Weight Estimation for Robot Interaction

Elizabeth Ottens
Stanford University
lottens@stanford.edu

James Van Kirk
Stanford University
jvk@stanford.edu

Abstract

Most computer vision applications related to object manipulation in robotics are focused on geometric shape recovery and pose estimation. While object pose and size are crucial for robot manipulation tasks, there are many more characteristics of objects that would be useful to recover using vision alone. By correlating visual appearance to weight and volume, we aim to enable robots without sophisticated or expensive force sensors to perform a more diverse range of tasks. In this work, we explore an end-to-end pipeline for the estimation of physical characteristics from visual imagery, with the overall goal of integrating this system into robotic platforms. Our proposed pipeline will take in visual data captured by cameras, detect and segment objects of interest within the field of view, and subsequently perform estimation of key physical characteristics of these objects such as weight and volume. This data may then be used by a robotic system to make informed decisions for grasping and manipulation tasks. Novel findings are the strong performance of multimodal large language models (M-LLMs) compared to prior state-of-the-art computer vision methods and human performance on the task of object weight or mass estimation on monocular RGB images.

1. Introduction

Most computer vision applications related to object manipulation in robotics are focused on geometric shape recovery and pose estimation. While object pose and size are crucial for robot manipulation tasks, there are many more characteristics of objects that would be useful to recover using vision alone. For example, extremely dense objects may need to be gripped differently than light, airy objects to prevent damage. At times a delicate balance of grip and force needs to be applied to sufficiently interact with an object without destruction, such as when picking oranges from a tree: each fruit must be squeezed sufficiently for a robotic arm to be able to detach each one off a branch, while not squeezing so much so to burst the fruit. During complex manipulation tasks such as catching or throwing, weight and density become even more important. Additionally, with accurate weight, volume, and density estimates, we can better estimate force limits and safety parameters for objects (i.e. extremely dense objects are less likely to be fragile).

In this work, we explore an end-to-end pipeline for the estimation of physical characteristics from visual imagery, with the overall goal of integrating this system into robotic platforms. Our proposed pipeline will take in visual data captured by cameras, detect and segment objects of interest within the field of view, and subsequently perform estimation of key physical characteristics of these objects such as weight and volume. By using visual appearance to estimate the size (length, width, height) and weight of objects from 2D images, robots can perform a wider range of tasks without relying on expensive and sophisticated force sensors. A computer vision based pipeline has the potential to enable weight or volume estimation of objects and support a wide variety of additional applications while minimizing the cost of single purpose sensors in robots. This data may then be used by a robotic system to make informed decisions for grasping and manipulation tasks. We aim not only to enhance the manipulation capabilities of robots but also to reduce the cost and complexity for robust robotic systems. This research can contribute to more versatile and cost-effective robotic solutions in various applications, from industrial automation to service robots in everyday environments.

We use two datasets in our results. First, we use the image2mass [9] household test set as a baseline comparisons. This is a manually collected dataset of 56 household objects with ground truth mass and size across numerous categories. Additionally, we collected a small dataset of 10 common objects (individually and in context) from overhead and angled side views. Using these datasets, our pipeline is used to estimate key properties of objects that enable robotic grippers to handle and manipulate objects more deftly. Performance is measured comparatively on items from the image2mass household test set and to ground truth on our custom dataset using weight and size estimation error. We show that multimodal large language models (M-LLMs) drastically outperform baseline results.

2. Related Work

Estimation of object parameters for robot manipulation is an active area of research. In DeepPhysNet [10], the authors present a method for learning the physical properties of objects through a series of dynamic interactions (i.e. sliding, rolling). Image2mass [9] provides an extensive dataset of images with size and weight information and explores weight estimation with 2D images. In particular, the Shape Aware model is a deep architecture for estimating the mass of an object using an RGB image and its dimensions. The model captures density and volume information, multiplying them to obtain weight. There are separate network towers for both volume and density. In addition to these two towers, Shape Aware has a module for estimating an object's 3D shape. At test time, the inputs to Shape Aware are an image (single object, white background) and the item's physical dimensions. While this approach performed well (in some cases on par with human performance), it's use cases are quite limited since it cannot handle images in context and it requires size information on input. In [5], the authors present an approach to estimating physical properties of objects from video using a physics engine and a correction estimator to compare physical and simulated videos. Various other weight and shape estimation methods are presented in [8], [6], [1]. There is very little research on the similar problem of volume estimation, with the only direct connection coming from medical imaging [2].

3. Datasets

3.1. Image2Mass Amazon Products Dataset

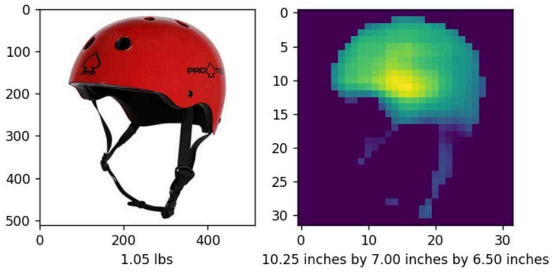


Figure 1. Visualization of samples from the Image2Mass Amazon Products Dataset. On the left is a sample raw image, and on the right is its estimated 3D reconstruction.

Image2Mass [9] consists of two primary datasets, the first being an extensive dataset containing approximately 150,000 products that have highly accurate absolute mass and size information, covering a multitude of object classes and was used to train the Shape Aware model (Figure 1). Item weights and dimensions were harvested from Amazon.com. Image2mass also contains a smaller, manually collected dataset of 56 household objects with ground truth mass and size. The image2mass data objects are stored as .pklz files and contain images, weights (in pounds), and volumetric dimensions (in inches). The household test set is a manually collected dataset of 56 household objects with ground truth mass and size across numerous categories.

3.2. Household Objects Dataset

The Household Objects dataset is manually collected set of 56 household objects with ground truth mass and size. In the original dataset, each object was captured from numerous different viewpoints to assist in weight determination. We use the image2mass [9] Household Objects Set for testing comparative and baseline testing with approaches that are capable of accurately estimating size and shape from a single image.

Figure 2. Examples from the Image2Mass Household Objects Test Set.

3.3. Our Home Objects Dataset

We collected a dataset of 10 home objects from overhead and from an angled side view and measure ground truth sizes and weights of all objects using a measuring tape and scale. Objects range in weight from 29g to 1.4kg and have dimensions ranging from 1cm to nearly 40cm.

Images were taken both individually and in context with other objects. All images were captured on an iPhone 12 Pro. For overhead images, the camera was positioned 60 inches above the table and parallel to the plane at a 2x zoom. For the angled images, the camera was placed 19 inches (straight line) from the blue marker in the image at an elevation of 15 inches at a 1x zoom.

To improve flexibility of our dataset, we built functionality using the Segment Anything Model [3] to segment and save individual objects or collections. For the small dataset, masks were generated for all objects and desired masks were manually selected and saved.

An excerpt of our Home Objects Dataset containing ground truth size and weight information is shown in Table 1.

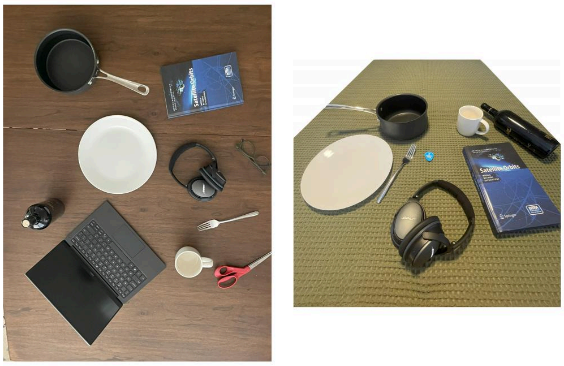


Figure 3. Top View (Left) and Angled View (Right).

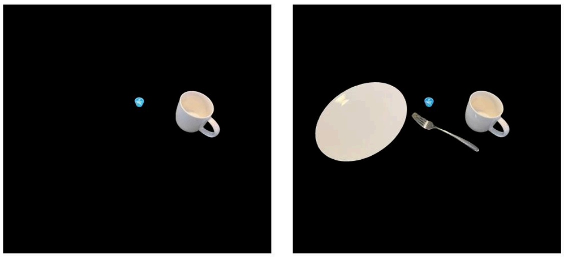


Figure 4. Individual Segmentation (Left) & Set Segmentation (Right).

Object	Weight (g)	Length (cm)	Width (cm)	Height (cm)
Glasses	29	13.00	40.00	3.0
Fork	45	20.00	2.50	1.0
Scissors	77	21.00	8.00	2.0
Headphones	196	19.00	15.00	6.0
Mug	315	8.50	12.00	9.0
Plate	366	26.00	26.00	2.0
Book	715	24.00	16.50	2.5
Pan	810	37.50	19.00	10.0
Laptop	1,172	30.00	20.00	2.0
Wine	1,365	8.25	8.25	30.0

Table 1. Object Measurements and Weight Estimates.

4. Methods

4.1. Multi-Stage Computer Vision Pipeline

In order to solve the weight estimation task for robot interaction with objects via end effectors without incorporating additional specific sensors, we propose a multi-stage computer vision pipeline consisting of detection, segmentation, and direct image to weight estimation via machine learning methods. Here, images from the camera are fed as input to an object detection model, segmented, and passed as input to a direct image to weight estimation model. We use off the shelf object detection and segmentation models, including Detectron2 and Segment Anything Model [3], and evaluate performance of weight estimation methods for the final stage of our pipeline.

4.2. Image Segmentation

Essential to our pipeline, the Segment Anything Model [3] provides a versatile model for various segmentation tasks. SAM was trained on the extensive SA-1B dataset, which contains over 1 billion segmentation masks across 11 million high-resolution, diverse, and privacy-protected images. This large-scale dataset ensures that SAM can handle a wide range of segmentation scenarios with high quality and diversity. SAM introduces the concept of promptable segmentation in which users can provide text, points, or bounding boxes. This flexibility allows for returning every mask in an image, or only those at in specific areas.

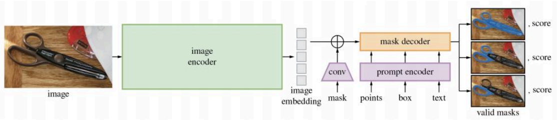


Figure 5. Segment Anything Model Diagram [3].

The model architecture has three main components: an image encoder based on a vision transformer, a prompt encoder, and a mask decoder (Figure 5). SAM also has an impressive zero-shot transfer capabilities to new segmentation tasks. The model performs well under various prompt qualities, although its segmentation stability can vary depending on the precision of the prompts.

Overall, SAM represents a significant advancement in the field of image segmentation, providing a robust, flexible, and scalable solution that can handle a wide range of tasks with high accuracy and real-time performance. The model's ability to generalize across tasks and its foundation on a vast and diverse dataset make it a powerful tool for both research and practical applications in computer vision. for example, the segmentation portion of our pipeline will leverage sample images from the SA-1B Dataset released with Segment Anything [4].

4.3. Machine Learning for Weight Estimation

Our baseline method is a state-of-the-art computer vision model architecture called the Image2Mass Shape-Aware Model. This model estimates the mass of an object from an input image by combining a density tower, a volume tower, and a geometry module in its architecture, whose architecture is visualized in detail in Figure 6). The geometry module is pre-trained using 3D models and takes as input an RGB image from which it produces a 2D thickness mask to estimate the object's thickness at each pixel to capture 3D shape. This mask, along with geometric features derived from it, is fed into both the volume and density towers. The density tower of the model architecture is a CNN-based network which extracts features related to the object's material, while the volume tower is a fully connected network that estimates the object's volume. These two towers are trained jointly, and their outputs are multiplied to produce the final mass estimate.

We use this model architecture as a baseline as it significantly outperforms several other baselines in the literature, including Xception k-NN and pure convolutional neural network (CNN) baselines, and is competitive with human performance on weight estimation tasks.

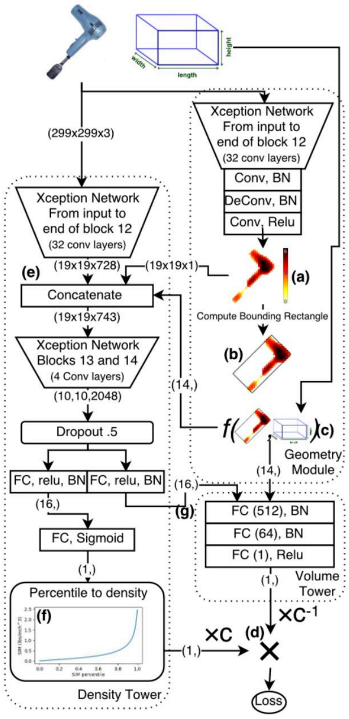


Figure 6. Baseline Image2Mass Shape Aware Model Architecture.

4.4. M-LLMs for Weight Estimation

Multimodal large language models (M-LLMs) are models that are capable of integrating and processing multiple types of data such as text and images to generate more comprehensive and semantically meaningful outputs. M-LLMs capable of understanding and reasoning across input modalities, allowing them to leverage the strengths of each type of data to improve their predictions and analyses.

In the context of an image to object weight prediction task, M-LLMs combine visual information from images taken at different viewpoints with relevant textual data to guide the model toward desirable output. Given the breadth of training data, the model has innate understanding of object material properties, context, and usage, leading to more accurate weight predictions. Additionally, multimodal models can draw on extensive pre-trained knowledge from diverse sources, allowing them to generalize better across different object types and scenarios.

Motivated by the failure cases of machine learning approaches for direct image to weight prediction for objects, we explore pre-trained M-LLMs and M-LLMs with instruction set fine tuning. We leverage GPT-4o [7], which is a state-of-the-art M-LLM trained on an extensive dataset consisting of billions of text tokens and images. The size of its training dataset and training process involving a mix of supervised learning and reinforcement learning from human feedback enhances its ability to understand content, context, and generate coherent and contextually relevant responses.

We also explore instruction tuning for language models through the customizable functionality of GPT-4o, creating an 'ObjectGPT' model to evaluate for the task of weight prediction. ObjectGPT may be provided with an image file with multiple objects as input for which the model predicts weight and size information. A link to ObjectGPT is provided here.

5. Experiments and Results

We evaluate our pipeline using weight estimation measurement error (%) for a diverse variety of object shapes, sizes and textures. In our experiments, we compare the Shape Aware Model which yields state-of-the-art performance on the Image2Mass datasets to M-LLMs. Since prior work found performance of the Shape Aware Model comparable to human-level performance on object weight estimation tasks, this gives us an implicit comparison of our methods to human-level performance on the same datasets as well. We compare results on both the Household Objects Dataset and our Home Objects Dataset and share key findings below, with complete results detailed in the Appendix.

5.1. Image2Mass Baseline

The baseline for evaluation is the Image2Mass Shape-Aware Model which estimates the weight of an object from its image by leveraging a density tower, a volume tower and a geometry module. We use this model architecture as a baseline as it significantly outperforms several other baselines in the literature, including Xception k-NN and pure convolutional neural network (CNN) baselines, and is competitive with human performance on weight estimation tasks.

On the Household Objects Test Set, the Shape Aware Model achieves 66% average weight estimation error. This error is too high for this approach to be considered for deployment in production robotics systems, and its similarity to human performance on weight estimation tasks yields insights into the difficulty of the task for humans as well.

Upon detailed failure case analysis, we found that cases with a significant amount of weight estimation error often correlated with particular material properties of the objects and their relationship to weight. For example, key categories from the Household Objects Dataset with unique material properties that have a significant impact on object weight are highlighted in Tables 3 and 4. These include cardboard such as in the Google Cardboard product, plastic in Lego Bricks, stuffing in Plush Toys, or metal in common objects.

This also revealed an insight into the challenges of approaching this task with machine learning-based approaches as there is often a significant amount of intra-class variation due to the material properties of objects. For example, chairs with identical geometrical forms and colors may have significantly different weights based on their material properties (e.g., whether they are made of plastics, aluminum, metal, or other), identifying a need for higher level contextual understanding of object context, form, function and material properties to more accurately predict their weight.

5.2. M-LLMs

M-LLMs have an innate understanding of object material properties, context, and usage, leading to more accurate weight predictions than other methods. Additionally, multimodal models can draw on extensive pre-trained knowledge from diverse sources, allowing them to generalize better across different object types and scenarios.

We find that M-LLMs drastically outperform our baseline method, particularly in cases where an understanding of the material properties of objects is essential or where objects are common products where the model was likely exposed to information about the object and its weight during training on internet data. M-LLMs yield an average weight estimation error of 38% on the Household Objects Dataset, drastically outperforming our baseline results on the object weight estimation task.

Method	Average Weight Estimation Error (%)
Shape Aware Model	66%
ObjectGPT M-LLM	38%

Table 2. Error Comparison on the Household Object Set

Shown in the Tables 3 and 4 are the ground truth and predicted weight values in addition to the weight prediction error (%) on the Household Test Set for select challenging object categories. The selected categories are particularly challenging for traditional approaches as they include objects with diverse materials such as cardboard, plastic, and stuffing. This leads us to conclude that semantic understanding of material properties and their relationship to object weight can greatly reduce estimation error. In other cases, general knowledge of the mass of objects may have been implicitly learned by LLMs when exposed to product information on the internet.

Category	M-LLM Error (%)	Shape-Aware Error (%)
Google Cardboard	41%	374%
Red Lego Brick	79%	256%
Green Lego Brick	79%	288%
Marshy the Elephant	4%	190%
Dust Pan	23%	429%
Hair Dryer	14%	153%

Table 3. Weight Prediction Error (%) for 6 Key Objects.

Category	Ground Truth (g)	M-LLM (g)	Shape-Aware (g)
Google Cardboard	71	100	335
Red Lego Brick	12	3	42
Green Lego Brick	12	3	46
Marshy the Elephant	209	200	605
Dust Pan	195	150	1034
Hair Dryer	599	680	1513

Table 4. Actual and Predicted Weight Values for 6 Key Objects.

Detailed results containing the ground truth and predicted weight values as well as the weight prediction errors (%) on the Household Test Set for all 56 object categories and approaches are detailed in the Appendix.

5.3. ObjectGPT on our Home Objects Dataset

Using images from our dataset, we prompted ObjectGPT with raw and segmented images from our dataset. For the individual items, all images were taken from the angled setup. For the group items, images were taken from the overhead setup. Images were segmented using SAM. Each approach was tested in two ways. First, we provided context on camera type, positioning, and zoom level. Next we provided the input images without any context in a new session to ensure the model did not repeat the same predictions from prior knowledge. Table 5 shows results without context prompting and table 6. The context prompts were

- **Angled Dataset Prompt**: The blue circular marker is 19 inches (straight line distance) from the lens of the camera. The camera is elevated 15 inches off of the table. Do not provide size/weight estimates for the blue marker, just the other objects in the image.

- **Overhead Dataset Preamble:** The camera is parallel to and 60 inches above the object plane.

Category	Weight Error	Length Error	Width Error	Height Error
Masked Single	2.93	0.14	-2.73	6.08
Raw Single	-7.30	0.14	16.98	0.80
Masked Group	-2.65	-4.95	6.29	4.44
Raw Group	-11.39	-8.81	-2.10	4.44

Table 5. Average Prediction Errors (%), With Context.

Category	Weight Error	Len Error	Width Error	Height Error
Masked Single	-2.16	-19.42	-15.43	-3.70
Raw Single	-0.68	-4.18	0.63	7.20
Masked Group	-2.75	2.77	-2.10	3.70
Raw Group	-0.59	-9.77	-1.33	2.96

Table 6. Average Prediction Errors (%), No Context.

In all cases, results were sufficiently close to be informative and useful in a robot vision system. All estimations were qualitatively reasonable and surprisingly accurate. When provided with context on the image frame and camera positioning, the segmented images provided significantly better results. However, with no context about camera pose, unsegmented images performed significantly better. This is likely due to the model implicitly contextualizing the objects when there is a clear background (and in the case of individual objects, horizon lines).

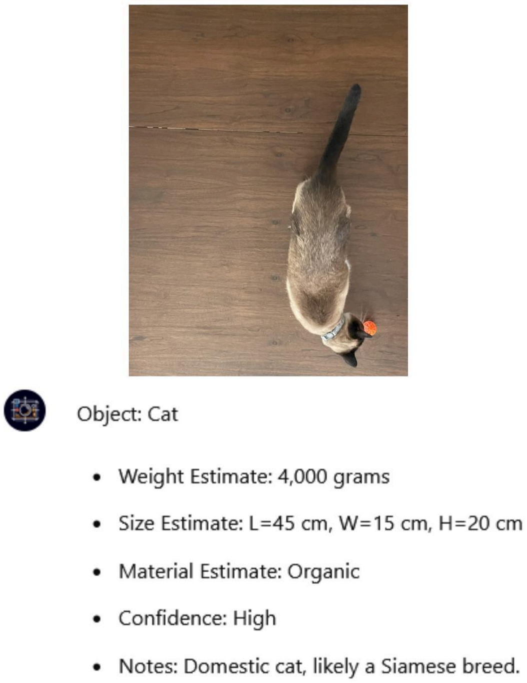
We captured a bonus image of of the authors' cat that came to investigate while capturing our dataset. As a test of generality, the image was provided to ObjectGPT without additional context. The model provided a weight within 10% of true weight and provided reasonable responses demonstrating the flexibility of this approach. The image and response is shown in Figure 7.

6. Conclusions and Future Work

Determination of object weight using computer vision systems is notoriously difficult. This work demonstrates the capability of leveraging M-LLMs for direct estimation of object weight, size and material properties from 2D images. We find that ObjectGPT M-LLM significantly outperforms prior state-of-the-art computer vision models and human level performance on the task of object weight estimation from images.

End-to-end computer vision pipelines including object detection, segmentation, and weight estimation benefit from using M-LLMs for the final object weight estimation pipeline stage given M-LLM's semantic understanding of object material properties, geometry, usage and their relationships with weight. Without the need for expensive sensors or significant computational overhead, vision pipelines on internet connected robots can leverage M-LLMs to extract rich object level data.

Future work includes fine-tuning M-LLMs with object material types, segmentation masks, depths and weight information and building a dataset with diverse camera viewpoints for validating performance across varied in-the-wild camera poses on robots. Additionally, we would like to explore building a drop-in ready vision system that provides real-time information for robot motion planning and manipulation tasks.



Object: Cat

- Weight Estimate: 4,000 grams
- Size Estimate: L=45 cm, W=15 cm, H=20 cm
- Material Estimate: Organic
- Confidence: High
- Notes: Domestic cat, likely a Siamese breed.

Figure 7. Generalization ability of ObjectGPT as shown through testing with an image of a live cat that was certainly not included in the Amazon Products or Household Objects Datasets.

References

[1] Tamas Aujeszky, Georgios Korres, Mohamad Eid, and Farshad Khorrami. Estimating weight of unknown objects using active thermography. *Robotics*, 8(4), 2019. 2

[2] Yikun Jiang, Peixin Li, Yungeng Zhang, Yuru Pei, Yuke Guo, Tianmin Xu, and Xiaoru Yuan. 3d volume reconstruction from single lateral x-ray image via cross-modal discrete embedding transition. In Mingxia Liu, Pingkun Yan, Chunfeng Lian, and Xiaohuan Cao, editors, *Machine Learning in Medical Imaging*, pages 322–331, Cham, 2020. Springer International Publishing. 2

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2, 3

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 3

[5] Martin Link, Max Schwarz, and Sven Behnke. Predicting physical object properties from video, 2022. 2

[6] Oier Mees, Maxim Tatarchenko, Thomas Brox, and Wolfram Burgard. Self-supervised 3d shape and viewpoint estimation from single images for robotics. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6083–6089, 2019. 2

[7] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. 4

[8] Prashanth Shivakumar Iyer, Rohan Soneja, and R Aarthi. Body weight estimation using 2d body image. *International Journal of Advanced Computer Science and Applications*, 12:304 – 320, 05 2021. 2

[9] Trevor Standley, Ozan Sener, Dawn Chen, and Silvio Savarese. image2mass: Estimating the mass of an object from its image. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 324–333. PMLR, 13–15 Nov 2017. 1, 2

[10] Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B. Tenenbaum, and Shuran Song. Densephysnet: Learning dense physical object representations via multi-step dynamic interactions, 2019. 2

Appendix

A.1. Prediction Results

Shown in the four tables below are the ground truth and predicted weight values as well as the weight prediction error (%) on the Household Test Set for all object categories and approaches.

Category	True Weight (g)	M-LLM Prediction (g)	Shape-Aware Model Prediction (g)
Big Elephant	582	300	540
Small Elephant	235	250	239
Big Bowl	566	500	504
Small Bowl	151	170	181
Hose Splitter	187	200	147
Kitchen Knife	104	200	134
Screwdriver	106	100	89
Google Cardboard	71	100	335
Wooden Spoon	60	70	127
Big Stapler	167	150	135
Little Stapler	45	100	62
Can Opener	178	250	183
Mouse	134	85	99
Presentation Remote	58	70	71
Plate	176	500	200
Lighter	78	50	107
Block	226	200	210
Toy Airplane	162	200	231
View Master	118	150	156
TV Remote	189	150	184
Pliers	217	200	199
Hammer	471	450	457
Drill	1208	1497	2591
Box of Soap	120	113	131
Bar of Soap	112	100	80
Chip Clip	61	50	113
Safety Razer	42	30	75
Tape Measure	194	230	115
Pillow	716	400	1557
Red Lego Brick	12	3	42
Green Lego Brick	12	3	46
Ivan's Phone	131	150	129
Ollie the Monkey	132	250	157
Marshy the Elephant	209	200	605
Bottle Brush	127	100	136
Hat	88	50	163
Airplane Clock	166	250	159
Sun Chips	207	198	392
Rope Bundle	66	150	106
Twine	85	200	147
Dust Pan	195	150	1034

Table 7. Ground Truth and Predicted Weight Values on the Household Test Set for the First 41 Objects.

Category	M-LLM Weight Prediction Error (%)	Shape-Aware Model Weight Prediction Error (%)
Big Elephant	48%	7%
Small Elephant	7%	2%
Big Bowl	12%	11%
Small Bowl	13%	20%
Hose Splitter	7%	21%
Kitchen Knife	92%	29%
Screwdriver	6%	16%
Google Cardboard	41%	374%
Wooden Spoon	17%	112%
Big Stapler	10%	19%
Little Stapler	123%	39%
Can Opener	41%	3%
Mouse	36%	26%
Presentation Remote	22%	24%
Plate	184%	13%
Lighter	36%	38%
Block	12%	7%
Toy Airplane	23%	42%
View Master	27%	32%
TV Remote	21%	3%
Pliers	8%	8%
Hammer	4%	3%
Drill	24%	114%
Box of Soap	6%	9%
Bar of Soap	10%	28%
Chip Clip	18%	85%
Safety Razer	28%	80%
Tape Measure	18%	41%
Pillow	44%	117%
Red Lego Brick	79%	256%
Green Lego Brick	79%	288%
Ivan's Phone	15%	1%
Ollie the Monkey	89%	19%
Marshy the Elephant	4%	190%
Bottle Brush	22%	7%
Hat	43%	85%
Airplane Clock	50%	4%
Sun Chips	4%	89%
Rope Bundle	127%	60%
Twine	135%	72%
Dust Pan	23%	429%

Table 8. Weight Prediction Error (%) on the Household Test Set for the First 41 Objects.

Category	True Weight (g)	M-LLM Prediction (g)	Shape-Aware Model Prediction (g)
Neck Pillow	311	200	490
Headphones	188	200	261
Boy Doll	177	181	325
Disk Drive	310	200	116
Keyboard	802	750	800
Dali Clock	621	300	764
Toaster	1077	1497	1026
Wooden Train Track	59	45	133
Baby Shoe	73	100	149
Back Scratcher	43	75	55
Hair Dryer	599	680	1513
Umbrella	651	300	917
15 Puzzle	73	150	43
Gardening Shears	269	300	215
Swiss Army Knife	217	150	65

Table 9. Ground Truth and Predicted Weight Values on the Household Test Set for 15 of 56 Objects.

Category	M-LLM Weight Prediction Error (%)	Shape-Aware Model Weight Prediction Error (%)
Neck Pillow	36%	58%
Headphones	7%	39%
Boy Doll	2%	83%
Disk Drive	36%	63%
Keyboard	6%	0%
Dali Clock	52%	23%
Toaster	39%	5%
Wooden Train Track	24%	125%
Baby Shoe	38%	105%
Back Scratcher	76%	30%
Hair Dryer	14%	153%
Umbrella	54%	41%
15 Puzzle	107%	41%
Gardening Shears	12%	20%
Swiss Army Knife	31%	70%

Table 10. Weight Prediction Error (%) on the Household Test Set for 15 of 56 Objects.