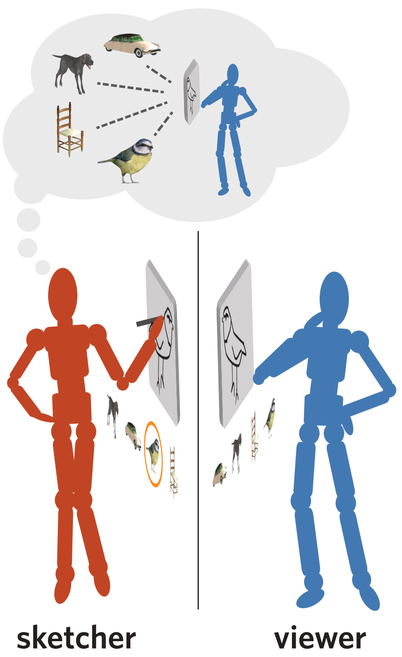
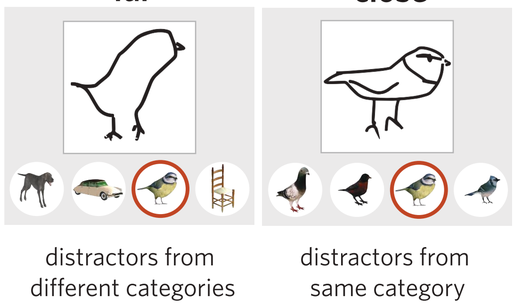
Supplemental for VSS 2018 submission: "Contextual flexibility in visual communication " (Fan, Hawkins, Wu, & Goodman)

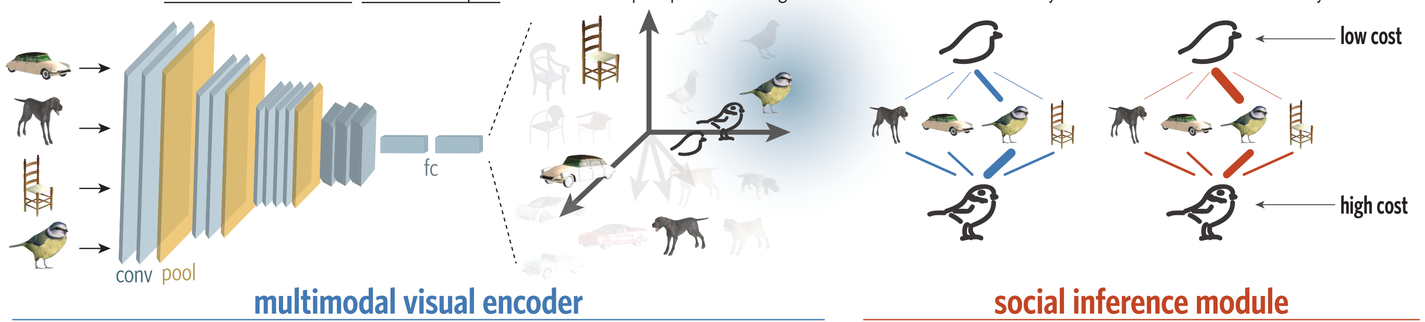
A visual communication task



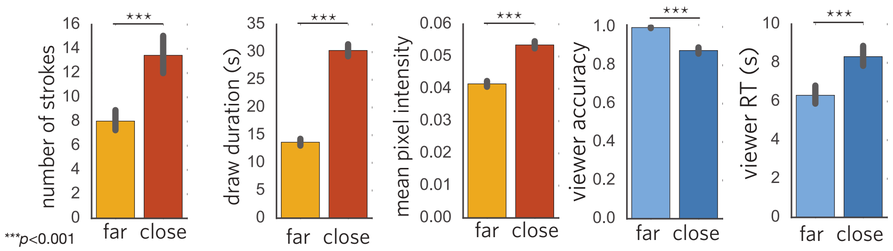
B context manipulation



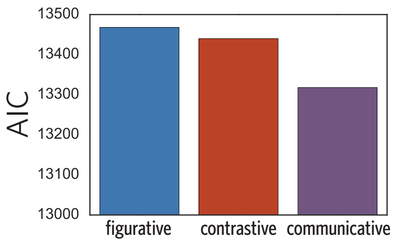
D model architecture



C behavioral results



E modeling results



Candidate sketcher models:

```
Communicative = Resemblance + Informativity - Cost
Contrastive   =               Informativity - Cost
Figurative    = Resemblance                 - Cost
```

A: Visual Communication Task: The sketcher's goal was to produce drawings so that the viewer could pick out a target object from a set of distractor objects. On each trial, both participants were shown an array of the same four objects, which appeared in different positions for each participant. Stimuli consisted of 3D renderings of 32 objects belonging to 4 categories (i.e., birds, chairs, cars, dogs), containing 8 objects each. **B: Context Manipulation:** For each pair, objects were grouped into eight quartets: Four contained objects from the same category (close); the other four contained objects from different categories (far). Each quartet was presented four times, such that each object in the quartet served as the target exactly once.

C: Behavioral Results: Participants (N=192) exploited information in common ground with their partner to efficiently communicate about the target: on far trials, sketchers achieved 99.7% recognition accuracy while applying fewer strokes, using less ink, and spending less time on their drawings than on close trials, where accuracy was still high (87.7%). **D: Model Architecture:** Our model combines a convnet visual encoder and Bayesian model of social reasoning during communication. The visual encoder maps images (3x224x224) to a fixed-length feature vector (1000-dimensional). To capture humans' ability to perceive resemblance across image domains, we adapt mid-level features from the pretrained VGG-19 convnet (Simonyan & Zisserman, 2014) to learn a multimodal feature embedding that captures image-level correspondences between drawings and photorealistic renderings of objects. This embedding minimizes the distance between matched renderings and drawings while preserving higher-order category structure, so we use distances in the embedding as a proxy for the perceptual similarity between images. The social reasoning module accepts a 4-tuple containing the distances between the human sketch and each of the four objects in the array, and outputs a probability distribution over drawings, reflecting the joint contribution of informativity, resemblance, and cost. **E: Modeling Results:** We instantiated three variants of the model: Figurative, which aims to produce drawings that are as faithful to the target as possible, ignoring the context; Contrastive, which aims to produce drawings that maximally distinguish the target from distractors, without regard for direct resemblance; and Figurative, which aims to produce drawings that both resemble the target and distinguish it from the distractors. Using Bayesian data analysis, we found that the Communicative sketcher provided a much stronger fit to the data than either of the Contrastive and Figurative sketcher models (lower AIC is better).