Leveraging Cognitive Science to Unravel the Complexities of Generative Models

Aida Nematzadeh

Google DeepMind

CMCL 2024

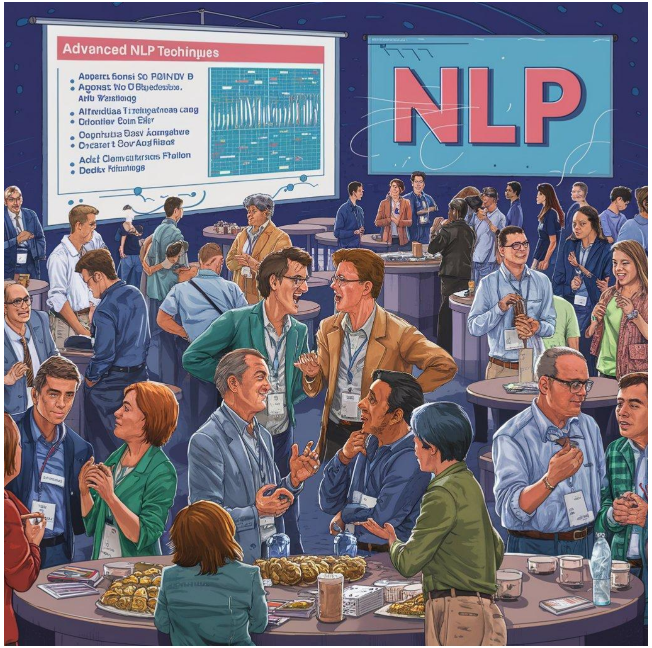
The Power of Generative Models

Write a poem about NLP conferences

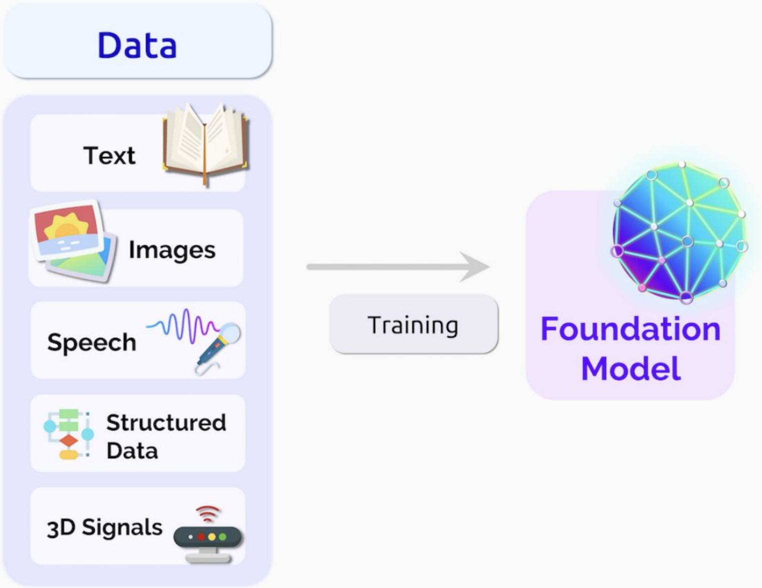
In tech halls, NLP minds unite,
Sharing knowledge, a brilliant light.
Neural nets and deep learning's art,
Push language's boundaries apart.
With posters, demos, and keynote spark,
NLP's future shines, ever dark.

ideogram

generate an image of an NLP conference



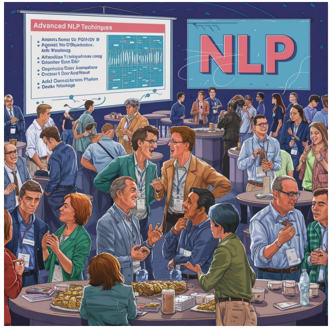
The details of data and model are often unknown.



[foundation models: Image source]

Generate

Observed Output



A Write a poem about NLP conferences

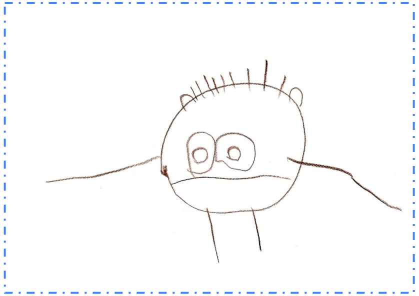
In tech halls, NLP minds unite,
Sharing knowledge, a brilliant light.
Neural nets and deep learning's art,
Push language's boundaries apart.
With posters, demos, and keynote spark,
NLP's future shines, ever dark.

The details of data and model are often unknown.

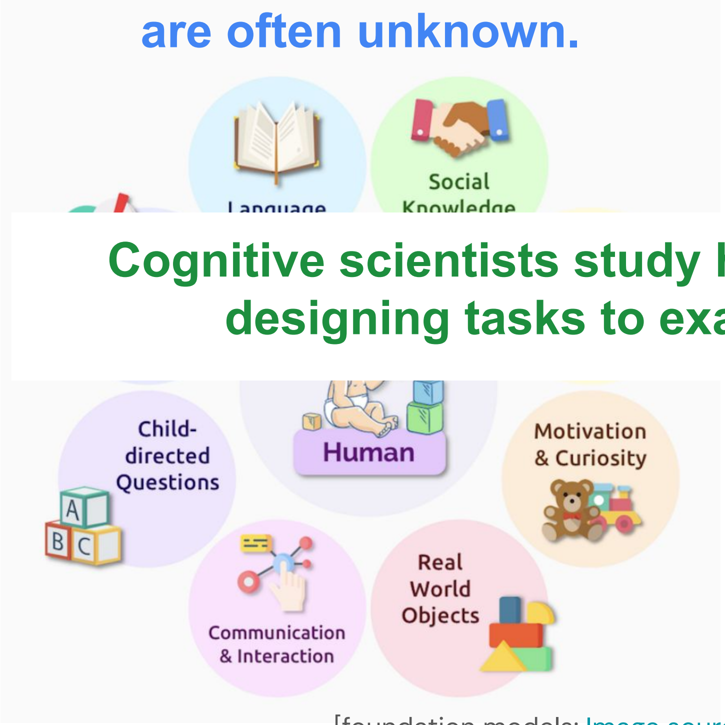
[foundation models: Image source]

Generate

Observed Output and Behavior



The details of data and model are often unknown.



[foundation models: Image source]

Observed Output and Behavior

Cognitive scientists study humans as a black box by designing tasks to examine their behavior.

Lessons from Cognitive Science

Collecting human data.

Controlled study of a specific phenomenon.

Lessons from Cognitive Science

Collecting human data.

Controlled study of a specific phenomenon.

Evaluating Multimodal Generative Models [Wiles et al, 2024]

(3) Develop metrics

Metrics

Human Evaluation

(2) Collect human data

measure alignment

Model 1

Model 2

Model 3

Generate images for a set of models

(1) Choose a prompt set

“cat with a pointer standing in front of a blackboard”

Evaluating Multimodal Generative Models

(3) Develop metrics

Metrics

Human Evaluation

(2) Collect human data

Model 1 Model 2 Model 3

Generate images for a set of models

(1) Choose a prompt set

“cat with a pointer standing in front of a blackboard”

Different Ways to Collect Human Data for Alignment

Prompt:
A dog is to the right of the cat.

(1) Likert

1 - <u>2</u> - 3 - 4 - 5

More consistent

(2) Word Level

A dog is to the right of the cat

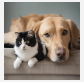
(3) DSG(H)

Q1: Is there a dog?
A: <u>Yes</u>, No
Q2: Is there a cat?
A: Yes, <u>No</u>
…

Absolute comparison

fine-grained annotations

(4) Preference (SxS)

 **VS** 

Relative comparison

There is no standardised way to collect human data across previous work.

Each Template Presents Its Own Challenges

Prompt:
A giraffe stands in the field.

Likert

Rater 1: 5 - Consistent
Rater 2: 5 - Consistent
Rater 3: 4 - Mostly consistent

Prompt:
A wood carving of an owl.

DSG(H)

Q1: Is there a church?
A: Yes, No
Q2: Is there a wood carving?
A: Yes, No
Q3: Is the wood carving made of wood?
A: Yes, No

No question relating owl and wood carving

Prompt:
A Nexus One is placed on a bench.

WL

Rater 1: A Nexus One is placed on a bench.

Rater 2: A Nexus One is placed on a bench.

Rater 3: A Nexus One is placed on a bench.

Raters disagree when rating words that are not relevant for the evaluation

Evaluating Human Templates: Data Quality

Measure the quality of the data across many conditions: compute overall inter annotator agreement with Krippendorff's α

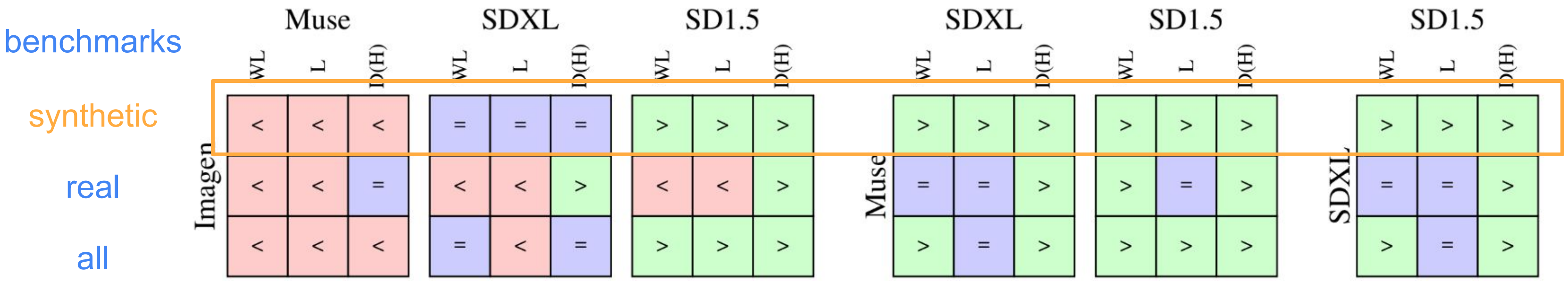
Agreement above chance levels for most generative models, with $\alpha > 0.5$.

image generation models	Word-Level	DSG(H)	Likert
Imagen	**0.81**	0.68	0.64
Muse	**0.82**	0.72	0.78
SDXL	0.75	0.57	**0.76**
SD1.5	**0.66**	0.66	0.36

Annotators agree more when fine-grained templates are used.

Evaluating Human Templates: Model Comparisons

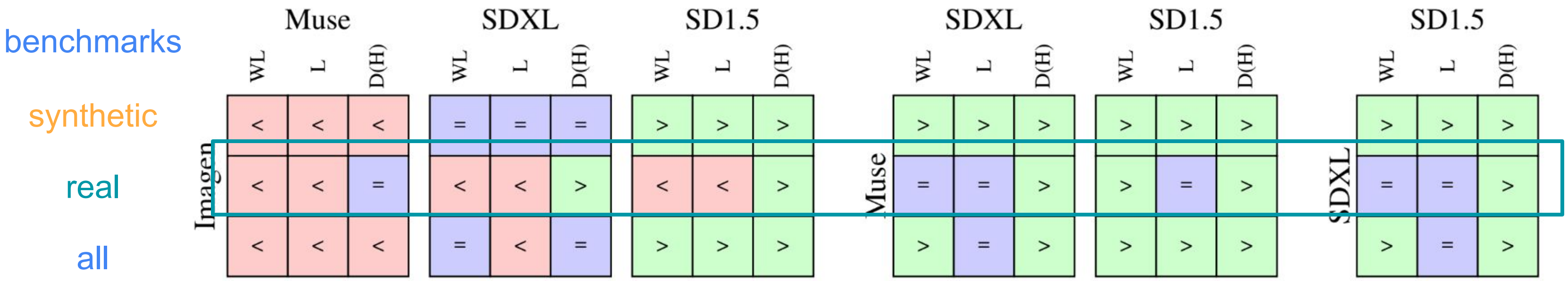
Test the statistical significance of differences in the scores for model pairs.



All templates agree on synthetic prompts.

Evaluating Human Templates: Model Comparisons

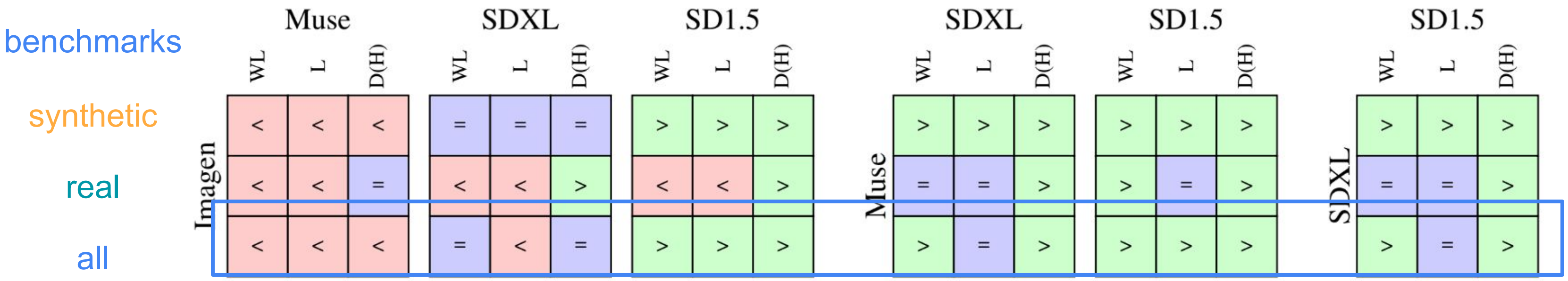
Test the statistical significance of differences in the scores for model pairs.



On real prompts, fine-grained templates (word-level and Likert) agree more.

Evaluating Human Templates: Model Comparisons

Test the statistical significance of differences in the scores for model pairs.



Looking at the full dataset, fine-grained templates agree but may disagree with Likert.

Evaluating Multimodal Generative Models

(3) Develop metrics

Metrics

Human Evaluation

(2) Collect human data

Model 1 | Model 2 | Model 3

Generate images for a set of models

(1) Choose a prompt set

“cat with a pointer standing in front of a blackboard”

Can We Reliably Replace Human Data?

generated image

VQA verification

Q1: Is there a dog?
A: yes ✔

Q2: Is there a cat?
A: yes ✔

Q2: Is the dog to the right of the cat?
A: No ✘

Text prompt
A dog is to the right of the cat.

QA generation
(prompt LLM)

Q1: Is there a dog?
C1: [**yes**, no]

Q2: Is there a cat?
C1: [**yes**, no]

Q3: Is the dog to the right of the cat?
C2: [**yes**, no]

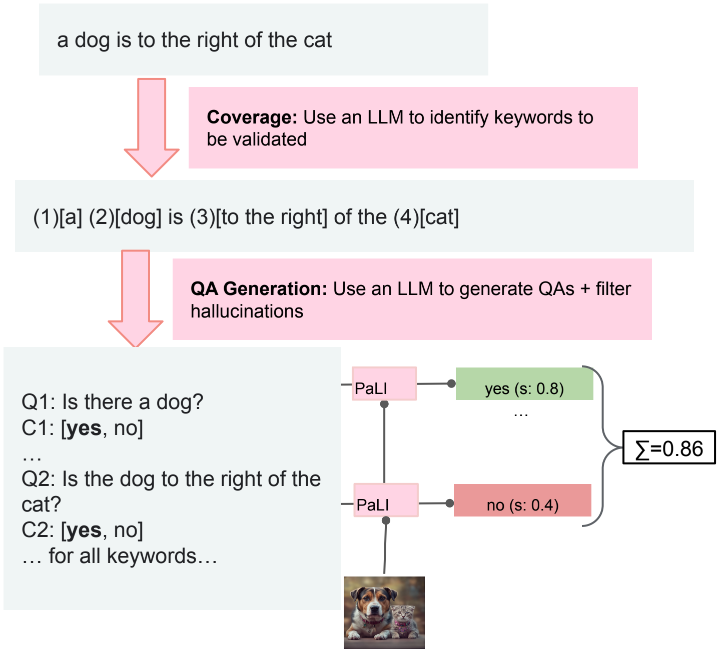
...

Use generative models as a proxy for humans

Gecko: An Automatic-Evaluation Metric for Alignment

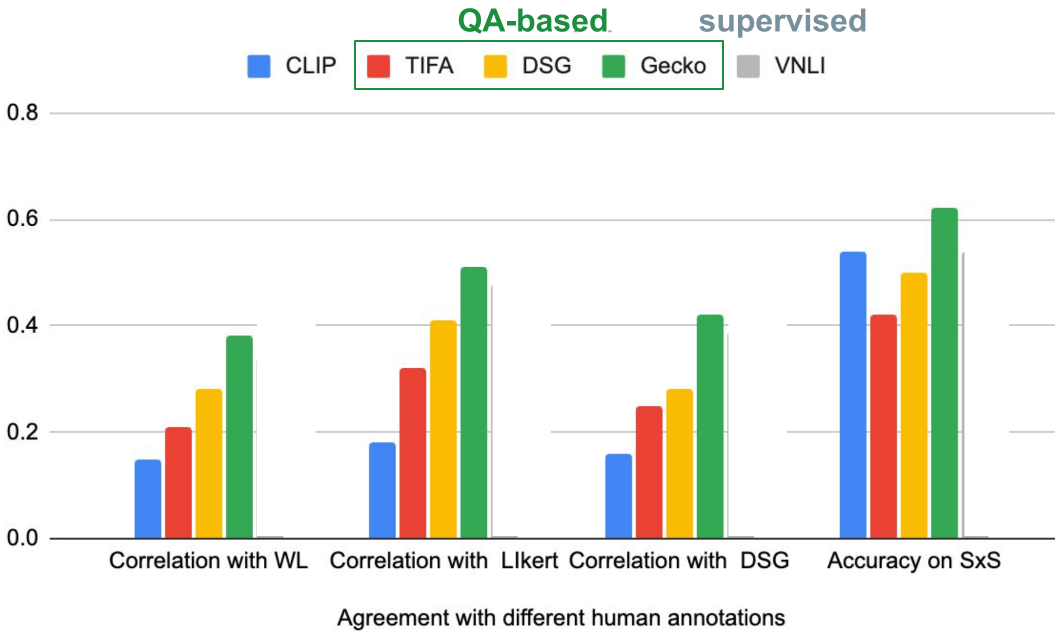
Replace human ratings with the score obtained by our metric.

Need to validate the metric to see how well it matches the human data.



Overview of Metric

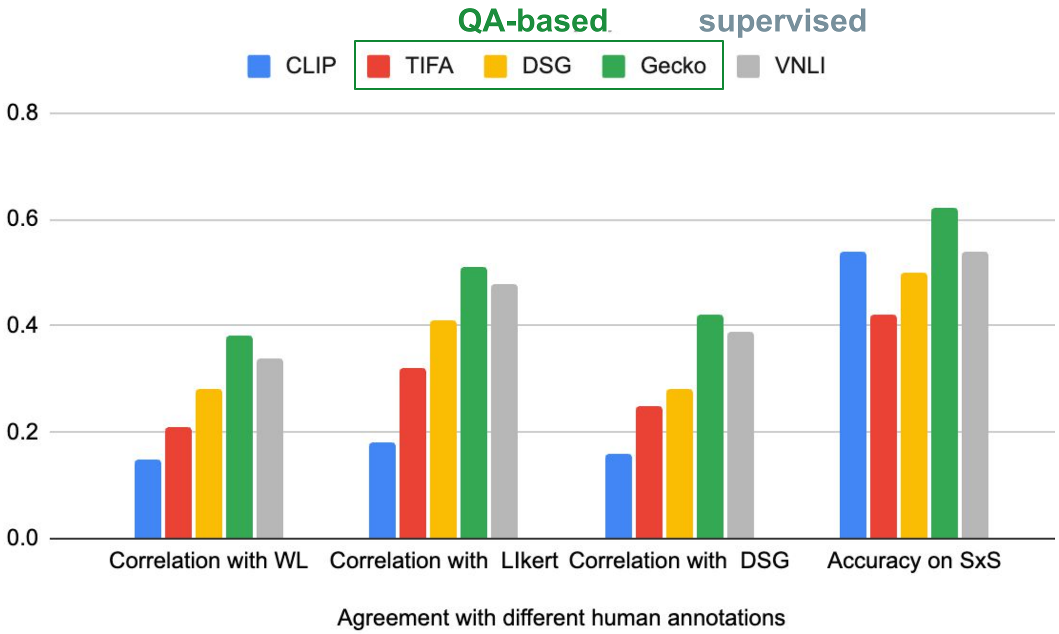
Automatic Evaluation Metrics Compared to Human Data



Gecko-R benchmark

QA-based approaches outperform CLIP→ fine-grained probing improves the result.

Automatic Evaluation Metrics Compared to Human Data

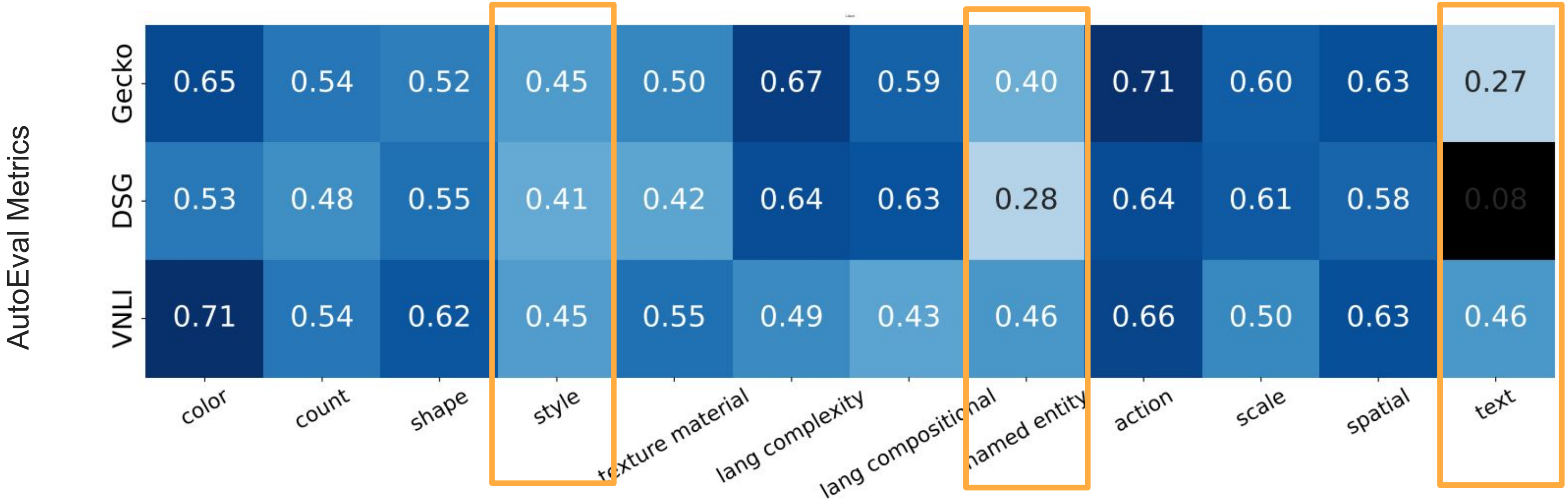


Gecko-R benchmark

QA-based approaches outperform CLIP→ fine-grained probing improves the result.

Gecko performs better than existing QA-based approaches and a supervised model.

What Categories are Challenging for the Metrics?



Measuring text, style, & named entity is hard for QA-based metrics→ Generative models fail answering these questions.

Lessons from Cognitive Science

Collecting human data.

- Finer-grained templates result in higher quality data (in terms of inter-annotator agreement) and more consistent model ordering.
- Automatic evaluation can replace humans if reliable models exist.

Controlled study of a specific phenomenon.

Lessons from Cognitive Science

Collecting human data.

- Finer-grained templates result in higher quality data (in terms of inter-annotator agreement) and more consistent model ordering.
- Automatic evaluation can replace humans if reliable models exist.

Controlled study of a specific phenomenon.

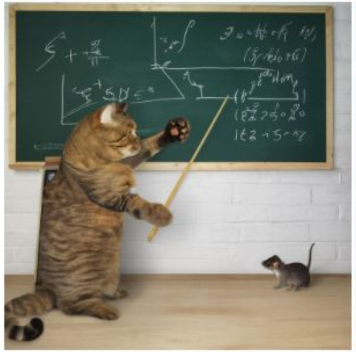
Evaluating Multimodal Generative Models

(3) Develop metrics

Metrics

Human Evaluation

(2) Collect human data



Model 1

Model 2

Model 3

Generate images for a set of models

(1) Choose a prompt set

"cat with a pointer standing in front of a blackboard"

Evaluating Multimodal Generative Models

(3) Develop metrics

Metrics

Human Evaluation

(2) Collect human data

numerical reasoning

Model 1 Model 2 Model 3

Generate images for a set of models

(1) Choose a prompt set

"cat with a pointer standing in front of a blackboard"

Probing for Numerical Reasoning [Kajić et al, 2024]

Task 1: Exact Quantities

Generate images containing an **exact** quantity

Task 2: Approximate Quantities

Interpret **approximate** quantities expressed linguistically

Task 3: Complex Reasoning

Understand more **complex** numerical concepts

How to Evaluate Numerical Reasoning?

1. Design a set of text prompts for each of the 3 tasks
   - Task 1: Exact Number Generation
   - Task 2: Approximate Number Generation
   - Task 3: Complex Reasoning
2. Generate images using 7* different text-to-image models
3. Annotate images with counts/descriptions of objects
4. Use annotations to evaluate model accuracy

Creating a Controlled Prompt Set

Task 1

Simple Numeric

- 3 cats.
- Two koalas.
- 7 cinnamon sticks.
- 1 okra.
- 6 paper clips.
- Ten flutes.

Sentence Numeric

- **An image showing** mushrooms.
- **There are** 5 mushrooms.
- **There are** 5 mushro **in this image.**

are **no** flowers in the vase.

Spatial Relationships

- There are four pistachios **to the right** of 4 flies.
- There are 2 mushrooms **above** 3 tables.
- There are **two** dogs **below 1** tree.

3

Part-whole

- There are **2** forks on the table, but one fork is broken into **two pieces**.
- There are **4** plates on the table, but one plate is broken into **two pieces**.

1386 Prompts

	Prompt Type	# of Prompts	Numbers
Task 1	numeric-simple	600	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
	attribute-color	160	1, 2, 3, 4
	numeric-sentence	100	1, 2, 3, 4, 5
	2-additive	100	1, 2, 3, 4, 5
	2-additive-color	100	1, 2, 3, 4, 5, 6, 7, 8
	3-additive	100	1, 2, 3, 4, 5
	attribute-spatial	100	1, 2, 3, 4, 5
Task 2	approx-1-entity	24	no, few, many
	approx-2-entity	45	fewer, as many as, more
Task 3	fractional-simple	36	1, 2, 3, 1/2, 1/3, 1/4, 1/5
	part-whole	15	1/2
	fractional-complex	6	1/3 + 2/3, 1/2

Results of Model Evaluation

	Task 1 Exact Number Generation	**Task 2** Approximate Number Generation and Zero	**Task 3** Conceptual Quantitative Reasoning
DALL·E 3	**45.2** $\pm$ 0.5 (+35.2%)	**48.7** $\pm$ 2.7 (+24.1%)	<u>48.8</u> $\pm$ 1.1 (−1.2%)
Imagen-A	26.3 $\pm$ 0.4 (+16.3%)	20.0 $\pm$ 2.2 (−4.6%)	41.1 $\pm$ 1.3 (−8.9%)
Imagen-B	27.0 $\pm$ 0.4 (+17.0%)	24.6 $\pm$ 2.3 (+0.0%)	42.9 $\pm$ 1.4 (−7.1%)
Imagen-C	<u>34.8</u> $\pm$ 0.4 (+24.9%)	27.0 $\pm$ 2.4 (+2.4%)	**50.6** $\pm$ 1.2 (+0.6%)
Imagen-D	28.5 $\pm$ 0.4 (+18.5%)	<u>28.7</u> $\pm$ 2.4 (+4.0%)	43.8 $\pm$ 1.3 (−6.2%)
Muse-A	34.8 $\pm$ 0.4 (+24.8%)	21.0 $\pm$ 2.2 (−3.6%)	45.1 $\pm$ 1.2 (−4.9%)
Muse-B	<u>39.8</u> $\pm$ 0.5 (+29.8%)	<u>24.6</u> $\pm$ 2.3 (+0.0%)	<u>46.2</u> $\pm$ 1.2 (−3.8%)
Random	10.0	24.6	50.0

Results of Model Evaluation

	Task 1 Exact Number Generation	Task 2 Approximate Number Generation and Zero	Task 3 Conceptual Quantitative Reasoning
DALL·E 3	**45.2** ± 0.5 (+35.2%)	**48.7** ± 2.7 (+24.1%)	48.8 ± 1.1 (−1.2%)
Imagen-A	26.3 ± 0.4 (+16.3%)	20.0 ± 2.2 (−4.6%)	41.1 ± 1.3 (−8.9%)
Imagen-B	27.0 ± 0.4 (+17.0%)	24.6 ± 2.3 (+0.0%)	42.9 ± 1.4 (−7.1%)
Imagen-C	34.8 ± 0.4 (+24.9%)	27.0 ± 2.4 (+2.4%)	**50.6** ± 1.2 (+0.6%)
Imagen-D	28.5 ± 0.4 (+18.5%)	28.7 ± 2.4 (+4.0%)	43.8 ± 1.3 (−6.2%)
Muse-A	34.8 ± 0.4 (+24.8%)	21.0 ± 2.2 (−3.6%)	45.1 ± 1.2 (−4.9%)
Muse-B	39.8 ± 0.5 (+29.8%)	24.6 ± 2.3 (+0.0%)	46.2 ± 1.2 (−3.8%)
Random	10.0	24.6	50.0

DALL.E 3 is the best performing model but there is a notable gap to best achievable performance.

Results of Model Evaluation

	Task 1 Exact Number Generation	**Task 2** Approximate Number Generation and Zero	**Task 3** Conceptual Quantitative Reasoning
DALL·E 3	**45.2** ± 0.5 (+35.2%)	**48.7** ± 2.7 (+24.1%)	48.8 ± 1.1 (−1.2%)
Imagen-A	26.3 ± 0.4 (+16.3%)	20.0 ± 2.2 (−4.6%)	41.1 ± 1.3 (−8.9%)
Imagen-B	27.0 ± 0.4 (+17.0%)	24.6 ± 2.3 (+0.0%)	42.9 ± 1.4 (−7.1%)
Imagen-C	34.8 ± 0.4 (+24.9%)	27.0 ± 2.4 (+2.4%)	**50.6** ± 1.2 (+0.6%)
Imagen-D	28.5 ± 0.4 (+18.5%)	28.7 ± 2.4 (+4.0%)	43.8 ± 1.3 (−6.2%)
Muse-A	34.8 ± 0.4 (+24.8%)	21.0 ± 2.2 (−3.6%)	45.1 ± 1.2 (−4.9%)
Muse-B	39.8 ± 0.5 (+29.8%)	24.6 ± 2.3 (+0.0%)	46.2 ± 1.2 (−3.8%)
Random	10.0	24.6	50.0

Task 3 is the hardest--all models perform close to chance. Task 2 is harder than task 1.

Results of Model Evaluation [Imagen3]

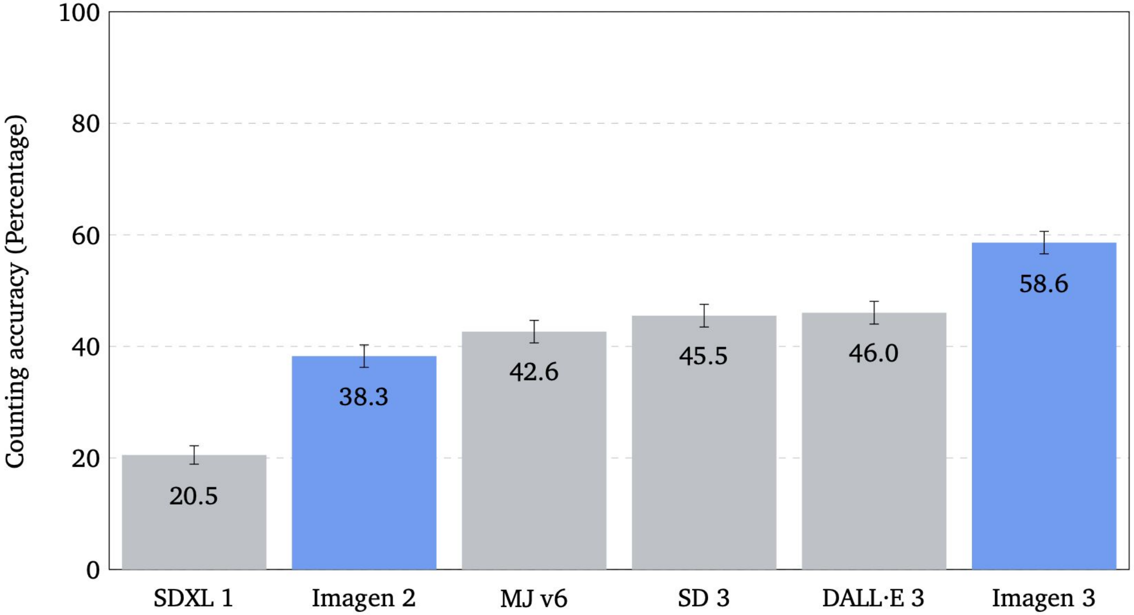
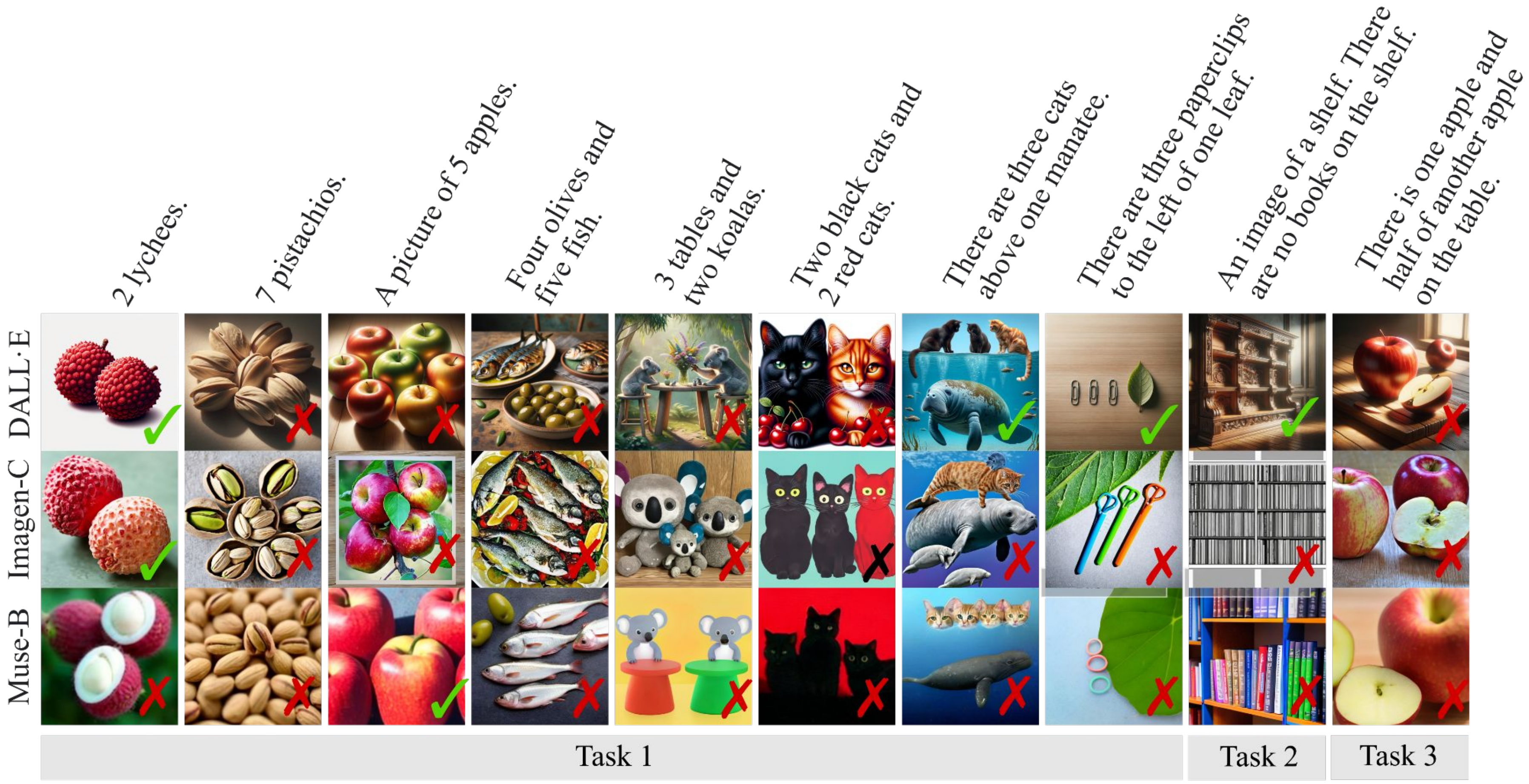
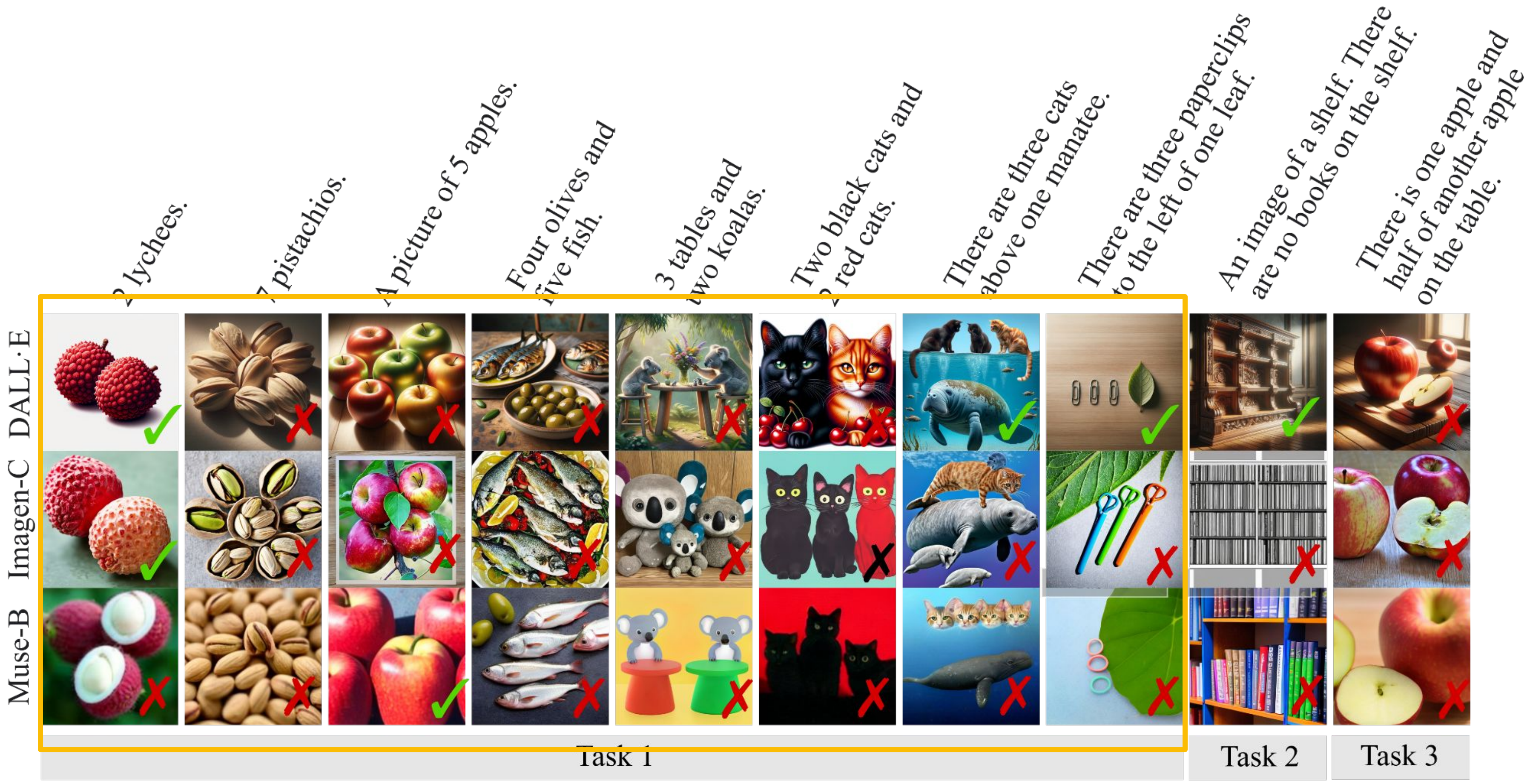
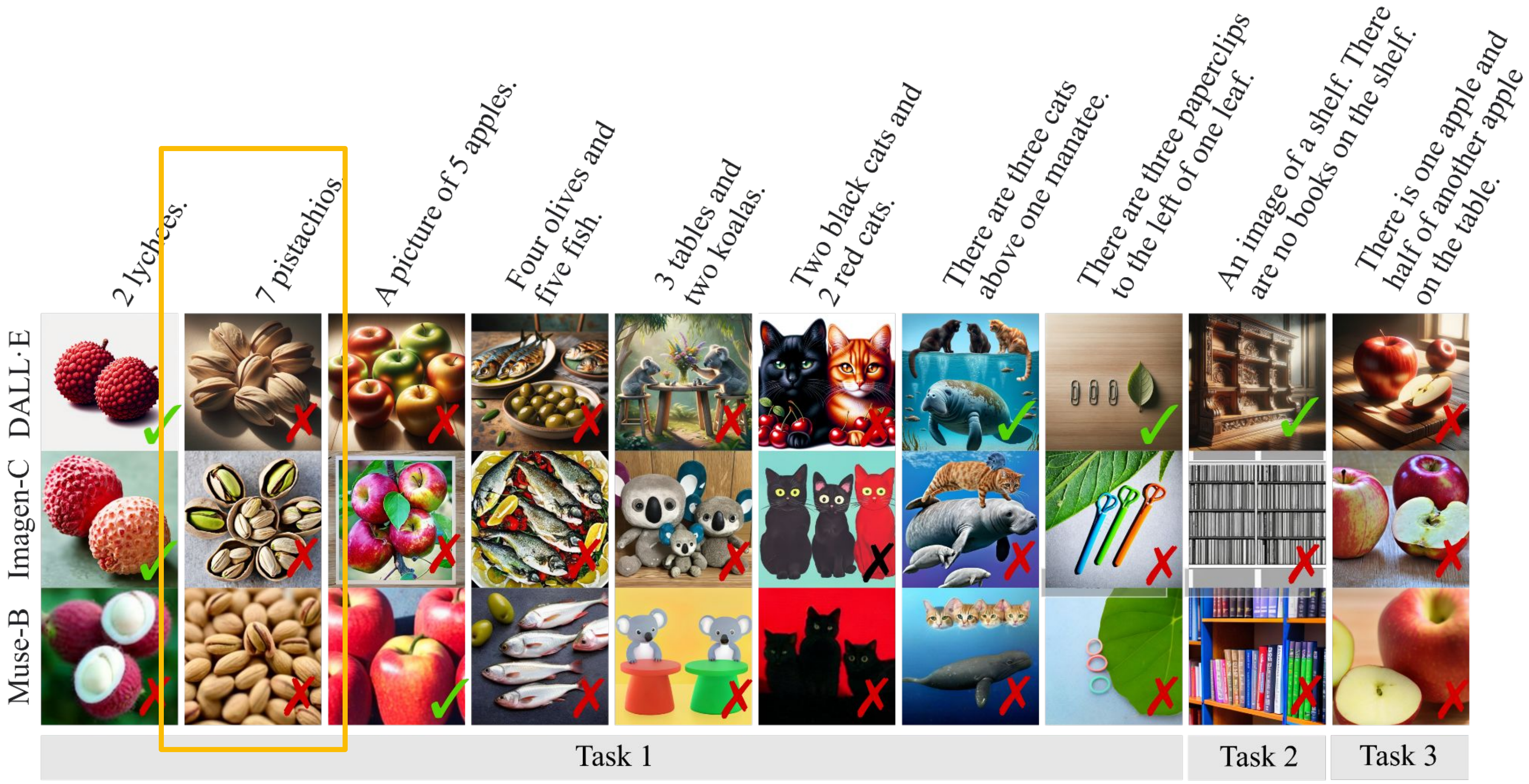


Imagen 3 (a more recent model) improves on Task 1 but there is still a notable gap to best achievable performance.

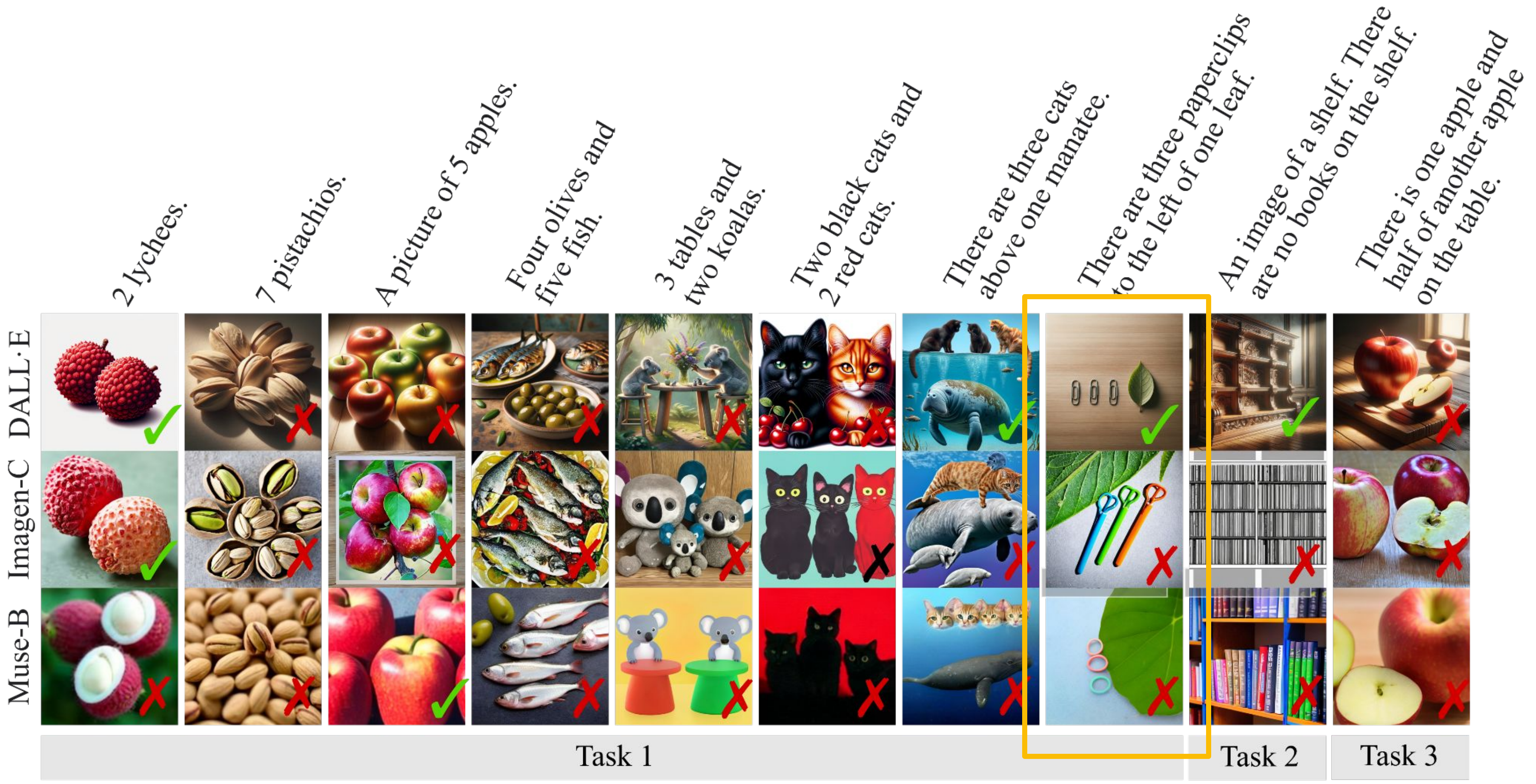

2 lychees.
7 pistachios.
A picture of 5 apples.
Four olives and five fish.
3 tables and two koalas.
Two black cats and 2 red cats.
There are three cats above one manatee.
There are three paperclips to the left of one leaf.
An image of a shelf. There are no books on the shelf.
There is one apple and half of another apple on the table.
DALL·E
Imagen-C
Muse-B
Task 1
Task 2
Task 3



lychees.
pistachios.
A picture of 5 apples.
Four olives and
ive fish.
3 tables and
wo koalas.
Two black cats and
red cats.
There are three cats
above one manatee.
There are three paperclips
to the left of one leaf.
An image of a shelf. There
are no books on the shelf.
There is one apple and
half of another apple
on the table.
DALL·E
Imagen-C
Muse-B
Task 1
Task 2
Task 3



2 lychees.
7 pistachios.
A picture of 5 apples.
Four olives and five fish.
3 tables and two koalas.
Two black cats and 2 red cats.
There are three cats above one manatee.
There are three paperclips to the left of one leaf.
An image of a shelf. There are no books on the shelf.
There is one apple and half of another apple on the table.
DALL·E
Imagen-C
Muse-B
Task 1
Task 2
Task 3


2 lychees.
7 pistachios.
A picture of 5 apples.
Four olives and five fish.
3 tables and two koalas.
Two black cats and 2 red cats.
There are three cats above one manatee.
There are three paperclips to the left of one leaf.
An image of a shelf. There are no books on the shelf.
There is one apple and half of another apple on the table.
DALL·E
Imagen-C
Muse-B
Task 1
Task 2
Task 3

Lessons from Cognitive Science

Collecting human data.

- Finer-grained templates result in higher quality data (in terms of inter-annotator agreement) and more consistent model ordering.
- Automatic evaluation can replace humans if reliable models exist.

Controlled study of a specific phenomenon.

- Reasoning about numbers, in particular, about approximate quantities and parts is challenging for image generation models.

Probing Representations for Verbs

Concrete nouns are **consistent** and **easily observable.**



classification

Verbs are less so, as they capture **relations**.



structured prediction

Zero-Shot Image Retrieval

Zero-shot image retrieval directly evaluates the goodness of **pretrained** representations.

Image Retrieval (IR)

"Grey haired man in black and yellow tie."

What Image Retrieval Tests

Order images with respect to their match to a sentence.

A person is riding a horse.

Subject Verb Object

Does not require fine-grained multimodal understanding.

What SVO-Probes Tests [Hendricks *et al.*, Findings of ACL 2021]

A person is **riding** a horse

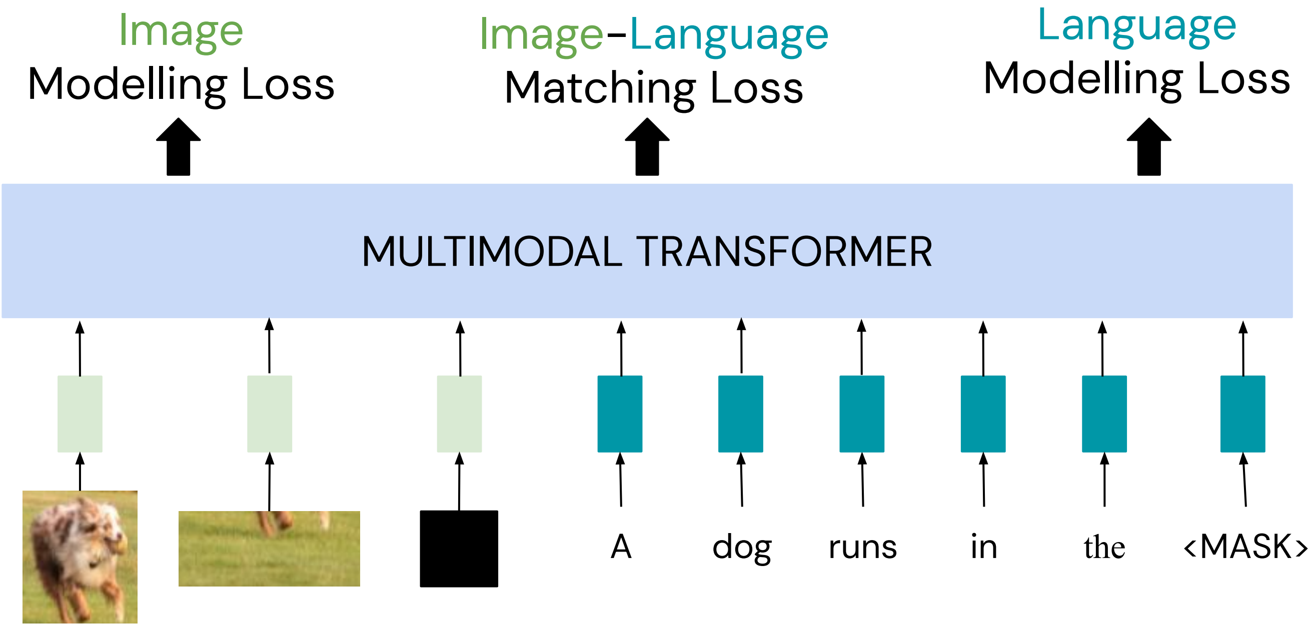
❌ 

✔ 

Correctly classify both the **positive** & **negative** examples.

We have released our dataset! 🎉🎉

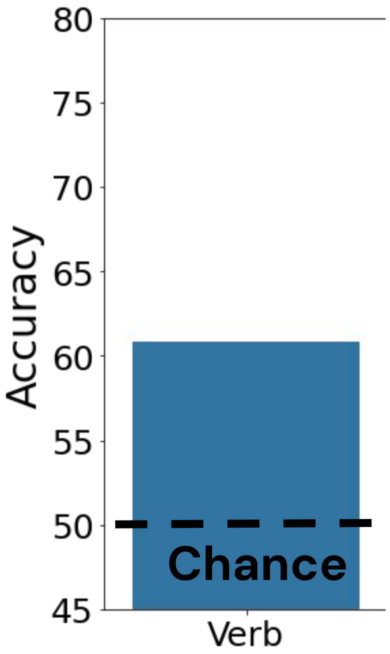
Multimodal Transformers (MMT)

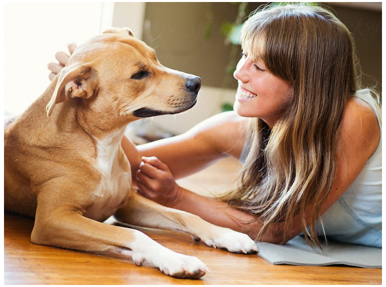


Similar architectures are widely adopted for multimodal pretraining [e.g, ViLBERT, LXMERT, UNITER].

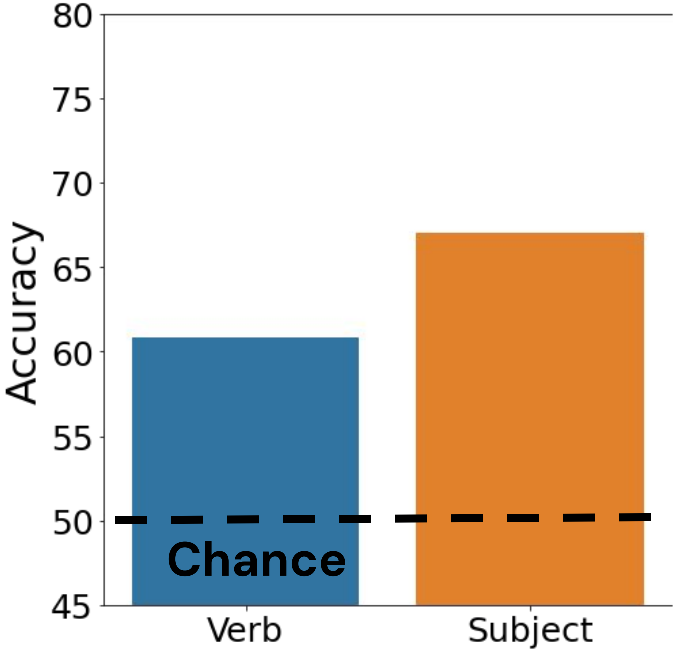
Do MMTs Have Fine-grained Verb Understanding?

A woman **lying** with a dog



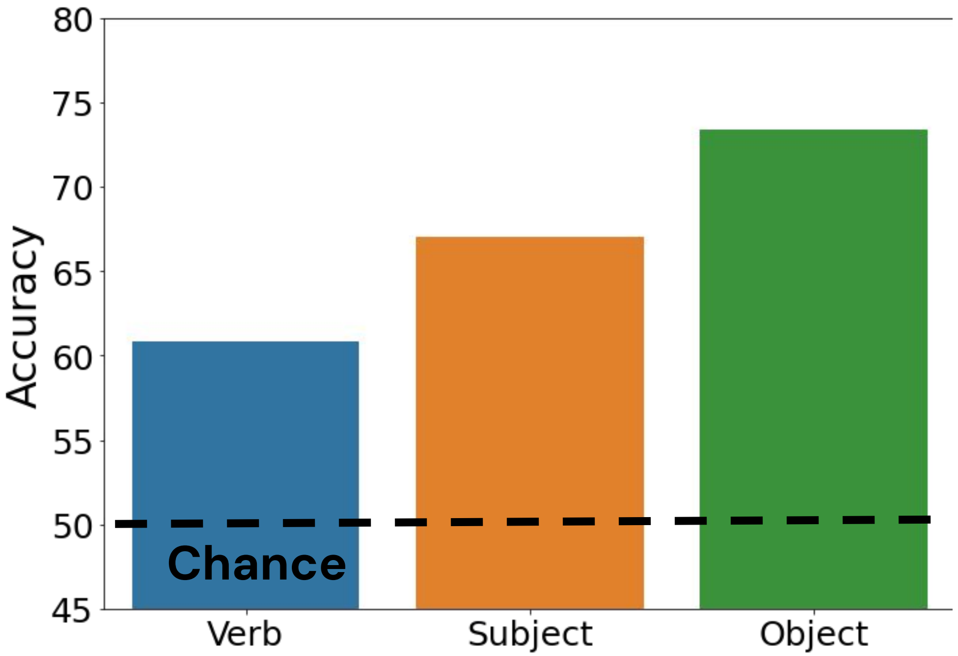


Do MMTs Have Fine-grained Verb Understanding?

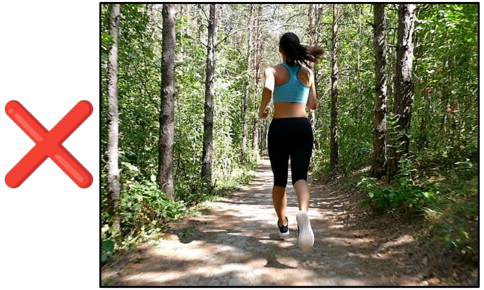


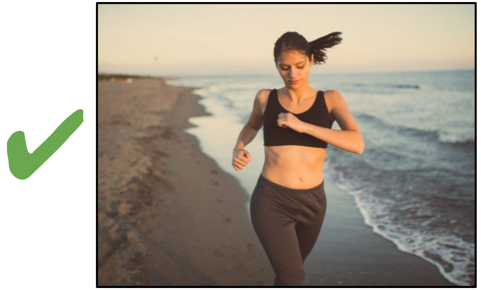
A **animal** lays in the grass

Do MMTs Have Fine-grained Verb Understanding?

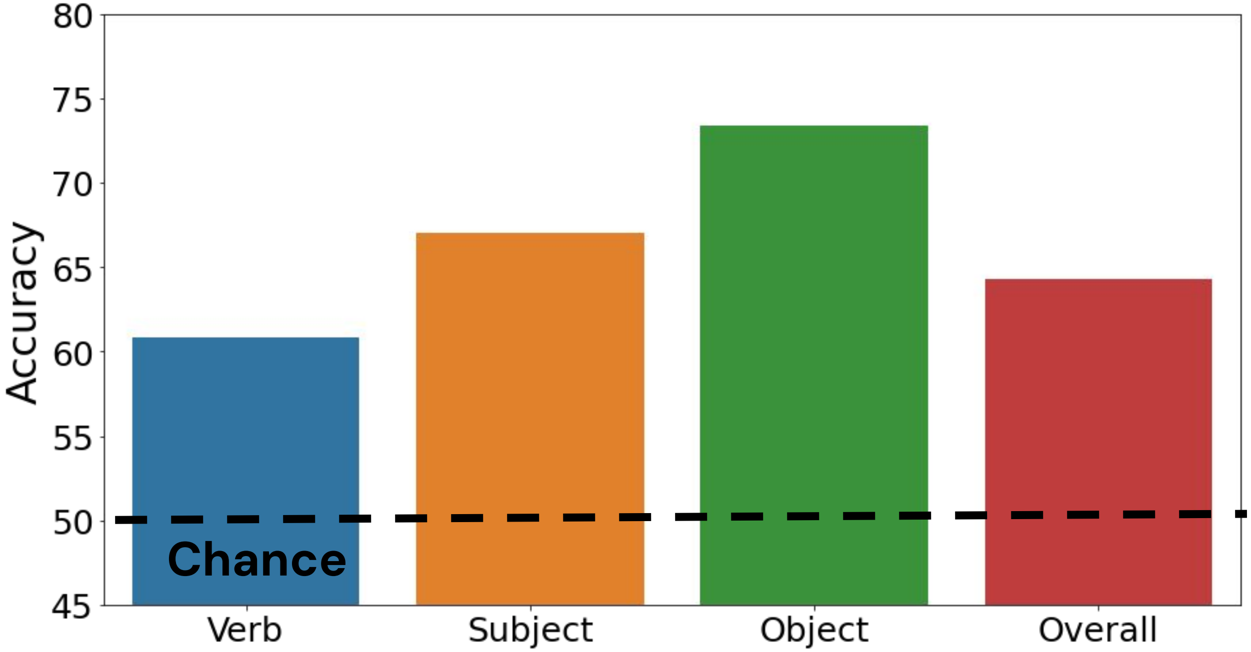


A woman jogs on the **beach**





Do MMTs Have Fine-grained Verb Understanding?



Overall MMT performance 64.3 -- lots of room for improvement!

Does the Training Dataset Impact Performance?

Conceptual Captions

"The scenic route through mountain ranges includes these unbelievably coloured mountains.

Large (3M images) ✔

Noisy (text might **not** describe the image)

Domain **matches** SVO-Probes ✔

MSCOCO

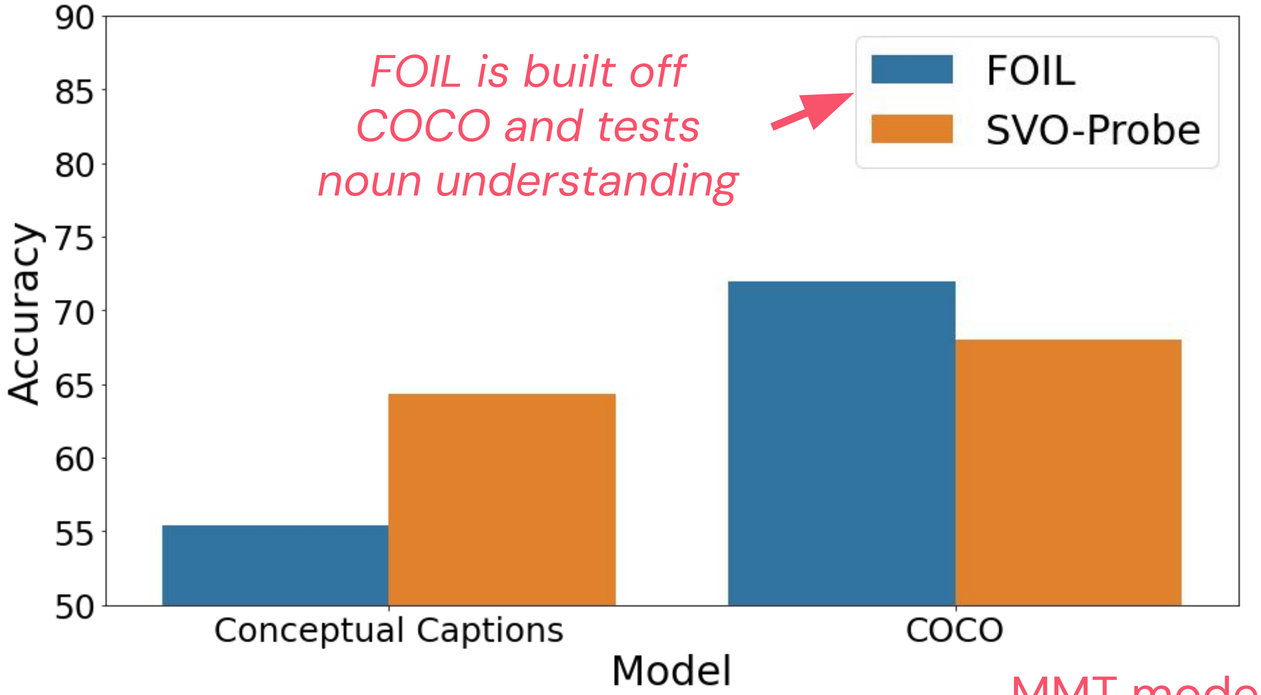
"The two people are walking down the beach."

Small (100K images)

Clean (manually-annotated) ✔

Domain **mismatch** from SVO-Probe

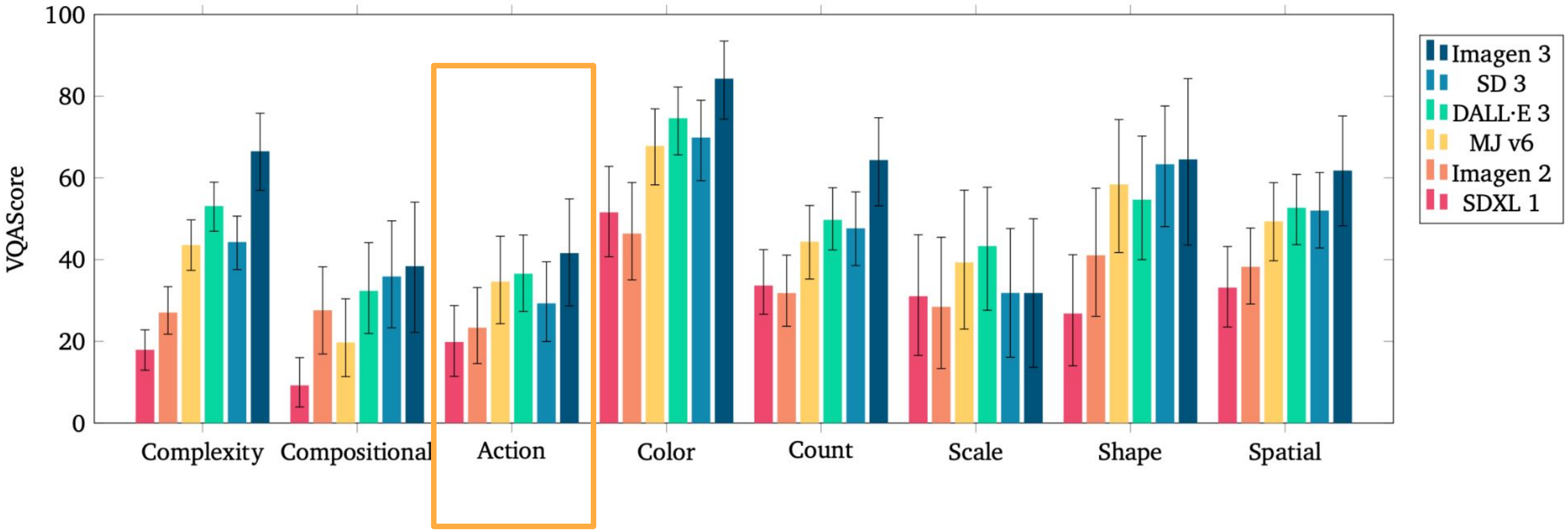
Does the Training Dataset Impact Performance?



Models trained with COCO perform better on probe datasets .

This could be because ***COCO data is less noisy***, meaning images match text better.

MMT models are not robust to noise.



100
80
60
40
20
0
VQAScore
Complexity
Compositional
Action
Color
Count
Scale
Shape
Spatial
Imagen 3
SD 3
DALL·E 3
MJ v6
Imagen 2
SDXL 1

Lessons from Cognitive Science

Collecting human data.

- Finer-grained templates result in higher quality data (in terms of inter-annotator agreement) and more consistent model ordering.
- Automatic evaluation can replace humans if reliable models exist.

Controlled study of a specific phenomenon.

- Reasoning about numbers, in particular, about approximate quantities and parts is challenging for image generation models.
- Reasoning about verbs is challenging for vision-language models.

Final Thoughts

Human data is the gold-standard for evaluating generative models---the evaluation and standardisation of human data templates is important to make reliable conclusions about models.

Given the power of recent generative models, probing for specific capabilities sheds lights on their strengths and identifies their shortcoming; this in turn can guide future modeling work.

Thanks!

Isabela Albuquerque

Lisa Anne Hendricks

Matthias Bauer

Emanuele Bugliarello

Ivana Kajic

Chris Knutsen

Ira Ktena

Yasumasa Onoe

Nelly Papalampidi

Jordi Pont-Tuset

Cyrus Rashtchian

Olivia Wiles

Chuhan Zhang

+
Su Wang