Supplementary Method: CLIP Evaluation Details

A zero-shot classification task was used to evaluate the performance of 4 pre-trained models of CLIP (Vit B/32, Vit B/16, Resnet 50X16, Resnet 50X4 model). All images for each of the modalities (Original, Outline, Dotted, Constellations) were collected from one of the main dataset folders (p = 0.002 or p=0.003) into separate folders. The class label for each image was obtained from the 'Top-down Category (manual selection)' field in things_concepts.tsv in folder main/ https://osf.io/jum2f/ (Things dataset). As the category is not provided in this field for each image, only images with a given category in this field were selected for this evaluation. Hence we used a total of 2566 images in each modality during this evaluation (The full dataset has 3533 image sets). Where more than 1 label was provided for an image, all labels were considered correct while evaluating the classification performance. Considering all these details, we were left with 41 labels in the final classification task. We reported a top-3 class accuracy i.e. considered the classification to be successful if one of the top 3 predictions of the model is a correct label.