Natural Language for Visual Reasoning

Alane Suhr, Mike Lewis, James Yeh, Yoav Artzi

lic.nlp.cornell.edu/nlvr/

Language and Vision

A small herd of cows in a large grassy field.

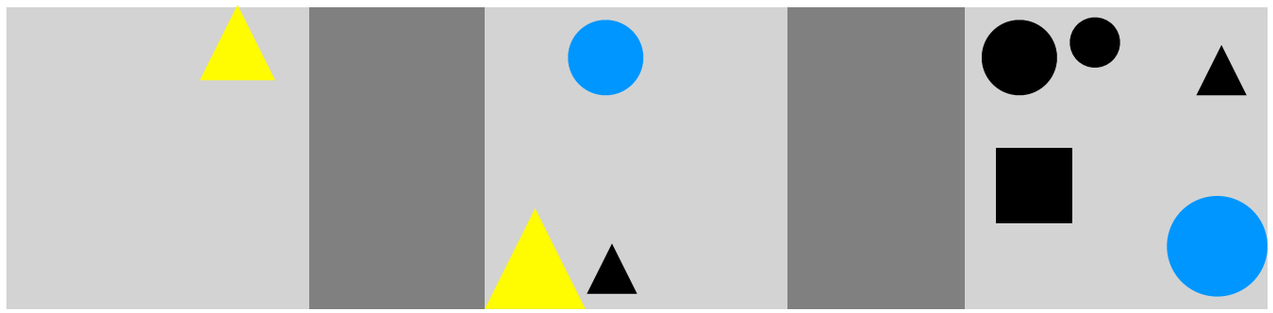
(Chen et al 2015)

What is the dog carrying?

(Agrawal et al 2015)

Our goal: more complex natural language

Natural Language for Visual Reasoning



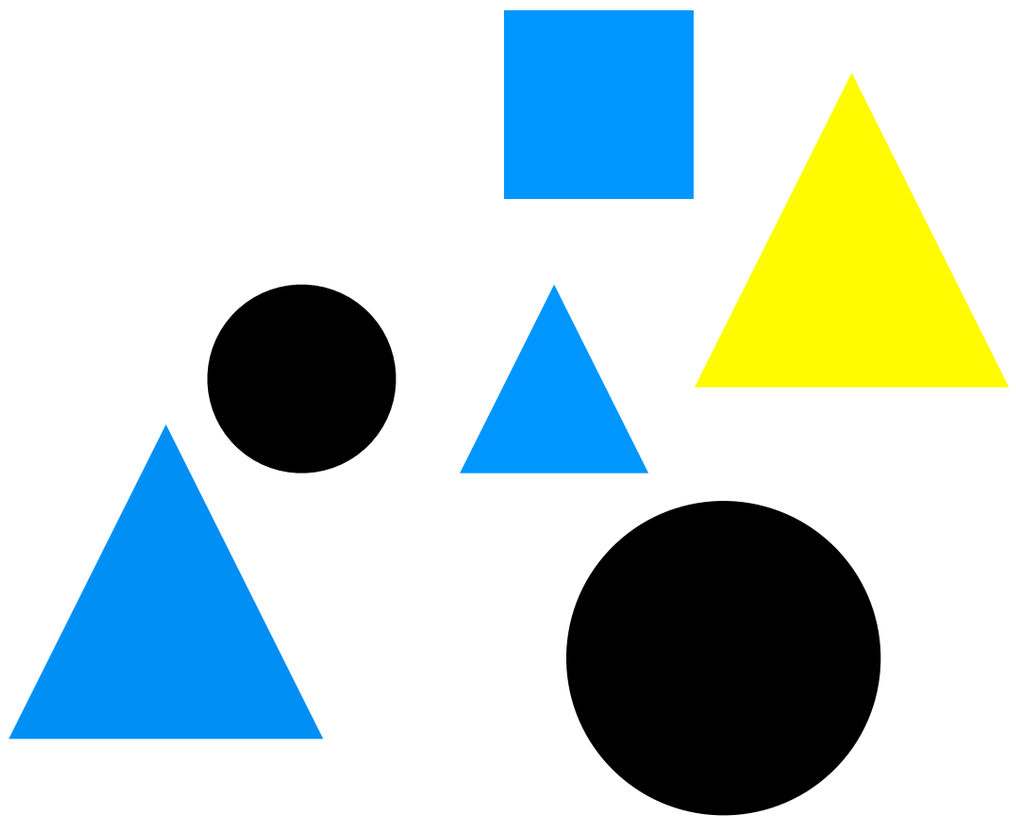
There is a box with 3 items of all 3 different colors.

TRUE

Task: determine whether the statement is true or false for the image.

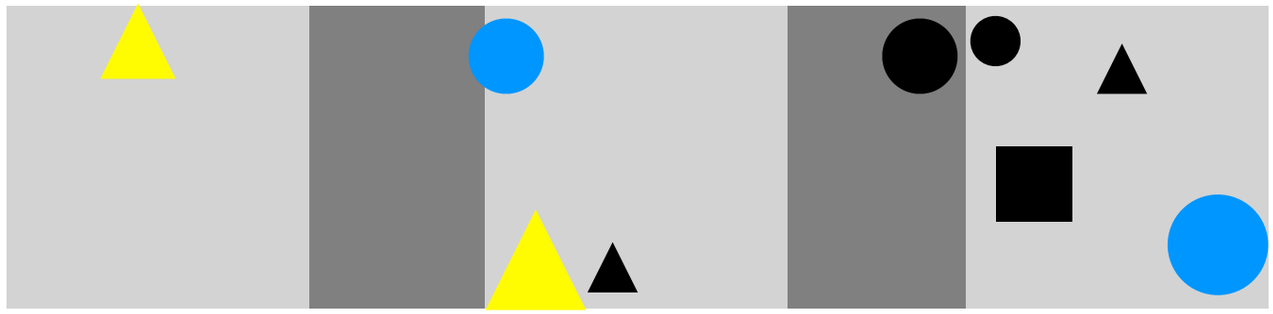
Outline

- Task and environments
- Data collection
- Analysis
- Baselines



Task and Environments

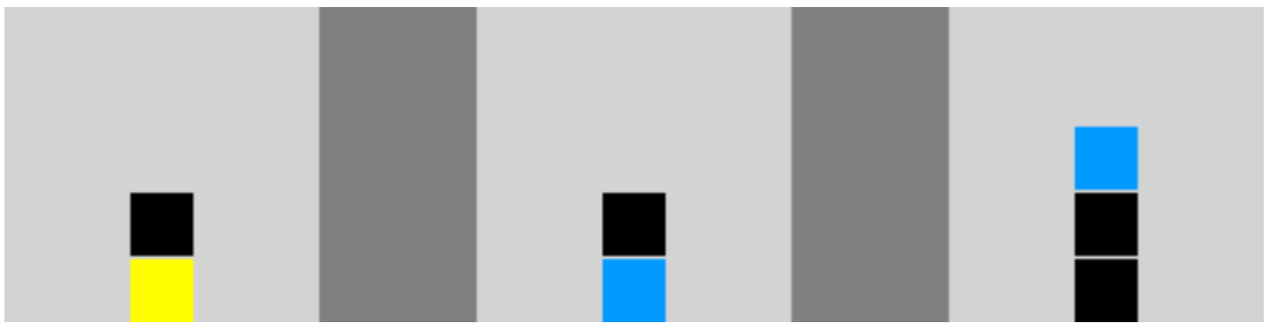
Scatter



There is a box with 3 items of all 3 different colors.

TRUE

Tower



There are only two towers which has the same base color.

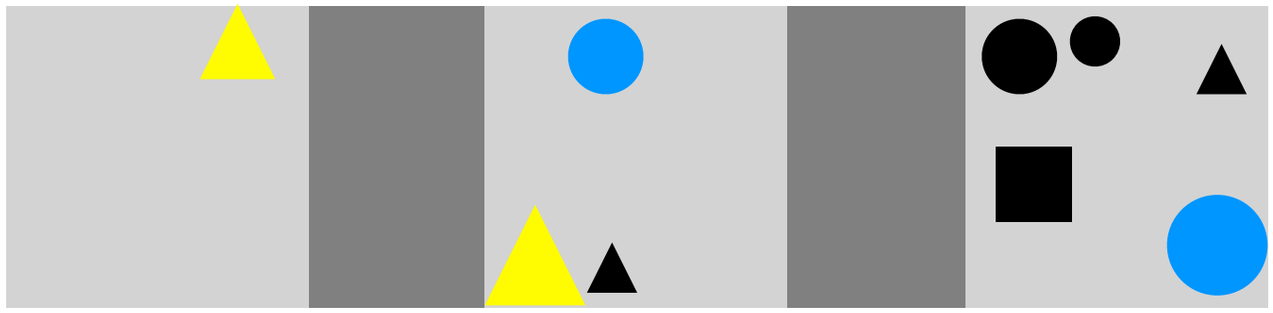
FALSE

Data collection

- **Goal:** collect complex natural language descriptions of images and true/false judgments
- Generate images
- Collect natural language sentences
- Label image/sentence pairs with truth judgments

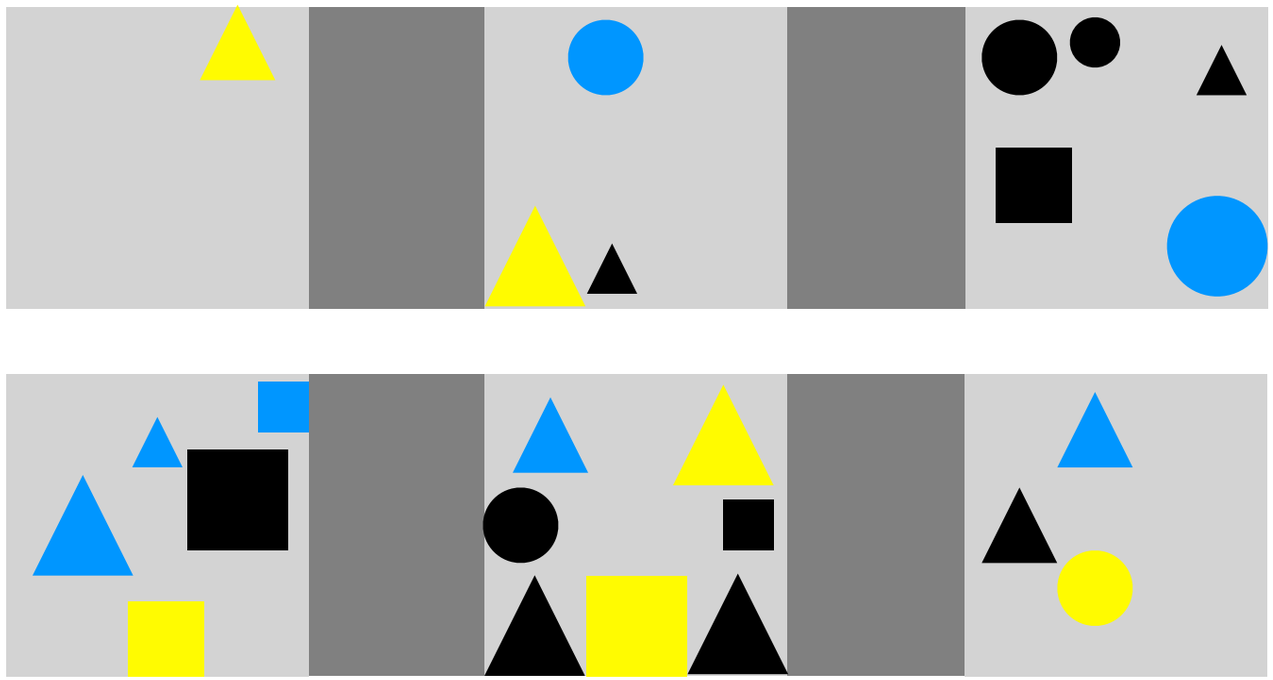
Image Generation

Image Generation



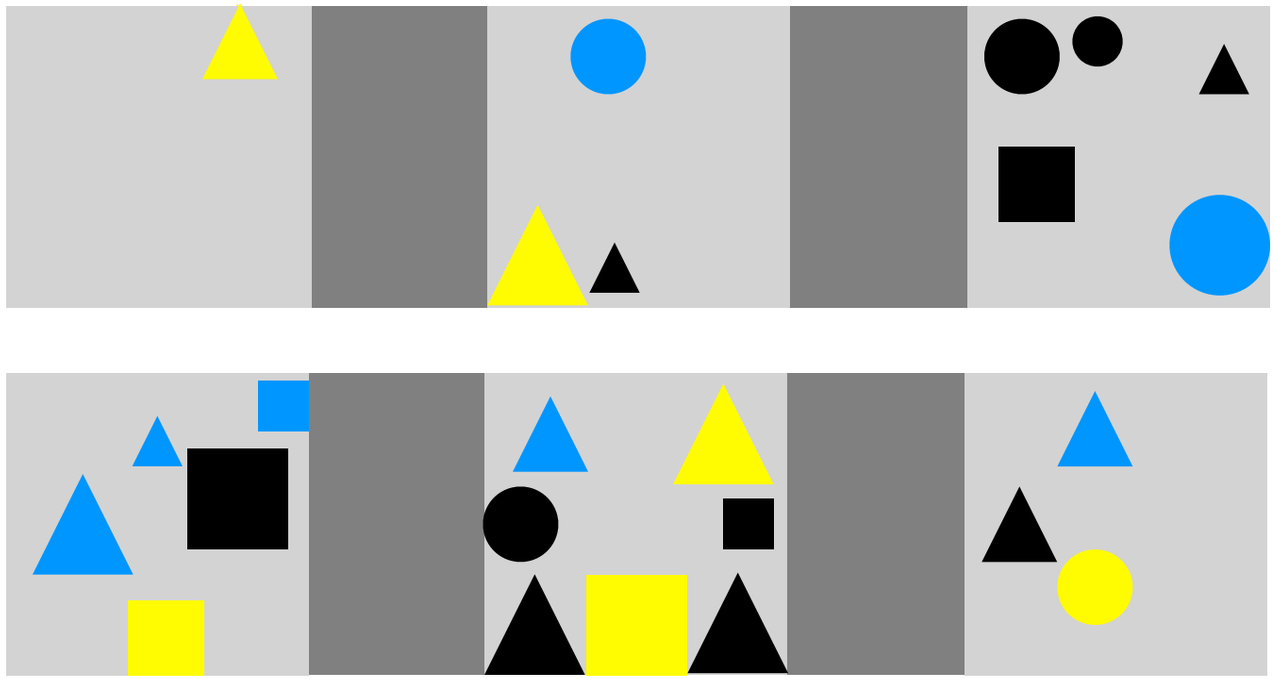
- Randomly choose number of items per box and item shapes, colors, sizes, and positions (without overlap)

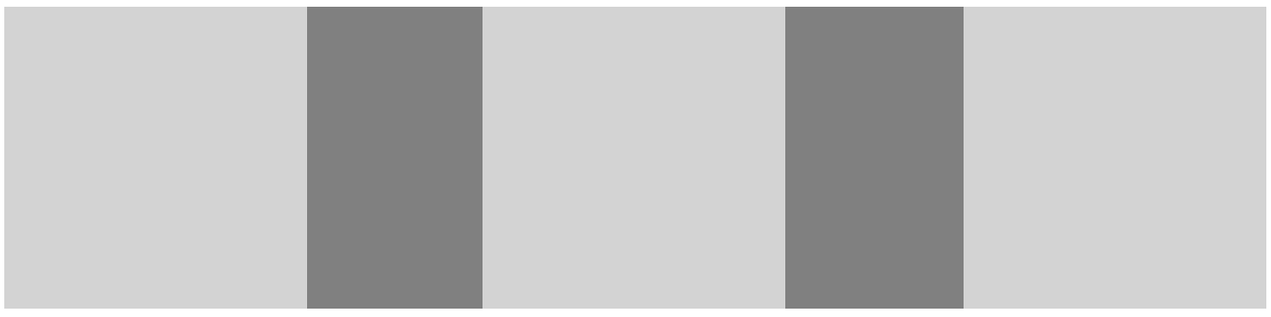
Image Generation



- Randomly choose number of items per box and item shapes, colors, sizes, and positions (without overlap)
- Construct second image with the same type

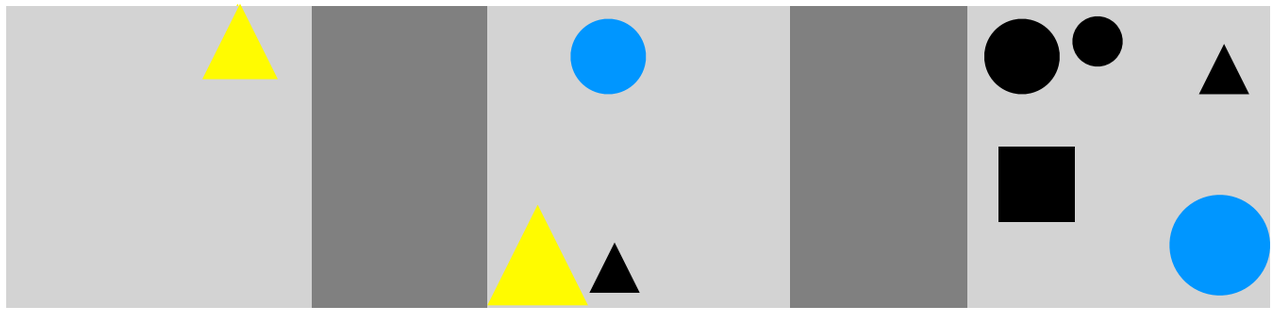
Image Generation

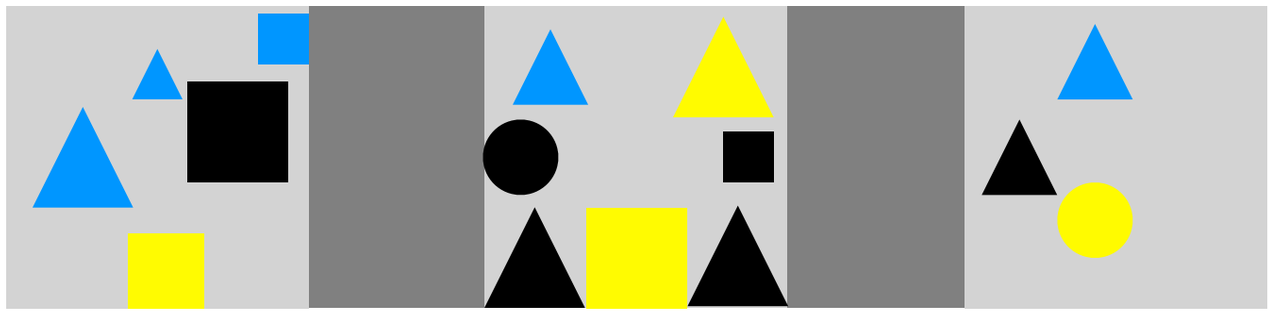


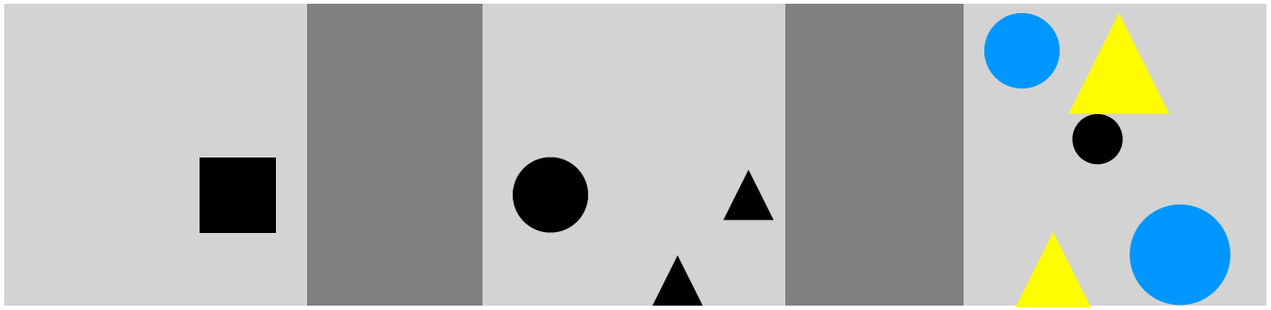


- Randomly choose number of items per box and item shapes, colors, sizes, and positions (without overlap)
- Construct second image with the same type

Image Generation

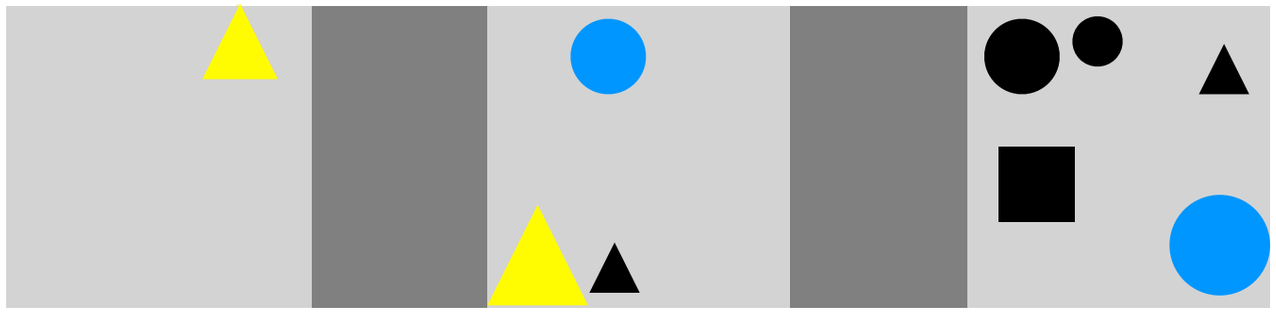


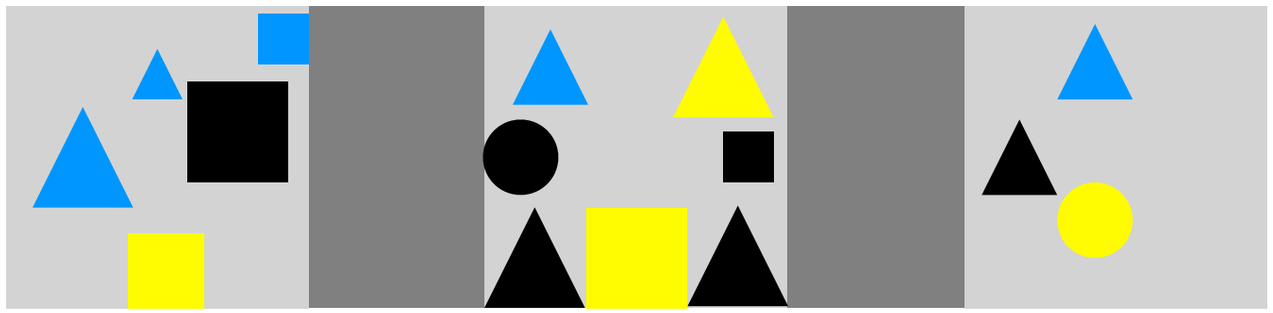


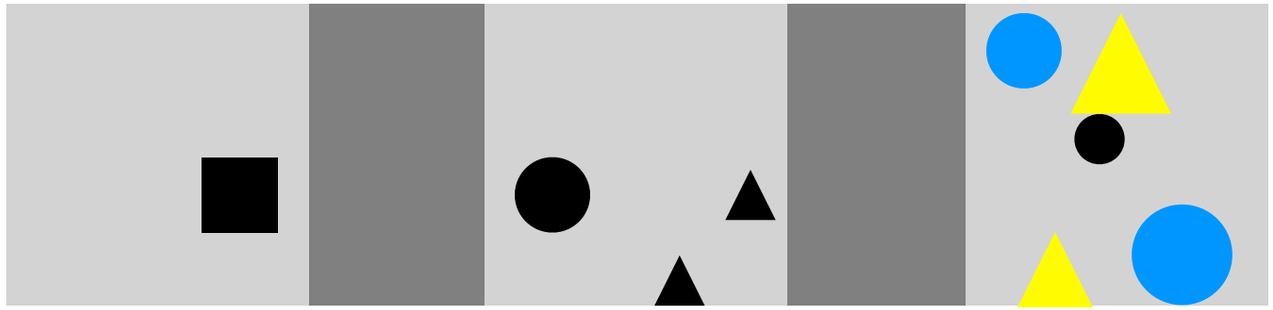


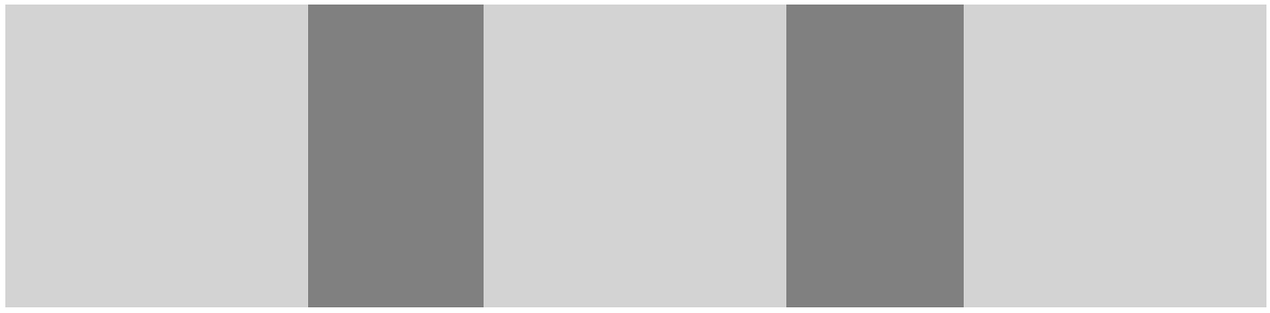
- Randomly choose number of items per box and item shapes, colors, sizes, and positions (without overlap)
- Construct second image with the same type
- Construct third image by shuffling items in the first image

Image Generation



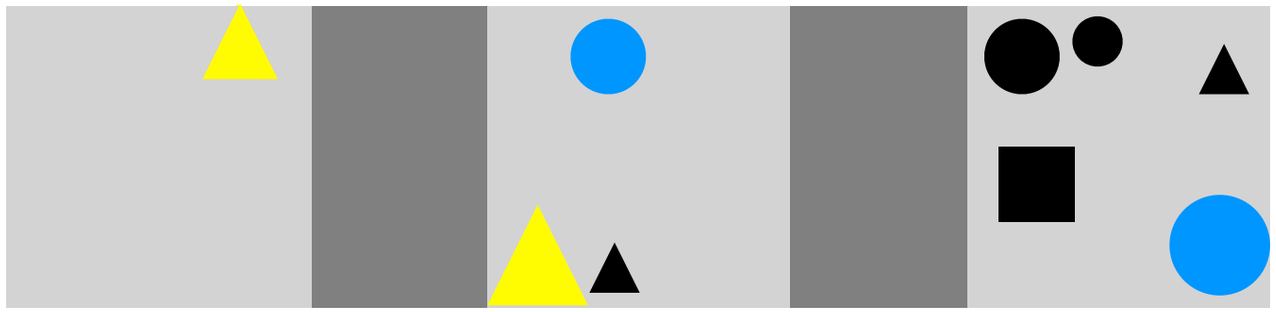


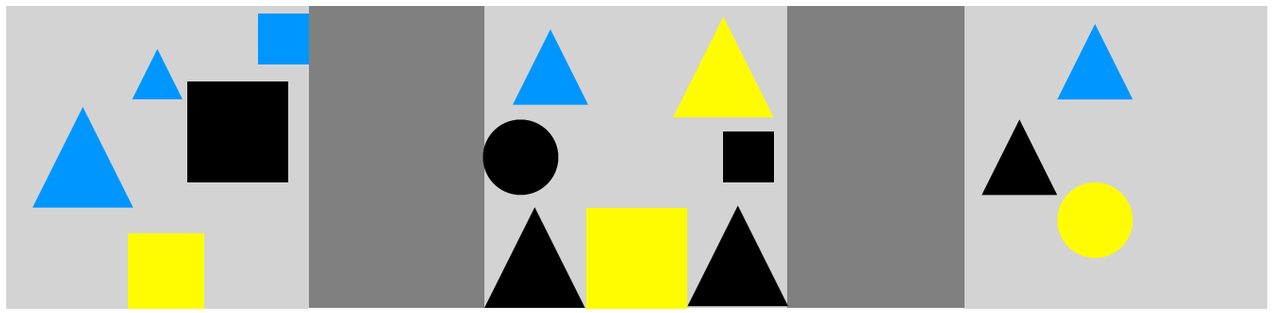




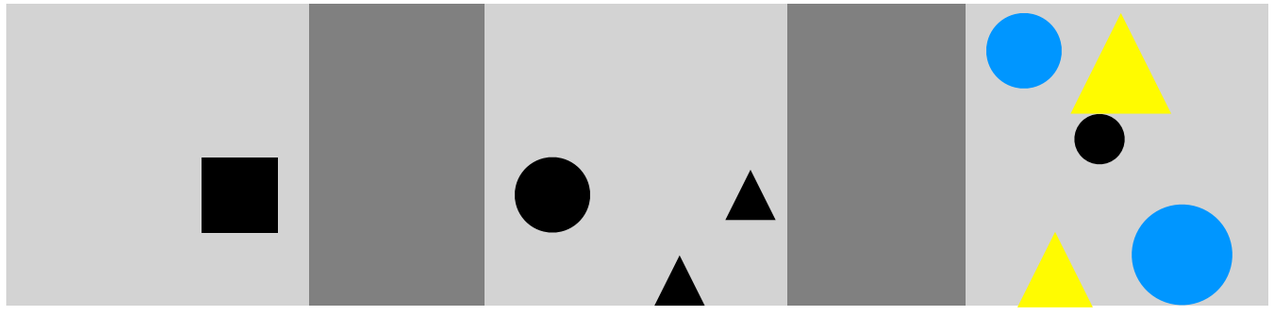
- Randomly choose number of items per box and item shapes, colors, sizes, and positions (without overlap)
- Construct second image with the same type
- Construct third image by shuffling items in the first image

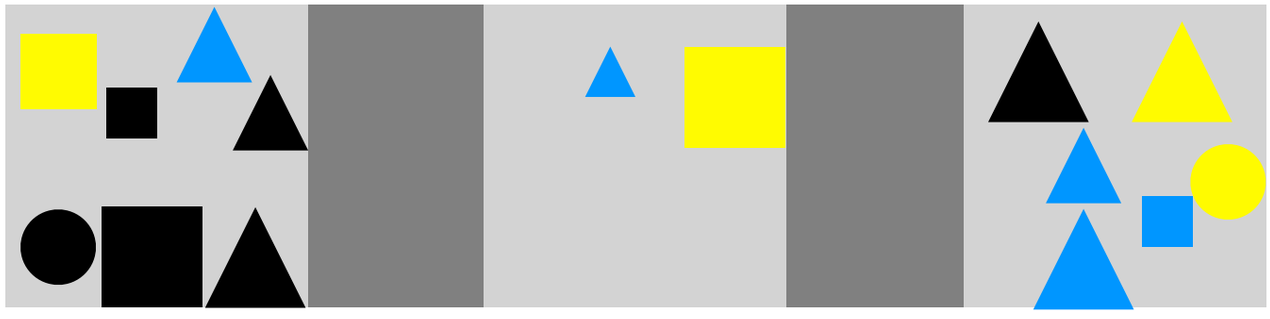
Image Generation





- Randomly choose number of items per box and item shapes, colors, sizes, and positions (without overlap)
- Construct second image with the same type
- Construct third image by shuffling items in the first image

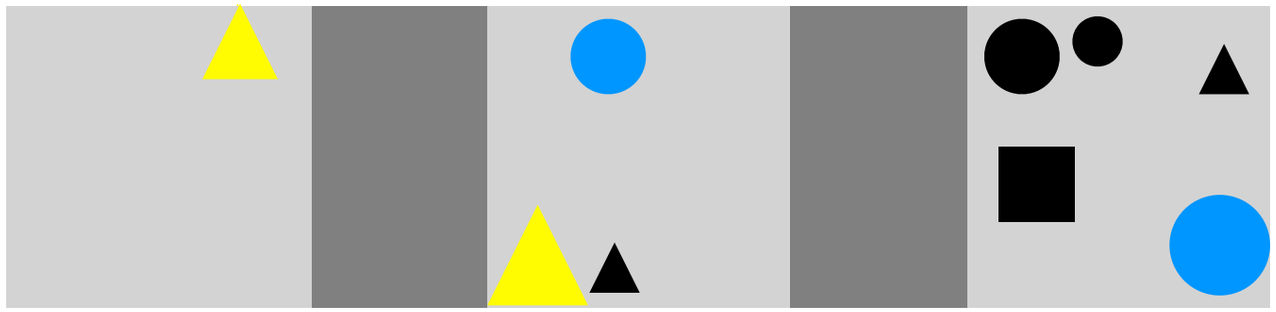


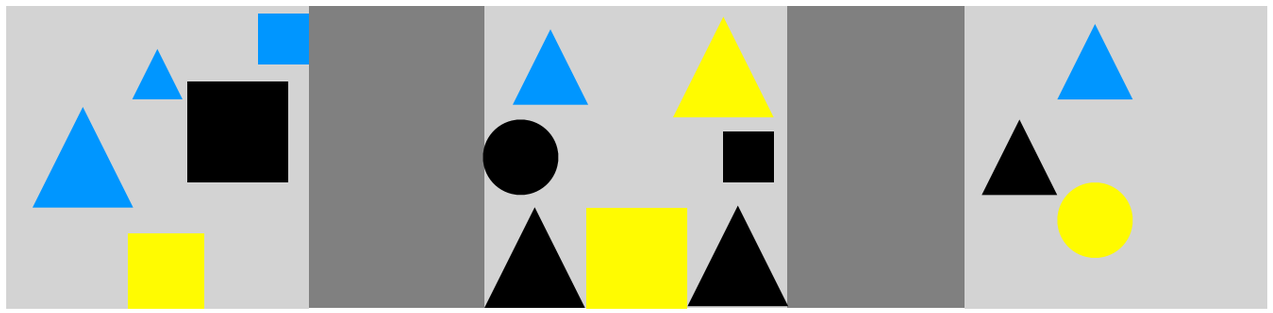


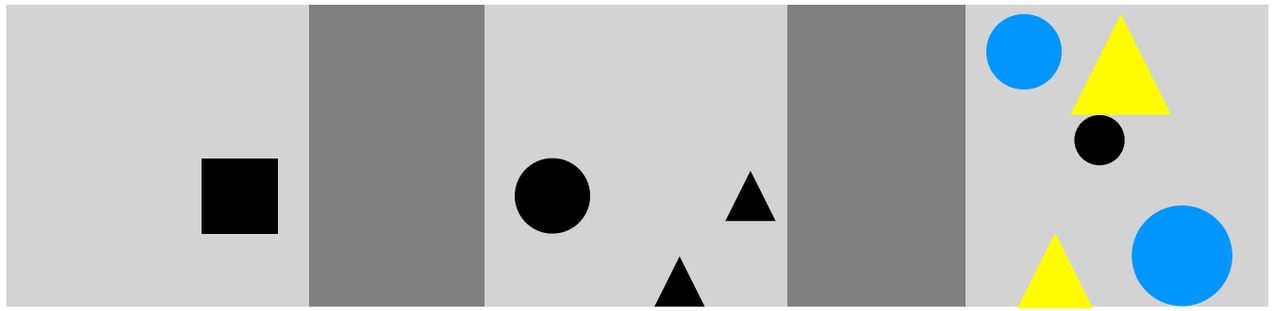
- Construct fourth image by shuffling items in the second image

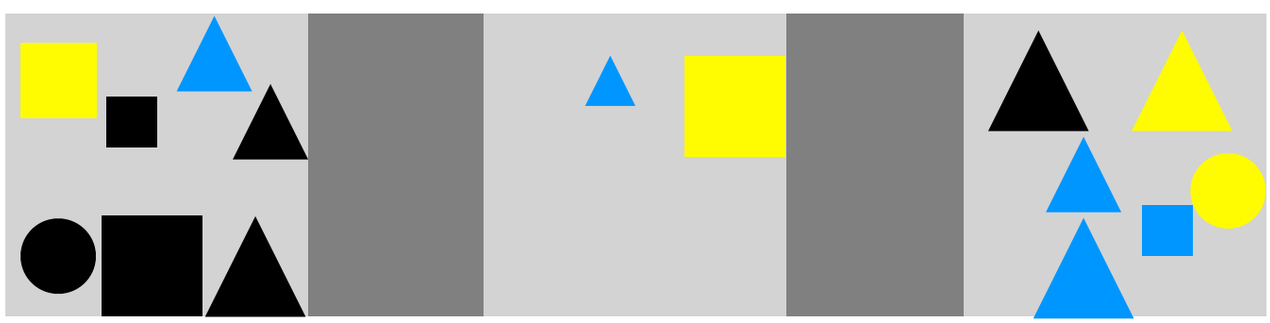
Generate two unique images and permute their items to create two other images

Sentence Writing







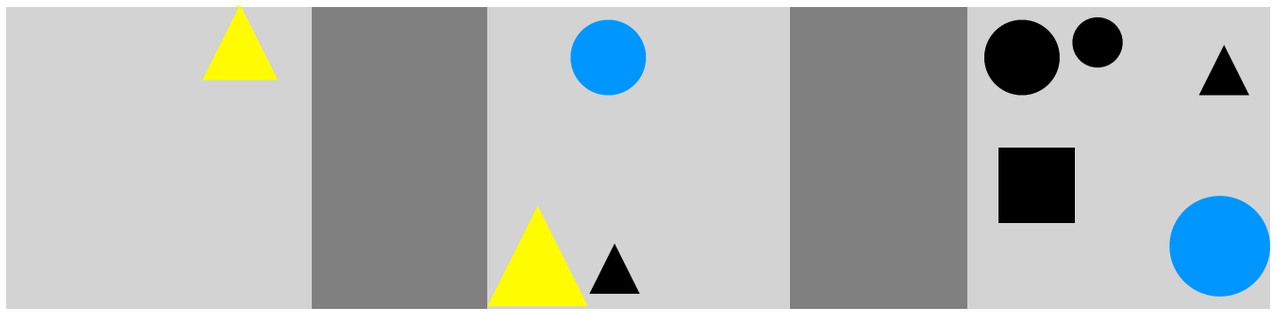


Write a sentence that is true about the top two images and false about the bottom two.

- Don't refer to the order of the images.
- Don't refer to the order of the boxes.

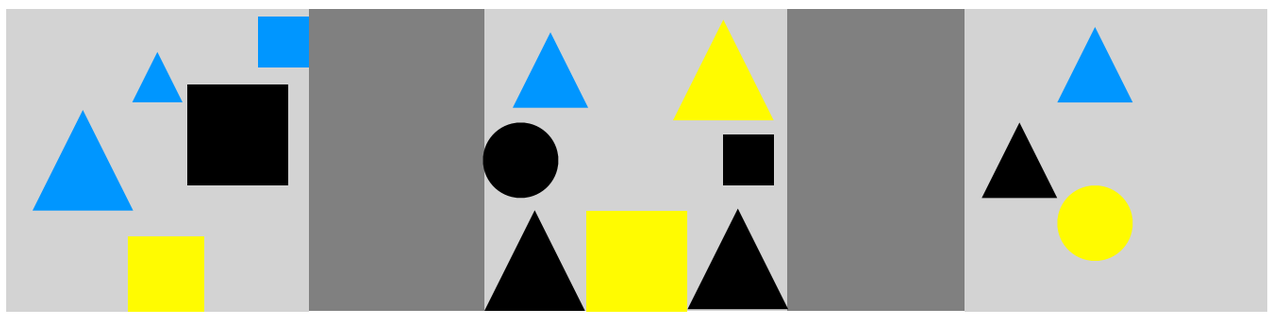
There is a box with 3 items of all 3 different colors.

Sentence Writing



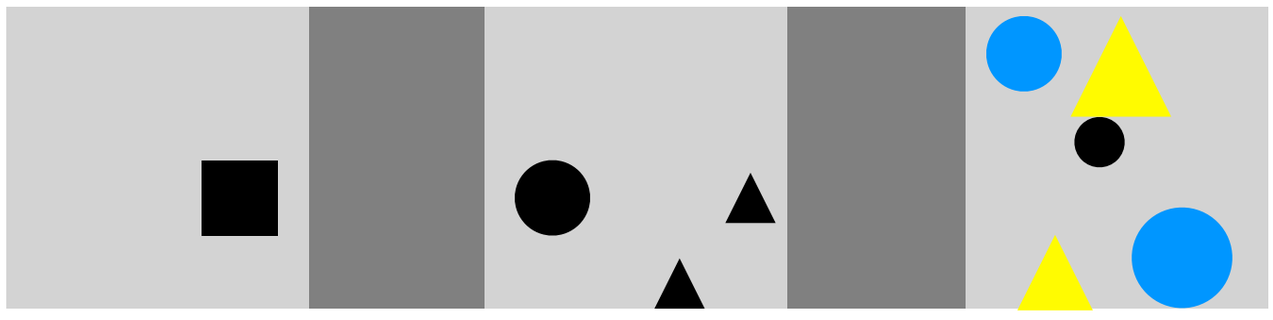
There is a box with 3 items of all 3 different colors.

T



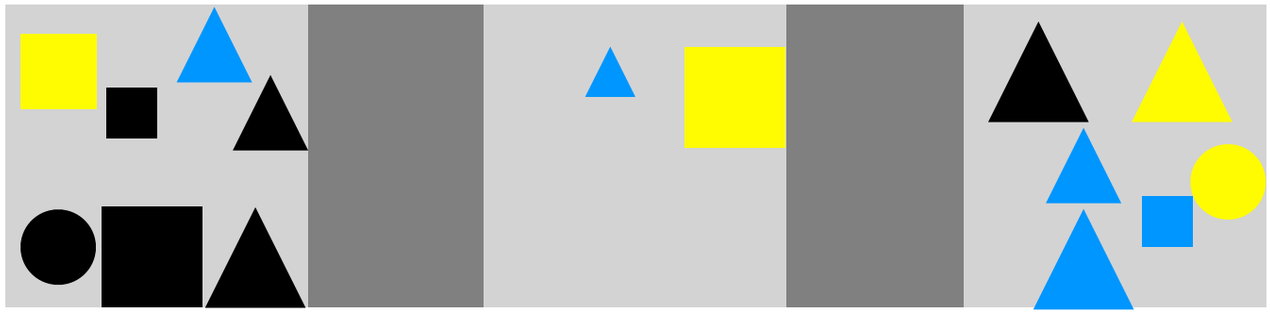
There is a box with 3 items of all 3 different colors.

T



There is a box with 3 items of all 3 different colors.

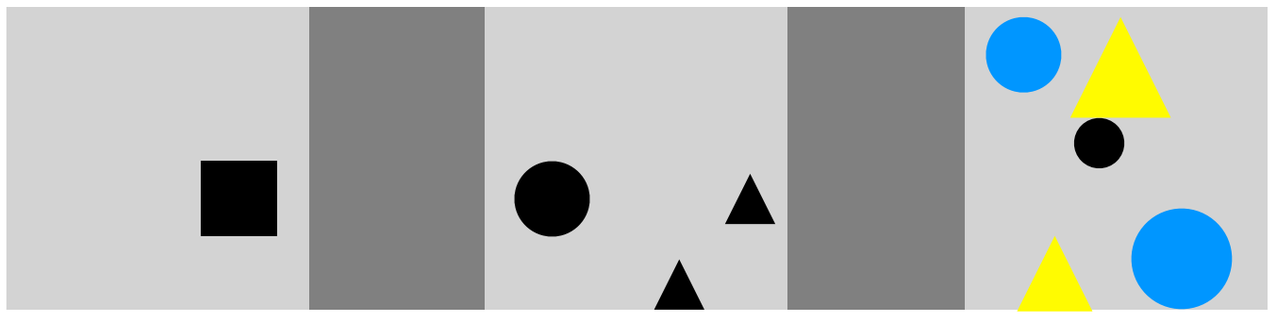
F



There is a box with 3 items of all 3 different colors.

F

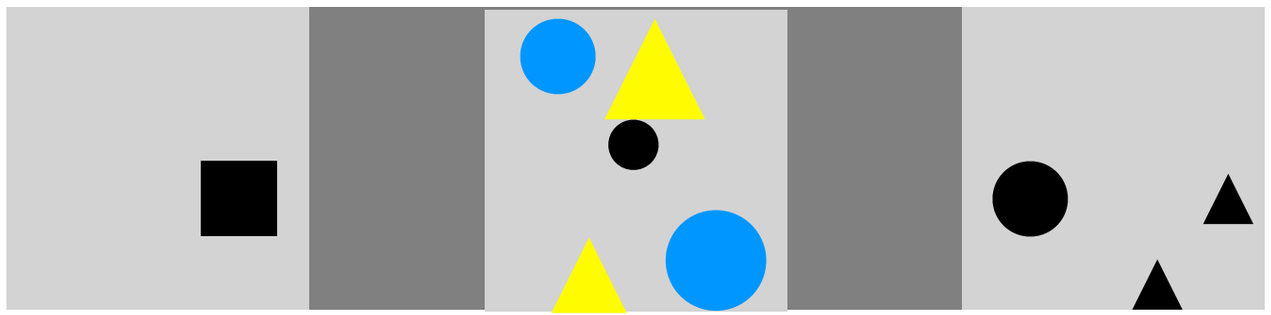
Validation



There is a box with 3 items of all 3 different colors.

- Higher-quality data
- Make sure workers followed the rules
- Recover examples by re-labeling them
- Disambiguate

Validation

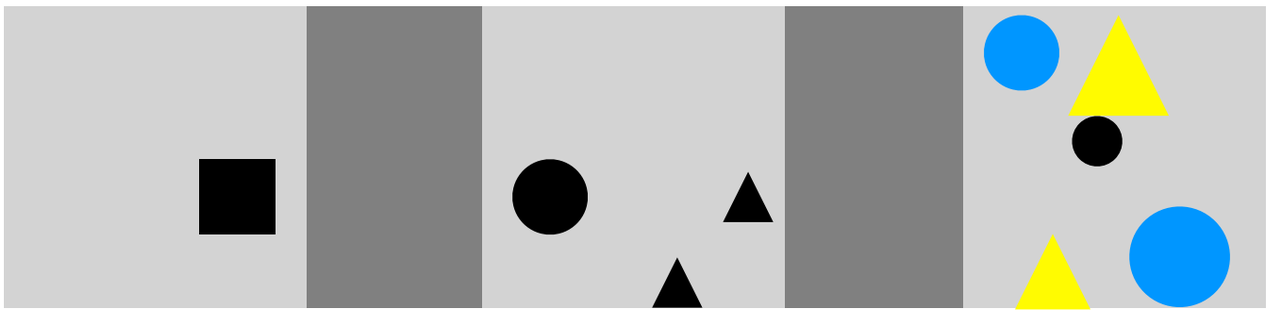


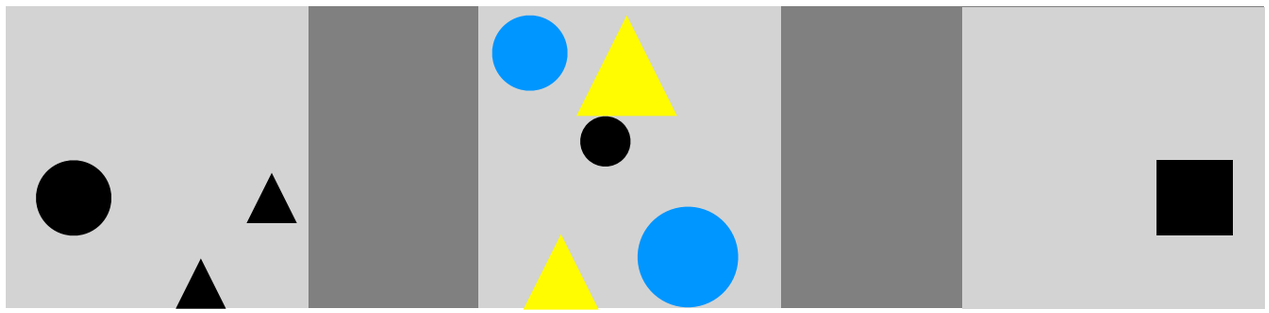
There is a box with 3 items of all 3 different colors.

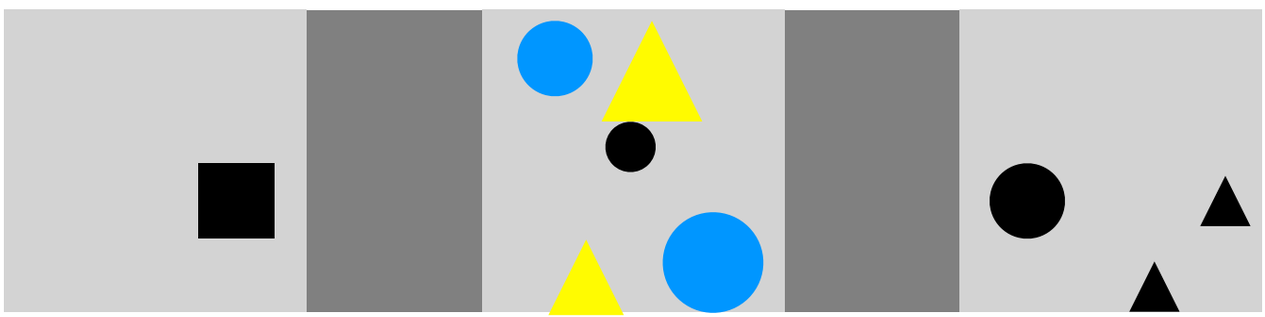
☐ TRUE

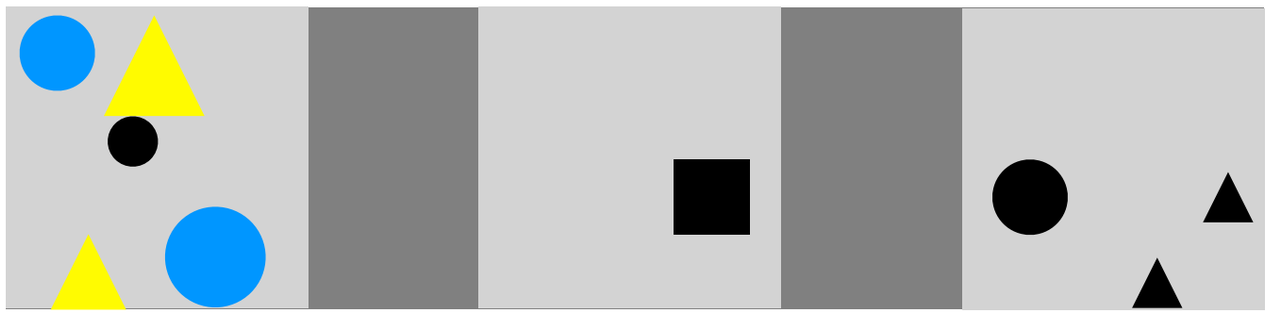
☑ FALSE

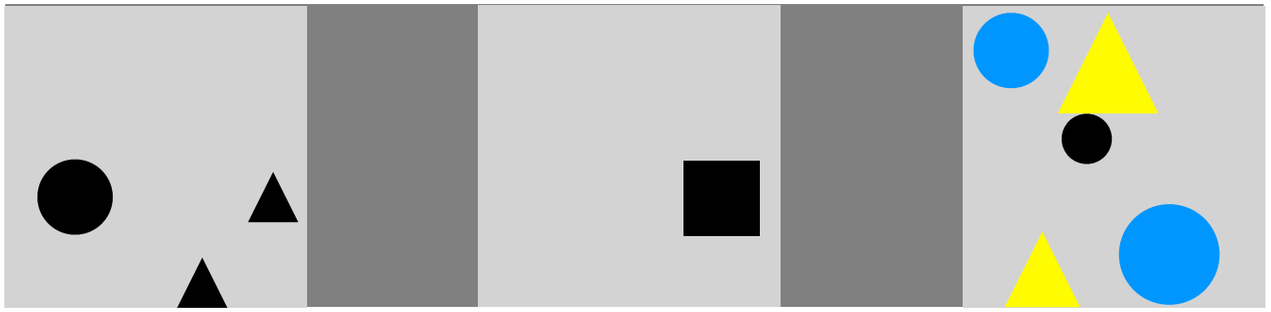
Permutation

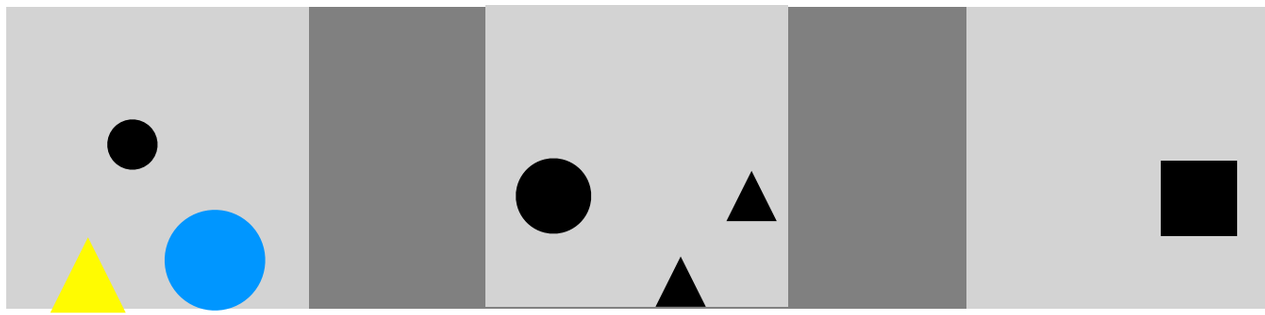












There is a box with 3 items of all 3 different colors.

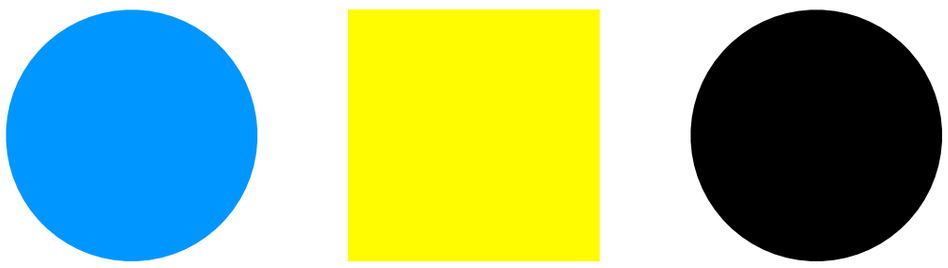
☐ TRUE

☑ FALSE

Corpus Statistics

- 92,244 examples
- 3,962 unique sentences
- Krippendorff's α: 0.831
- Fleiss' κ: 0.808
- 262 words in the vocabulary
- Average sentence length of 11.2

- Four data splits
  - 80.7% training
  - 6.4% development
  - 6.4% public test
  - 6.4% unreleased test



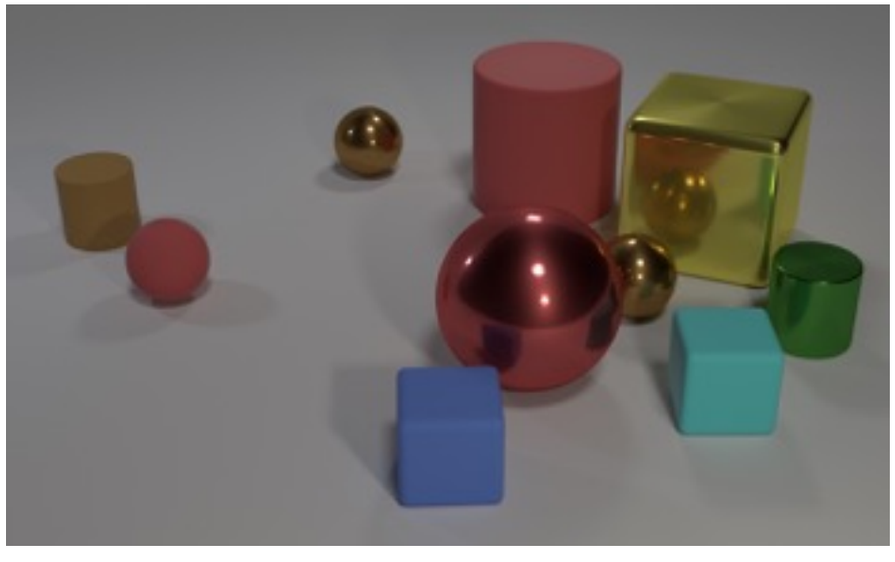
lic.nlp.cornell.edu/nlvr

Related Corpora

A small herd of cows in a large grassy field.

Microsoft COCO
Chen et al 2015

- Image captions
- Photographs
- Natural language



Are there an equal number of large things and metal spheres?

CLEVR
Johnson et al 2016

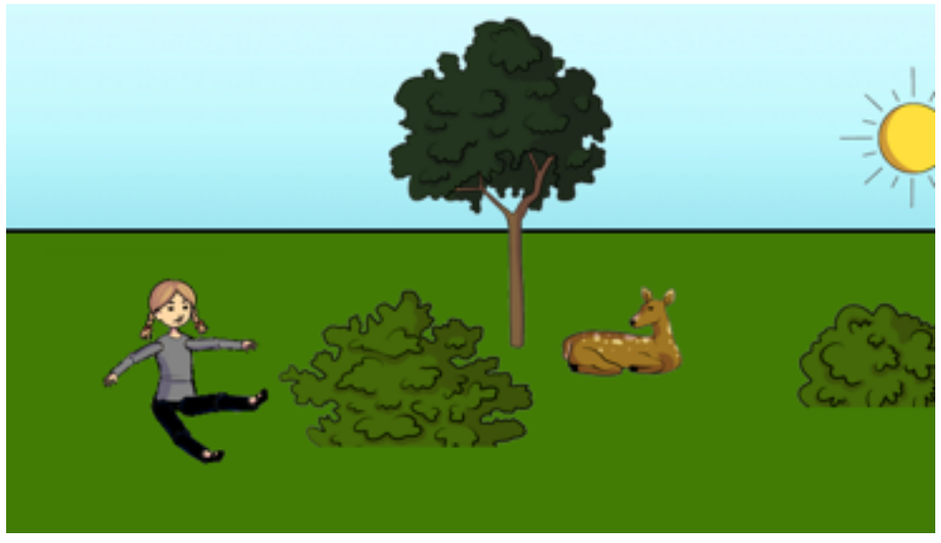
- Open-ended questions
- Synthetic images
- Synthetic language

Related Corpora

What is the dog carrying?

VQA Real Images
Agrawal et al 2015

- Open-ended questions
- Photographs
- Natural language

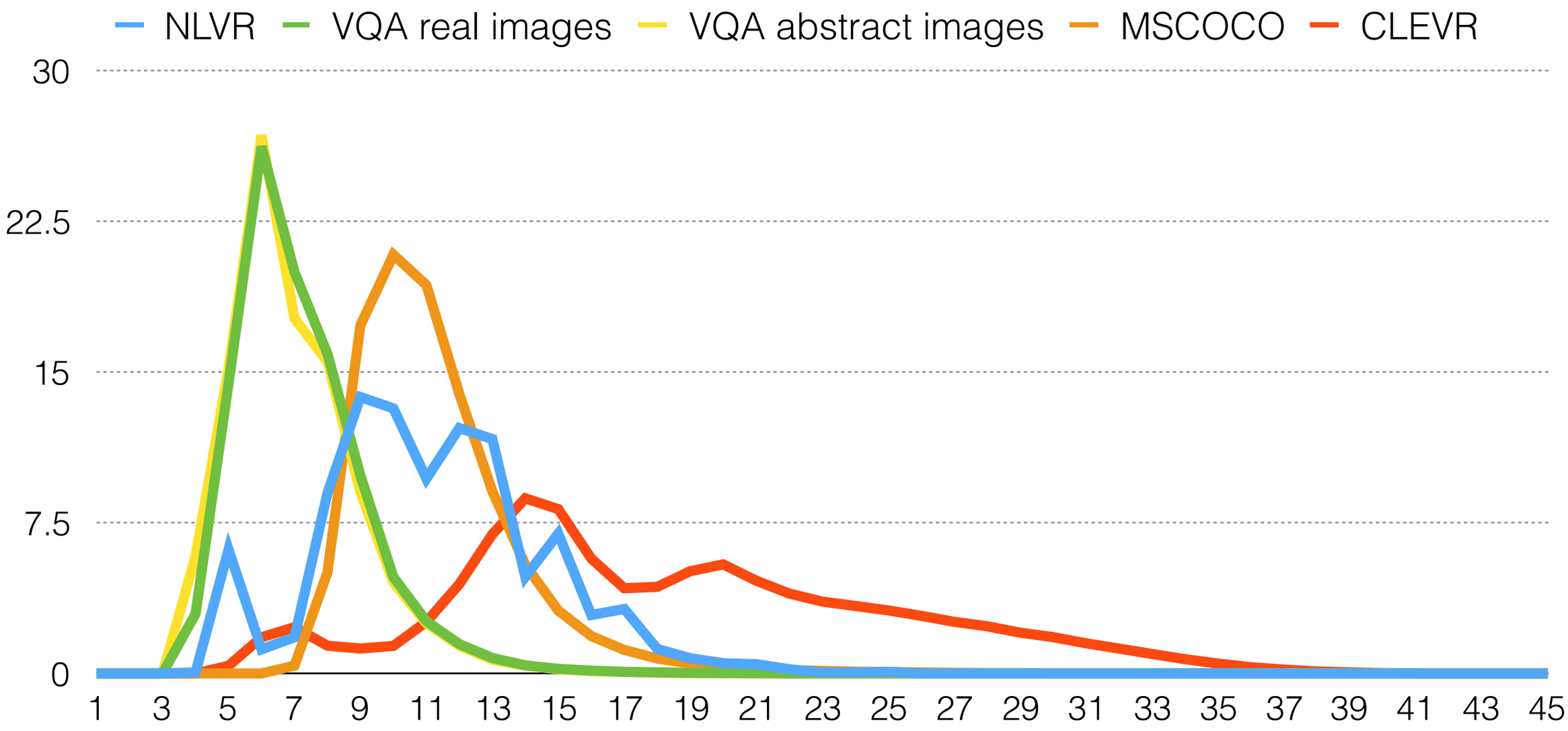


Is the deer standing?

VQA Abstract Images

- Open-ended questions
- Synthetic images (scenes)
- Natural language

Lengths

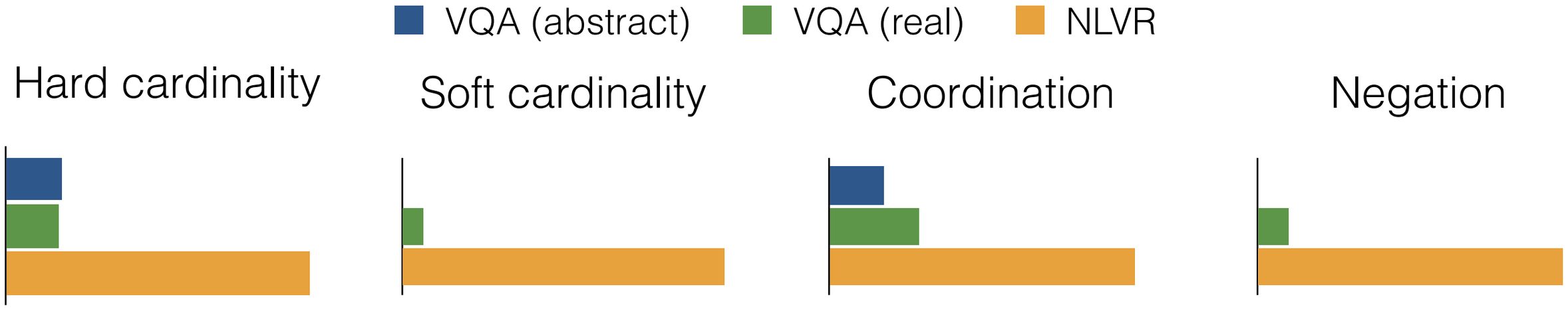


Longer than VQA

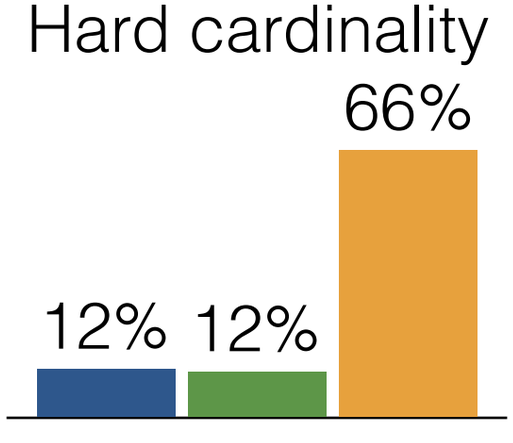
Similar to MS COCO, **easier to evaluate**

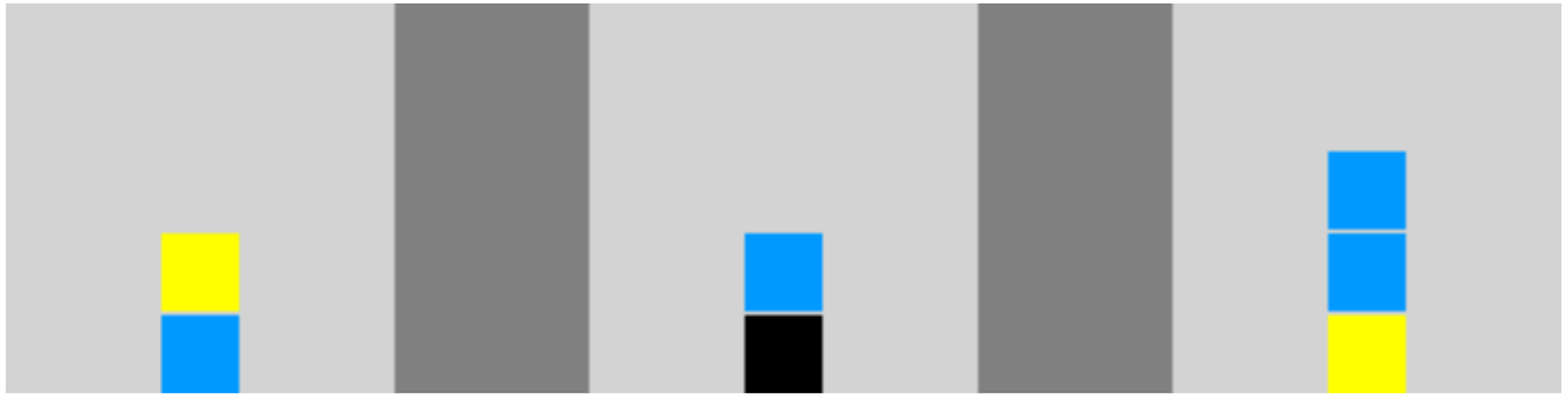
Linguistic Analysis

Analyzed 200 random development sentences.

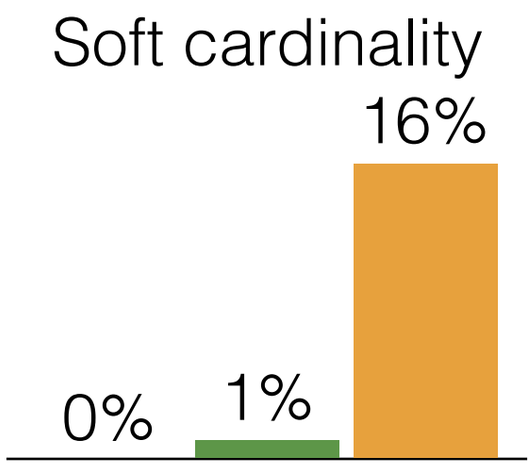


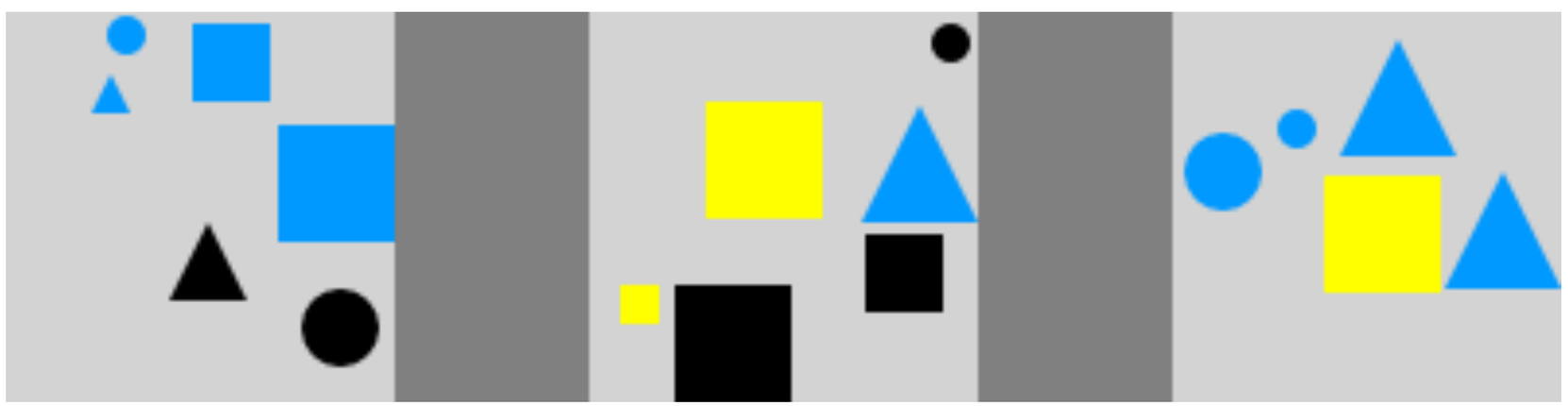
Numerical Expressions





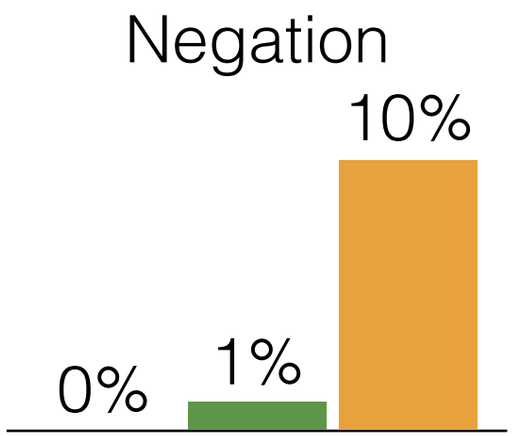
There is a tower with **exactly three** blocks,
and it has a yellow block and **two** blue blocks.

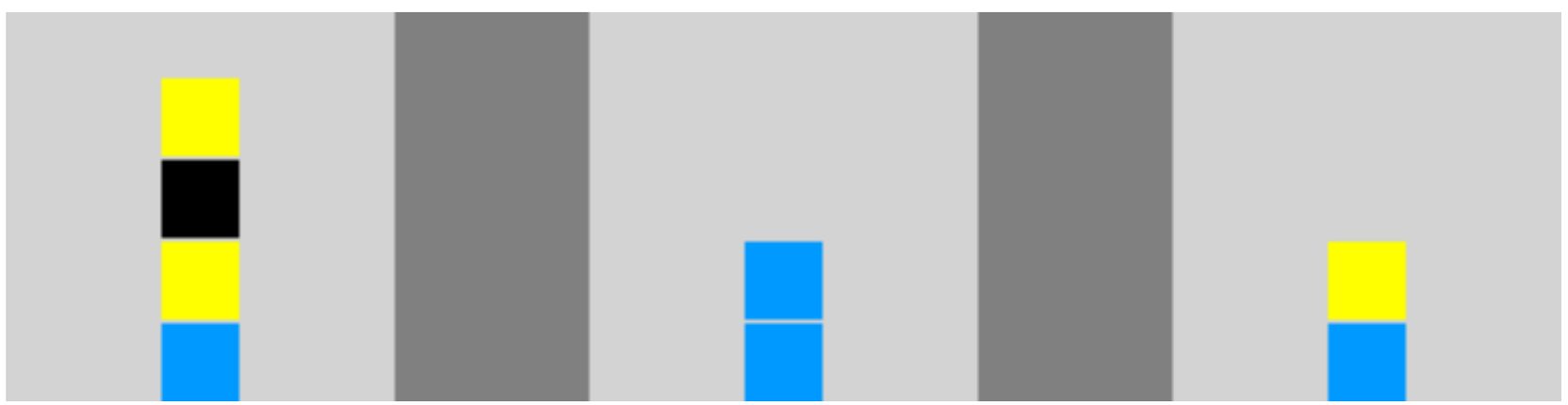




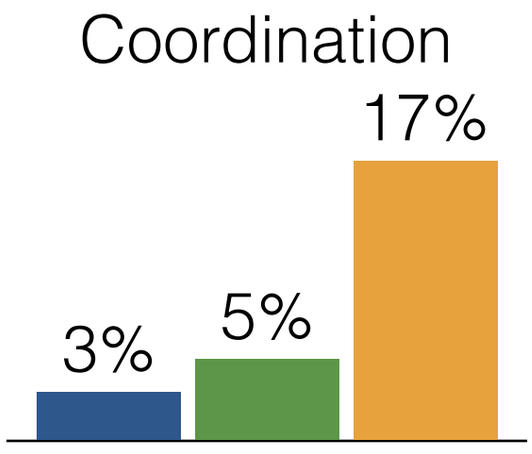
there are **at least two** yellow squares not
touching any edge

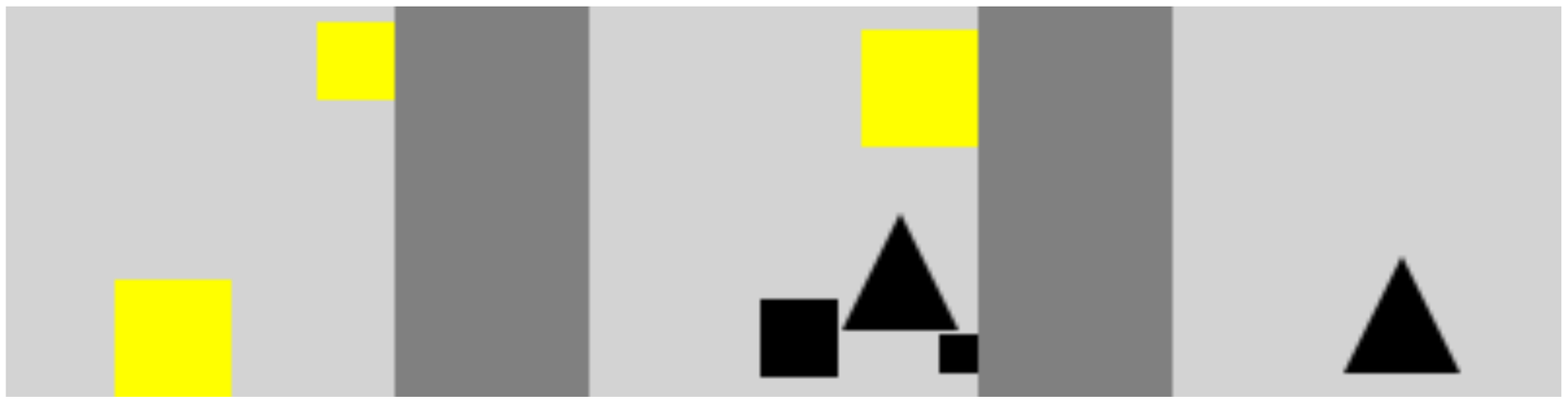
Negation and Coordination





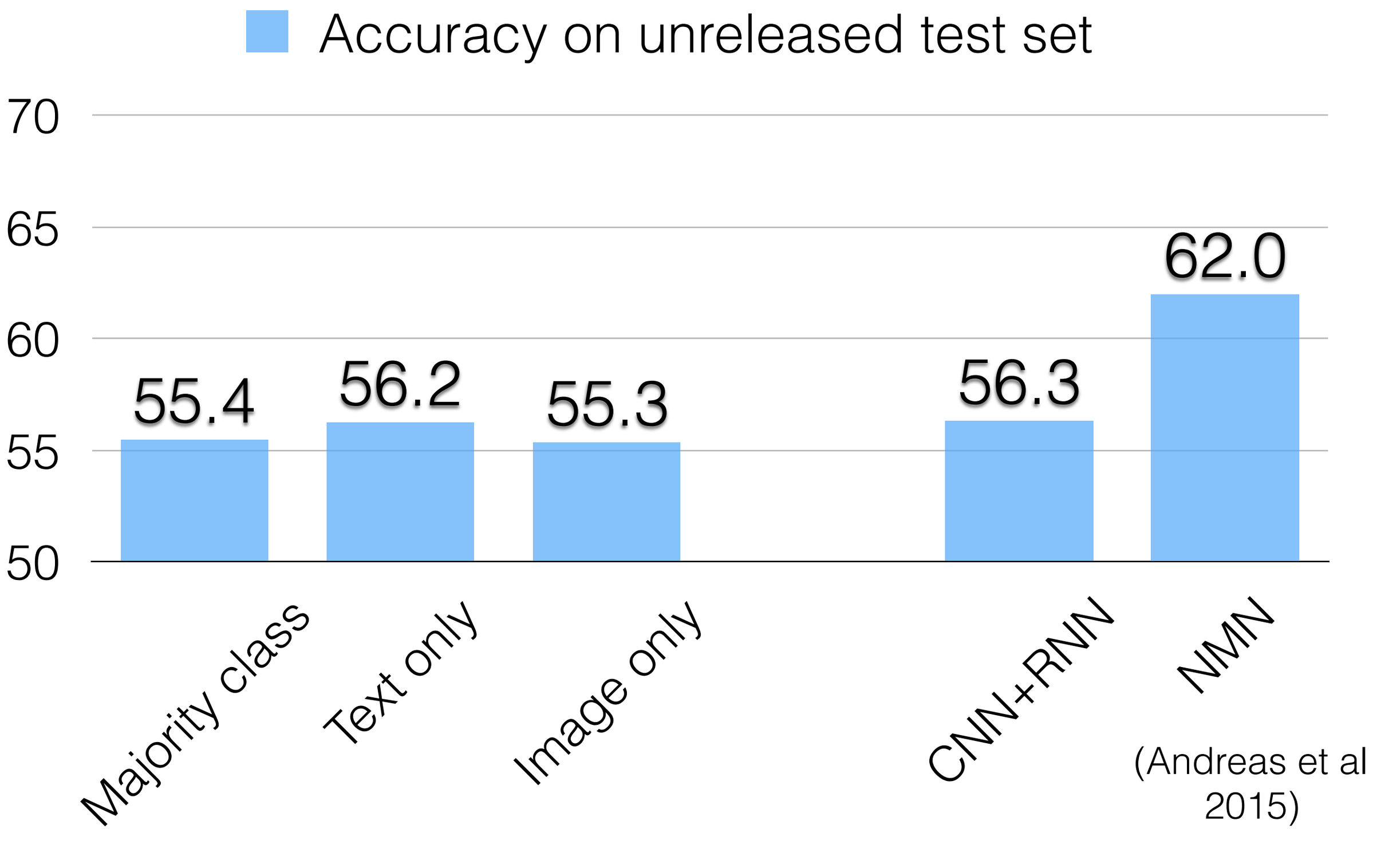
There is a box with a black item between 2 items of the same color and **no item on top of that.**





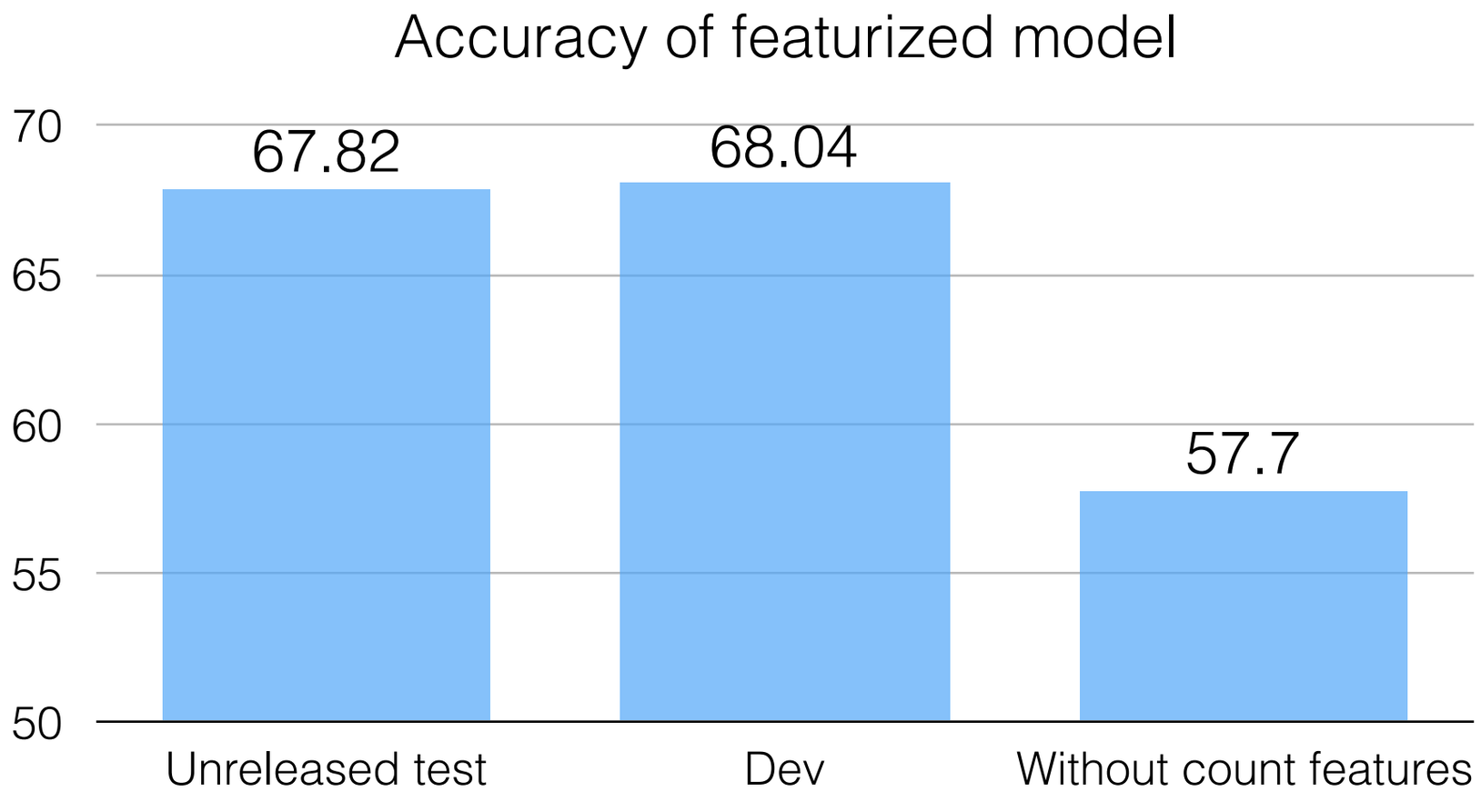
There is a box with a yellow item **and** three black items.

Baselines


Accuracy on unreleased test set
70
65
60
55
50
55.4
56.2
55.3
56.3
62.0
Majority class
Text only
Image only
CNN+RNN
NMN
(Andreas et al 2015)

Feature-based Analysis

- Features **text** and **structured representation**
- Use **maximum entropy model**
- Achieve accuracy of **67.6%** and **67.8%**
- Most influential feature: **count-based features**



http://lic.nlp.cornell.edu/nlvr/

Thank you!