Zero-shot Capabilities of Visual Language Models with Prompt Engineering for Images of Animals

Andrea Daniela Tejeda Ocampo
Department of Electrical Engineering
University of California - Riverside
Riverside, CA 92521
andrea.tejedaocampo@email.ucr.edu

Eric C. Orenstein
MBARI
7700 Sandholdt Road
Moss Landing, CA 95039
eorenstein@mbari.org

Kakani Katija
MBARI
7700 Sandholdt Road
Moss Landing, CA 95039
kakani@mbari.org

Abstract

Visual Language Models have exhibited impressive performance on new tasks in a zero-shot setting. Language queries enable these large models to classify or detect objects even when presented with a novel concept in a shifted domain. We explore the limits of this capability by presenting Grounding DINO with images and concepts from field images of marine and terrestrial animals. By manipulating the language prompts, we found that the embedding space does not necessarily encode scientific taxonomic organism names, but still yields potentially useful localizations due to a strong sense of general objectness. Grounding DINO struggled with objects in a challenging underwater setting, but improved when fed expressive prompts that explicitly described morphology. These experiments suggest that large models still have room to grow in domain use-cases and illuminate avenues for strengthening their understanding of shape to further improve zero-shot performance. The code to reproduce these experiments is available at: https://github.com/bioinspirlab/deepsea-foundation-2023.

1 Introduction

Recent advances in large foundation models represent a substantial leap in our ability to process image data from new domains. The zero-shot capabilities of large foundation models like SegmentAnything are impressive, particularly when deployed in an interactive fashion with human prompting [1]. Visual Language Models (VLMs) like CLIP and GLIP are likewise exceeding expectations on zero-shot tasks when prompted with language queries [2, 3]. Among many other uses, these tools have huge potential for domain specific tasks like medical imaging [4, 5].

Biologists and ecologists are likewise eager to use these frameworks. Marine scientists in particular are in need of tools to speed the annotation of new data, particularly for detection and segmentation tasks; there is a huge amount of raw underwater image data but few publicly accessible annotated datasets [6, 7]. Creating a high-quality, taxonomically correct set of labeled data for training models remains an extremely time consuming task [8]. Highly trained annotators must spend 100s of hours examining images and footage to identify a sufficient number of animals to appropriately tune fully supervised, domain specific models [9]. Even high performing supervised models struggle when

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

applied in out-of-distribution target environments, thus requiring continuing manual annotation efforts as scientists seek to work in new regions or with different tools [10, 11].

The ocean science community is hoping to leverage foundation models to enrich existing marine annotated datasets, assist in future manual labeling efforts, and enable robust model deployment in dynamic natural environments that are subject to extreme distribution shifts on short time and spatial scales. But the morphology of underwater biological targets is often quite different from the types of objects found in the large image sets used to train foundation models. Likewise, the Latinate names used in rank-based species classification are atypical in the bodies of text usually used for training VLMs. Indeed, the space of classes of marine organisms is itself poorly constrained due to sampling bias and underexploration of the deep sea [12].

The promise of VLMs for marine science lies in their ability to generalize well in dynamic natural environments by manipulating prompts rather than retraining an entire model. In general, the zero-shot generalization ability of VLMs is highly-dependent on a well-designed prompt [13]. In this work, we probe the concept embeddings of Grounding DINO with manual prompt tuning for localizing images of marine and terrestrial animals to explore the limits of its zero-shot capabilities [14]. The goal is to better understand how to leverage the existing semantic space without fine tuning. The results both underscore how domain scientists could effectively start intuitively using VLMs and illuminate an interesting grey area in the semantic space of existing models.

Related work Most papers outlining VLM approaches include an evaluation of zero-shot performance on a suite of tasks. VLMs typically perform well on zero-shot tasks that share semantic commonalities with the corpus of text and image pairs used for training the base model. Radford et al. [2] noted that while zero-shot CLIP worked well on many benchmarks that it is "quite weak on several specialized, complex, or abstract tasks" like satellite and medical imagery. Li et al. [3] reported high zero-shot performance on 13 small datasets available on Roboflow, but the model struggled on two small test sets that contained marine organisms even after manual prompt tuning. Grounding DINO likewise exhibited relatively weak zero-shot performance on datasets that included underwater images or limited number of marine animal concepts [14]. In both cases, the marine-adjacent datasets did not cover actual field images of marine organims. For example, Shellfish-OpenImage[1] is largely composed of crustaceans in kitchens at various stages of food preparation. Likewise, the Aquarium[2] dataset is composed of images collected in constrained, human-made habitats that are often taken through tank walls. To our knowledge, there have been no systematic experiments to quantify the zero-shot capabilities of VLMs for underwater imagery, especially of diverse biological concepts collected in their natural habitats.

2 Method

2.1 Datasets

COCO The 2017 validation split of COCO, one of the canonical computer vision competition datasets, is used here as a baseline to compare our results against [15]. We randomly selected 992 images from COCO. If the image contained several localizations, we randomly selected just one in the frame. This procedure resulted in testing 78 of the classes in COCO.

iNat2017 The iNaturalist Classification and Detection Dataset (iNat2017) is a fine-grained object detection dataset drawn from the entire iNaturalist labeled image repository [16]. The dataset is comprised of over 560,000 manually created bounding boxes of 5,000 species. Each species is additionally sorted into a supercategory. 88% of the images only contain a single instance. For these experiments, we split the iNat2017 dataset into two subsets. iNat-Marine (iNat-M) selected all available imagery associated with the 199 marine animals in iNat2017 for a total of 2854 images. iNat-Terrestrial (iNat-T) took a random subset of 57 of the non-marine organism concepts for a total of 1030 images.

FathomNet FathomNet is a large-scale database for sorting and working with annotated and localized underwater images [17, 18]. The images are associated with metadata including: the

[1] https://public.roboflow.com/object-detection/shellfish-openimages

[2] https://public.roboflow.com/object-detection/aquarium

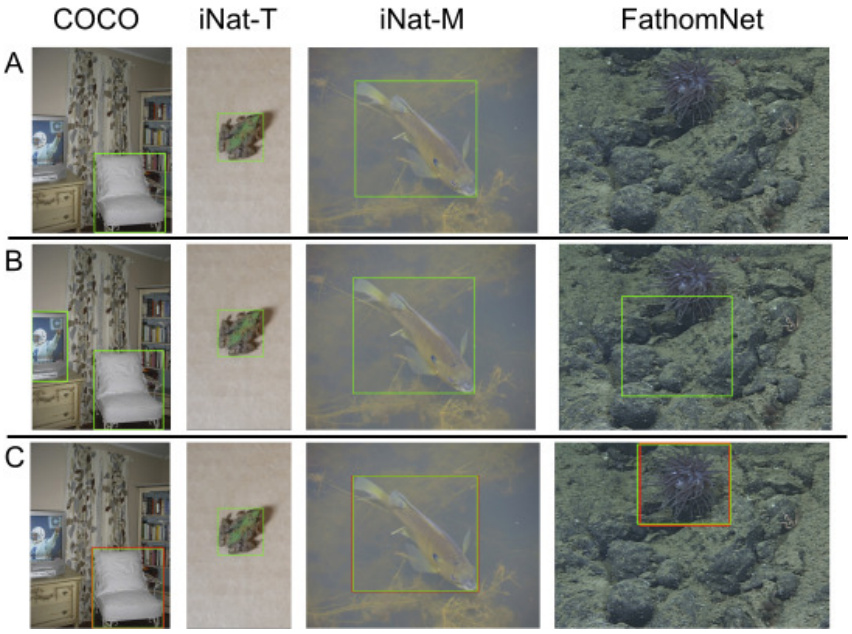

Figure 1: Example output on each dataset, indicated by column headings, from Grounding DINO. The target concept for each dataset was 'chair', '*Acris crepitans*' (frog), '*Lepomis macrochirus*' (fish), and '*Liponema brevicorne*' (anemone) for COCO, iNat-T, iNat-M, and FathomNet respectively. The prompts given to the model are: (A) 'piano', (B) 'crocodile', and (C) the target concept name. Green boxes are model output. Red boxes in row (C) are the ground truth associated with the concept name.

full taxonomic hierarchy associated with a concept; geographic coordinates; and environmental information. The database currently contains nearly 2400 fine-grained concepts of biologically and morphologically diverse animals. The images, especially those collected along the sea floor, often contain several localizations. The targets are usually small relative to the full frame image. For these experiments, we randomly selected 30 concepts from FathomNet imaged off the coast of Central California. From that subset, we picked 1009 random bounding boxes for an approximately uniform distribution of concepts.

2.2 Implementation details

Model Grounding DINO is a transformer-based VLM with grounded pre-training that can identify objects specified by human language inputs [14]. The model is a dual-encoder-single-decoder architecture that uses an image and text backbone with a variety of task-dependent fusion layers. For these experiments, we used the Grounding DINO-B checkpoint provided by the model developers that leverages a Swin transformer as the image backbone [19]. The checkpoint is pretrained with COCO, O365, GoldG, Cap4M, OpenImage, ODinW-35 and RefC [15, 20, 21, 3, 22–24].

Prompts We experimented with several prompts to explore the semantic space learned by Grounding DINO for animal localization (Figure 1). All four datasets were tested with the concept name, 'piano', and 'crocodile', and '*Caiman crocodilus*'. The concept name is the label associated with a given bounding box localization. 'piano' is a common human-made object, but not a category included in COCO dataset. 'crocodile' is a common animal class not found in FathomNet or our subsets of iNat2017. '*C. crocodilus*' is the Latinate scientific classification of one particular species in the order Crocodilia.

We also fed Grounding DINO rich prompts for three FathomNet concepts that the model struggled with in the first round of experiments: '*Swiftia kofoidi*', a deep sea coral; 'Caridea', an order of shrimp; and '*Liponema brevicorne*', a pom-pom shaped anemone. These three particular organisms were selected from all the FathomNet concepts Grounding DINO struggle with for their morphological

Table 1: mIOU and the proportion of detected objects for Grounding DINO output for each prompt on the target datasets. The proportion of detected objects is specified in parenthesis.

	Prompt			
Dataset	Concept	'piano'	'crocodile'	'C. crocodilus'
COCO 2017 val	0.85 (1.0)	0.44 (0.42)	0.43 (0.87)	0.42 (0.98)
iNat-T	0.88 (1.0)	0.90 (0.63)	0.89 (0.99)	0.89 (0.99)
iNat-M	0.87 (0.99)	0.89 (0.90)	0.88 (0.95)	0.88 (0.99)
FathomNet	0.49 (0.94)	0.52 (0.43)	0.56 (0.76)	0.53 (0.96)

and taxonomic diversity. We used sentences of the form "[CONCEPT] which is [FEATURE] and has [FEATURE]" to form the prompts:

- Swiftia kofoidi which is red or orange and has branches
- Caridea which is a red or orange and is elongated
- Liponema brevicorne which is globe shaped and has tentacles

Such expressive prompts have proven useful for improving the detectability of novel concepts in zero-shot settings.

Metrics To evaluate the results of each prompt, we computed the Intersection Over Union (IOU) between the region proposal from Grounding DINO and localizations identified by a human annotator. For each frame, the proposals resulting from a given prompt are compared to all human generated boxes associated with the target concept and the highest IOU is retained. The mean IOU (mIOU) is computed for each box generated by a prompt across all concepts in each dataset (Table 1). We note that the mIOU includes bounding boxes that are generated from prompts that do not match the target concept. For example, if 'piano' resulted in a bounding box that overlapped with a localization in an iNat-M image, the score will count toward the mIOU computation. The idea is to evaluate the model's ability to find salient objects even if the actual target concept's name is not present in the embedding space. The proportion of detections is used to quantify how often Grounding DINO found an object in the frame regardless of the prompt ($\#\ detection/\#\ boxes$).

3 Results

Grounding DINO was most successful on the in-distribution test done on the subset of the COCO 2017 val split (Table 1). When prompted with the appropriate concept name, the model returned a high mIOU and did not miss any localizations at all (Figure 1C). The model ignored the prompts for 'piano' more successfully than on the other datasets (Figure 1A). While the mIOU was low for both 'crocodile' and '*C. crocodilus*', prompting with both a common and scientific animal name returned a high number of detections (Figure 1B).

The mIOU was consistent across all prompts for the two iNat splits. The model found a localization when prompted with the appropriate scientific name for every iNat-T image. Notably, Grounding DINO returned fewer detections on iNat-T when prompted with 'piano'(Figure 1C). Otherwise, the model found a localization for every prompt that often overlapped with the real object(Table 1).

The model yielded low mIOU for all prompts when operating on FathomNet images, but also exhibited a low propensity toward finding objects that do not exist in the image (Figure 1). Grounding DINO produced better results on the three challenging classes when given expressive prompts, particularly the anemone *L. brevicorne* (Table 2).

4 Discussion

Our experiments used a variety of prompts to test the zero-shot capabilities of a popular VLM when applied to images of animals with a focus on marine species. The prompts included the scientific name of the object, a common object, and the scientific and common name of animal not present

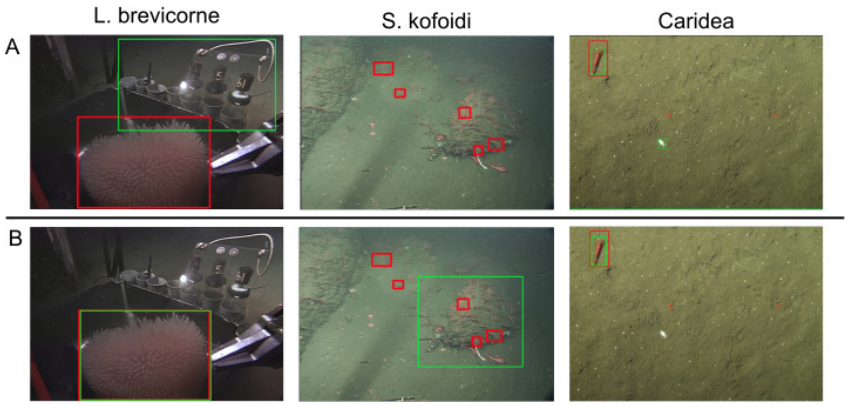

Figure 2: Using richer prompts resolved ambiguity in Grounding DINO output for the three FathomNet concepts named in each column. (A) Grounding DINO output using just the concept name and (B) with richer instructions based on the organism's morphology. Red boxes are ground truth and green boxes are model output.

Table 2: mIOU for Grounding DINO output on three FathomNet concepts with both basic and richer prompts.

	Prompt	
Concept	Species	Species+plain language
Swiftia kofoidi	0.04	0.05
Caridea	0.07	0.10
Liponema brevicorne	0.07	0.22

in the given zero-shot image set. The results underscore the strength of Grounding DINO's general understanding of objectness, regardless of the prompt, even when presented with images that are quite unlike the original training data.

In iNat images, where there is usually a single animal that stands out from the background, the model appears to have taken the language prompt to mean "find the foreground object" even if the word occupies a different semantic space than the actual target. This held true for all prompts in both iNat-T and iNat-M suggesting that using the scientific name for the target did not activate a different region of the embedding space.

FathomNet imagery posed a challenge for Grounding DINO, regardless of the prompt. The concepts in the dataset are morphologically strange objects that are often small relative to the frame, imaged through water instead of air, and referred to with scientific terminology not well-represented in most natural language datasets. This result is consistent with the observation that VLMs do not appear to generalize well to specialized tasks [2].

Expressive prompts improved the model's detection ability, particularly for the anemone *L. brevicorne*. In particular, manual inspection of the model output revealed the additional semantic information allowed the model to better distinguish between equipment and the target animal (Figure 2). The rich prompts used in these experiments were limited to additional color and shape information. Further experimentation should be done to determine what sorts of descriptors are most informative for VLMs in zero shot deployments on scientific image datasets. Likewise, richer prompts could be designed based on more complete descriptions of these organisms. However, collecting the necessary plain text information for many marine concepts is quite challenging; these creatures often do not have descriptions in free-content resources like Wikipedia or domain-specific repositories like the World Register of Marine Species.

Overall, these experiments suggest that VLMs like Grounding DINO are not globally applicable in zero-shot scenarios for domain tasks like animal localization in images. They are, however, potentially useful for generating saliency maps when prompted with apparent non-sequitur labels to assist human

annotators. The tests with expressive prompts are intriguing, suggesting that VLMs understand shape in a manner that extends to diverse body morphologies of marine organisms. More experimentation is needed to establish how effective such prompts are in new domains using such compositional understanding [25]. In particular, rigorous tests to understand which morphological descriptors are most informative to the VLM would provide valuable insght into the model's embedding space and suggestions for operators attempting to deploy them in a zero-shot capacity. While models like Grounding DINO can be fine-tuned with appropriately annotated domain specific data, figuring out how to use them with minimal additional training will be extremely valuable for scientists working in dynamic environments like the ocean that experience dramatic distribution shifts.

Acknowledgments and Disclosure of Funding

The authors wish to thank the program chairs of the I Can't Believe It's Not Better Workshop and the four anonymous reviewers for their invaluable feedback and suggestions. E.O. thanks Suzanne Stathatos, Benjamin Woodward, and Geneviève Patterson for the many discussions about the issues treated here.

The authors are grateful for the funding support we have received through the NSF Convergence Accelerator (ITE #2137977 and #2230776) and the David and Lucile Packard Foundation.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[3] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[4] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023.

[5] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616 (7956):259–265, 2023.

[6] Jean-Olivier Irisson, Sakina-Dorothée Ayata, Dhugal J Lindsay, Lee Karp-Boss, and Lars Stemmann. Machine learning for the study of plankton and marine snow from images. *Annual Review of Marine Science*, 14, 2021.

[7] Qimin Chen, Oscar Beijbom, Stephen Chan, Jessica Bouwmeester, and David Kriegman. A new deep learning engine for coralnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3693–3702, 2021.

[8] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[9] Alex J Hughes, Joseph D Mornin, Sujoy K Biswas, Lauren E Beck, David P Bauer, Arjun Raj, Simone Bianco, and Zev J Gartner. Quanti.us: a tool for rapid, flexible, crowd-based annotation of images. *Nature Methods*, 15(8):587–590, 2018.

[10] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI 15*, pages 456–473. Springer, 2018.

[11] Eric C Orenstein, Kasia M Kenitz, Paul LD Roberts, Peter JS Franks, Jules S Jaffe, and Andrew D Barton. Semi- and fully supervised quantification techniques to improve population estimates from machine classifiers. *Limnology and Oceanography: Methods*, 2020.

[12] Alice C Hughes, Michael C Orr, Keping Ma, Mark J Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu, and Huijie Qiao. Sampling biases shape our view of the natural world. *Ecography*, 44(9): 1259–1269, 2021.

[13] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[14] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[16] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[17] Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, et al. FathomNet: A global image database for enabling artificial intelligence in the ocean. *Scientific Reports*, 12(1):15914, 2022.

[18] Eric Orenstein, Kevin Barnard, Lonny Lundsten, Geneviève Patterson, Benjamin Woodward, and Kakani Katija. The FathomNet2023 competition dataset. *arXiv preprint arXiv:2307.08781*, 2023.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[20] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019.

[21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The Open Images Dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[23] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022.

[24] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 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, 2016.

[25] Madeline Chantry Schiappa, Michael Cogswell, Ajay Divakaran, and Yogesh Singh Rawat. Probing conceptual understanding of large visual-language models. *arXiv preprint arXiv:2304.03659*, 2023.