Supplementary Material for
When Pigs Fly: Contextual Reasoning in Synthetic and Natural Scenes

Philipp Bomatter[1,*], Mengmi Zhang[2,3,*], Dimitar Karev[4], Spandan Madan[3,5],
Claire Tseng[4], and Gabriel Kreiman[2,3]

[1]ETH Zürich
[2]Children's Hospital, Harvard Medical School
[3]Center for Brains, Minds and Machines
[4]Harvard College, Harvard University
[5]School of Engineering and Applied Sciences, Harvard University
[*]Equal contribution

Address correspondence to gabriel.kreiman@tch.harvard.edu

List of Supplementary Figures

S1. Synthetic Out-of-context Dataset (OCD)

S1.1. Environment setup for various contextual conditions:

We leveraged the VirtualHome environment [3] developed in the Unity simulation engine to synthesize images in indoor home environments within 7 apartments and 5 rooms per apartment. A maximum of 9 fixed view angles are captured for each target object and location. Azimuth angles range from 0 to 320 degrees in steps of 40 degrees, fixed elevation angle of 19.5 degrees and radius of 1.5 meters. In some special cases, the elevation angle is set to -19.5 degrees to prevent the camera location from penetrating the ceiling.

In the gravity condition, the elevation angle of the camera is adjusted to center the view on the floating target object. This causes the surrounding context to vary slightly compared with the same configuration in the normal context condition. The small difference in the number of images between the normal context and the gravity condition is due to the removal of some images because of invalid camera positions or target collisions with other objects after the targets are lifted up. In other out-of-context conditions, same object and camera configurations remain as the normal context conditions.

S1.2. Performance Evaluation

The object-to-context ratio is critical [5] — context plays a larger role for smaller objects. Therefore, we split images into two groups: according to the target object sizes in degrees of visual angle (dva), based on the psychophysics experiments (**Figure 3d**)): (i) images $\leq$ 2 dva, and (ii) images $>$ 2 dva. The pixel to dva conversion is based on the human experiment setups of a display with 1024×1280 pixels and distance of 0.5 meters (actual size varies in MTurk depending on viewing conditions).

S1.3. Training on Synthetic (OCD) Data

If we train the models with natural images from COCO-Stuff [1] and then evaluate them on synthetic images from our OCD dataset, we face a domain gap. This domain gap might play a role in influencing the comparison of the same model across different context conditions and between different models. To evaluate the ability of closing the domain gap for our CRTNet model, we first train CRTNet on the synthesized training set in normal context conditions, and then we test CRTNet in the normal condition in the test set as elaborated in the main paper.

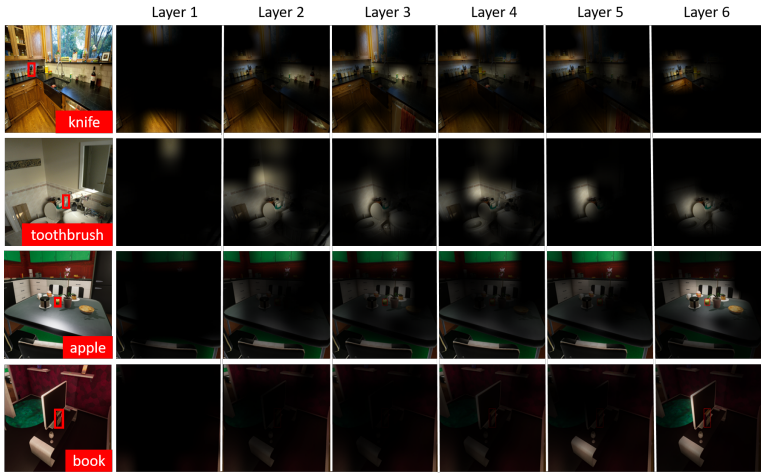
To synthesize the training set in normal context conditions, for each object belonging to the 36 object classes, its possible locations are uniformly arranged in a grid over the entire supporting surfaces where these surfaces are typically determined according to the co-occurrence statistics. The grid size is chosen relative to the target object size (5 times the target object size) because a small location shift introduces very little view variance for a larger object, *e.g.*, the view difference by moving a microwave 0.2 meters away is small compared with a cellphone moved by 0.2 meters. For each target object location, we randomly sampled 12 camera views with the following parameters: azimuth angle from [0, 360] degrees with a step size of 2 deg, elevation angle from [-35, 90] degrees with a step size of 2 degrees, and radius of [0.5, 5] meters with a step size of 0.5 meters. To introduce variations and avoid overfitting during training, we replace the target object's texture with a random texture from [4]. Collision checking and camera ray casting are enabled to prevent object collisions and occlusions. We used 5 out of the 7 VirtualHome apartment scenes as training set and the images from the remaining two apartments as validation set. CRTNet achieves an accuracy as high as 80% in the normal context condition on our test set as elaborated in the paper, compared to 88% accuracy on the validation set. This implies that our CRTNet is capable of closing the domain gap.

S2. Cut-and-paste Dataset

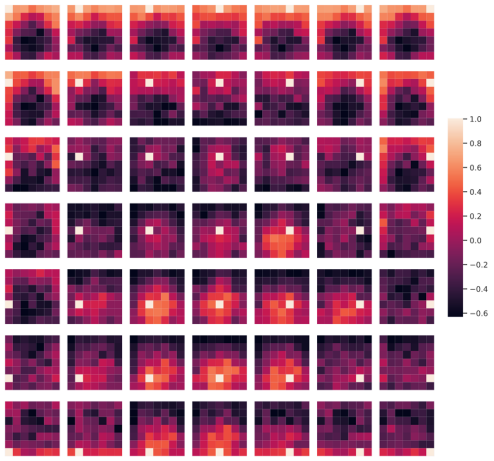
The Cut-and-paste dataset [5] is based on images from the COCO dataset [2]. In addition to normal context and minimal context (rectangular bounding box enclosing the object and grey pixels outside the box) conditions, the target objects were cut from a given image and pasted onto another one with either a congruent context (context contains an object of the same class label) or incongruent context (context taken from an image with different class label). The images are grouped into four bins based on the target object sizes in degrees of visual angle (dva): Size 1 dva [16-32 pixels], Size 2 dva [56-72], Size 4 dva [112-144], and Size 8 dva [224-288].

S3. Visualization of Attention Maps

Visualizations of attention maps on example images (Supp Fig. S1a) show that CRTNet globally attends to image regions that are semantically relevant for classification and that it narrows its focus as the information progresses over the hierarchy of the network layers. For example, in row 1 where the target object is a knife, the attention map starts from the kitchen floor in layer 1, slowly expands to table surfaces and eventually narrows down on the knife's location.



(a) Visualization of attention maps over hierarchical transformer decoding layers



(b) Similarity of position embeddings of CRTNet

Figure S1: **CRTNet predicts meaningful attention maps and learns reasonable positional embeddings for feature tokens.** (a) Visualization of attention maps on four example images (one example per row). The ground truth label of the target object (red box, column 1) is in the bottom right. Over six transformer decoding layers (6 columns), we show the attention map averaged over all attention heads within the same layer and overlaid the attention map on the original image. The two top rows show examples from the test set of COCO-Stuff dataset [1] and the two bottom rows show examples from the test set of our OCD dataset. (b) Each tile shows the cosine similarity between the position embeddings of the patch with the indicated row and column and the position embeddings of all other patches. See color bar on the right for cosine similarity values.

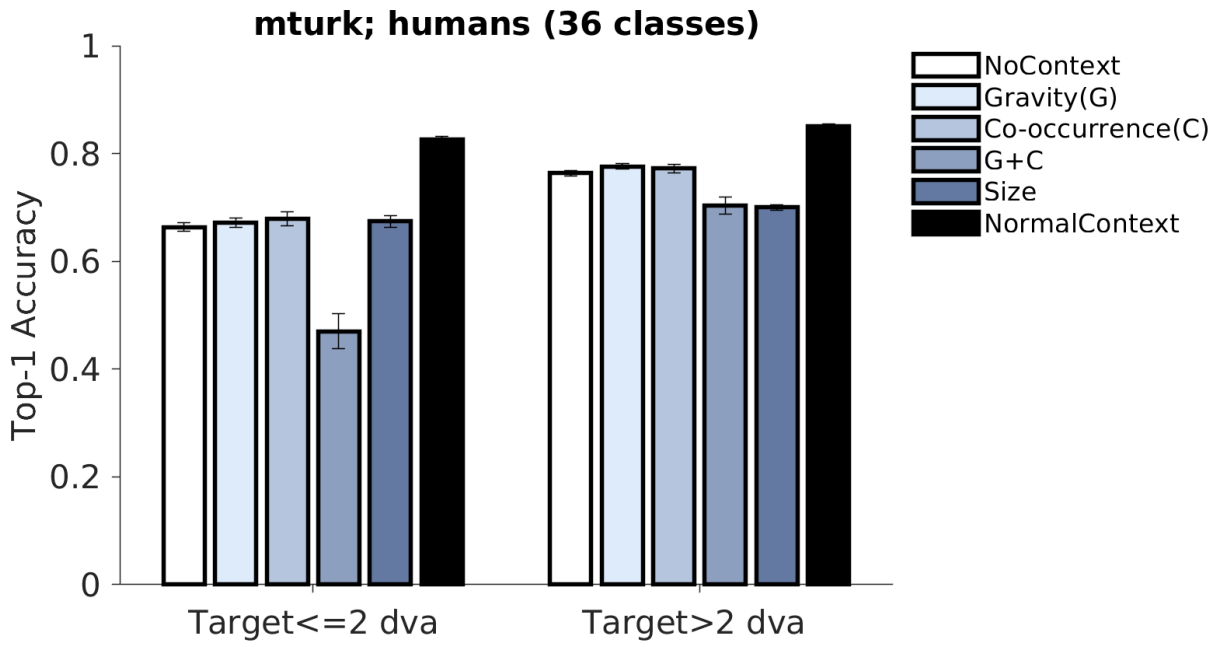


Figure S2: **Human performance across context conditions for 33 object classes.** Different colors denote context conditions (Sec. 4.2.1, Fig. 1). The trials are divided into two groups based on target object sizes in degrees of visual angle (dva). Error bars denote standard errors of the mean (SEM). This figure expands the results shown in the main text, which corresponds to only the 16 classes that overlap with COCO-Stuff.

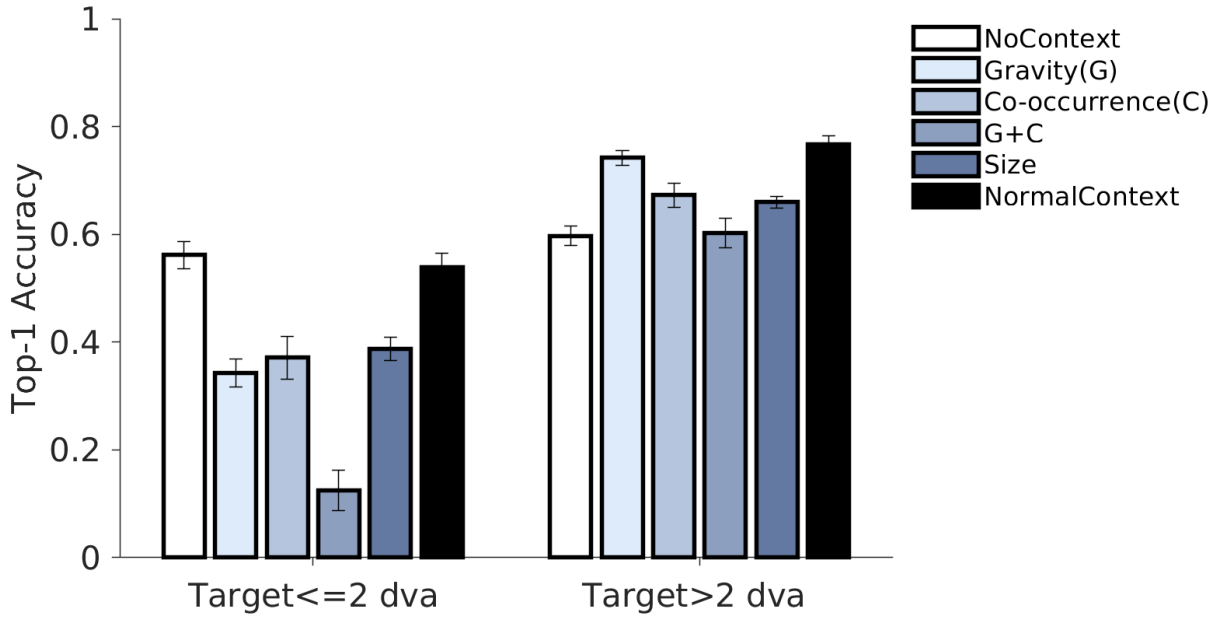


Figure S3: **Ablation - shared encoder.** In this ablation, we enforced weight sharing between the two encoders $E_t(\cdot)$ and $E_c(\cdot)$. Different colors denote context conditions (Sec. 4.2.1, Fig. 1). Conventions and format follow **Figure S2**.

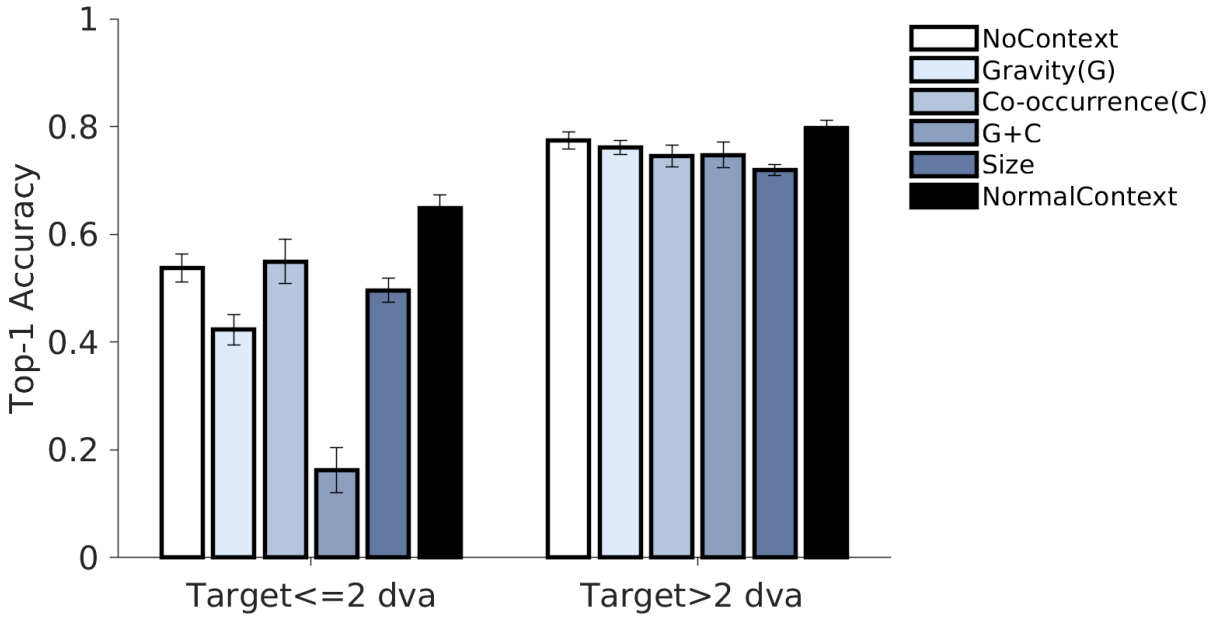


Figure S4: **Ablation - target object only.** In this ablation, we used y_t (based only on target information) instead of y_p as the final prediction. Different colors denote context conditions (Sec. 4.2.1, Fig. 1). Conventions and format follow **Figure S2**.

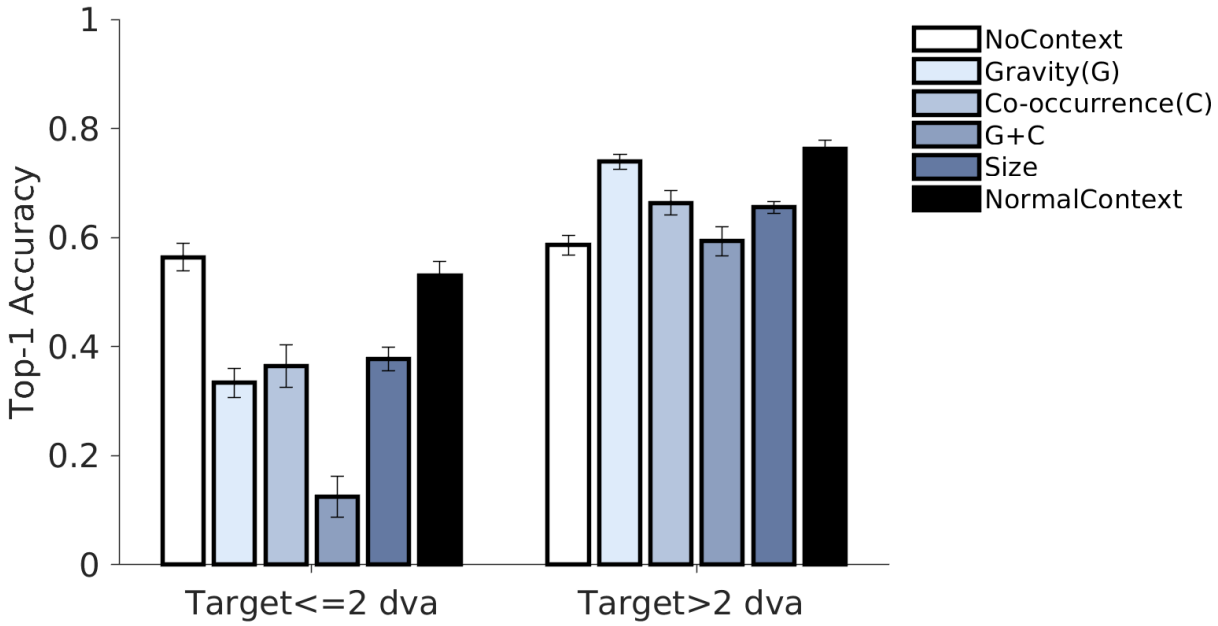


Figure S5: **Ablation - contextualized only.** In this ablation, we used the contextualized prediction $y_{t,c}$ instead of the weighted prediction y_p as the final prediction. Different colors denote context conditions (Sec. 4.2.1, Fig. 1). Conventions and format follow **Figure S2**.

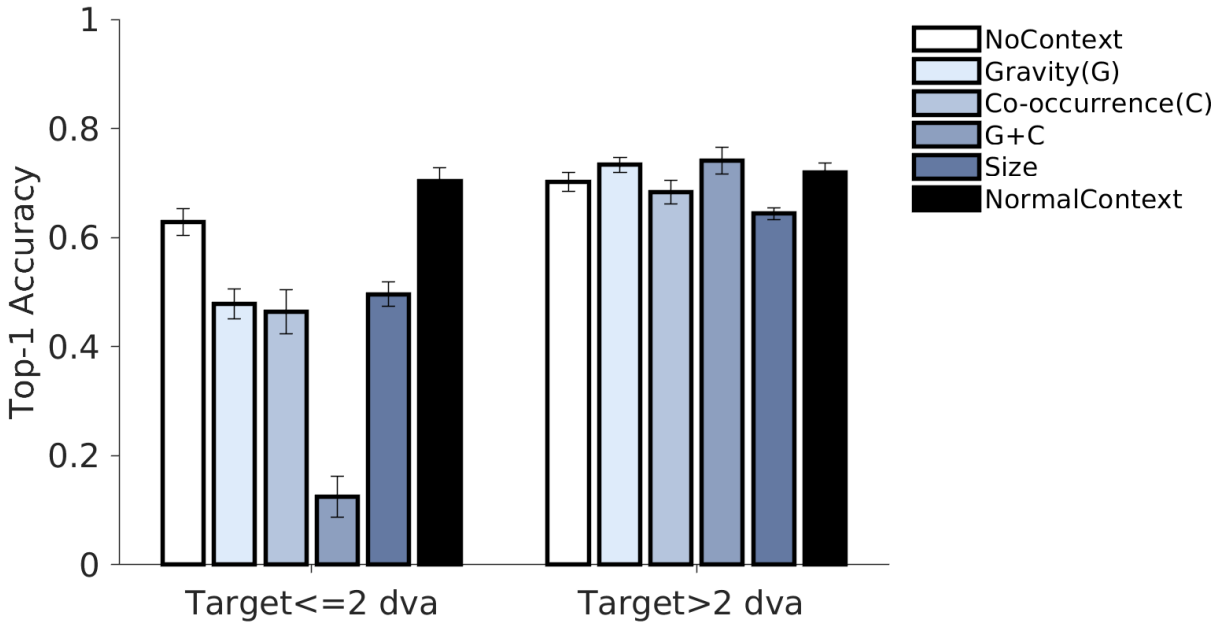


Figure S6: **Ablation - joint training.** In this ablation, we do not detach the gradient corresponding to the cross-entropy loss with respect to y_t, therefore allowing it to influence training of the target encoder $E_t(\cdot)$. Different colors denote context conditions (Sec. 4.2.1, Fig. 1). Conventions and format follow **Figure S2**.

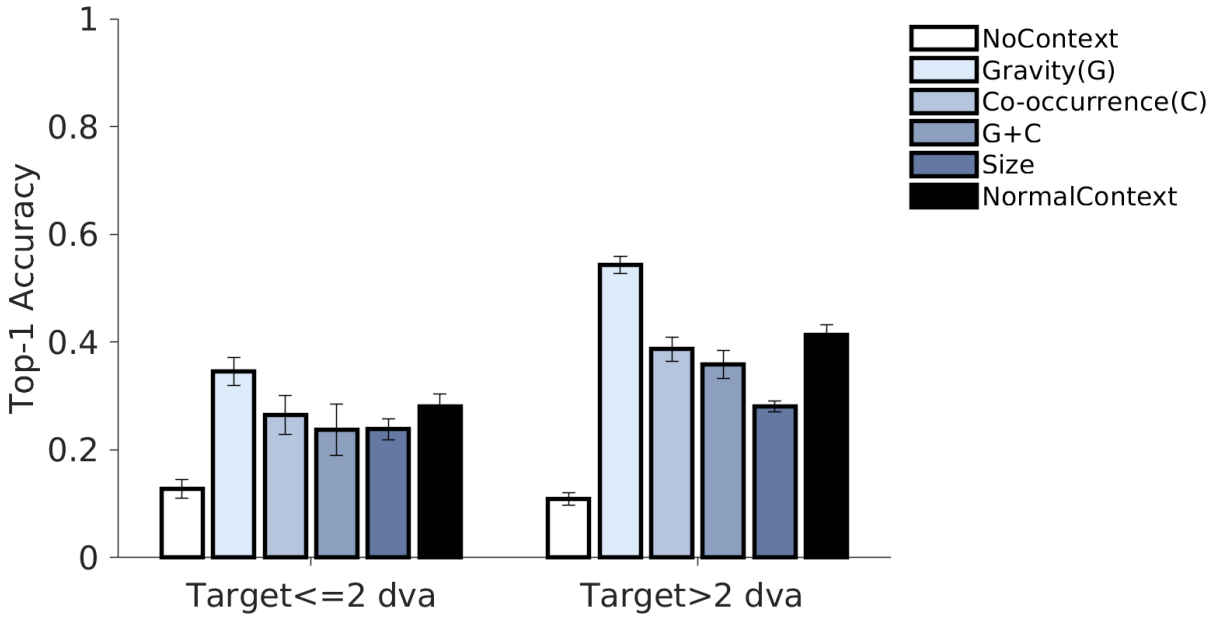


Figure S7: **CATNet performance across context conditions.** Different colors denote context conditions (Sec. 4.2.1, Fig. 1). Conventions and format follow **Figure S2**.

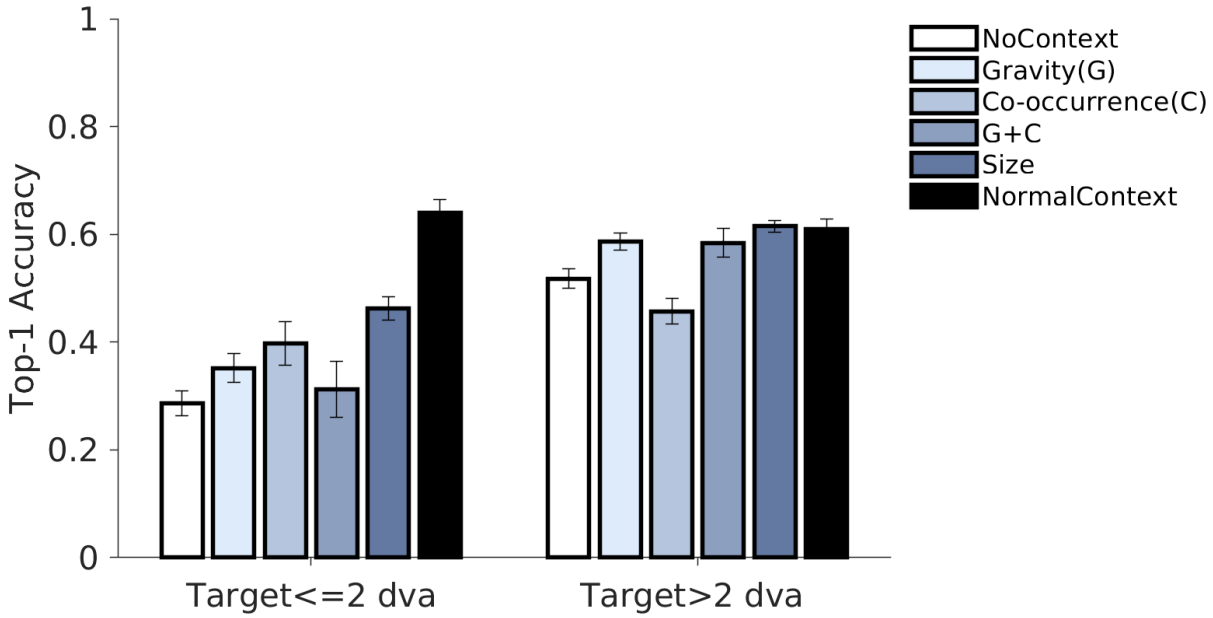


Figure S8: **Faster R-CNN performance across context conditions.** Different colors denote context conditions (Sec. 4.2.1, Fig. 1). Conventions and format follow **Figure S2**.

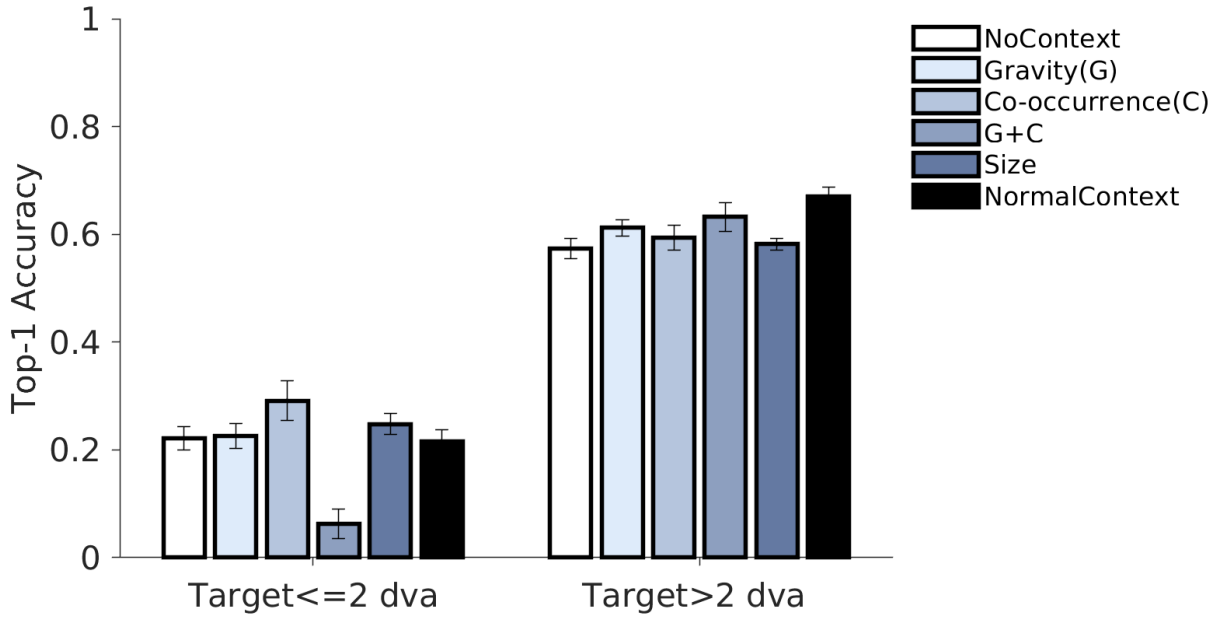


Figure S9: **DenseNet performance across context conditions.** Different colors denote context conditions (Sec. 4.2.1, Fig. 1). Conventions and format follow **Figure S2**.

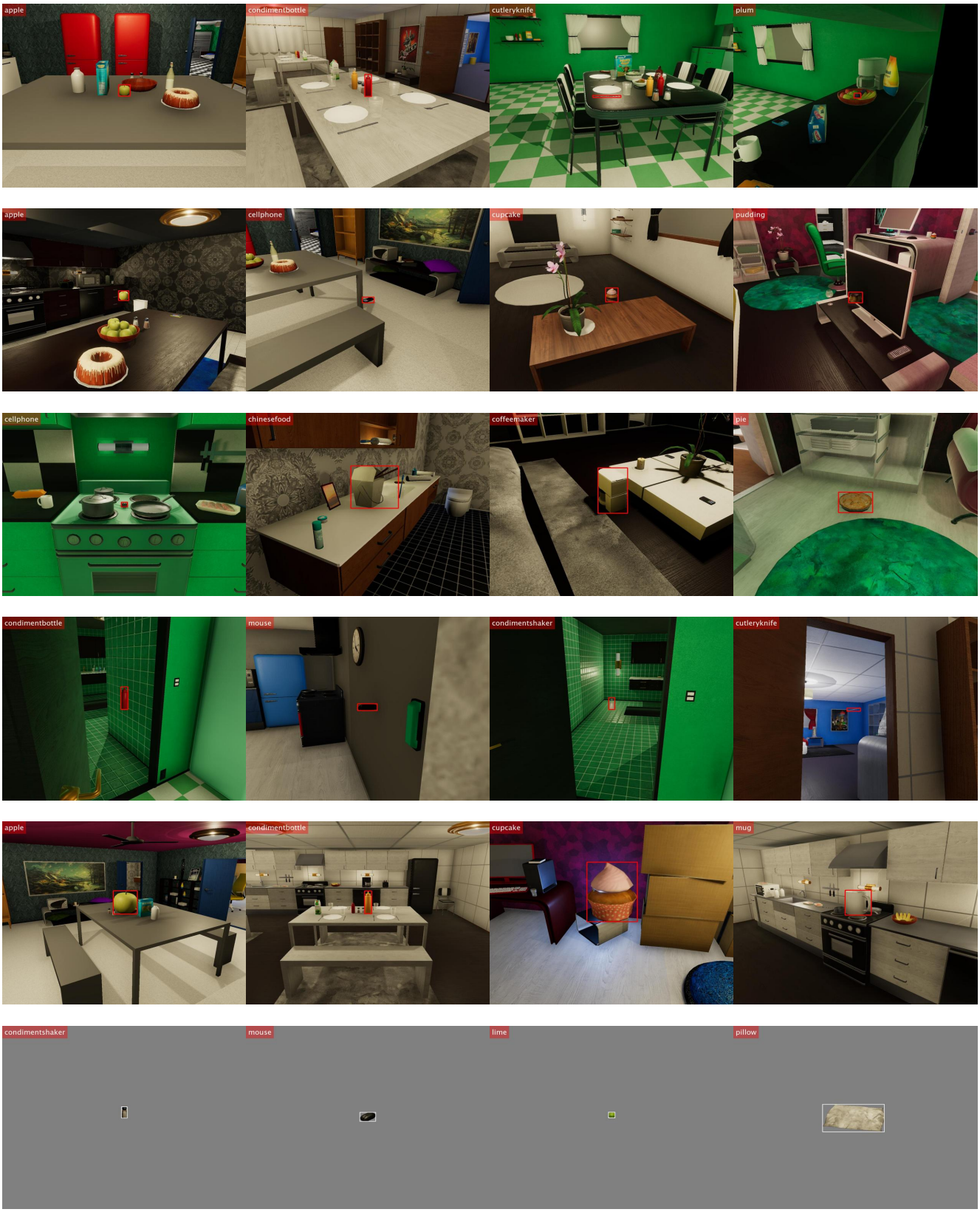


Figure S10: **OCD example images.** Each row contains examples for one of the six context conditions (normal, gravity, co-occurrence, co-occurrence+gravity, size, no context).

Figure S11: **Failure Examples for Psychophysics Experiments.** The examples were randomly sampled from the trials with lowest human recognition accuracy. Each row shows examples for one of the six context conditions (normal, gravity, co-occurrence, co-occurrence+gravity, size, no context).

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2, 3

[2] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[3] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. 1

[4] Lavanya Sharan, Ruth Rosenholtz, and Edward Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8):784–784, 2009. 2

[5] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020. 2