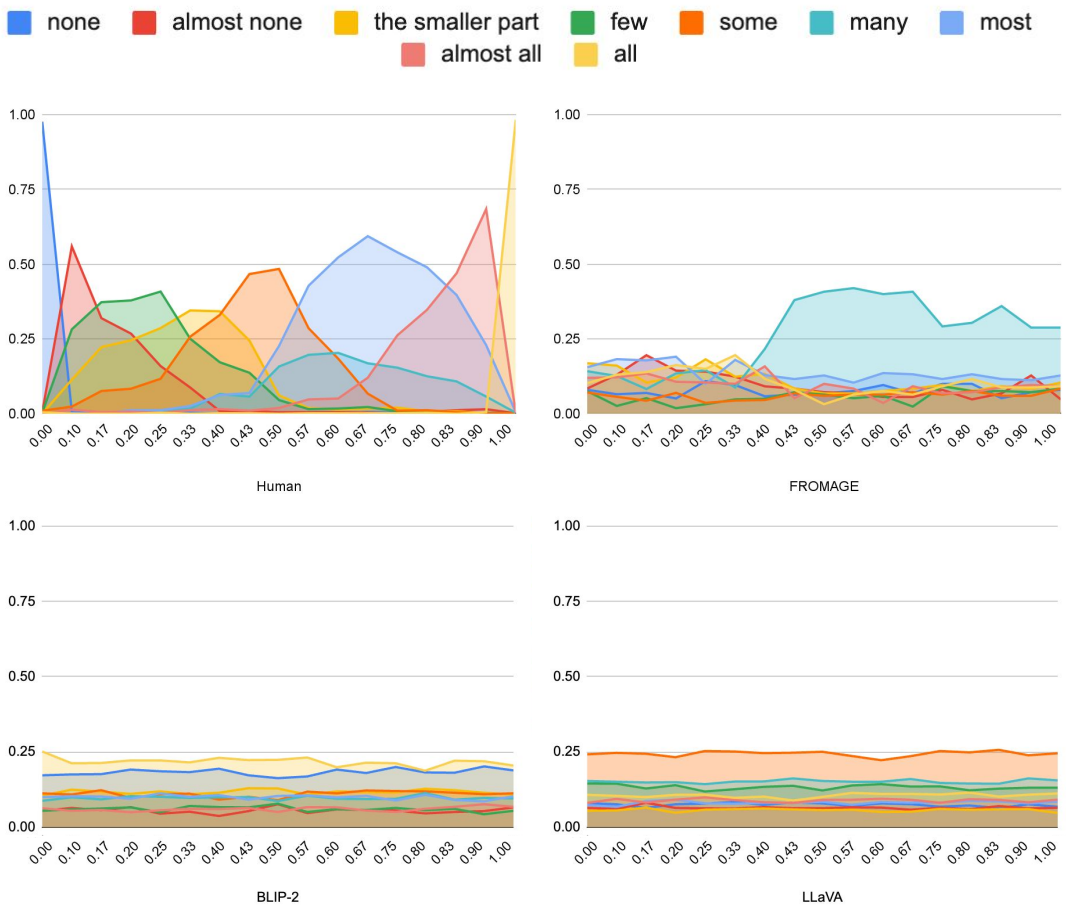


Figure 8: Density plot reporting the frequency distribution of responses for the 9 quantifiers (y-axis) against the proportion of targets in the scene (x-axis).

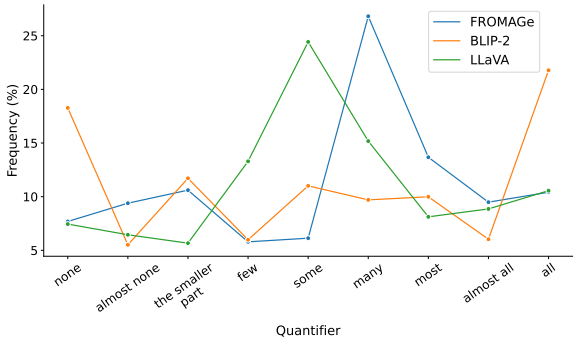


Figure 9: How often each quantifier (x-axis) is selected by the model (y-axis, expressed as %), regardless of the proportion of targets (i.e., animals) in the image.

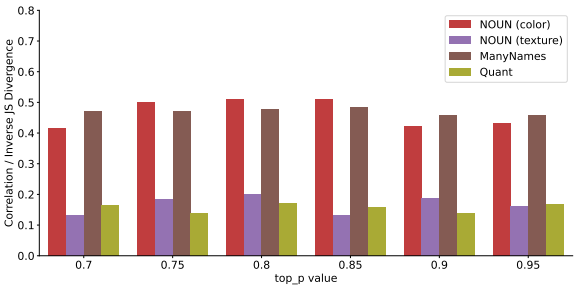


Figure 10: The role of different *top_p* values for nucleus sampling decoding using the LLaVa model.

How many animals are there in the image?

Correct:	**0**	**1**	**2**	**7**
FROMAGe:	3	15	'many'	'many'
BLIP-2:	1	5	7	8
LLaVA:	0	1	2	4
Correct:	**2**	**6**	**4**	**14**
FROMAGe:	3	'many'	4	'many'
BLIP-2:	6	12	10	70
LLaVA:	3	4	4	12

Figure 11: How many animals are there in the image? All models fail to successfully count the number of animals in the image. Note that models generally output a number but sometimes FROMAGe outputs the quantifier 'many' which, interestingly, is the quantifier the model is strongly biased towards as illustrated in Figure 8 and discussed in Section 4.