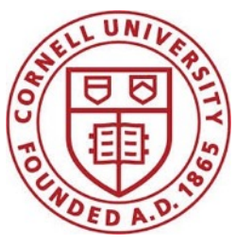
Natural Language for Visual Reasoning **for Real**:

NLVR^2

Alane Suhr*, Stephanie Zhou*, Iris Zhang, Huajun Bai, Yoav Artzi

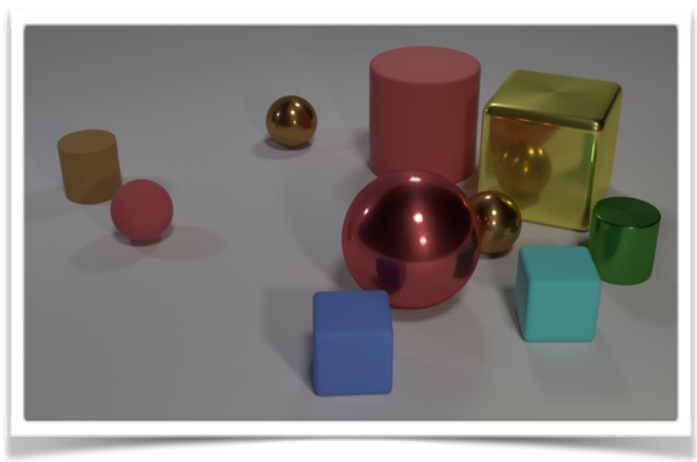
*Equal contribution



Reasoning about Language and Vision

What is the dog carrying?

(**VQA**, Agrawal et al. 2015)

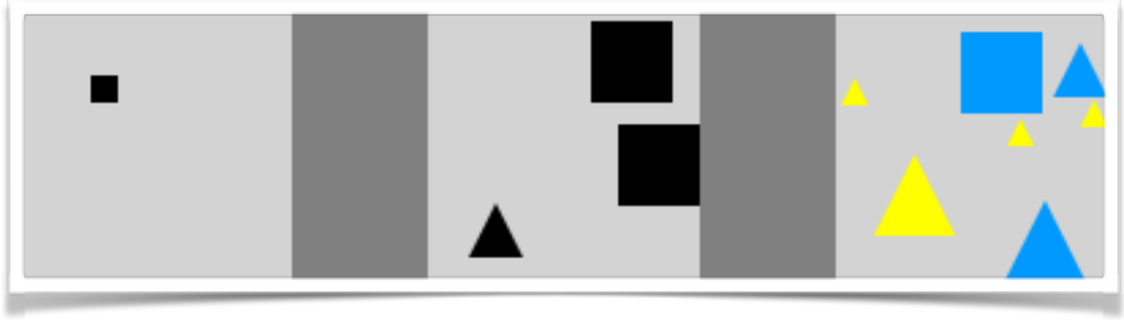


Are there an equal number of large things and metal spheres?

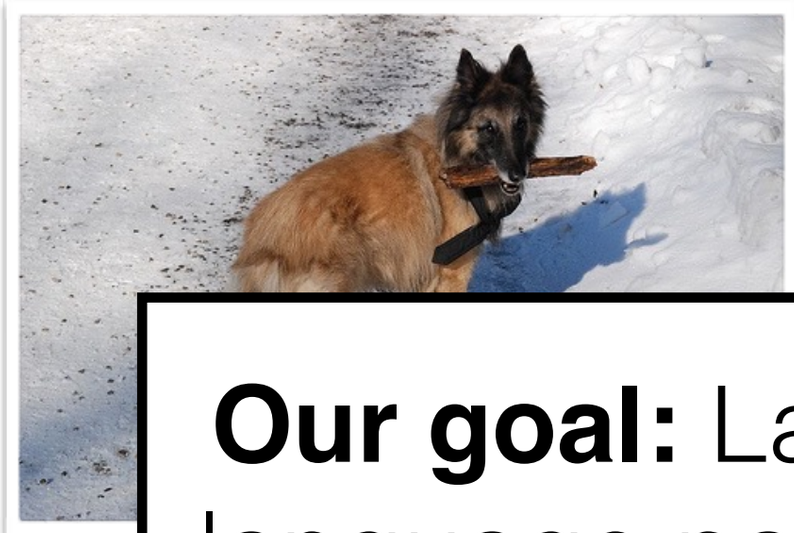
(**CLEVR**, Johnson et al. 2017)

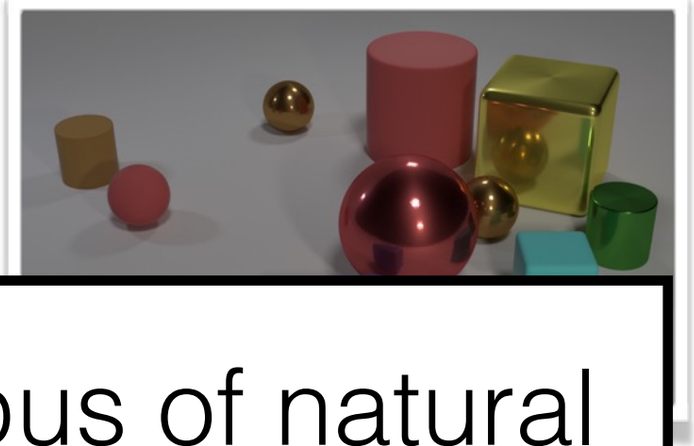
there are exactly three squares not touching any edge

(**NLVR**, Suhr et al. 2017)



Reasoning about Language and Vision





Our goal: Large corpus of natural language paired with photographs focusing on a diverse set of linguistic phenomena

Who

(**VQ**

there a

not touching any edge

(**NLVR**, Suhr et al. 2017)

of large

es?

17)

Natural Language for Visual Reasoning **for Real (NLVR^2)**

All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.

Task: Determine whether the sentence is true or false about the pair of images.

Natural Language for Visual Reasoning **for Real (NLVR^2)**

All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.

TRUE

Task: Determine whether the sentence is true or false about the pair of images.

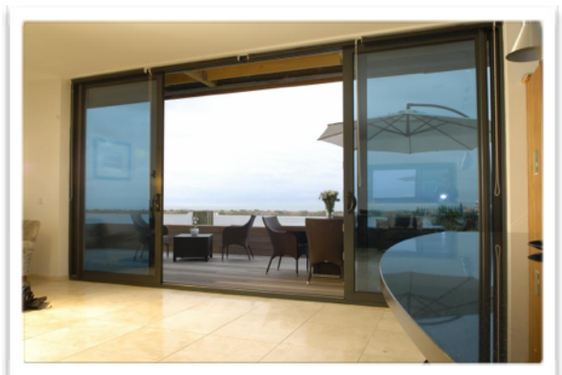
Outline

1. Task
2. Data collection
3. Analysis
4. Results

Task

All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.

TRUE



The left and right image contains the same number of white framed door window panes on at least one door way.

FALSE

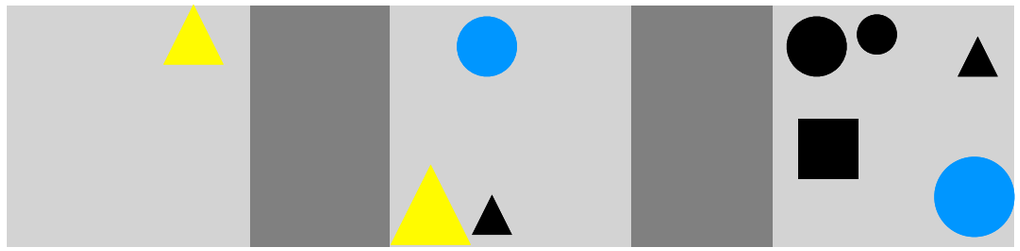
Data Collection

Our goal: Natural language descriptions of photographs and truth judgments

1. Background: NLVR data collection
2. Image collection
3. Sentence writing
4. Validation

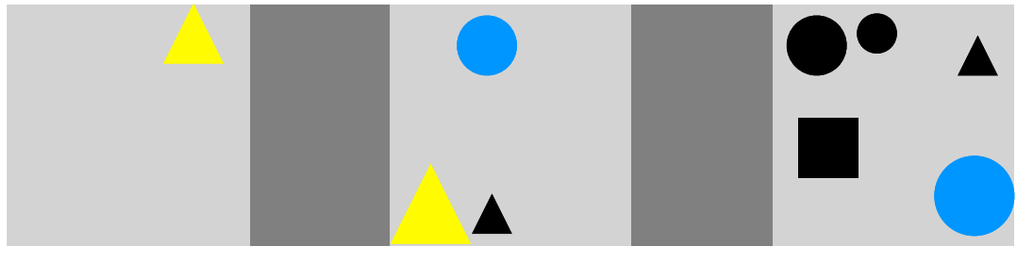
Background:
NLVR Image Generation

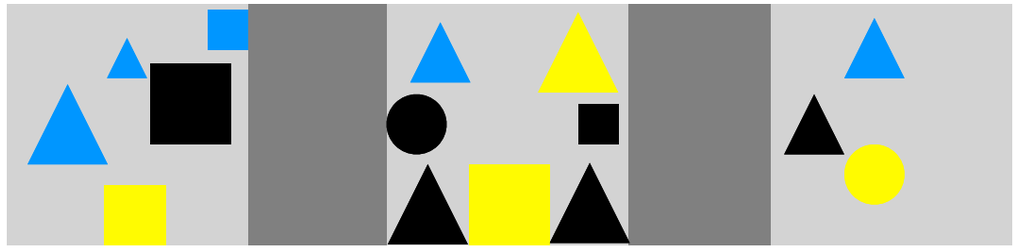
Background:
NLVR Image Generation



- Randomly generate a single image

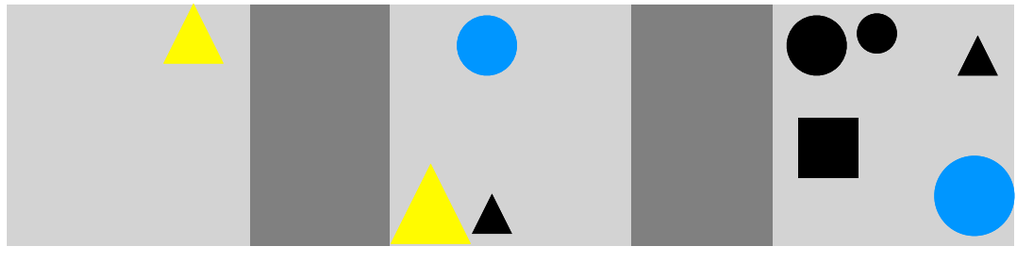
Background: NLVR Image Generation

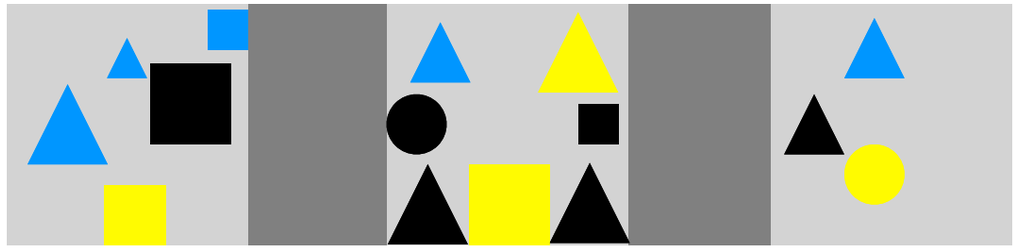


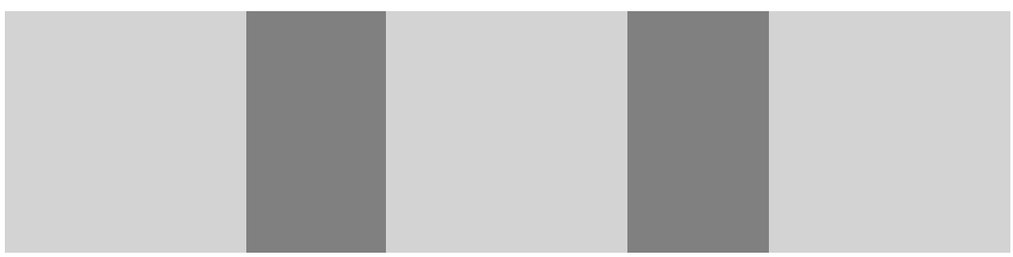


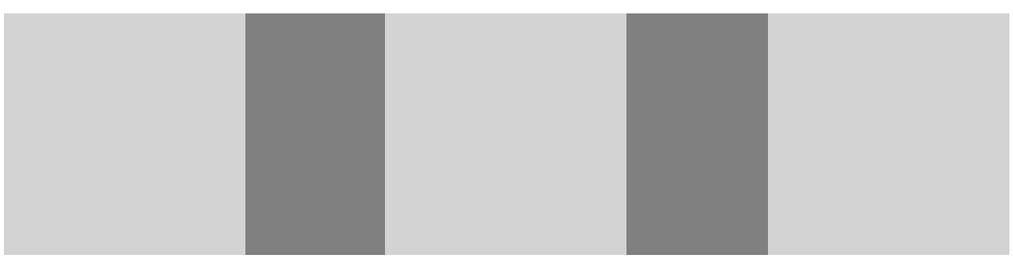
- Randomly generate a single image
- Randomly generate another image

Background: NLVR Image Generation



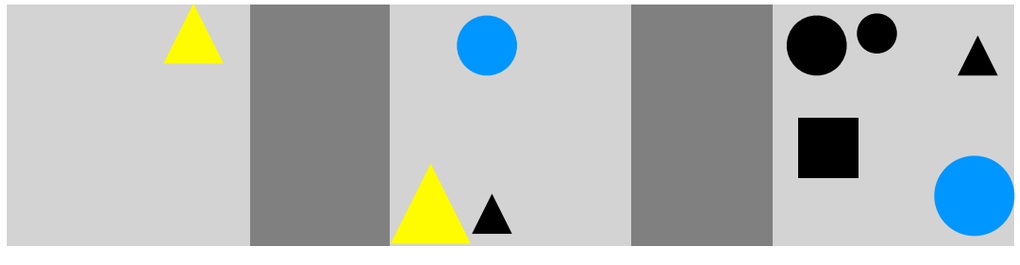


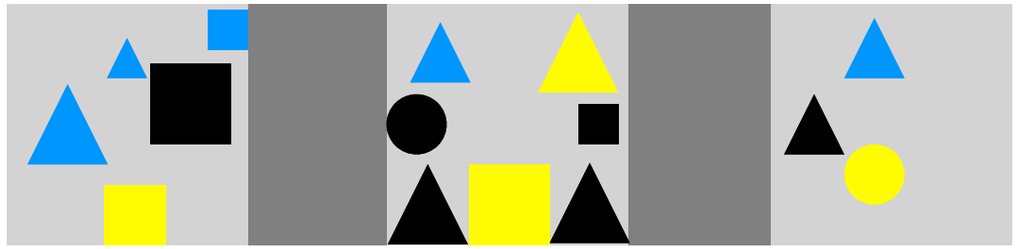


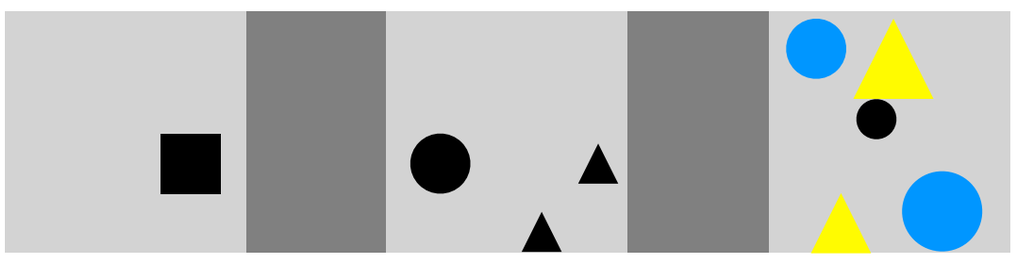


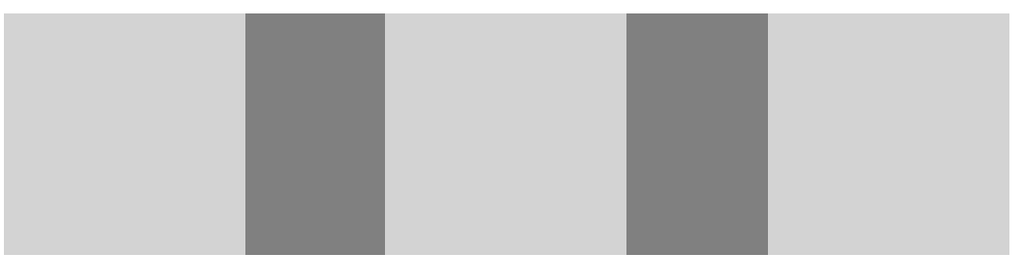
- Randomly generate a single image
- Randomly generate another image

Background: NLVR Image Generation



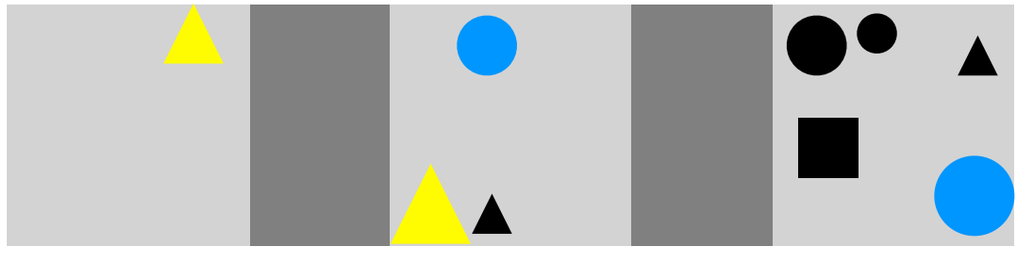


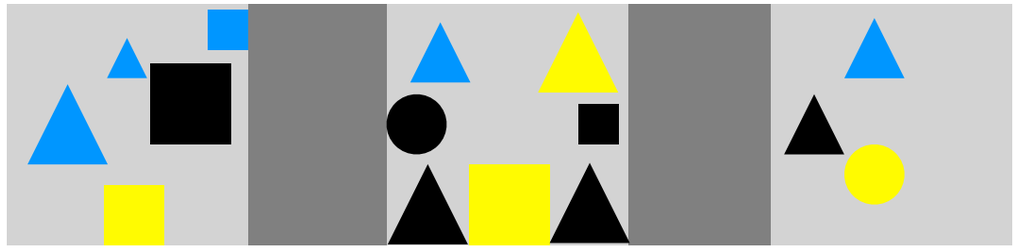


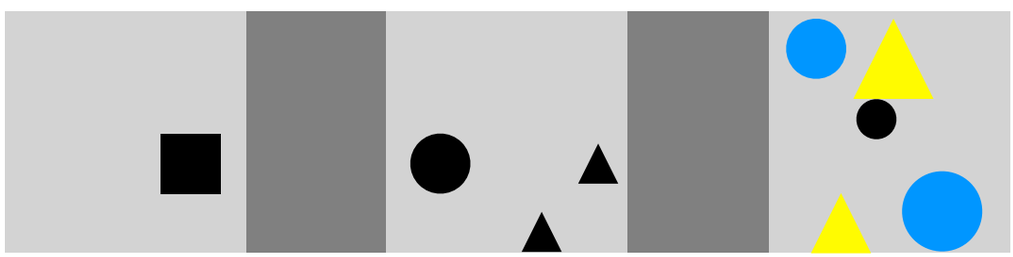


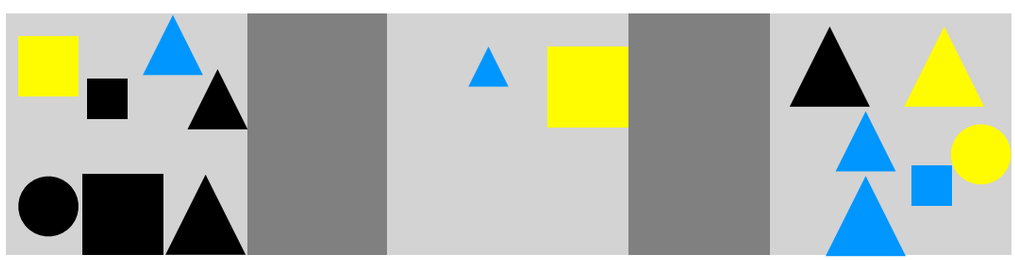
- Randomly generate a single image
- Randomly generate another image
- Generate a third image, using objects from top image

Background: NLVR Image Generation



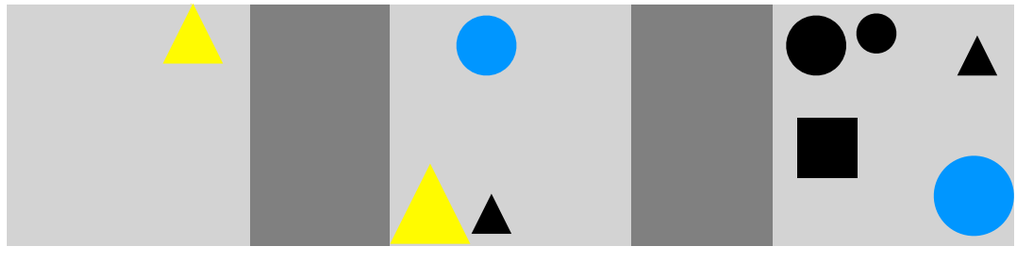




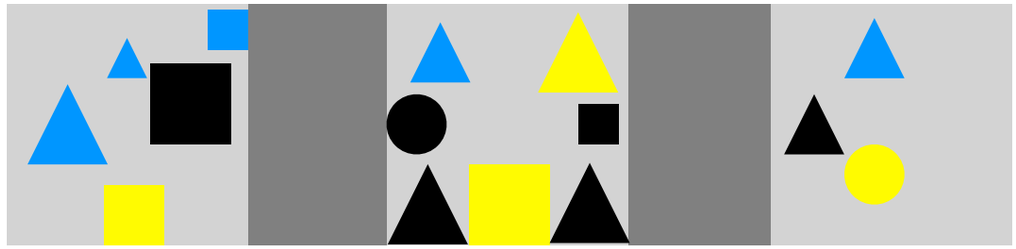


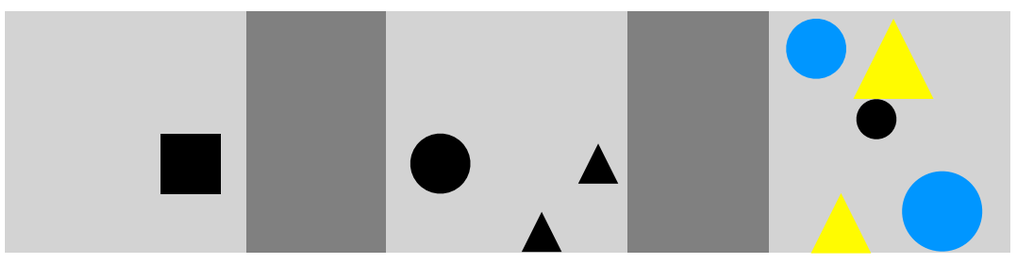
- Randomly generate a single image
- Randomly generate another image
- Generate a third image, using objects from top image
- Generate a fourth image similarly

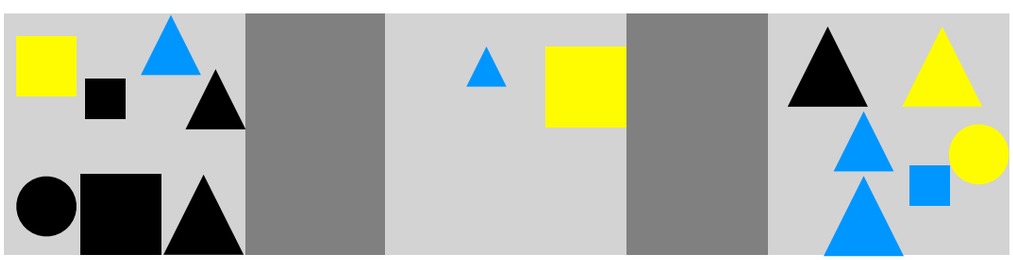
Background:
NLVR Sentence Writing



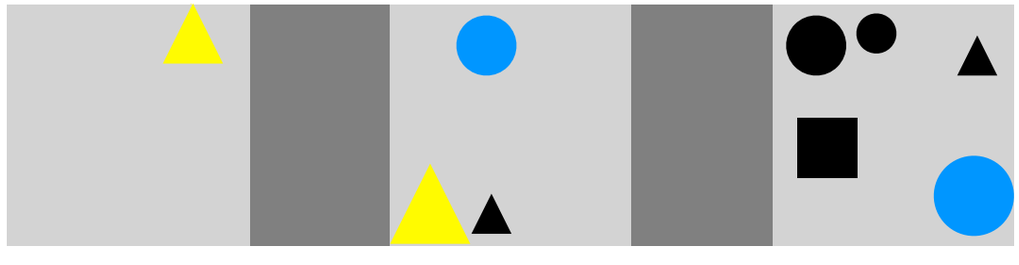
There is a box with 3 items of all 3 different colors.





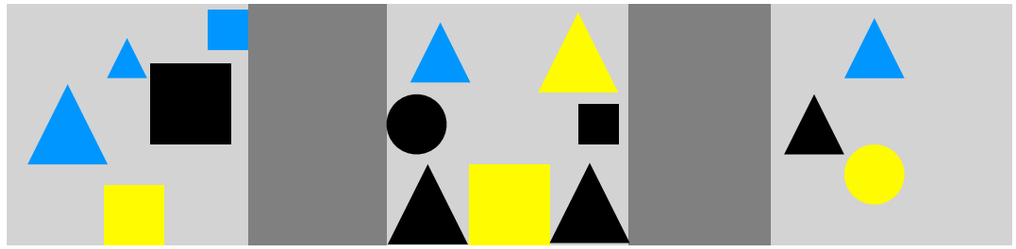


Background:
NLVR Sentence Writing



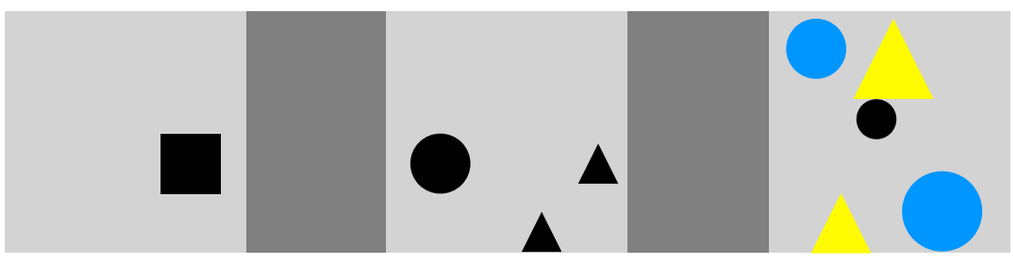
There is a box with 3 items of all 3 different colors.

TRUE



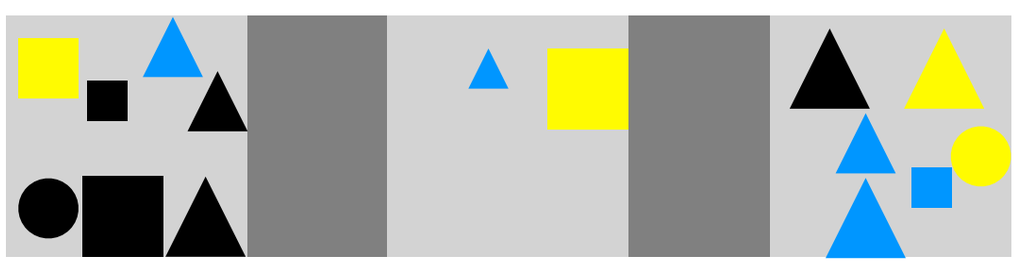
There is a box with 3 items of all 3 different colors.

TRUE



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

FALSE



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

FALSE

Key Data Collection Challenges

- In NLVR, can control image generation to enable complex reasoning

- Can’t generate real images

- How can we use real images but ensure images are visually complex?
  - Queries that elicit complex images
  - Similar Image tools

Image Collection

1. **Pick 124 synsets from ImageNet**
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump

   - Allows use of ImageNet models and tools
   - Allows for weak annotation of image content

Image Collection

1. **Pick 124 synsets from ImageNet**
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump

   - Allows use of ImageNet models and tools
   - Allows for weak annotation of image content

Image Collection

1. **Pick 124 synsets from ImageNet**
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump
2. **Generate and execute search queries**
   Combined synset names with numerical phrases, hypernyms, and similar words

🔍 two acorns

Image Collection

1. **Pick 124 synsets from ImageNet**
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump
2. **Generate and execute search queries**
   Combined synset names with numerical phrases, hypernyms, and similar words

🔍 two acorns

Image Collection

1. **Pick 124 synsets from ImageNet**
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump
2. **Generate and execute search queries**
   Combined synset names with numerical phrases, hypernyms, and similar words

Image Collection

3. **Remove low-quality images**

   Don’t contain synset, drawings, inappropriate content

Image Collection

3. **Remove low-quality images**

   Don’t contain synset, drawings, inappropriate content

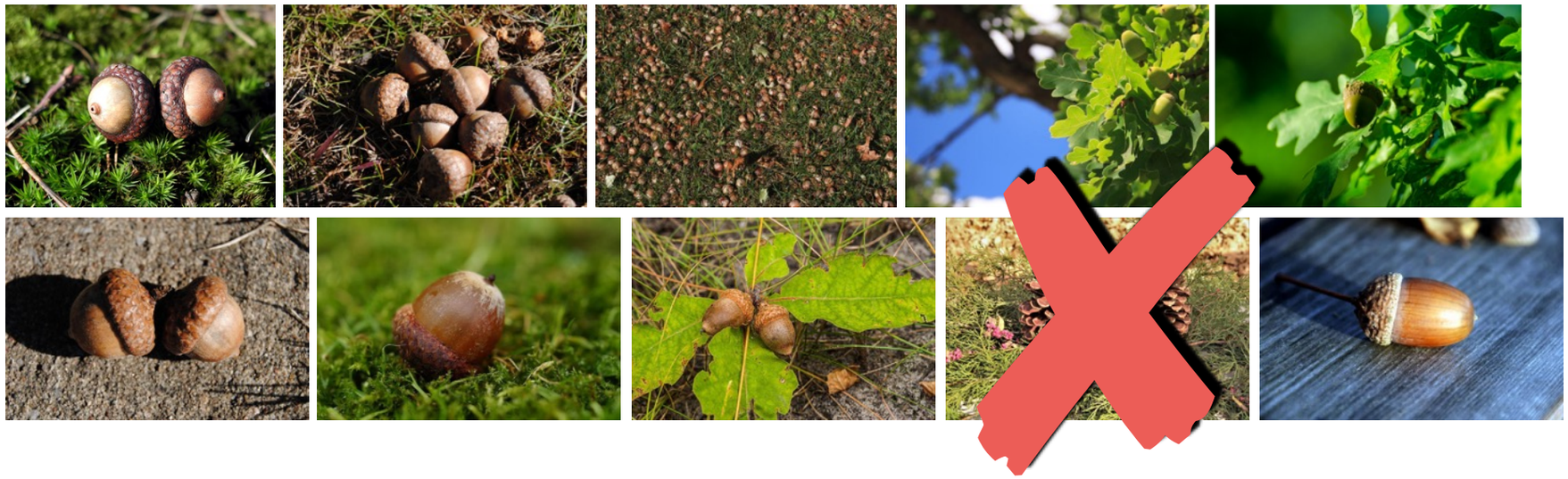


Image Collection

3. **Remove low-quality images**

   Don’t contain synset, drawings, inappropriate content

4. **Construct sets of eight images**

   Each set must contain at least three *interesting* images (e.g., multiple objects)

Image Collection

3. **Remove low-quality images**

   Don’t contain synset, drawings, inappropriate content

4. **Construct sets of eight images**

   Each set must contain at least three *interesting* images (e.g., multiple objects)

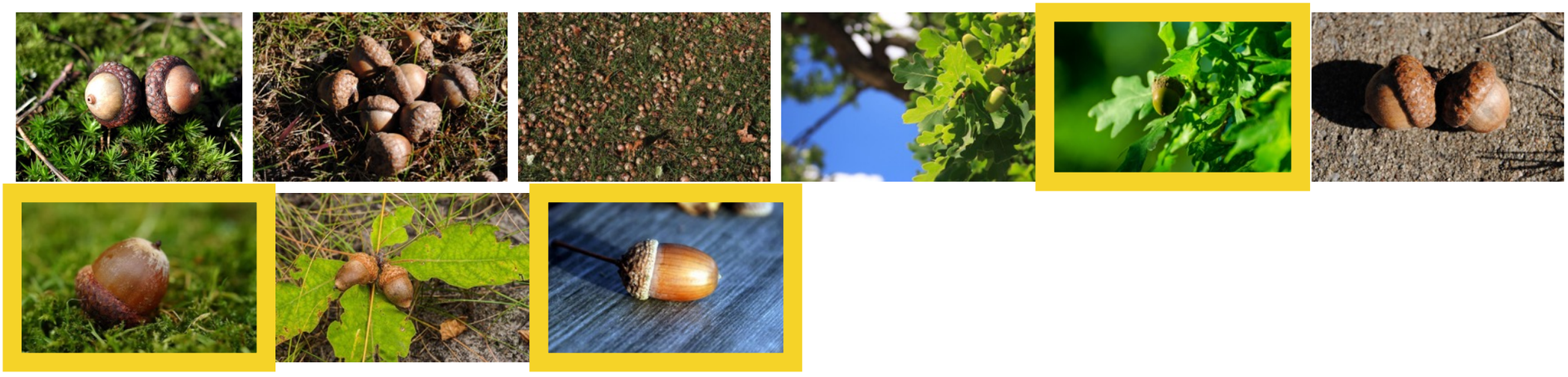


Image Collection

3. **Remove low-quality images**

   Don’t contain synset, drawings, inappropriate content

4. **Construct sets of eight images**

   Each set must contain at least three *interesting* images (e.g., multiple objects)

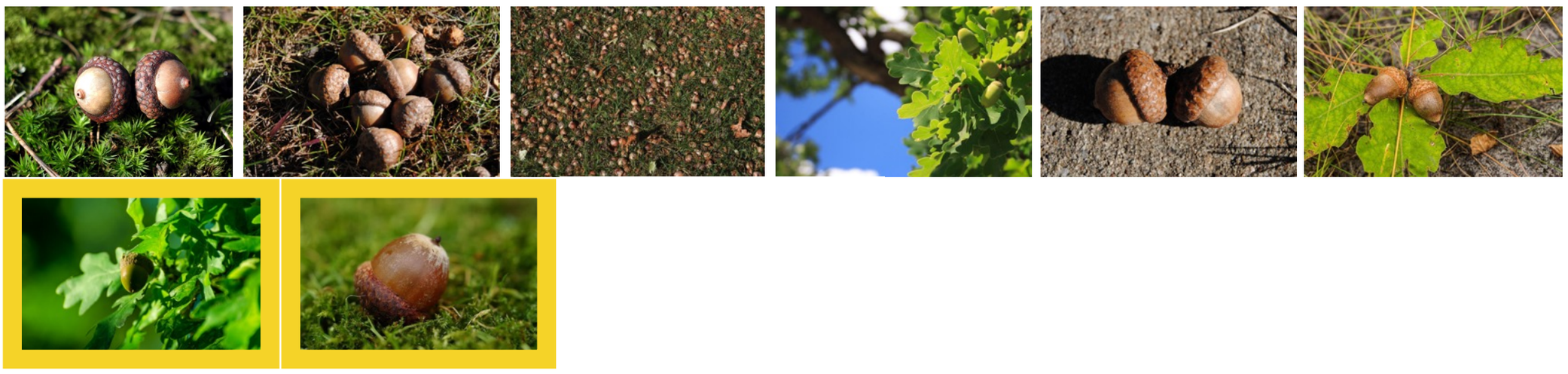
Image Collection

3. **Remove low-quality images**

   Don’t contain synset, drawings, inappropriate content

4. **Construct sets of eight images**

   Each set must contain at least three *interesting* images (e.g., multiple objects)

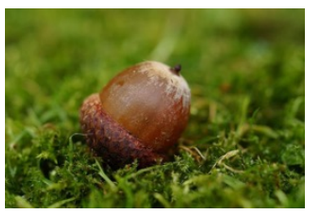


Sentence Writing

5. Display a set of randomly paired images





Sentence Writing

5. **Display a set of randomly paired images**

6. **Ask workers to select two pairs**

Sentence Writing

5. **Display a set of randomly paired images**

6. **Ask workers to select two pairs**

7. **Workers write a sentence true about the selected pairs, but false about the others**

One image shows exactly two brown acorns in back-to-back caps on green foliage.

Validation

8. **Show each image/sentence pair to another work and ask them to label it**

One image shows exactly two brown acorns in back-to-back caps on green foliage.

True ☐

False ☐

Validation

8. **Show each image/sentence pair to another work and ask them to label it**

One image shows exactly two brown acorns in back-to-back caps on green foliage.

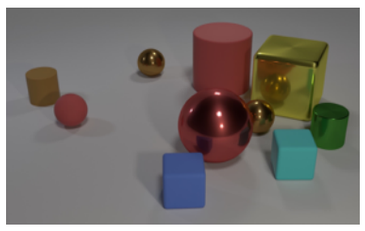
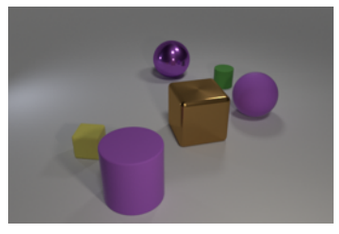
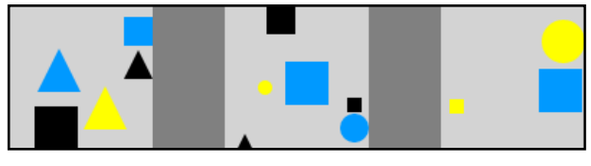
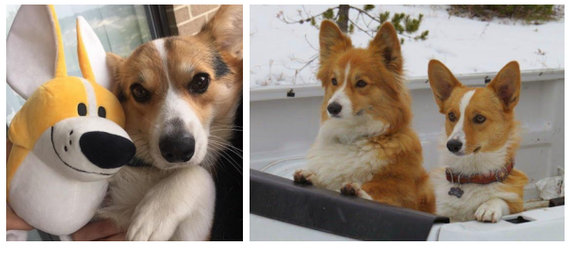
True ☐

False ☑

Data Collection and Corpus Statistics

- **107,296 total examples**
  - 29,680 unique sentences
  - 127,506 unique images
  - 80% train, 20% evenly split among dev and two test sets
- **Agreement:** near perfect ($\alpha = 0.912$, $\kappa = 0.889$)
- **Total cost:** $19,282.99
- **Average sentence length:** 14.8 tokens
- **Vocabulary size:** ~7,500 word types

Related Corpora

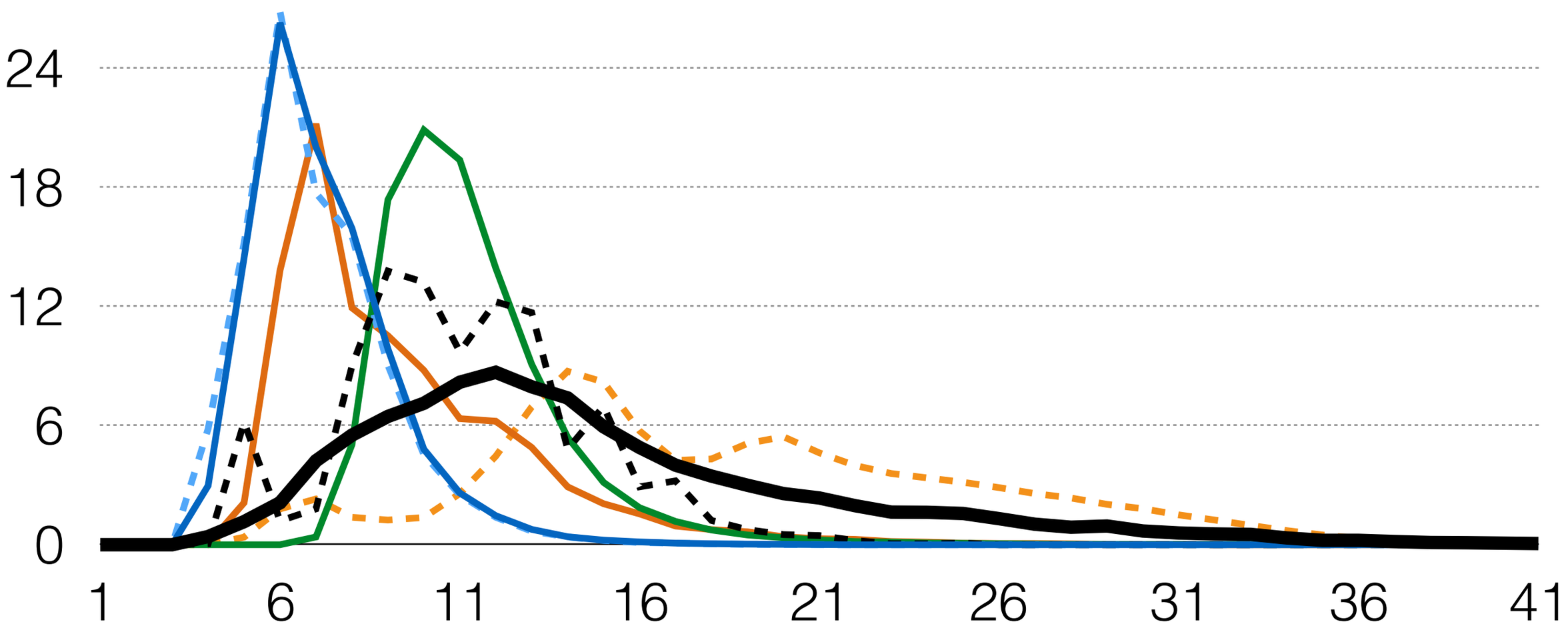
	Task	Examples	
MSCOCO (Chen et al 2015)	**Caption generation**		*A small herd of cows in a large grassy field.*
VQA (Agrawal et al 2015)	**Question answering**		*What is the dog carrying?*
CLEVR (Johnson et al 2017a)	**Question answering**		*How many objects are either small cylinders or red things?*
CLEVR-Humans (Johnson et al 2017b)	**Question answering**		*How many objects are not purple and not metallic?*
NLVR (Suhr et al 2017)	**Binary classification**		*there are exactly three blue objects not touching any edge*
NLVR²	**Binary classification**		*All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.*

Related Corpora

	Task	Real images?	Natural language?
MSCOCO (Chen et al 2015)	**Caption generation**	✔	✔
VQA (Agrawal et al 2015)	**Question answering**	✔	✔
CLEVR (Johnson et al 2017a)	**Question answering**	✗	✗
CLEVR-Humans (Johnson et al 2017b)	**Question answering**	✗	✔
NLVR (Suhr et al 2017)	**Binary classification**	✗	✔
NLVR²	**Binary classification**	✔	✔

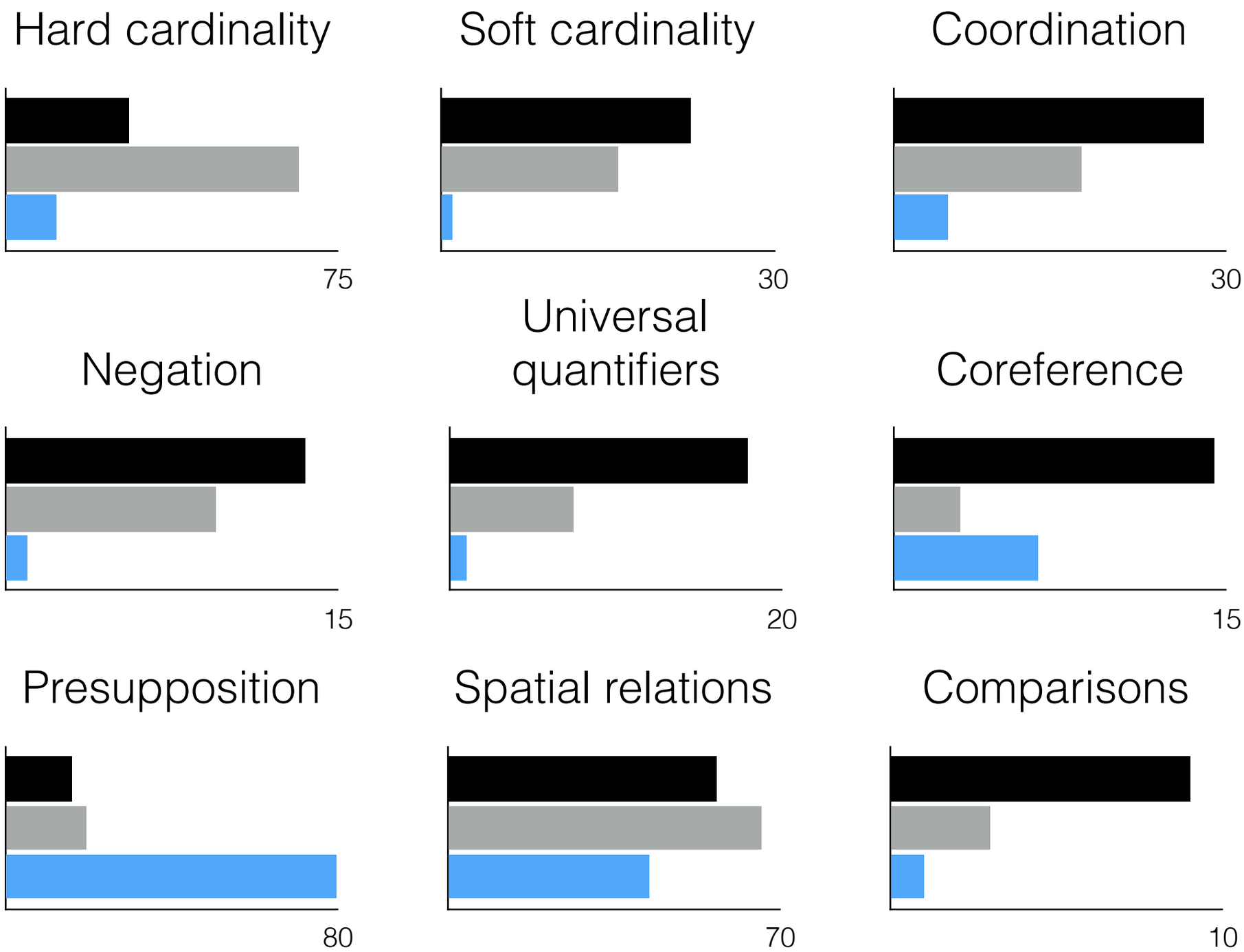
Sentence Lengths

- **NLVR² (ours)**
- NLVR
- MSCOCO
- VQA real images
- VQA abstract images
- CLEVR-Humans
- CLEVR

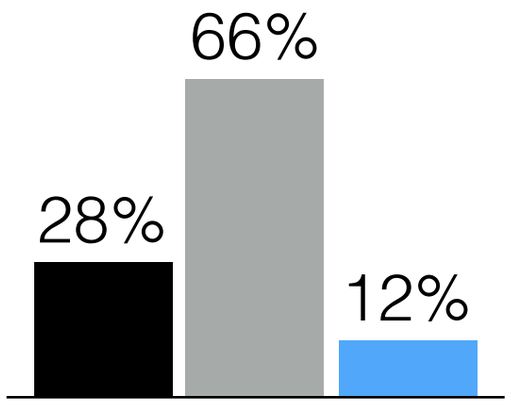


Linguistic Analysis

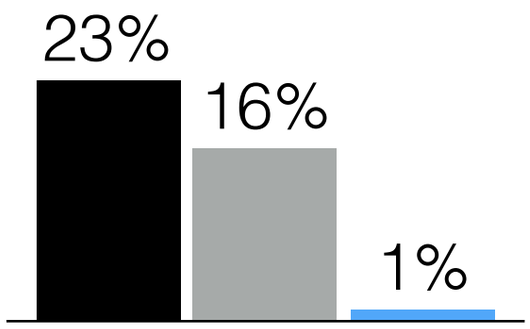
NLVR²
NLVR
VQA (real)



Hard Cardinality



Soft Cardinality



There are <u>two</u>, and only <u>two</u>, people.

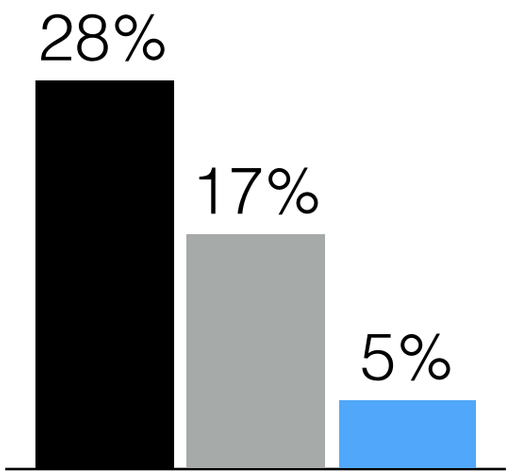
NLVR²

NLVR

VQA (real)

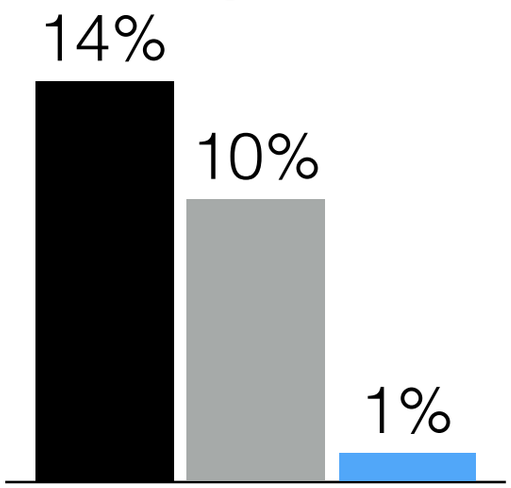
There are <u>no more than eight</u> bottles in total.

Coordination



Each image contains just one bird, <u>and</u> the wires of a cage are behind the bird in one image.

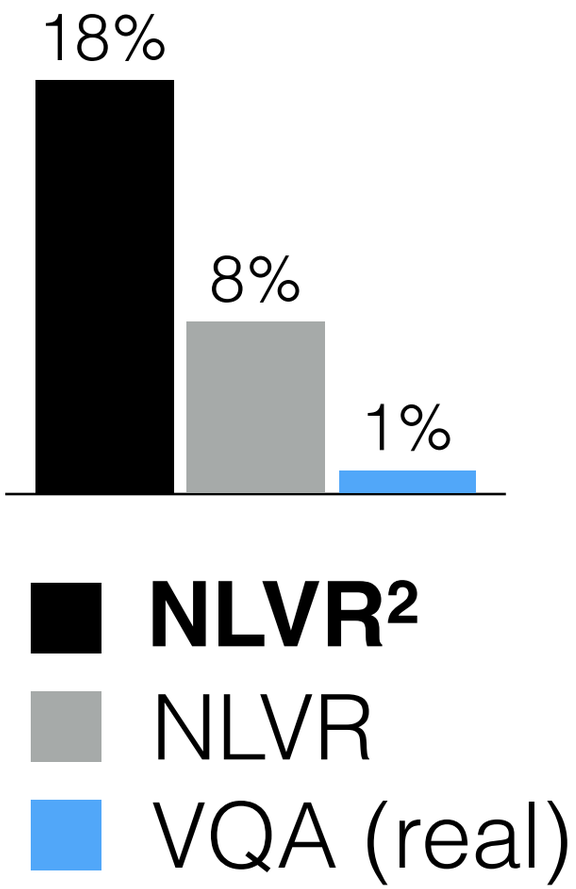
Negation



NLVR²
NLVR
VQA (real)

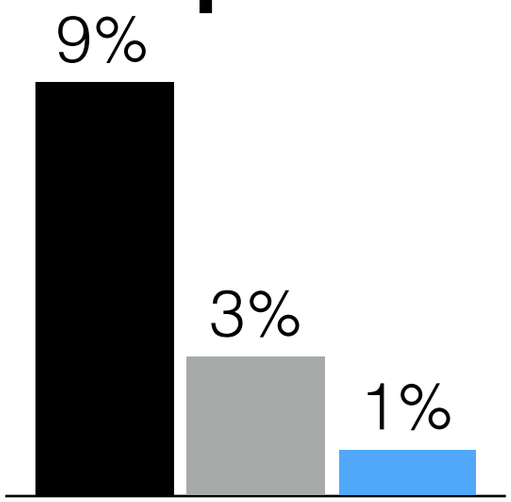
A mitten is being worn in one image and the mittens are not being worn in the other image.

Universal Quantifiers



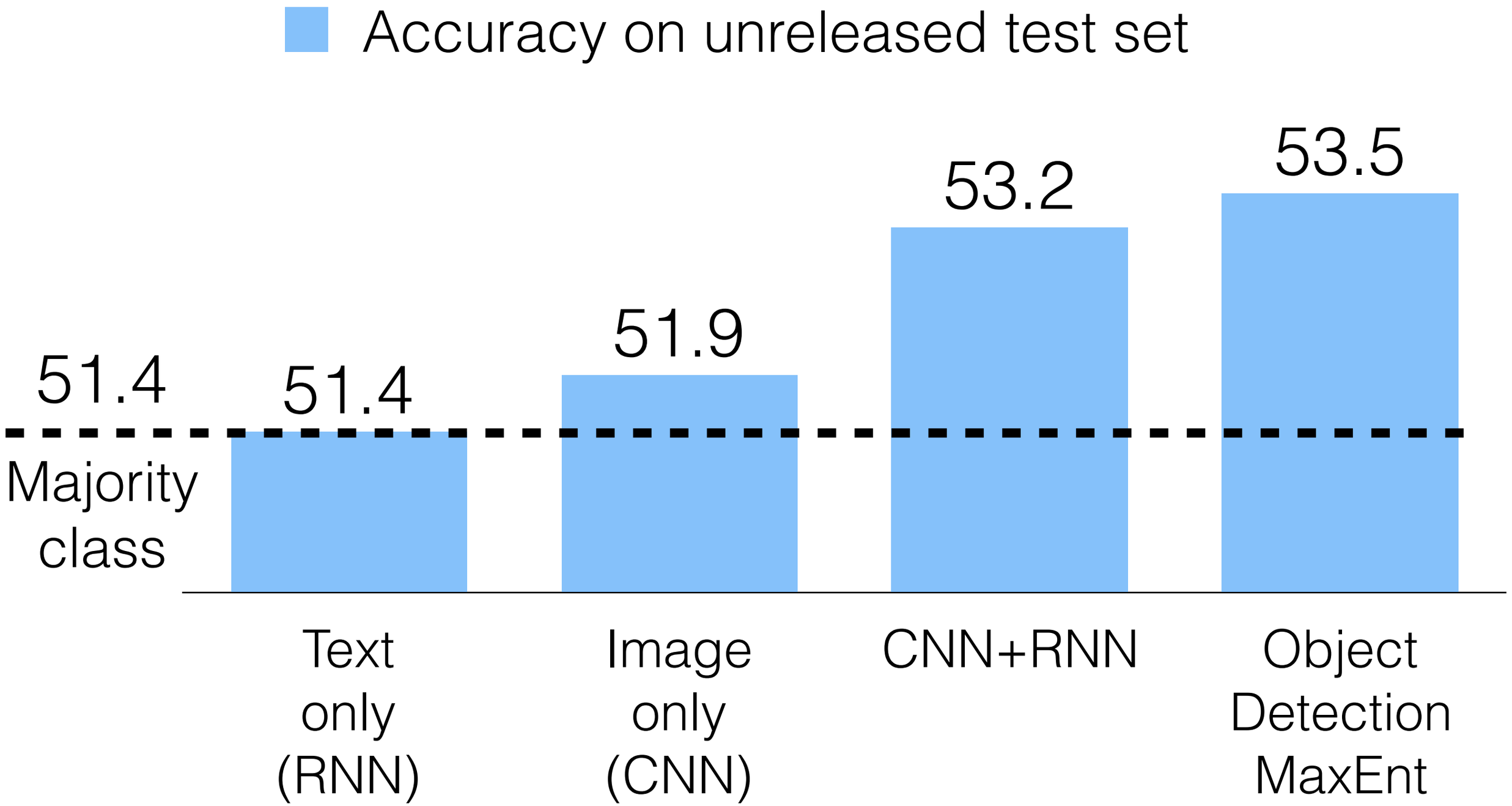
<u>Both</u> images shows a silver pail being used as a flower vase.

Comparisons

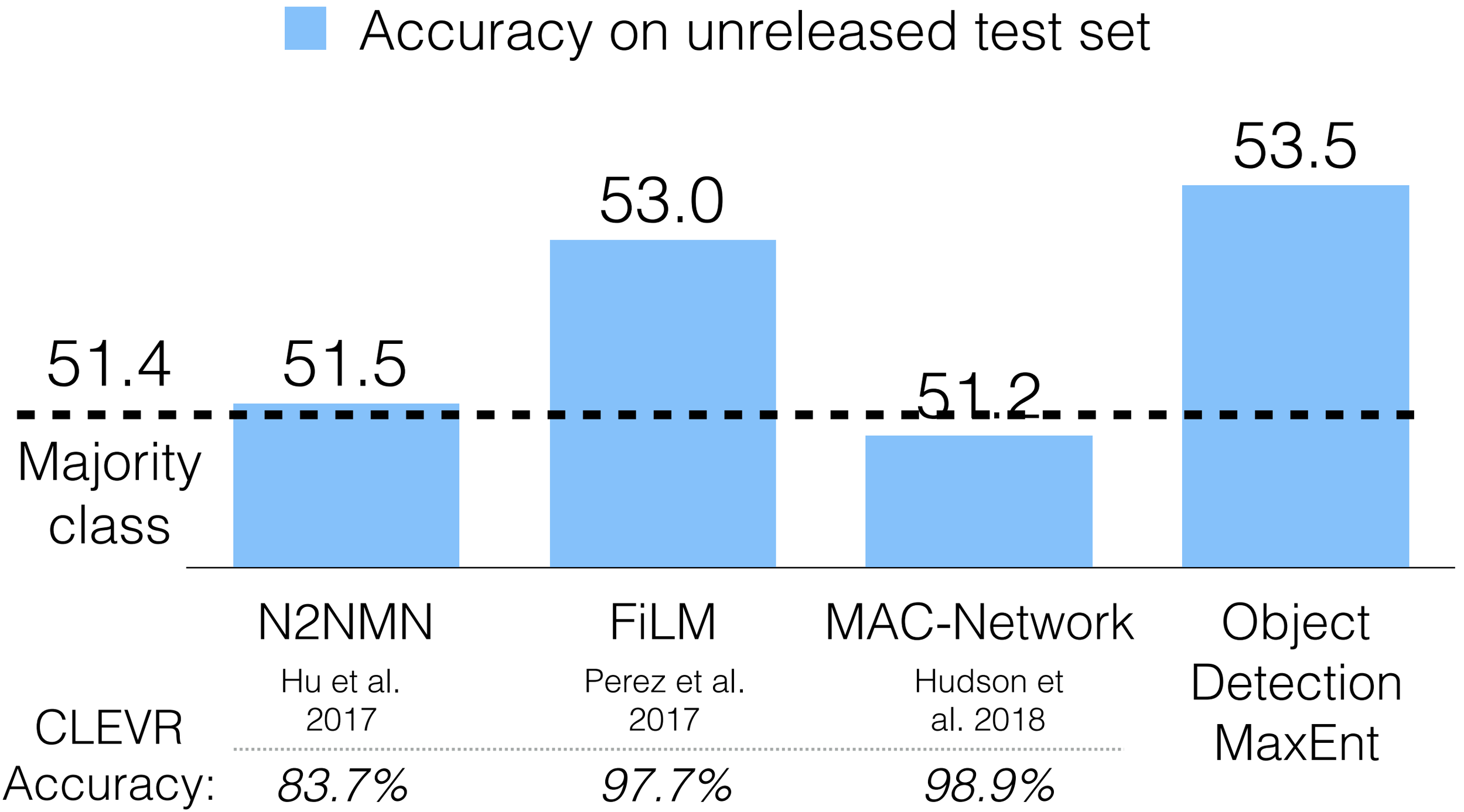


the left image has 4 balloons
of all <u>different</u> colors

Baselines


Accuracy on unreleased test set
51.4
51.4
51.9
53.2
53.5
Majority class
Text only (RNN)
Image only (CNN)
CNN+RNN
Object Detection MaxEnt

SOTA Visual Reasoning



Paper, Data, Code

- Paper, data, and code will be released soon
- Will integrate with ParlAI
- **Thank you for attending**
- **<u>Thanks to ParlAI for funding!</u>**