# Foundation Models need to be culturally fine-tuned

Jose Alfredo Garcia Alvarado[1*], Floris Erich[2], Tomohiro Motoda[2], Abdullah Mustafa[2], Yukiyasu Domae[2] and Ixchel Ramirez-Alpizar[2]

***Abstract*—This paper investigates the adaptability of Vision-Language Models (VLM) use in environments other than in the West, exemplified by CLIP. While models like CLIP exhibit commendable performance on established image datasets, their effectiveness in recognizing objects within specific cultural contexts remains an open question. Our experiments, conducted in a simulated environment, reveal noteworthy performance disparities between Western and Japanese datasets. Additionally, we explore the integration of a segmentation model to obtain segmentation masks with language-aligned features. By addressing these crucial gaps, our study provides insights into the nuanced challenges of cross-cultural recognition within the vision-language paradigm. These findings contribute to informed and unbiased model development for practical applications across diverse cultural domains.**

## I. Introduction

Since the emergence of token-based learning, as in [1], language acquisition has been gaining more ground over time. The release of Vision Transformers [2], a class of models that uses transformer architectures originally designed for natural language processing, has significantly influenced various fields. Vision Transformers have introduced transformer-based architectures to image data, demonstrating notable advancements in tasks such as image recognition. Progress in natural language learning has notably advanced image recognition, with models like Flamingo [3], Align [4], and CLIP [5] showcasing impressive performance. Their benchmark scores surpass conventional results significantly. While these models have excelled with well-known image datasets, their generalization to specific concepts or objects associated with certain cultural spheres remains an open question. Can these models replicate their remarkable performance for writing systems or objects outside the scope of their training? Our previous studies of such Foundation Models show signs of performance benchmarks inclining towards said Western objects. This paper conducts experiments to unveil the performance disparity between a Western product dataset and a Japanese dataset, focusing on isolating specific cultural domains. Products in both domains present themselves in different kinds of packaging, where imagery and language are crucial for visual recognition, as shown in Fig. 1 The experiments compare the performance within a simulation, set up in a robot arm manipulation environment.

$$Similarity(A, B) = A \cdot B / \|A\| \cdot \|B\| \tag{1}$$

* Corresponding author, reachable at jalfredo.gaal@gmail.com.
[1]Tecnologico de Monterrey, Monterrey, Mexico.
[2]National Institute of Advanced Industrial Science and Technology, Tokyo, Japan.


Fig. 1. Samples of object from both dataset used. First row corresponding to the YCB dataset. Second row corresponding to the Japanese convenience store products dataset.

## II. Preliminaries

### A. Zero-shot image-text similarity

The zero-shot recognition prowess of Constrastive Language-Image Pre-Training [5] enables it to generalize across modalities, effortlessly associating natural language descriptions with corresponding images without the need for specific training on paired data. This unique feature enhances the adaptability of the model and widens its applicability, positioning it as a powerful tool in various interdisciplinary domains where seamless cross-modal understanding is essential. CLIP integrates both a text encoder and an image encoder, each contributing to the generation of feature vectors. These normalized vectors can subsequently be compared through cosine similarity, as shown in Eq. 1, facilitating effective cross-modal matching. The model underwent training on ResNets and Vision Transformers (ViT) of varying sizes, further enhancing its performance.

### B. Image segmentation

Segment Anything Model [6] signifies a notable leap in automatic segmentation mask generation, rooted in a ViT-based architecture. SAM achieves precise and detailed segmentation through its distinctive positive/negative points-based segmentation methodology. SAM also excels in object delineation without manual annotation, and showcases remarkable accuracy in segmenting diverse objects across scenes. This automatic segmentation prowess is crucial for tasks like image understanding, scene parsing, and object recognition, providing a versatile solution for precise delineation and

localization of visual elements in images. SAM incorporates a semi-automated mask generation approach, strategically positioned between fully automated and assisted manual generation, employing a bounding box-based method. This nuanced feature aligns seamlessly with SAM's overarching objective of providing flexibility and adaptability in diverse segmentation scenarios.

## III. APPROACH

The main goal of this research is to set a formal precedent for the leap in performance of Large Language Models and Vision-Language Models in different cultural domains, and to highlight potential challenges stemming from model training biases. It is crucial to acknowledge that these models may excel primarily in English or Roman alphabet text-based pairs. The potential under-representation of other writing systems and cultural-specific contexts emphasizes the need for evaluating zero-shot recognition models across diverse datasets.

### A. Object Datasets

The YCB Object Set [7] serves as a designed benchmark for robotic manipulation, featuring a diverse array of daily life objects with varying shapes, sizes, textures, weight, and rigidity. Additionally, the dataset includes widely used manipulation tests and provides a rich set of high-quality online scans, enhancing its suitability for benchmarking in the experiments. In contrast, the Japanese object set, referred to as "Konbini" (the Japanese translation for "convenience store"), consists of common objects typically found in Japanese convenience stores, offering a culturally specific perspective.

### B. Environment

For the purposes of our experiments, a simulation was used to easily build controlled scenarios. To allow for robot integration we utilized the RAVENS Gym API [8]. RAVENS is an environment for robot task manipulation, based on PyBullet [9]. The PyBullet environment, while pre-loaded with various tasks, required adjustments to recreate a scenario closely aligned with object manipulation. These tweaks aimed to closely replicate a scenario relevant to object manipulation via a robotic arm, ensuring a controlled setting for evaluating zero-shot recognition performance. Fig. 2 shows an example of the configuration inside the simulation, using YCB Objects and Japanese convenience store objects.

### C. Text Labels

To prevent potential biases, external parties unrelated to the experiment provided the object descriptions used for zero-shot identification. Said external parties consisted of a native Japanese speaker and a proficient English speaker. This approach ensures that the descriptions remain untainted by any intentional tweaks that a person within the experiment might make to influence the model's recognition, thereby promoting a more accurate comparison.

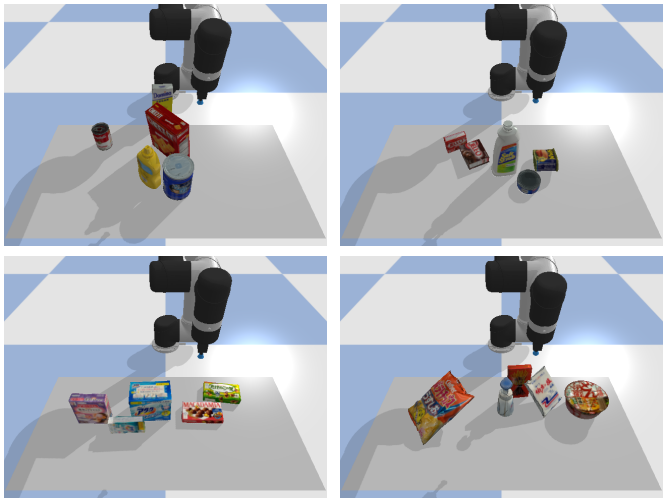

Fig. 2. Objects inside simulation from the camera point of view. Upper images are objects from the YCB dataset and bottom images are from the Japanese convenience store dataset.

## IV. EXPERIMENTS

We ran two sets of experiments to evaluate the effectiveness of applying Foundation Models to Western and Japanese products. The first experiment focuses on zero-shot image identification using a Vision-Language Model. The second experiment focuses on segmentation performance using a Foundation Model for image segmentation.

### A. Stand alone zero-shot image identification

Individual object masks from ground truth, generated by the PyBullet simulator, provide the best score for similarity of image and text tokens. This sets a fair comparison between the different sets of texts embeddings, since we are using the same object patch.

We asked the labelers to provide two sets of descriptions: (1) Natural descriptions that could include naming the product brand, and (2) descriptions in which naming the brand should be avoided. In addition, for Japanese objects, as the Japanese written language does not use roman characters, we asked the labeler to provide a description written using Japanese characters (hiragana, katakana and kanji).

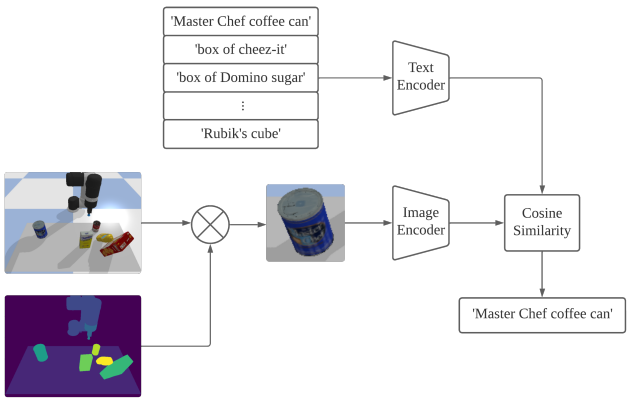


Fig. 3. General process for image and text data to perform object identification via CLIP. Both are encoded to then approximate similarity and output a best match.

*1) Setup:* We take 5 object out of the 15 objects per dataset, and arrange them into the scene. Each object undergoes five iterations to get a identification using the zero-shot recognition. The first of the iterations places the object in the workspace by itself, aiming to get the highest score in a clear view of the stand alone object. As for the rest of the four iterations, five objects are placed in the scene in random orientations and positions within the workspace visible to the camera view. The top 1 label for each object is considered for analysis. Fig. 3 shows the general process of the image and text data in each experiment iteration.

*2) Results analysis:* The scores of the five iterations per object are averaged, with the consideration of them being the correct match for the object according to which patch was sent to the CLIP model. In the opposite case, the score is stored as zero to represent a mismatch. The general performance was obtained through the average of all the objects. According to the performance reports of [5], the accuracy for the YCB dataset was expected to be around 80%. In Table I, we can see that as for the evaluations of the Japanese data set, the accuracy drops considerably from the score of the YCB. Given that the highest accuracy corresponds to the English labels considering the branding, it is possible to relate this occurrence with the poor Japanese characters recognition capability of the model. This statement is supported by the gap of almost 60% between the YCB descriptions and the Japanese descriptions, both considering branding, showing impact of text into performance. Among the Konbini labels (Japanese and English), when comparing the best and worst performing, it is possible to isolated the issue even further. The former highlights the capability of the model to recognize the general shape when prompted with the brand-less English description. In contrast, the later showcases a weak identification of the same general shape prompted in Japanese text. Even so, the results for the Japanese description considering branding indicate the relevance of the object names written in Japanese.

TABLE I
AVERAGE ACCURACY OF EACH DATASET.

| | W/o brand | W/brand |
|---|---|---|
| **YCB** | 79.16 % | 82.48% |
| **Konbini** (English) | 63.33 % | 48.78% |
| **Konbini** (Japanese) | 10.87% | 24.39% |

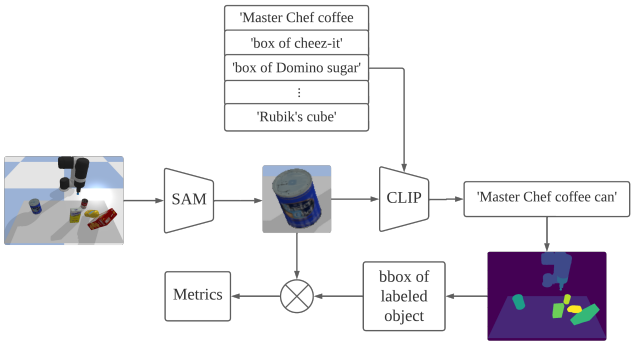


Fig. 4. General process for image and text data to identify objects via SAM and CLIP integration. SAM outputs a segmentation mask to match with a text label, the resulting label is used to obtain the ground truth for that object and both values are weighted against each other.

## B. Segmentation model identification against ground truth

By using a segmentation model such as SAM in combination with a Vision-Language model such as CLIP, we can obtain segmentation masks with language-aligned features. Since we are using the simulation, it is possible to automatically determine the ground truth of the object identification and evaluate it against the SAM patch with the top 1 score. The intersection over union (IoU) of these two segments can showcase the accuracy of the model for recognition of the better image samples. Fig. 4 shows the general process of this experiment, mainly in contrast to the previous one.

*1) Setup:* For each object in the data set, there is an iteration of the simulation. In each of these iterations an image is rendered and sent to the SAM model. Using the automated mask generation feature, patches above a 0.9 stability score threshold are inputted to the recognition model and the best scored patch is kept as a match of a correct bounding box. Since the environment is controlled and just includes one of each object, there can just be one object of each label in the scene. Therefore, when a high enough recognition precedes an existent one, the new one overwrites the previous.

*2) Results analysis:* Since not many filters or considerations are done besides those exposed in the setup section. Some matching can be noisy in the sense of getting a larger than needed area, which can dull the final Intersection over Union (IoU) area.

The results in this experiment, included in Table. II, follow a similar trend as the previous one, since the best score in all metrics but one is the YCB dataset with the branding consideration and the lowest is the Konbini dataset

TABLE II
QUANTITATIVE RESULTS OF SAM AND CLIP INTEGRATION.

| Dataset | W/o brand | | | | | W/brand | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1-score | IoU | Accuracy | Precision | Recall | F1-score | IoU | Accuracy | Precision | Recall |
| YCB | 0.75 | 0.31 | 0.75 | 0.75 | 0.75 | 0.76 | 0.32 | 0.76 | 0.75 | 0.78 |
| Konbini (Eng) | 0.57 | 0.25 | 0.64 | 0.70 | 0.49 | 0.46 | 0.19 | 0.60 | 0.73 | 0.34 |
| Konbini (Jap) | 0.26 | 0.06 | 0.55 | 0.72 | 0.16 | 0.41 | 0.12 | 0.59 | 0.76 | 0.28 |

in Japanese without the branding. IoU is considered as the main metric for success in each event, obtained by using the patch from [6] against the simulation ground truth of the supposed identified object. Additionally, other metrics related to the performance of the model were calculated, such as F1-score, based on the success of [6] and [5] to coincide on an object detection (be it a stable enough mask or a high enough score, respectively). F1-scores follow the previous experiment trend closely, except the scores of the Konbini dataset in Japanese are considerably higher. While the performance may seem elevated as a whole, a closer look at the other metrics, like IoU or Recall, shows the lack of relevant items retrieved in each evaluation for the Konbini dataset objects.

## V. CONCLUSIONS

The experiment comparing western and Japanese datasets revealed performance disparities in the model performance, raising the question of how these models generalize and the specific assets considered to build this capability. The approach emphasized the need to set a formal precedent for model performance in diverse cultural domains, acknowledging potential biases and limitations, particularly in English or Roman alphabet text-based pairs. Using the YCB Object Set as a benchmark, the study lays the groundwork for understanding challenges in recognizing objects within a cultural context. To better understand the implications, it is essential to consider how these findings might affect related projects like Robotics Transformer [10] or SayCan [11]. This highlights the need for guidelines to inform future studies, specifically exploring the impact of cultural context on recognition. Looking ahead, further experiments are needed. Examining a wider range of models and applying our findings to real robot manipulation scenarios will offer valuable insights. This exploration will contribute to a richer understanding of model performance, guiding the integration of projects like [10] or [11] in practical applications, specially in term of localization to a certain region. As for fine-tuning Foundation Models to have higher performance in foreign cultures, recently LoRA [12] has arisen as a promising technique, which we will explore in the future.

## APPENDIX

As mentioned earlier, third parties unrelated to the experiment provided text description in natural language for each of the objects used. The exact descriptions are as shown in Table III and Table IV.


## ACKNOWLEDGMENT

Jose Alfredo Garcia Alvarado extends his sincere gratitude to the AIST for providing him with the opportunity to serve as a technical trainee. He also thanks Luis Alberto Muñoz Ubandu for his exceptional tutelage and unwavering support. His mentorship has been instrumental in shaping Alfredo's understanding and contributing to the success of this endeavor.


## REFERENCES


[1] A. Vaswani, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Long Beach, CA, USA, Dec 2017, pp. 5998–6008.

[2] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. on Learning Representations (ICLR)*, Virtual Conference, May 2021.

[3] J.-B. Alayrac, "Flamingo: a visual language model for few-shot learning," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, New Orleans, LA, USA, Dec 2022, pp. 23 716–23 732.

[4] C. Jia, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. on Machine Learning (ICML)*, vol. 139, Virtual Conference, Jul 2021, pp. 4904–4916.

[5] A. Radford, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. on Machine Learning (ICML)*, vol. 139, Virtual Conference, Jul 2021, pp. 8748–8763.

[6] A. Kirillov, "Segment anything," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Paris, France, Oct 2023, pp. 4015–4026.

[7] B. Calli, "Benchmarking in manipulation research: Using the yale–cmu–berkeley object and model set," *IEEE Robotics and Automation Magazine*, vol. 22, no. 3, pp. 36–52, Sep 2015.

[8] A. Zeng, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Proc. Conf. on Robot Learning (CoRL)*, vol. 155, Cambridge, MA, USA, Nov 2020, pp. 726–747.

[9] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," http://pybullet.org, 2016–2021.

[10] A. Brohan, "Rt-1: Robotics transformer for real-world control at scale," in *Proc. Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, Jul 2023.

[11] M. Ahn, "Do as i can, not as i say: Grounding language in robotic affordances," in *Proc. Conf. on Robot Learning (CoRL)*, vol. 205, Auckland, New Zealand, Dec 2022, pp. 48–61.

[12] E. J. Hu, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. on Learning Representations (ICLR)*, Virtual Conference, May 2022.

TABLE III
YCB DATASET DESCRIPTIONS

| W/o Brand | W/Brand |
|---|---|
| blue can of coffee | Master Chef coffee can |
| red box of cookies | box of cheez-it |
| white and yellow box of sugar | box of Domino sugar |
| red and white can of soup | Campbell condensed tomato soup |
| yellow bottle of mustard | French's mustard bottle |
| little blue tin of tuna | StarKist Tuna tin |
| brown box of chocolate jelly | box of Jell-o jelly |
| red box of strawberry jelly | Jell-o box of strawberry jelly |
| blue and yellow tin of ham | SPAM tin |
| white bottle of cleanser | bottle of Soft Scrub cleaner |
| red and navy blue spray bottle | Windex cleaner bottle |
| rectangular irregular green object | sponge covered with green plastic |
| white and black cylindric marker | Black Expo marker |
| blue and black pointed object | black and blue Phillips screwdriver |
| cube of different square colors | Rubik's cube |

TABLE IV
KONBINI DATASET DESCRIPTIONS

| English Labels | | Japanese Labels | |
|---|---|---|---|
| W/o Brand | W/Brand | W/o Brand | W/Brand |
| green box of chocolate cookies | Takenokonosato box | 黄緑色のお菓子の箱 | たけのこの里 |
| big blue box of bleach | Attack bleach box | 青い洗剤の箱 | アタック リセットパワーの箱 |
| white and red box of chocolates | Macadamia chocolate box | チョコレートの箱 | MACADAMIAチョコレートの箱 |
| blue and white small detergent box | Attack Neo detergent box | 洗剤の箱 | アタックの箱 |
| purple box of eye masks | MegaRhythm box | 紫色のアイマスクのパッケージ | 紫色のめぐりズムのパッケージ |
| red rectangular chocolate cookie box | chocolate Pocky box | 赤い箱 | 赤いポッキーの箱 |
| hand soap white bottle | Biore soap | ハンドソープのボトル | ビオレのボトル |
| big potato chips bag | Calbee chips bag | ポテトチップスの袋 | ポテトチップスうすしお味 |
| soba red bowl | Nissin soba bowl | カップめん | どん兵衛 |
| sea salt bag | Hakatanoshio salt bag | 塩の袋 | 伯方の塩 |
| small soy sauce white bottle | kikkoman soy sauce bottle | 醤油のボトル | キッコーマンの醤油ボトル |
| green bottle of green tea | Oi Ocha bottle | 緑色のペットボトル | おーいお茶のペットボトル |
| bag of corn flavored chips | Toomoriko chips | 黄色いスナック菓子の袋 | トウモリコの袋 |
| blue bag of peanuts | tabekiri peanut bag | 水色のスナック菓子の袋 | 海味鮮というスナック菓子 |
| yellow box of apple and honey curry | Baamoto Curry box | カレールーの箱 | バーモントカレーの箱 |