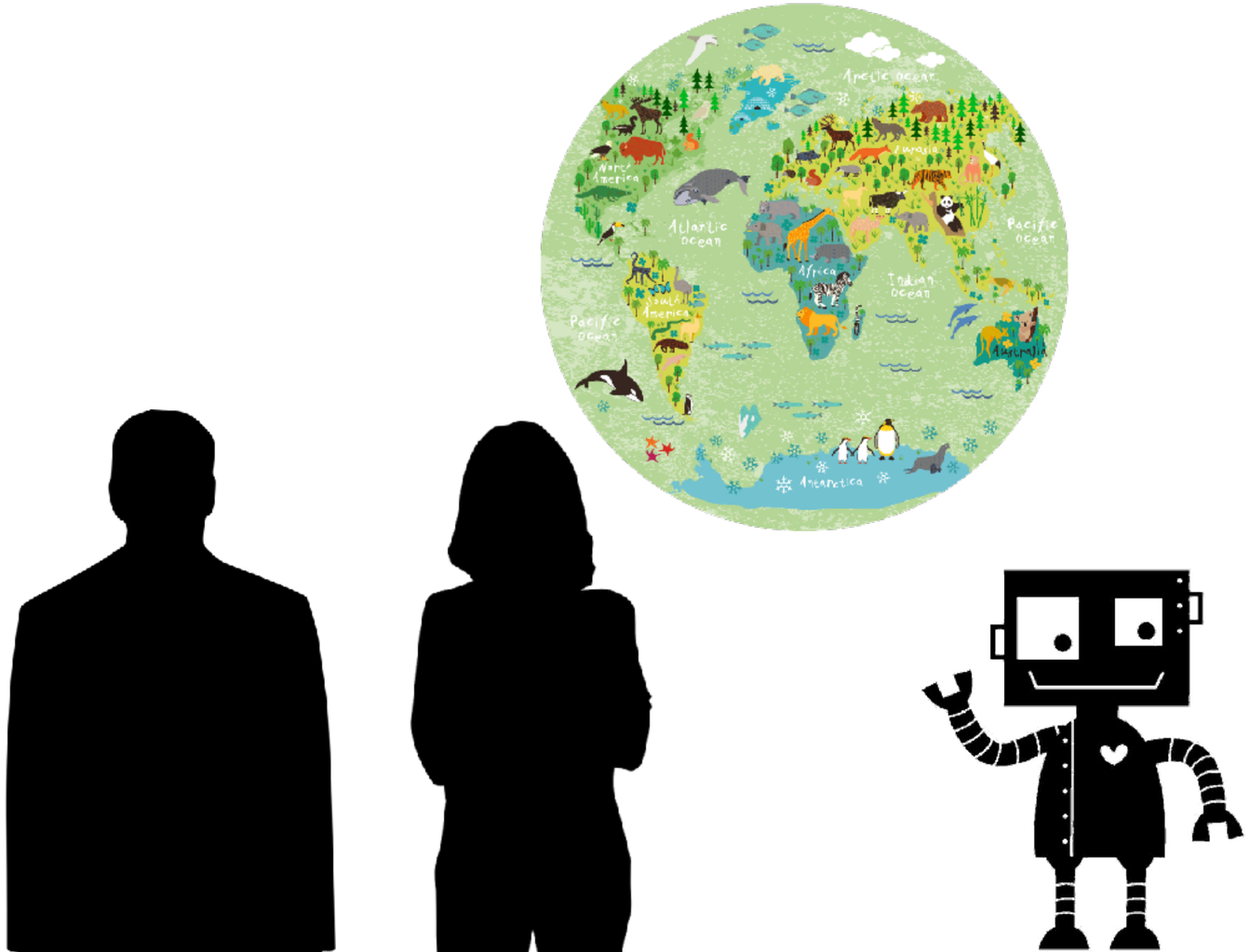
Towards Pragmatic Visual Description Generation

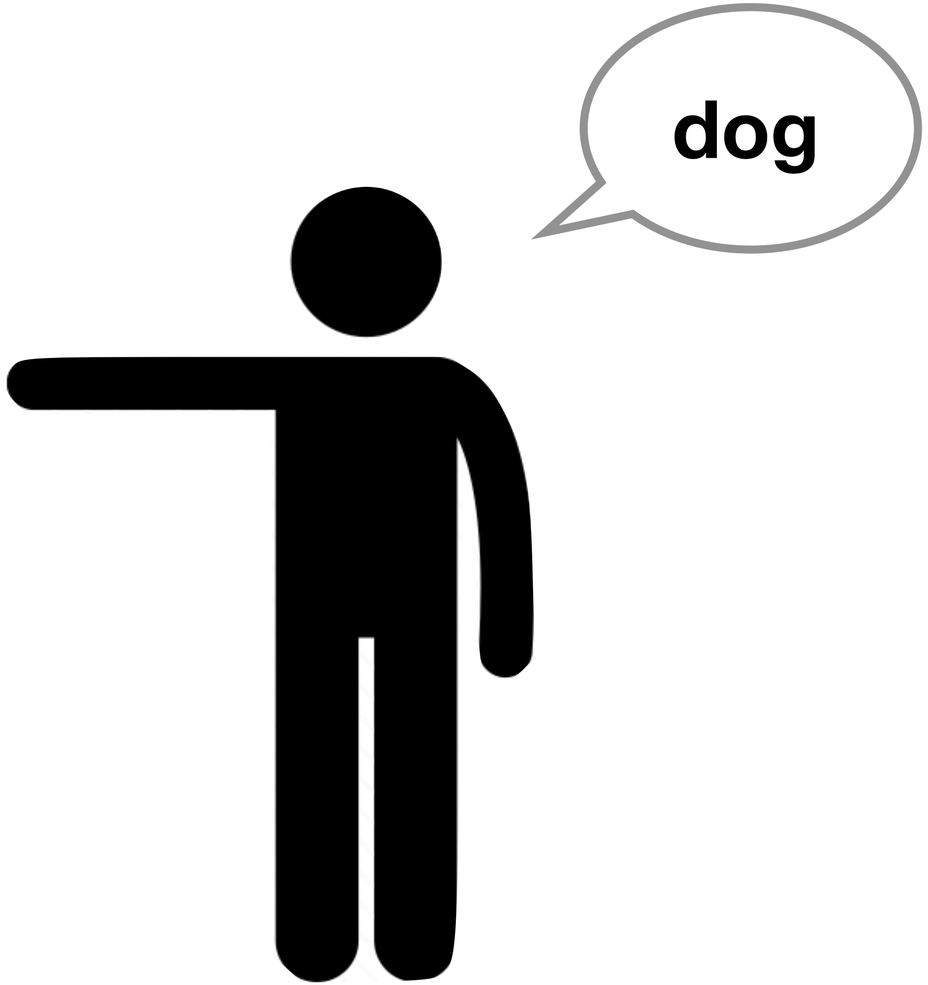
Elisa Kreiss
Department of Communication
UCLA



CLASP Seminar
Apr 2 2025

Communicating about a visual world

Communicating about a visual world

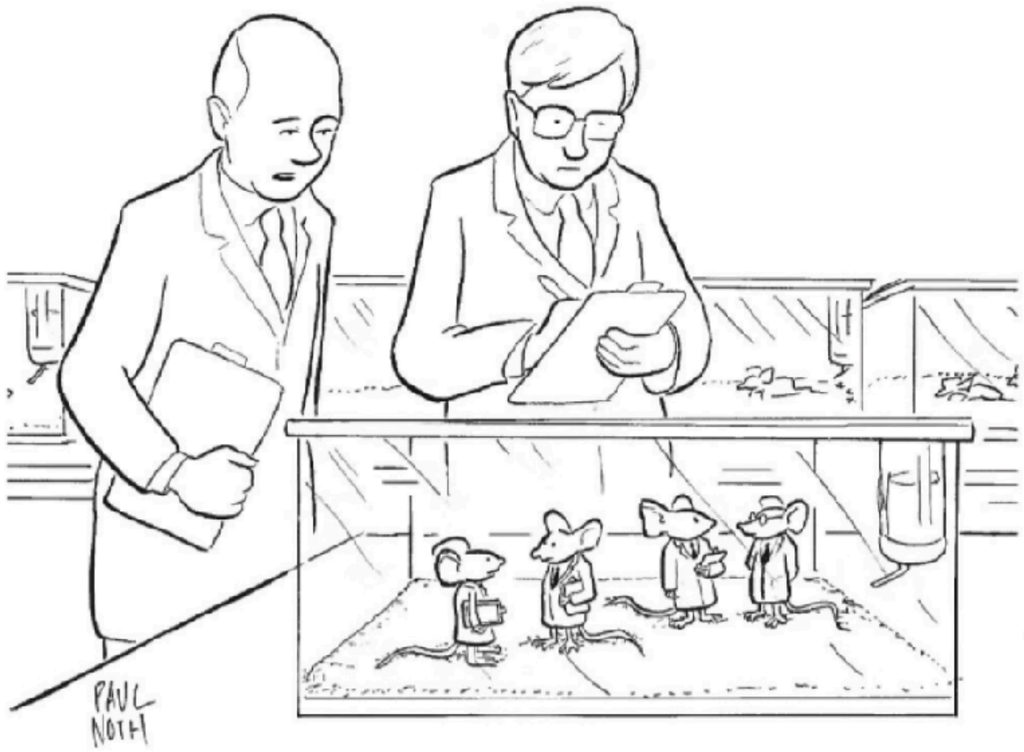


Communicating about a visual world

Communicating about a visual world

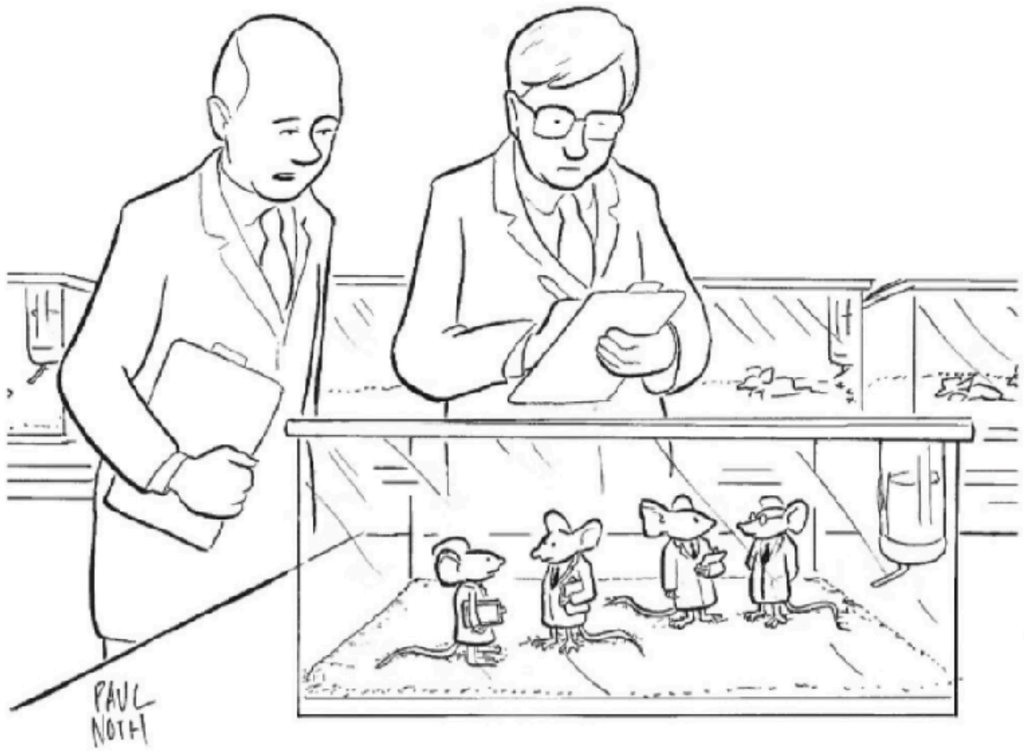
"De Materie" opera show in New York, starring 100 sheep.

Communicating about a visual world



New Yorker Captioning challenge

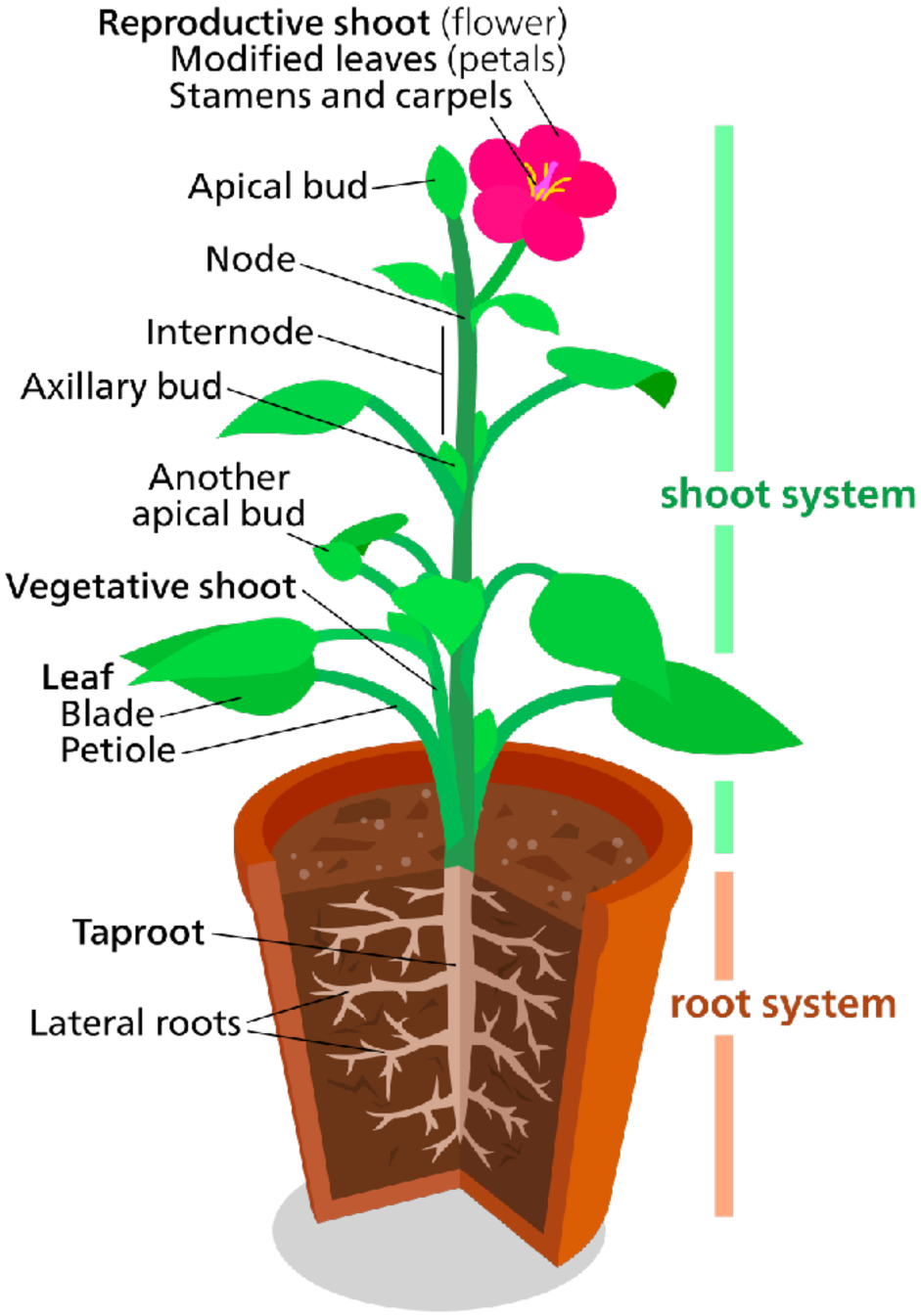
Communicating about a visual world



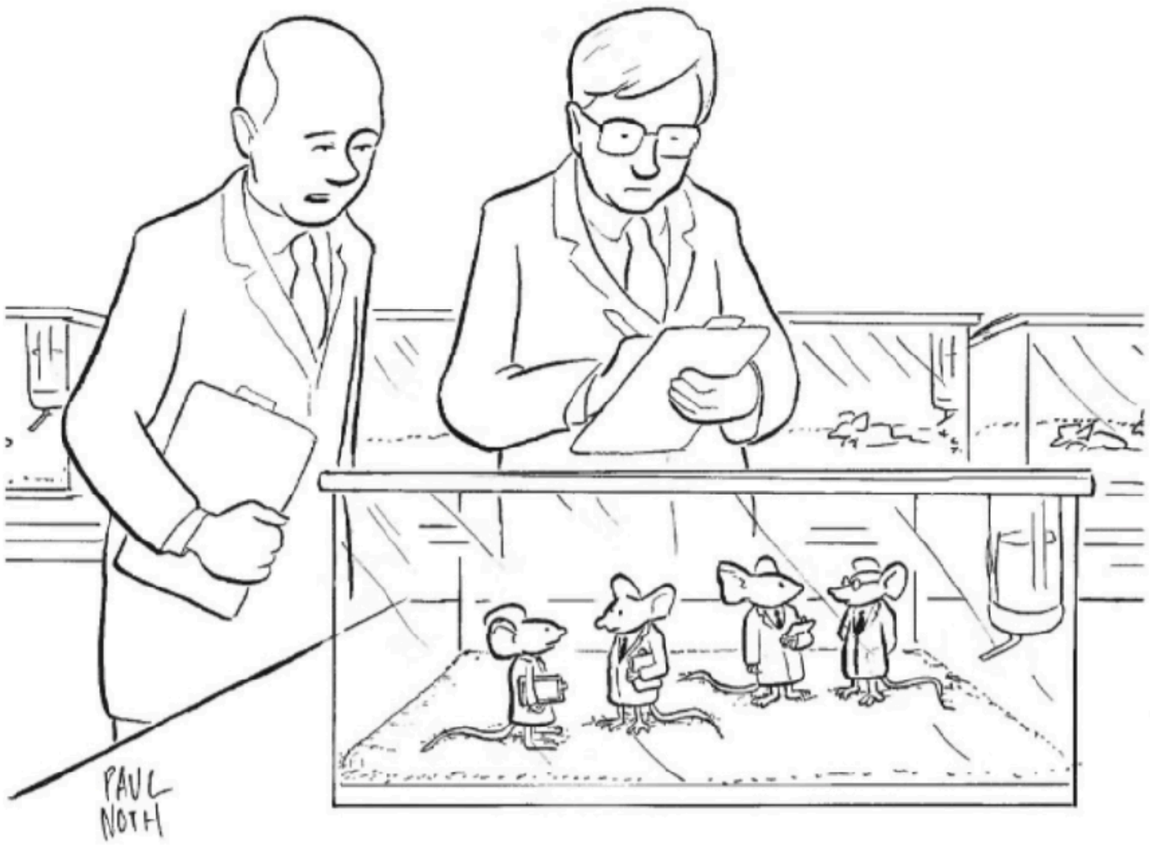
"O.K., let's slowly lower in the grant money."
Todd Bearson, Arlington, Massachusetts.
2009

New Yorker Captioning challenge

Education


Reproductive shoot (flower)
Modified leaves (petals)
Stamens and carpels
Apical bud
Node
Internode
Axillary bud
Another apical bud
Vegetative shoot
Leaf
Blade
Petiole
Taproot
Lateral roots
shoot system
root system

Entertainment



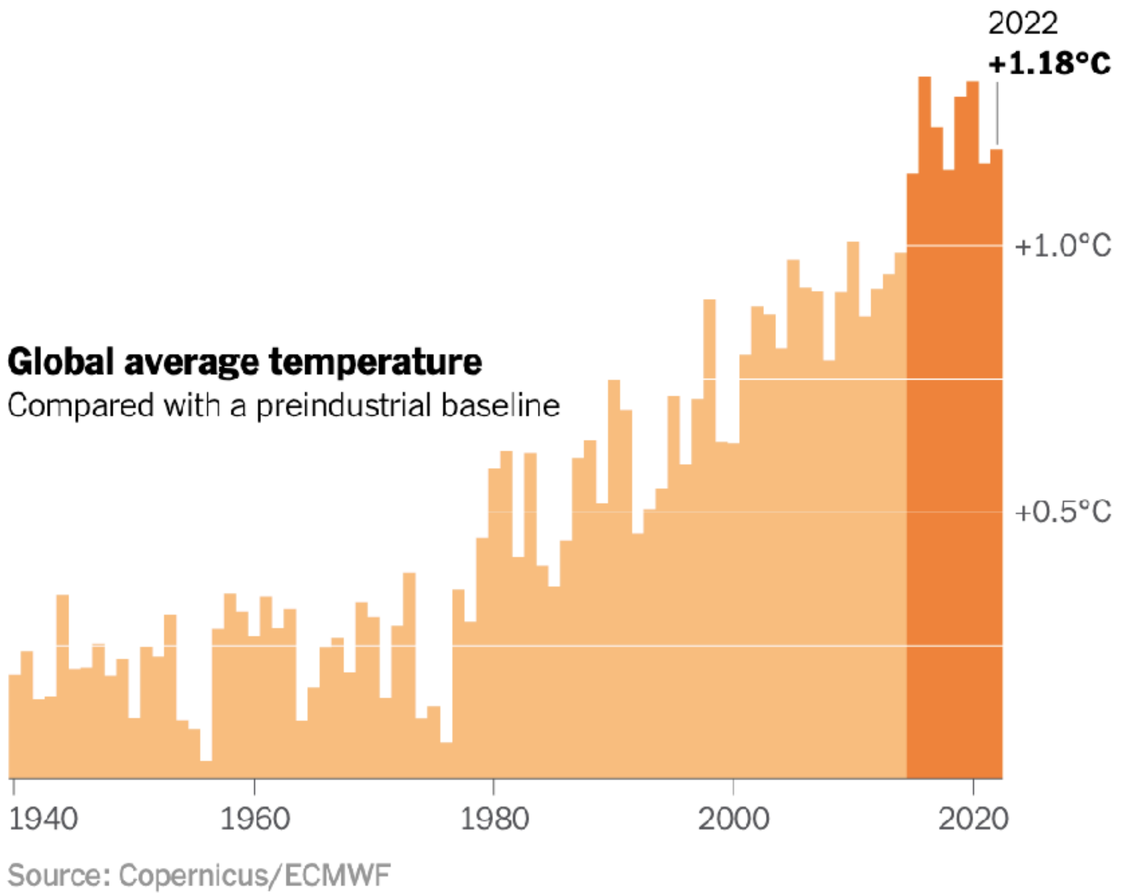
"O.K., let's slowly lower in the grant money."
Todd Bearson, Arlington, Massachusetts.
2009

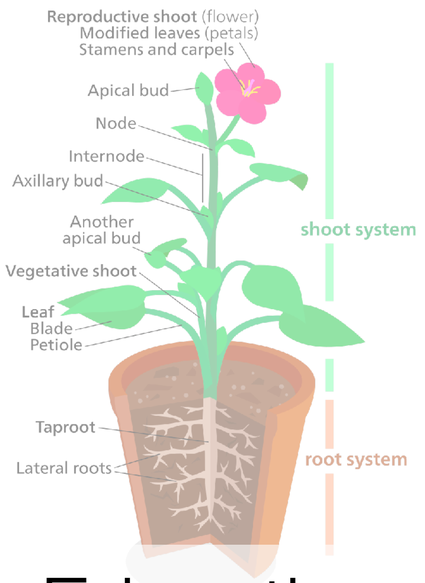
News

The Last 8 Years Were the Hottest on Record

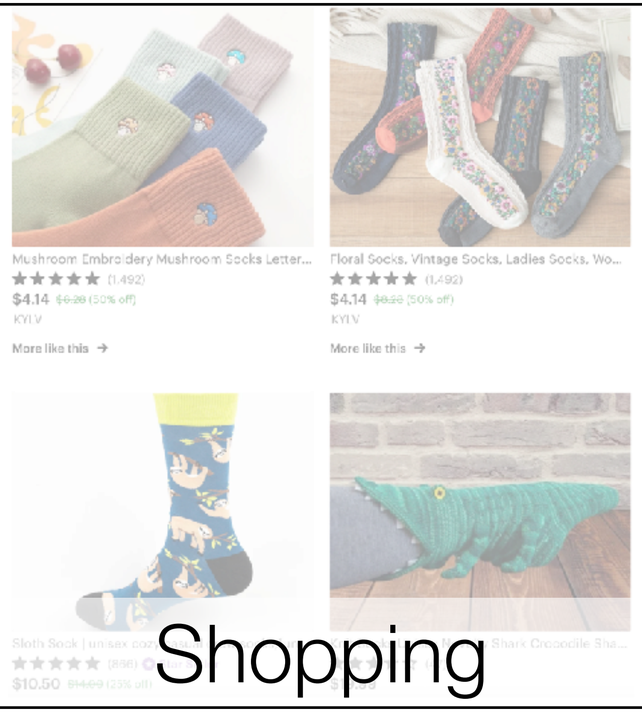
By Henry Fountain and Mira Rojanasakul Jan. 10, 2023

The world remained firmly in warming's grip last year, with extreme summer temperatures in Europe, China and elsewhere contributing to 2022 being the fifth-hottest year on record, European climate researchers said on Tuesday.

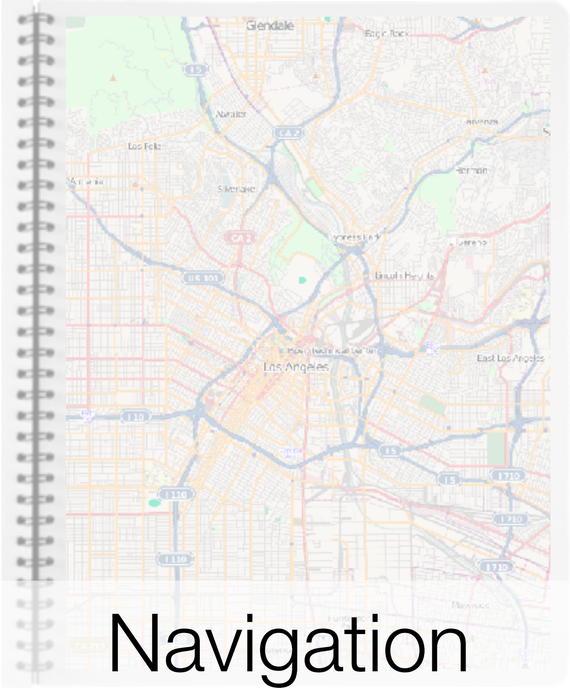




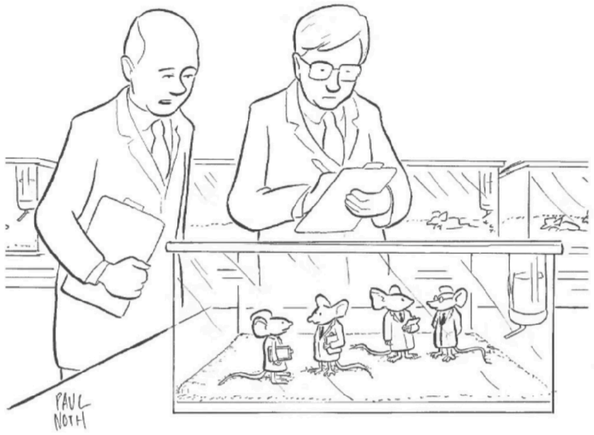
Education



Shopping



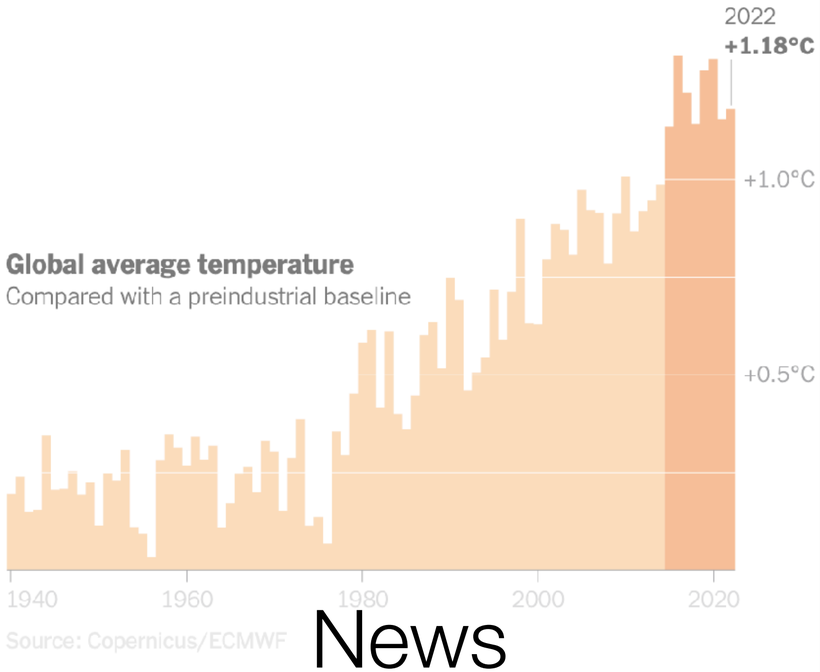
Navigation



"O.K., let's slowly lower in the grant money."
Todd Bearson, Arlington, Massachusetts.
2009

Entertainment

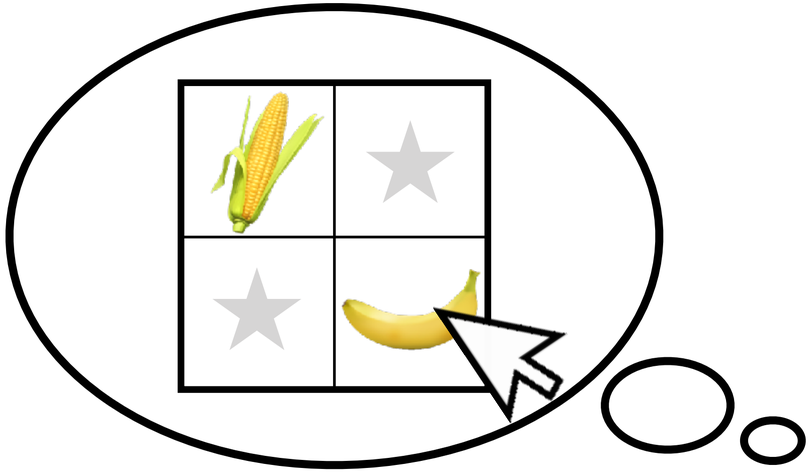
Communication happens in complex visual worlds



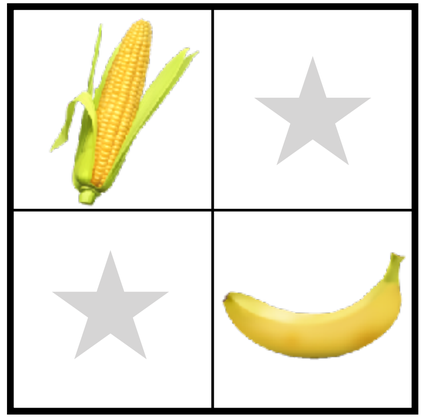
News

Communication happens in complex visual worlds

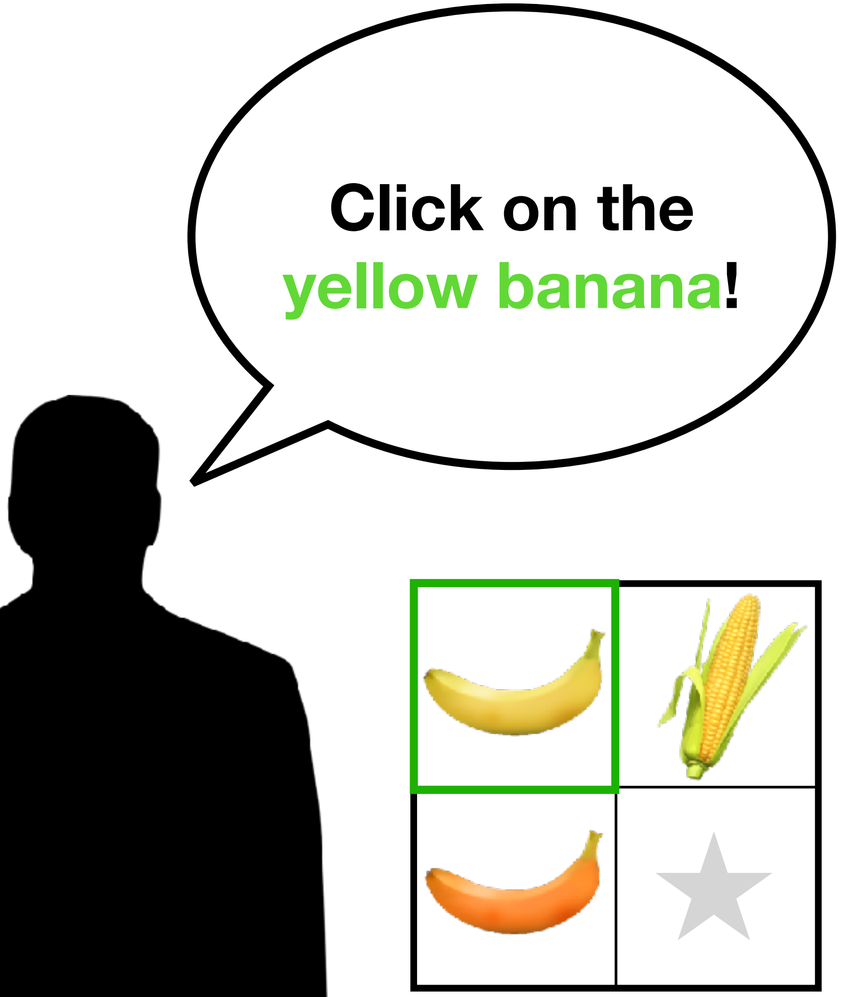
display from Kreiss & Degen 2020, following paradigms from Sedivy et al. 1999, Westerbeek 2015

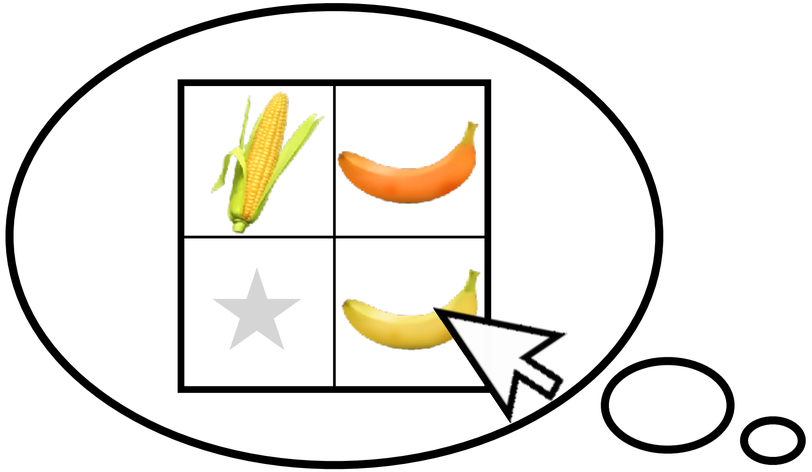


Communication happens in complex visual worlds

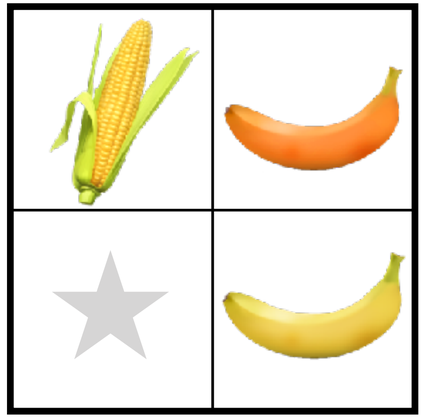


display from Kreiss & Degen 2020, following paradigms from Sedivy et al. 1999, Westerbeek 2015





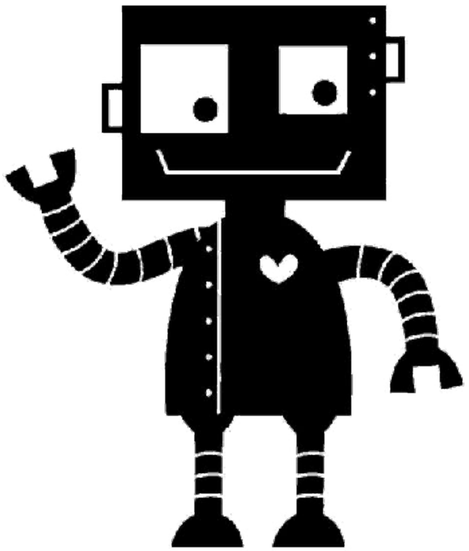
Communication happens in complex visual worlds



Communication happens in complex visual worlds

→ **CAT** or **DOG**

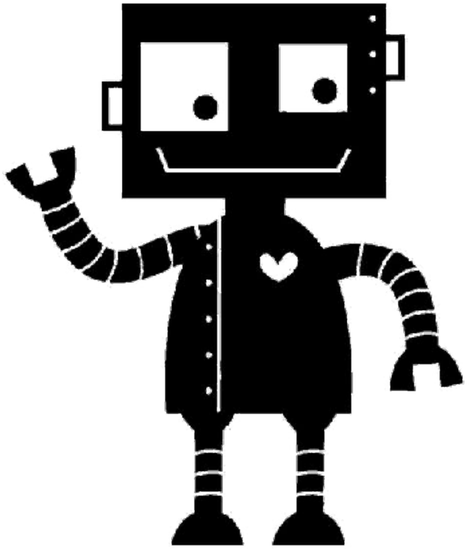
Communication happens in complex visual worlds



facilitated by datasets such as CIFAR-10: Krizhevsky et al. 2012

"a dog"

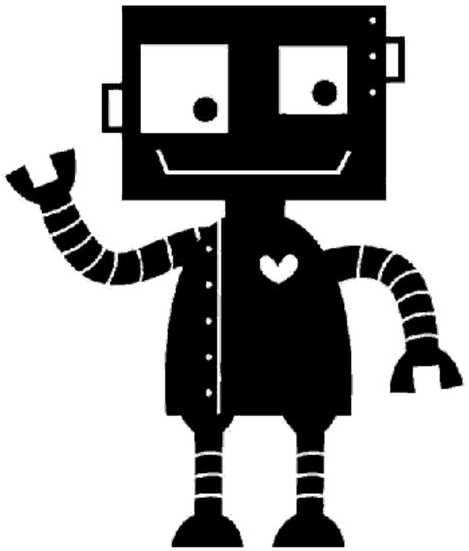
Communication happens in complex visual worlds



inspired by Karpathy & Fei Fei 2015

there is a brown dog that is lying on a couch

Communication happens in complex visual worlds

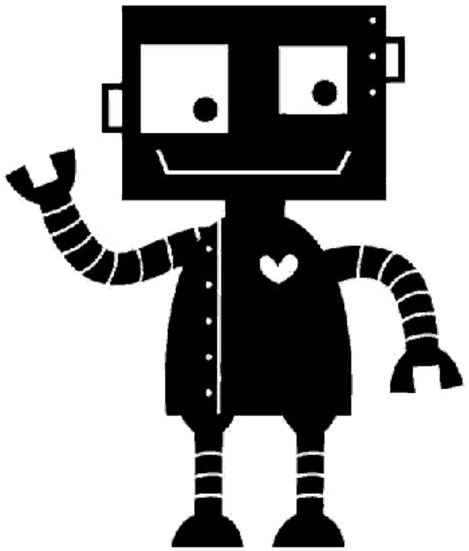


e.g., BLIP: Li et al. 2023, OpenFlamingo: Awadalla et al. 2023

Core task

→

there is a brown dog that is lying on a couch



Core task ...

→

there is a brown dog that is lying on a couch

... situated in a communicative context

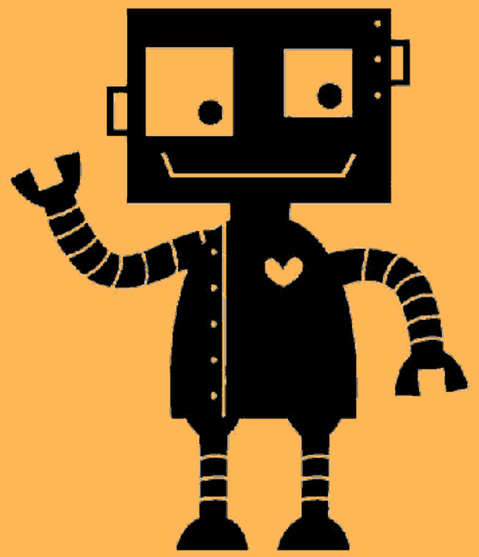
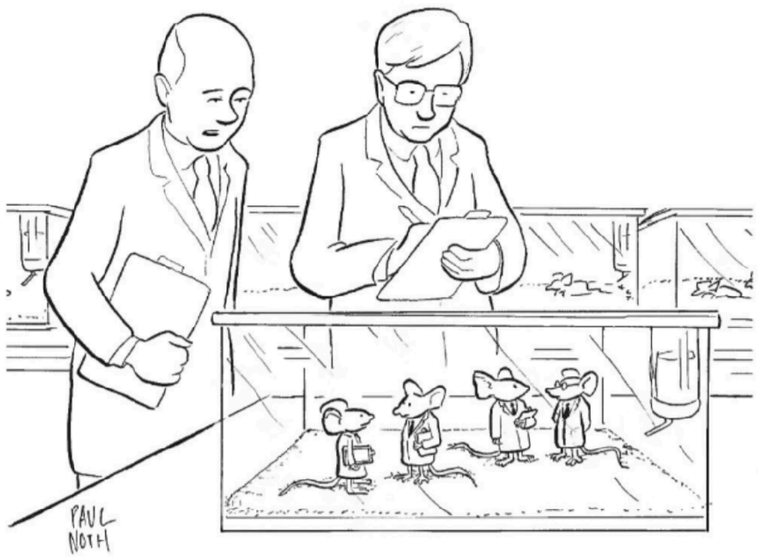


Image Accessibility

An Opportunity and Challenge for AI

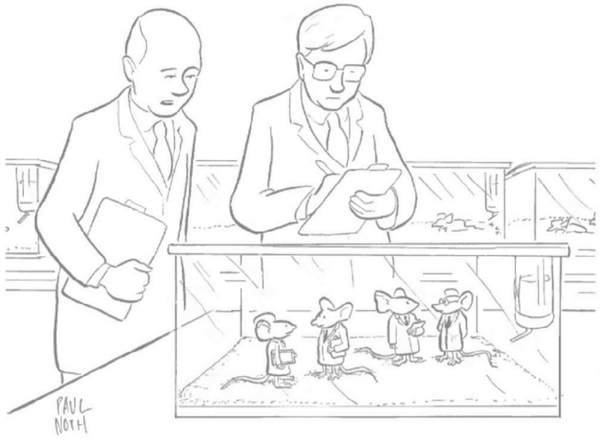
Globally, at least 2.2 billion people have a near or distance vision impairment.
World Health Organization, 2022



"O.K., let's slowly lower in the grant money."
Todd Bearson, Arlington, Massachusetts.
2009

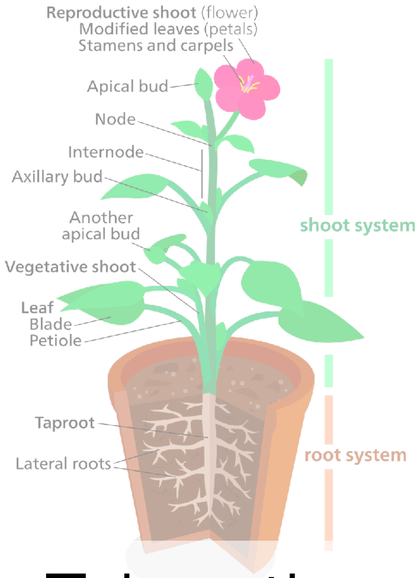
Globally, at least 2.2 billion people have a near or distance vision impairment.
World Health Organization, 2022

The promise of the internet: The leveling playing field for **equal access**.

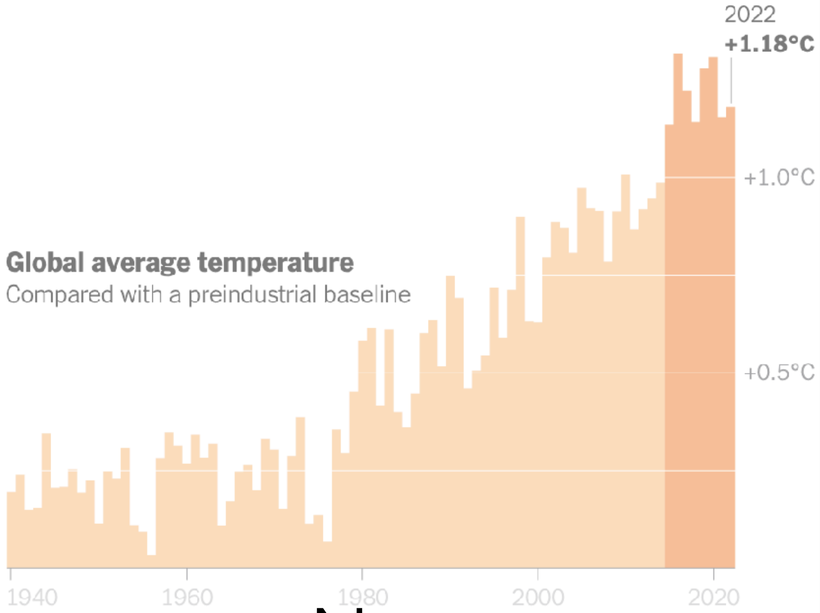


"O.K., let's slowly lower in the grant money."
Todd Bearson, Arlington, Massachusetts.
2009

Entertainment



Education

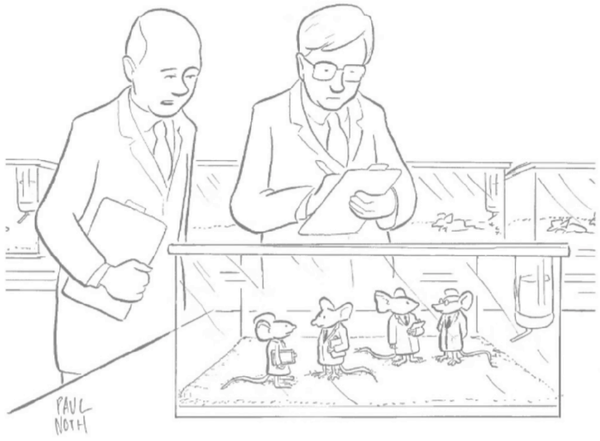


News

Globally, at least 2.2 billion people have a near or distance vision impairment.
World Health Organization, 2022

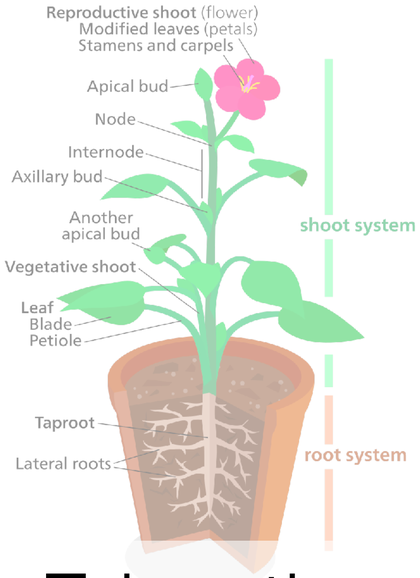
The promise of the internet: The leveling playing field for **equal access**.

The reality of the internet: Those who can't access, are **disadvantaged**.

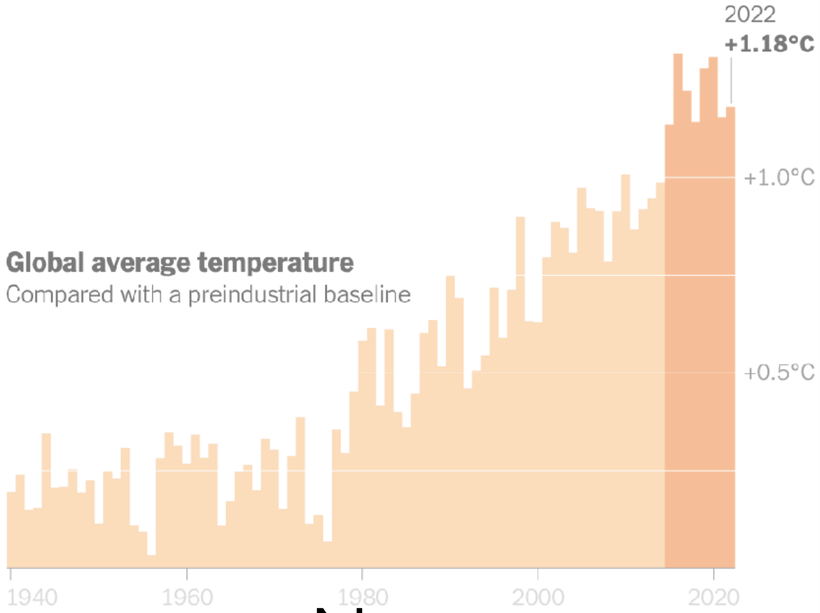


"O.K., let's slowly lower in the grant money."
Todd Bearson, Arlington, Massachusetts.
2009

Entertainment



Education



News

Accessing the Web When You Can't See

Screen reader: *software program that allows blind or visually impaired users to read the text that is displayed on the computer screen with a speech synthesizer or braille display* (American Foundation of the Blind, 2022)

I'm so excited for the talk today!

Accessing the Web When You Can't See

Screen reader: *software program that allows blind or visually impaired users to read the text that is displayed on the computer screen with a speech synthesizer or braille display* (American Foundation of the Blind, 2022)

Accessing the Web When You Can't See

Screen reader: *software program that allows blind or visually impaired users to read the text that is displayed on the computer screen with a speech synthesizer or braille display* (American Foundation of the Blind, 2022)

???

Accessing the Web When You Can't See

Screen reader: *software program that allows blind or visually impaired users to read the text that is displayed on the computer screen with a speech synthesizer or braille display* (American Foundation of the Blind, 2022)

Cute, small dog sitting on a sidewalk, looking up with big eyes. Ears are propped up.

Gleason et al. 2019

Accessing the Web When You Can't See

Screen reader: *software program that allows blind or visually impaired users to read the text that is displayed on the computer screen with a speech synthesizer or braille display* (American Foundation of the Blind, 2022)

Cute, small dog sitting on a sidewalk, looking up with big eyes. Ears are propped up.

very rare
e.g., only 0.1% of Twitter images have alt text

Opportunity for AI to help!

Gleason et al. 2019

Visual Accessibility: **An Opportunity for AI**

Vision-Language models translate images into text.

→ model → image-based text

Visual Accessibility: **An Opportunity for AI**

Image-based text generation

Vision-Language models translate images into text.

→ model → image-based text

Visual Accessibility: **An Opportunity for AI**

Image-based text generation

A man in a suit gestures while speaking, with the U.S. flag in the background.
(GPT-4V, OpenAI 2023)

Visual Accessibility: **An Opportunity for AI**

There are many texts that can go with every single image.

Barack Obama giving his last state of the union address.

...

A hand with an index finger that points up, seemingly attached to a person who has their mouth open. Teeth are visible.

A microphone facing away from the viewer. A flag is in the background. A man stands between the microphone and the flag.

A man in a suit gestures while speaking, with the U.S. flag in the background.
(GPT-4V, OpenAI 2023)

Visual Accessibility: **An Opportunity for AI**

Challenge: What makes an accessibility description useful?

Barack Obama giving his last state of the union address.

...

A hand with an index finger that points up, seemingly attached to a person who has their mouth open. Teeth are visible.

A microphone facing away from the viewer. A flag is in the background. A man stands between the microphone and the flag.

A man in a suit gestures while speaking, with the U.S. flag in the background.
(GPT-4V, OpenAI 2023)

Visual Accessibility: **An Opportunity for AI**

Consequence: Even sophisticated model can miss accessibility goals.

[**See also:** MacLeod et al., 2017; Bennett et al., 2021; Herskovitz et al., 2023]

Example

Be My AI: OpenAI / Be My Eyes

(GPT-4V(ision) System Card, 2023)

A man in a suit gestures while speaking, with the U.S. flag in the background.
(GPT-4V, OpenAI 2023)

Visual Accessibility: **An Opportunity for AI**

Consequence: Even sophisticated model can miss accessibility goals.

[**See also:** MacLeod et al., 2017; Bennett et al., 2021; Herskovitz et al., 2023]

Example

Be My AI: OpenAI / Be My Eyes

(GPT-4V(ision) System Card, 2023)

Barack Obama $\xrightarrow{?}$ A man in a suit gestures while speaking, with the U.S. flag in the background.

(GPT-4V, OpenAI 2023)

Reframing Image-Based Text Generation

What **can** we say about an image?

What **should** we say about an image?

Reframing Image-Based Text Generation

What **can** we say about an image? → What **should** we say about an image?

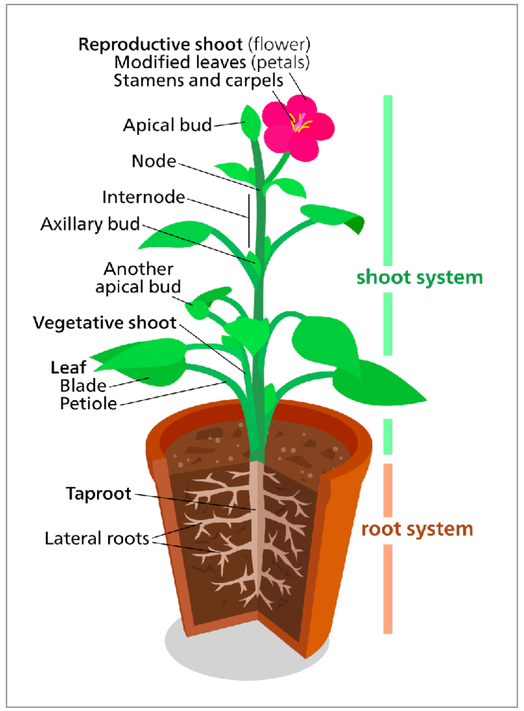
ground-truth description → **pragmatically useful** description

Image-based text generation depends on ...

1 the **image-based text**'s communicative goal.

2 the **image**'s communicative goal.

The Communicative Goal of the Image-Based **Text**



Concadia: Towards image-based text generation with a purpose; **Kreiss**, Fang, Goodman, Potts (EMNLP 2022)

The Communicative Goal of the Image-Based **Text**

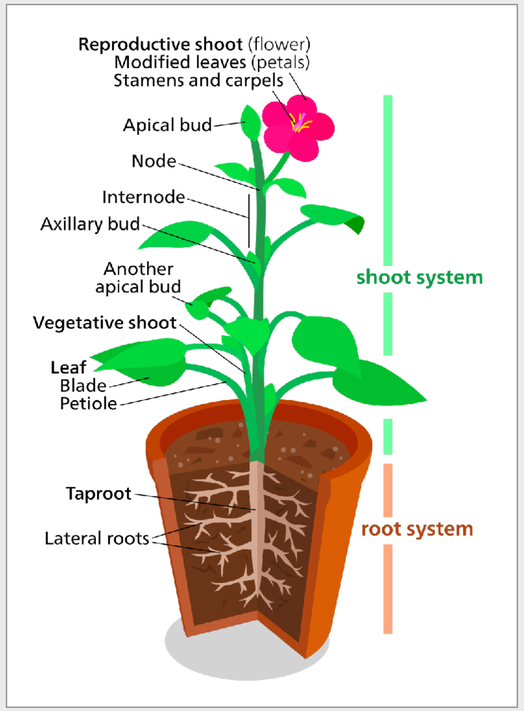
WIKIPEDIA
The Free Encyclopedia

Multimodal Pedagogy

Multimodal pedagogy is an approach to the teaching of writing that implements different modes of communication.[1][2] Multimodality refers to the use of visual, aural, linguistic, spatial, and gestural modes in differing pieces of media, each necessary to properly convey the information it presents.[3][4]

The visual mode conveys meaning via images and the visible elements of a text such as typography and color. The aural mode refers to sound in the form of music, sound effects, silence, etc. The linguistic mode includes written and spoken language. The spatial mode focuses on the physical arrangement of elements in a text. The gestural mode refers to physical movements such facial expressions and how these are interpreted. A multimodal text is characterized by the combination of any two or more modes to express meaning.[5]

Multimodality as a term was coined in the late 20th century,[6] but its use predates its naming, with it being used as early as Egyptian hieroglyphs and classical rhetoric.[7] Compositionists and writing theorists have been exploring how the five modes of communication interact with each other and how multimodality can be used in the teaching of writing since the 20th century.[8]



Example of complementing linguistic and visual information leading to learning benefits over unimodal approaches.

Concadia: Towards image-based text generation with a purpose; **Kreiss**, Fang, Goodman, Potts (EMNLP 2022)

The Communicative Goal of the Image-Based **Text**

WIKIPEDIA
The Free Encyclopedia

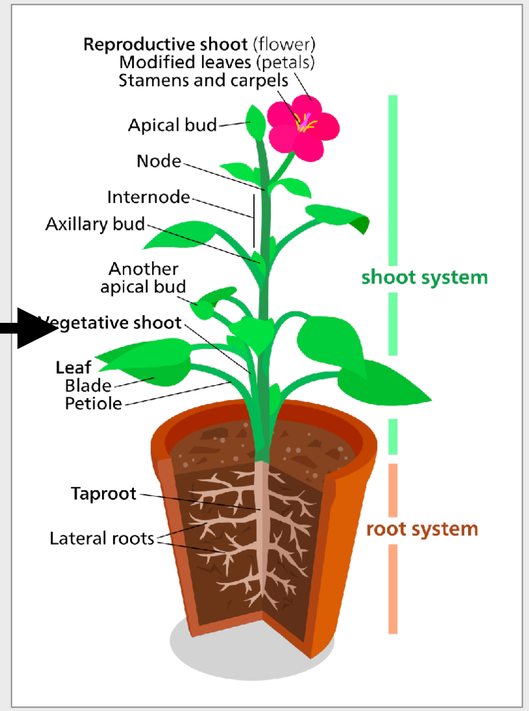
Multimodal Pedagogy

Multimodal pedagogy is an approach to the teaching of writing that implements different modes of communication.[1][2] Multimodality re[illegible]s in differing pieces of media, each ne[illegible]

The visual mode conveys meaning [illegible] color. The aural mode refers to sound in [illegible]des written and spoken language. The [illegible]t. The gestural mode refers to physical m[illegible]ltimodal text is characterized by the combin[illegible]

Multimodality as a term was coined i[illegible]ing used as early as Egyptian hieroglyphs and classical rhetoric.[7] Compositionists and writing theorists have been exploring how the five modes of communication interact with each other and how multimodality can be used in the teaching of writing since the 20th century.[8]

An educational sketch of a plant, illustrating the shoot system which is above the soil, and the root system which is below the soil. The different parts of the plant are labeled to point out, e.g., the plant's reproductive shoot (i.e., the flower) or the lateral roots.



Example of complementing linguistic and visual information leading to learning benefits over unimodal approaches.

Concadia: Towards image-based text generation with a purpose; **Kreiss**, Fang, Goodman, Potts (EMNLP 2022)

The Communicative Goal of the Image-Based **Text**

WIKIPEDIA
The Free Encyclopedia

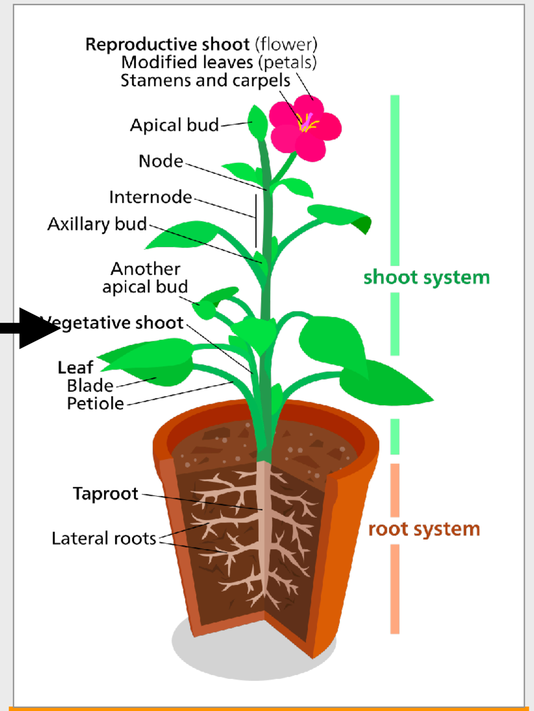
Multimodal Pedagogy

Multimodal pedagogy is an approach to the teaching of writing that implements different modes of communication.[1][2] Multimodality re[illegible]s in differing pieces of media, each ne[illegible]

The visual mode conveys meaning [illegible] color. The aural mode refers to sound in [illegible]des written and spoken language. The [illegible]t. The gestural mode refers to physical m[illegible]ltimodal text is characterized by the combin[illegible]

Multimodality as a term was coined [illegible]ing used as early as Egyptian hieroglyphs and classical rhetoric.[7] Compositionists and writing theorists have been exploring how the five modes of communication interact with each other and how multimodality can be used in the teaching of writing since the 20th century.[8]

An educational sketch of a plant, illustrating the shoot system which is above the soil, and the root system which is below the soil. The different parts of the plant are labeled to point out, e.g., the plant's reproductive shoot (i.e., the flower) or the lateral roots.



Example of complementing linguistic and visual information leading to learning benefits over unimodal approaches.

Concadia: Towards image-based text generation with a purpose; **Kreiss**, Fang, Goodman, Potts (EMNLP 2022)

The Communicative Goal of the Image-Based **Text**

WIKIPEDIA
The Free Encyclopedia

Multimodal Pedagogy

Multimodal pedagogy is an appr

Description

writing that implements different modes of communication.[1][2] Multimodality re

s in differing pieces of media, each ne

The visual mode conveys meaning

color. The aural mode refers to sound in

des written and spoken language. The

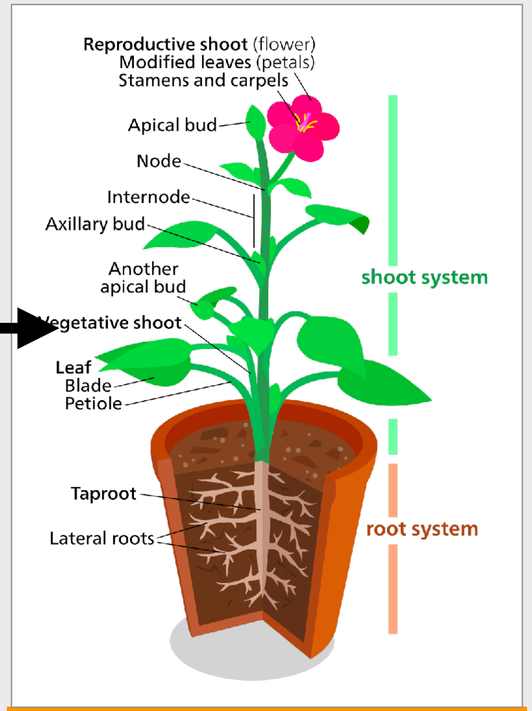
t. The gestural mode refers to physical m

ltimodal text is characterized by the combin

Multimodality as a term was coined i

ng used as early as Egyptian hieroglyphs and classical rhetoric.[7] Compositionists and writing theorists have been exploring how the five modes of communication interact with each other and how multimodality can be used in the teaching of writing since the 20th century.[8]

An educational sketch of a plant, illustrating the shoot system which is above the soil, and the root system which is below the soil. The different parts of the plant are labeled to point out, e.g., the plant's reproductive shoot (i.e., the flower) or the lateral roots.

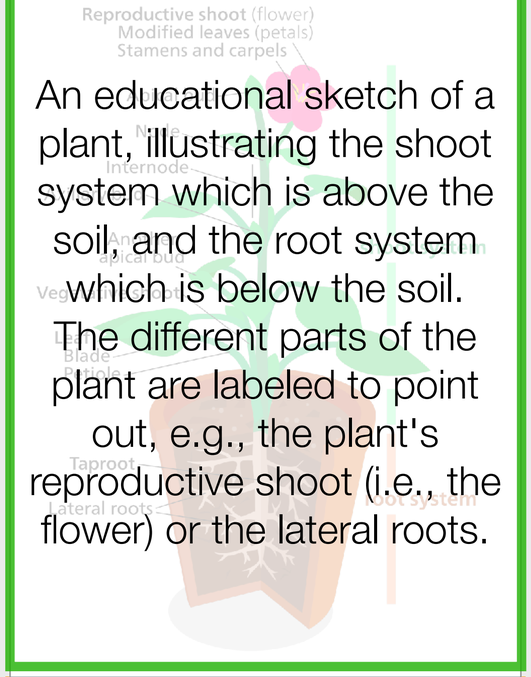


Caption

Example of complementing linguistic and visual information leading to learning benefits over unimodal approaches.

Concadia: Towards image-based text generation with a purpose; **Kreiss**, Fang, Goodman, Potts (EMNLP 2022)

The Communicative Goal of the Image-Based **Text**



An educational sketch of a plant, illustrating the shoot system which is above the soil, and the root system which is below the soil. The different parts of the plant are labeled to point out, e.g., the plant's reproductive shoot (i.e., the flower) or the lateral roots.

Example of complementing linguistic and visual information leading to learning benefits over unimodal approaches.

Description

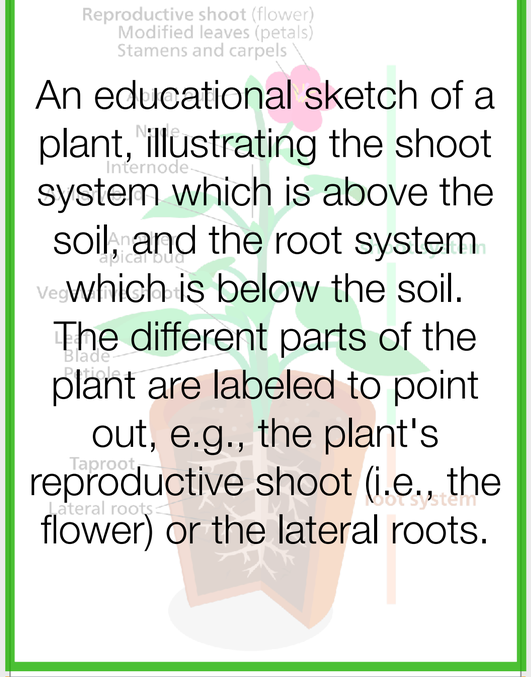
- accessibility
- replaces image

Caption

- complements image

Concadia: Towards image-based text generation with a purpose; **Kreiss**, Fang, Goodman, Potts (EMNLP 2022)

The Communicative Goal of the Image-Based **Text**



An educational sketch of a plant, illustrating the shoot system which is above the soil, and the root system which is below the soil. The different parts of the plant are labeled to point out, e.g., the plant's reproductive shoot (i.e., the flower) or the lateral roots.

Example of complementing linguistic and visual information leading to learning benefits over unimodal approaches.

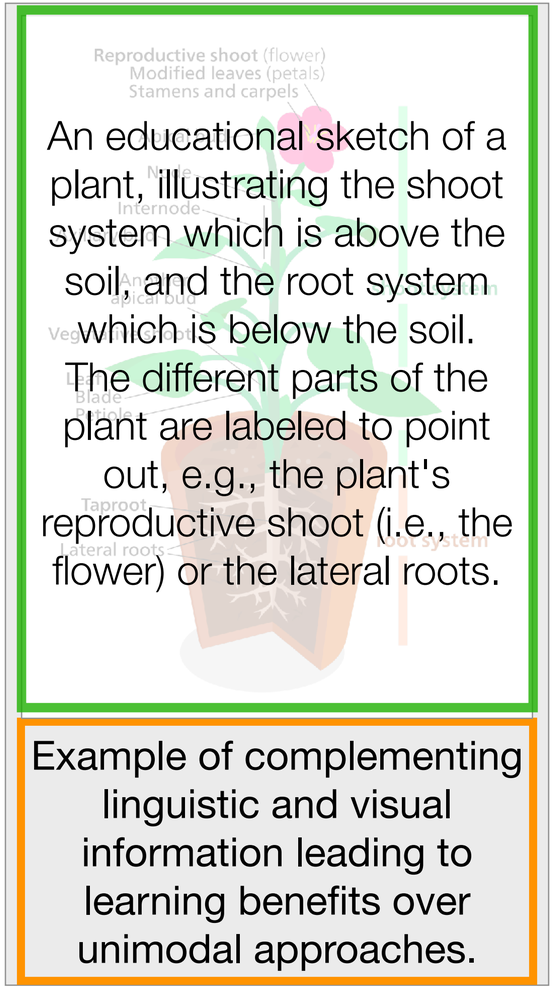
Description
accessibility
replaces image

Caption
complements image

Captions and descriptions are both image-based, but are produced in complementary **settings**.
(Goodwin & Duranti; 1992)

Concadia: Towards image-based text generation with a purpose; **Kreiss**, Fang, Goodman, Potts (EMNLP 2022)

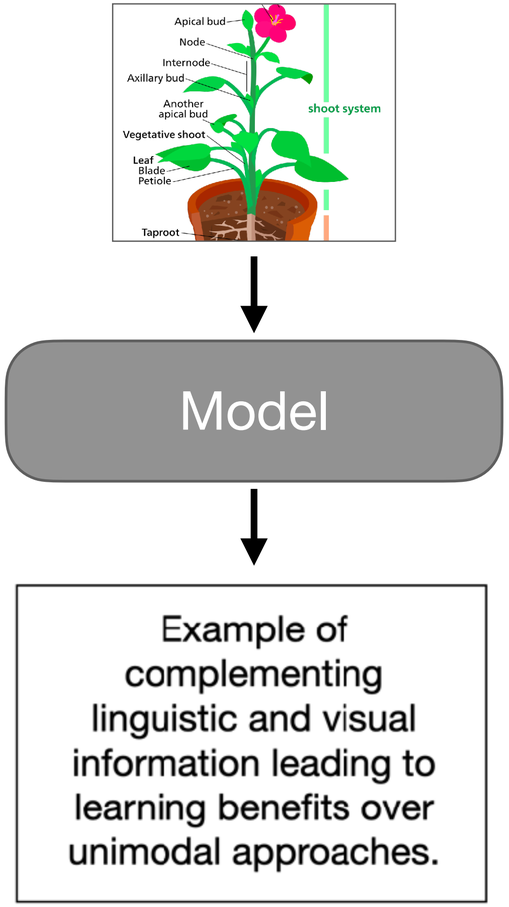
The Communicative Goal of the Image-Based **Text**



Description
accessibility
replaces image

Caption
complements image

In AI, we tend to reduce them to the same problem.

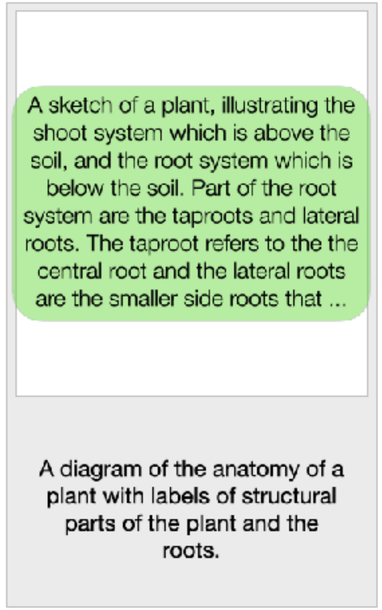


Concadia: Towards image-based text generation with a purpose; **Kreiss**, Fang, Goodman, Potts (EMNLP 2022)

Image-based text generation depends on ...

1 the **image-based text**'s communicative goal.

→ description ≠ caption



2 the **image**'s communicative goal.

Finding the Image-Based **Text**'s Communicative Goal

Concadia: A naturalistic image-based text dataset from Wikipedia

96,918 images with captions, alt descriptions and surrounding paragraph

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Finding the Image-Based **Text**'s Communicative Goal

Concadia: A naturalistic image-based text dataset from Wikipedia

96,918 images with captions, alt descriptions and surrounding paragraph

Wikipedia-Article on **Banana**

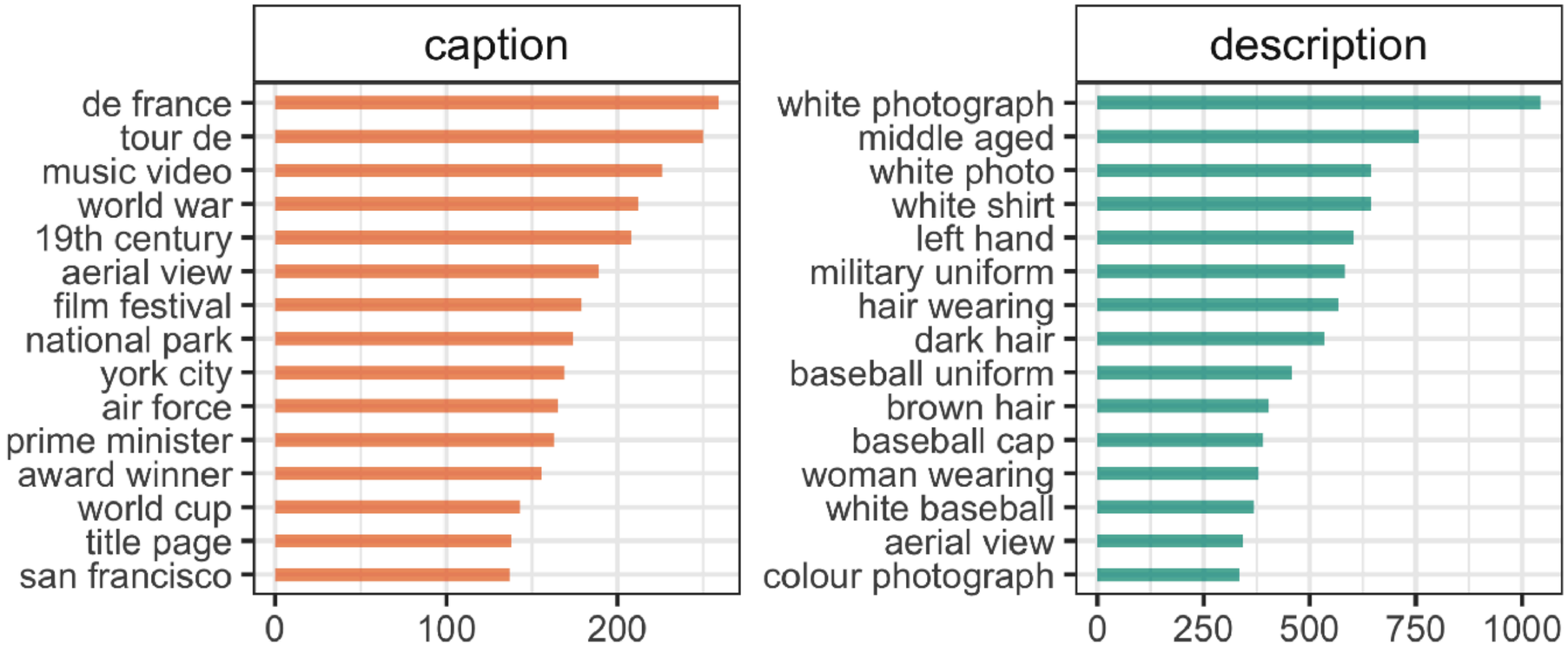
image context: In global commerce in 2009, by far the most important cultivars belonged to the triploid AAA group of Musa acuminata, commonly referred to as Cavendish group bananas. They accounted for the majority of banana exports, despite only coming into existence in 1836. The cultivars Dwarf Cavendish and Grand Nain (Chiquita Banana) gained popularity in the 1950s after the previous mass-produced cultivar, Gros Michel (also an AAA ...

(Accessibility) **Description:**
Grocery store photo of several bunches of bananas

(Contextualizing) **Caption:**
Cavendish bananas are the main commercial banana cultivars sold in the world market.

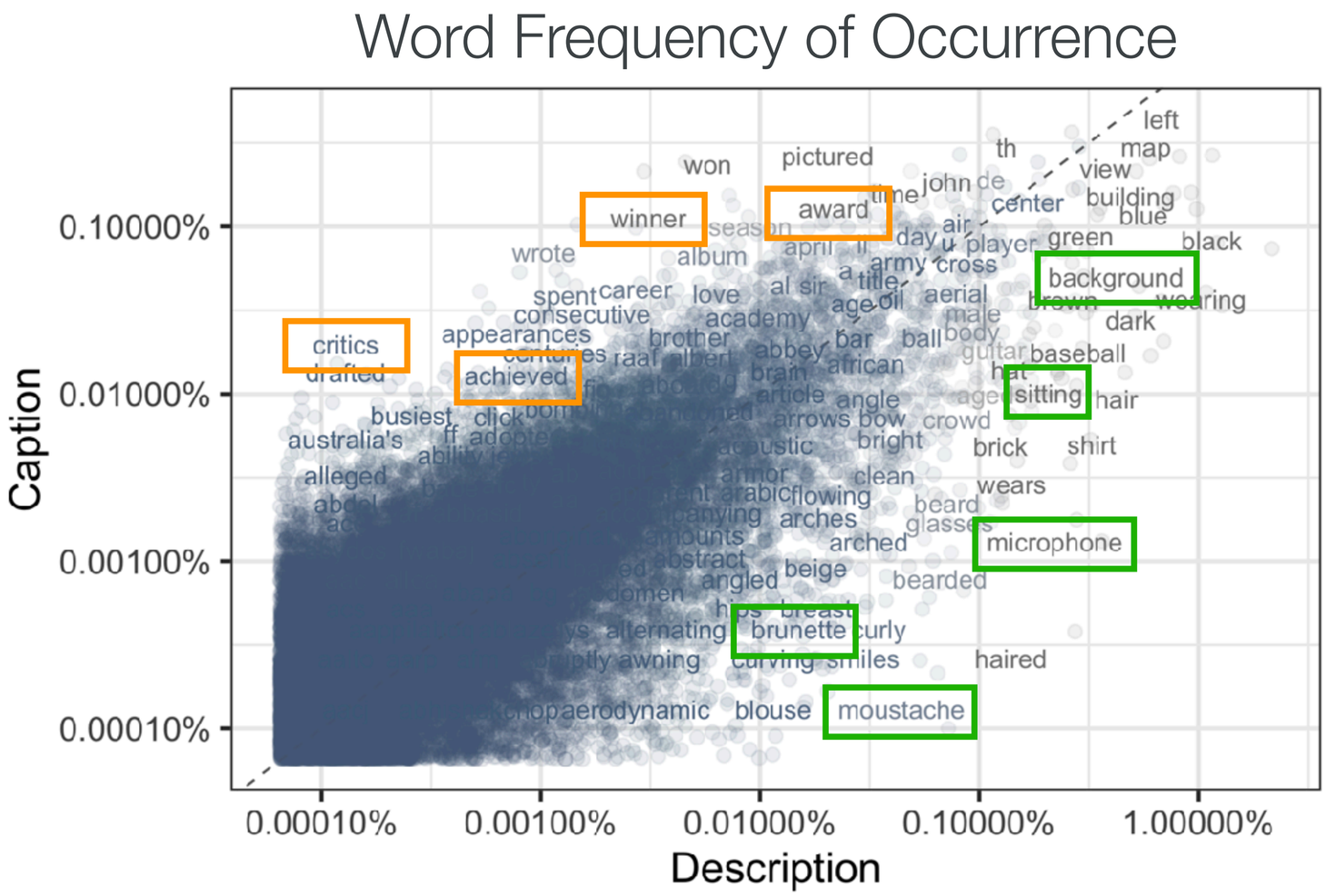
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Concadia: A naturalistic image-based text dataset



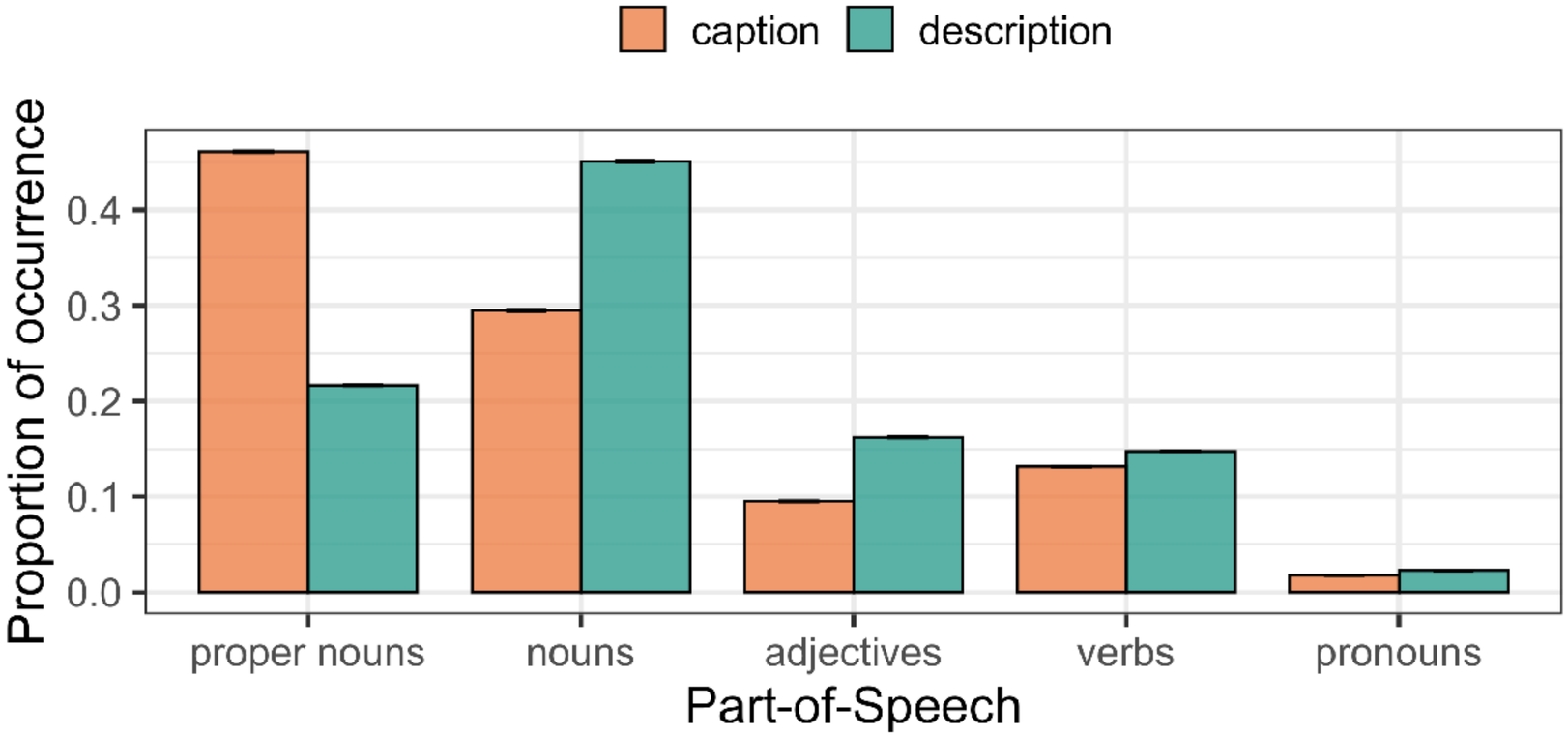
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Concadia: A naturalistic image-based text dataset



Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Concadia: A naturalistic image-based text dataset

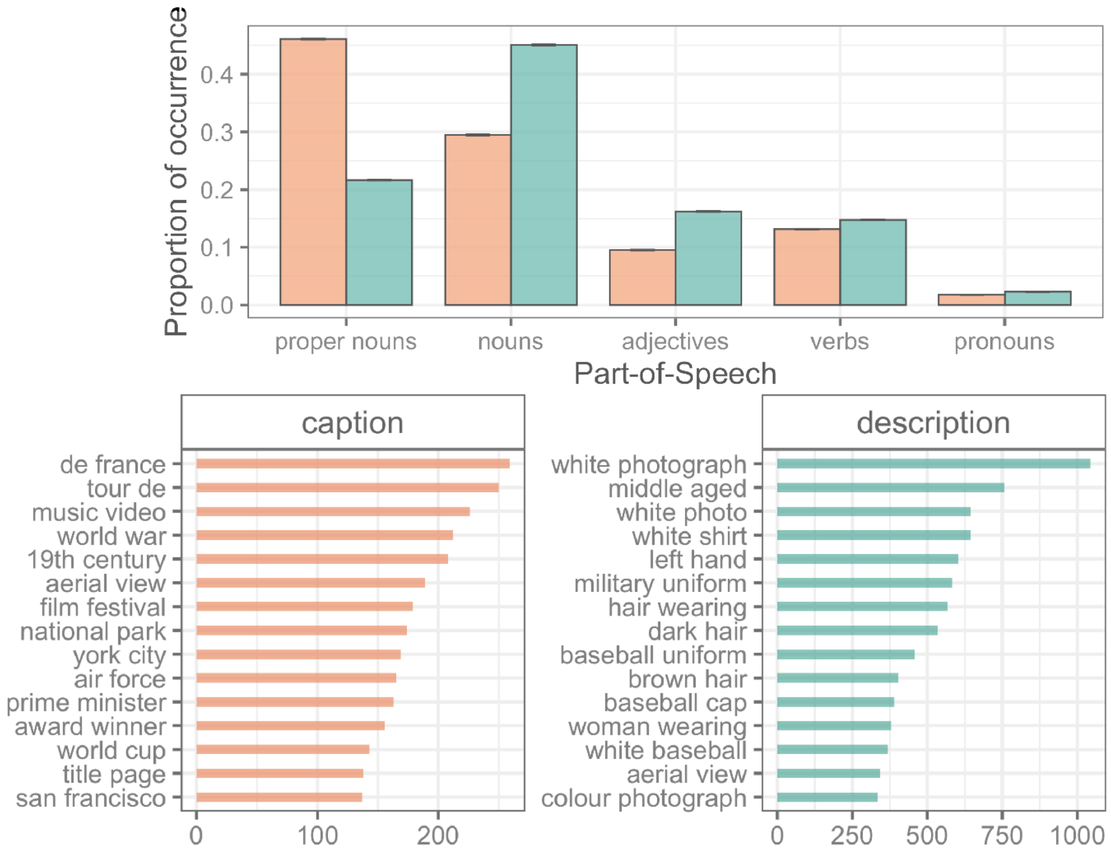


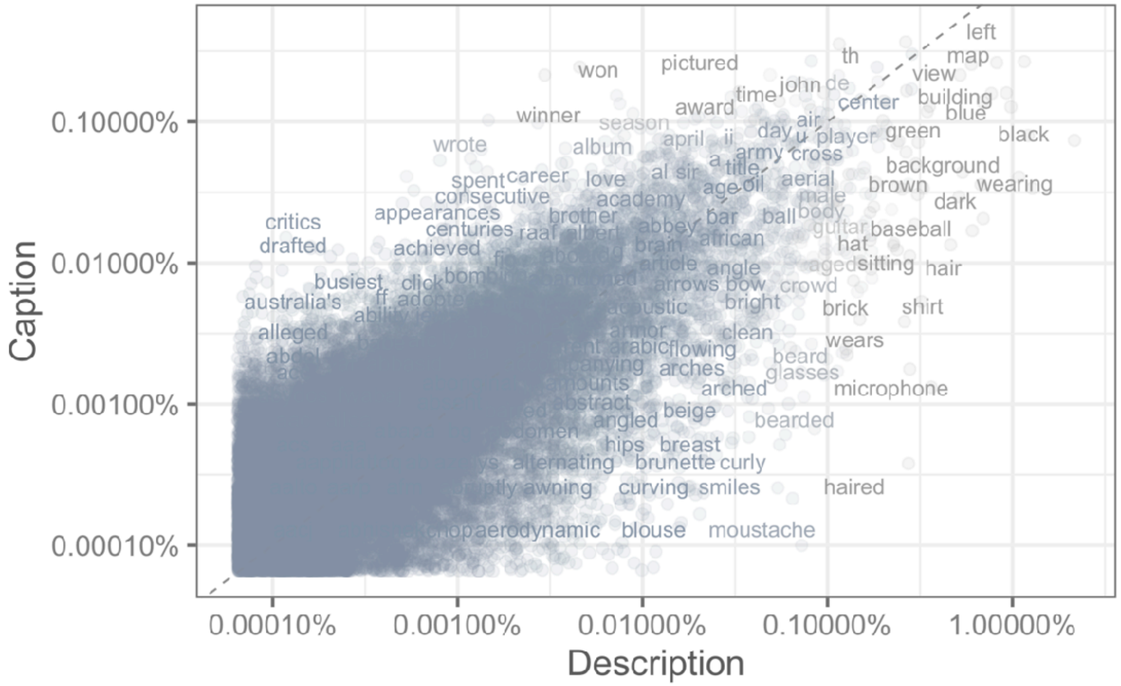
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

The Image-Based **Text**'s Communicative Goal

Large-scale analysis of naturalistic data:

The content of alt descriptions and captions differs in structured ways.





Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

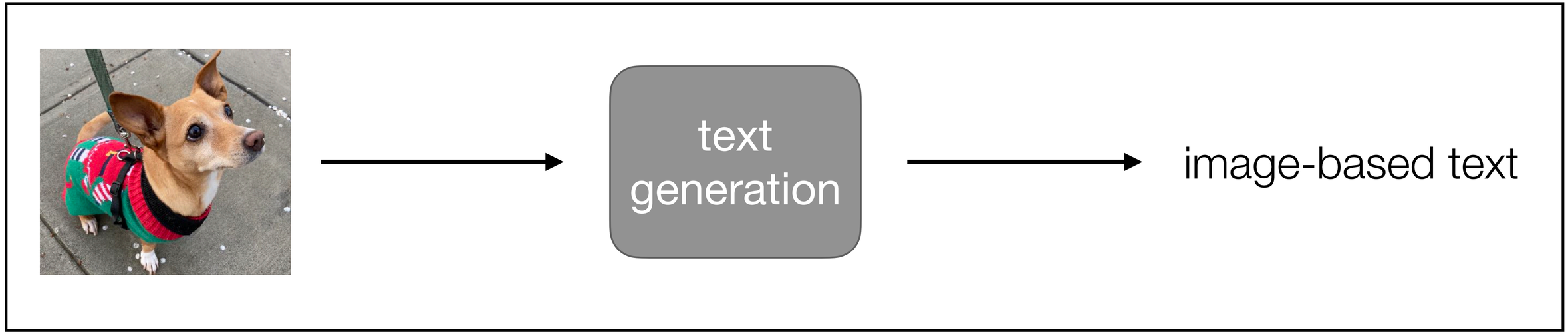
Generating Image-Based Text with a Purpose

Architecture

ResNet + LSTM

DenseNet + LSTM

OSCAR (VinVL)



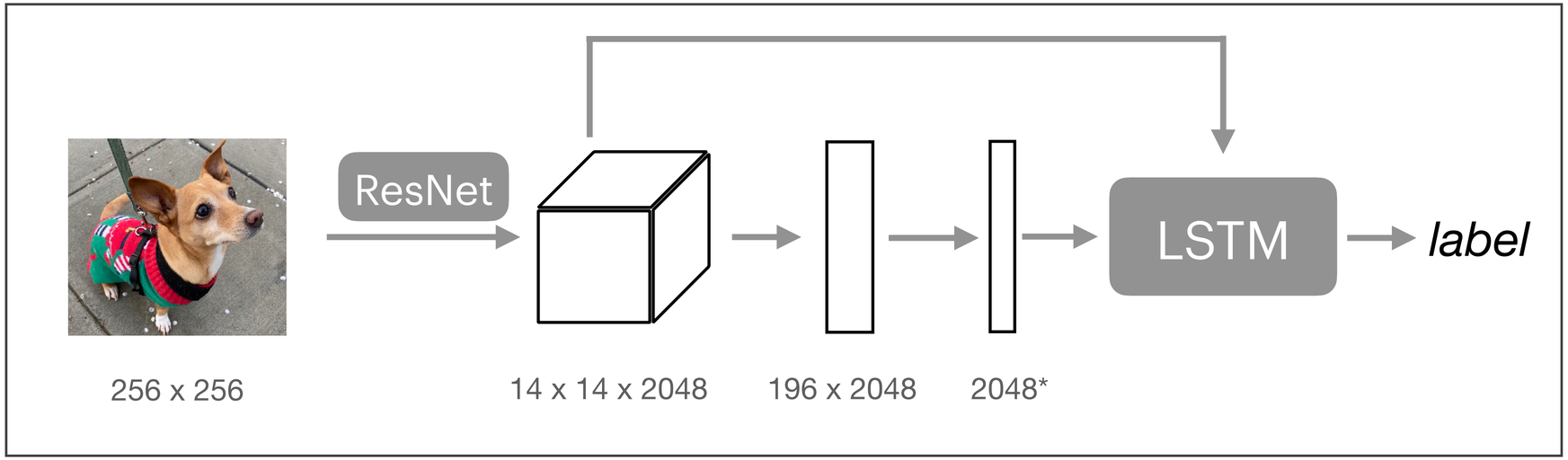
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Generating Image-Based Text with a Purpose

Architecture

ResNet + LSTM DenseNet + LSTM OSCAR (VinVL)

Input Encoding Attention Decoding Output



*simplified but will be expanded later in the talk

Based on Xu et al., 2015, He et al., 2016

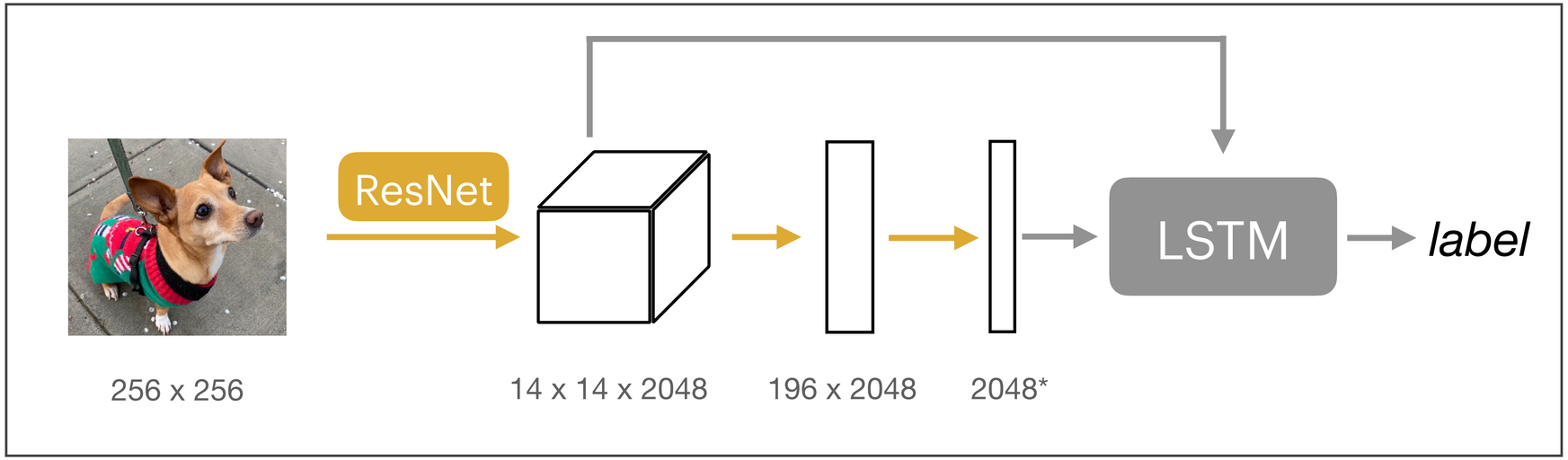
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Generating Image-Based Text with a Purpose

Architecture

ResNet + LSTM DenseNet + LSTM OSCAR (VinVL)

Input Encoding Attention Decoding Output



*simplified but will be expanded later in the talk

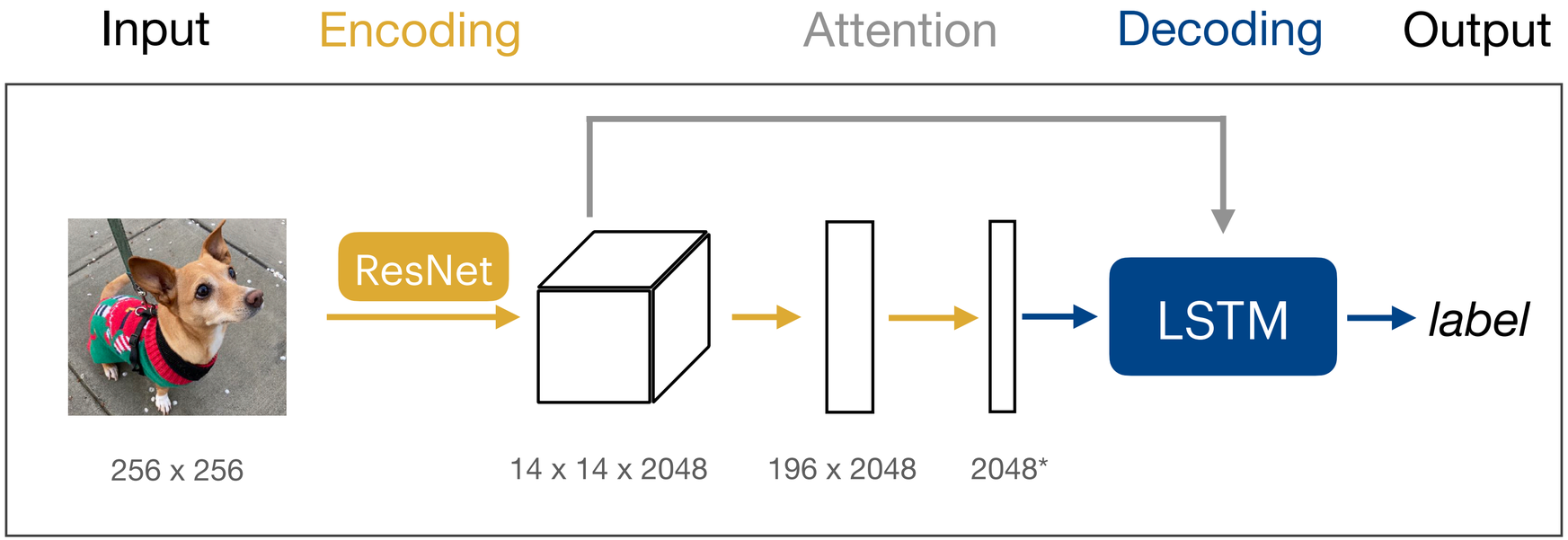
Based on Xu et al., 2015, He et al., 2016

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Generating Image-Based Text with a Purpose

Architecture

ResNet + LSTM DenseNet + LSTM OSCAR (VinVL)



*simplified but will be expanded later in the talk

Based on Xu et al., 2015, He et al., 2016

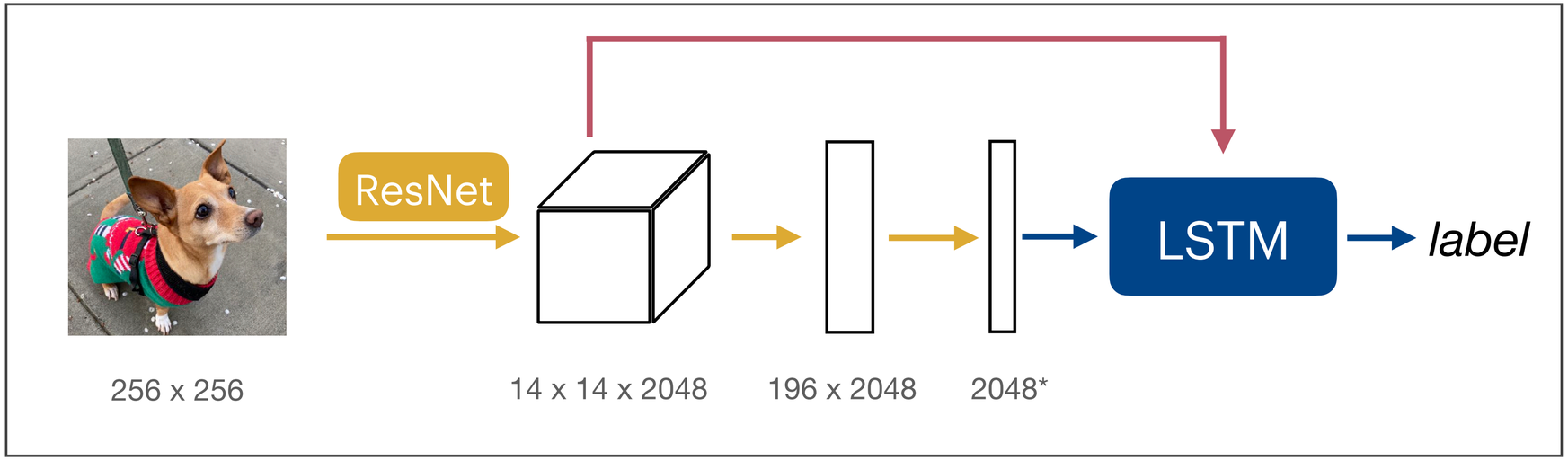
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Generating Image-Based Text with a Purpose

Architecture

ResNet + LSTM DenseNet + LSTM OSCAR (VinVL)

Input Encoding Attention Decoding Output



*simplified but will be expanded later in the talk

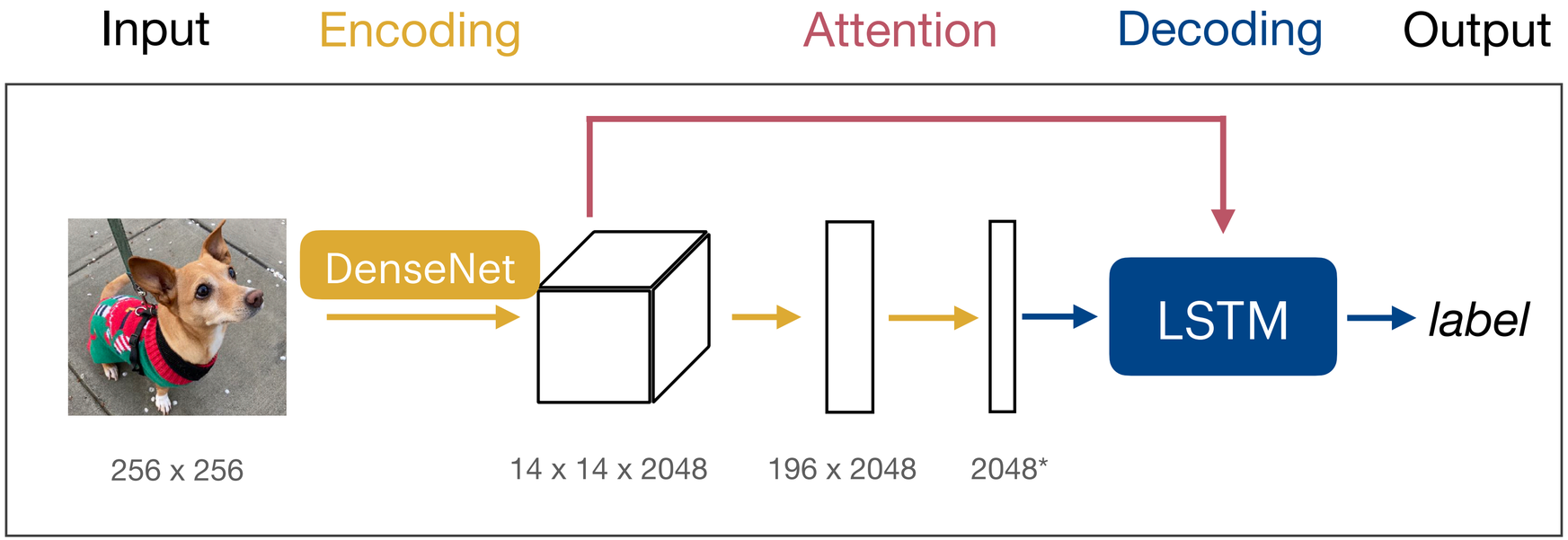
Based on Xu et al., 2015, He et al., 2016

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Generating Image-Based Text with a Purpose

Architecture

ResNet + LSTM | **DenseNet + LSTM** | OSCAR (VinVL)



*simplified but will be expanded later in the talk

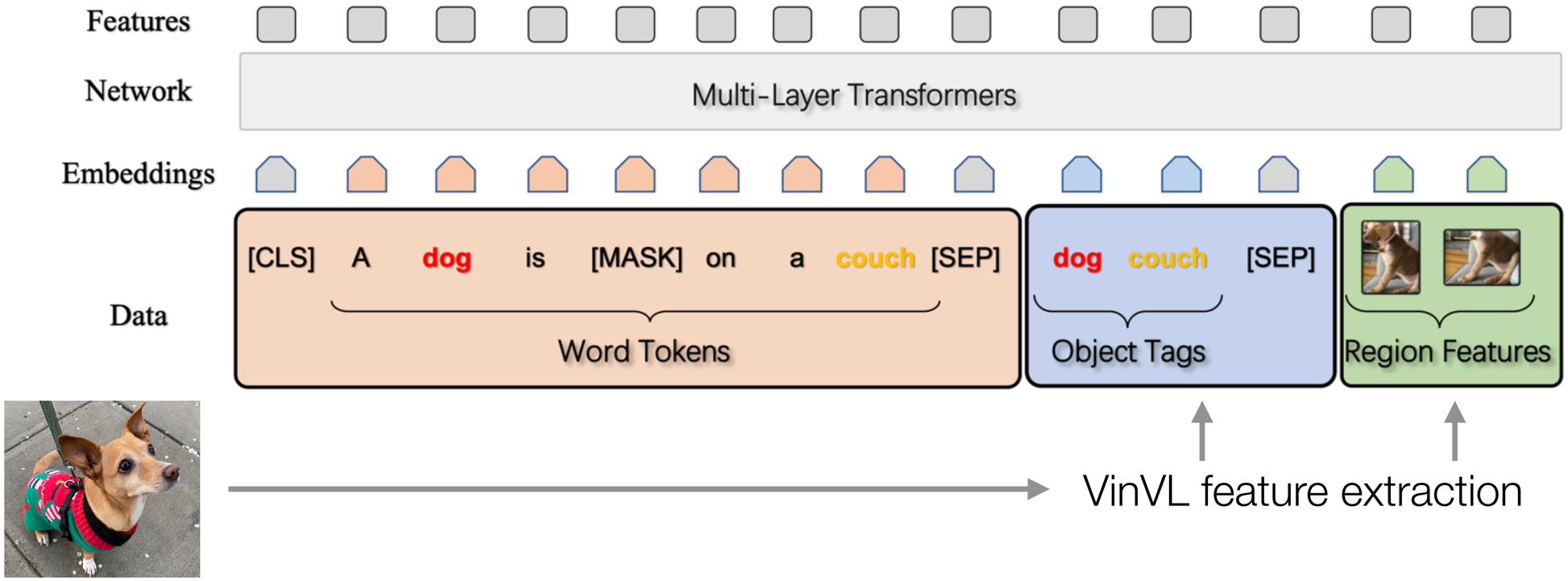
Based on Deng et al., 2020; Hossain et al., 2021

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Generating Image-Based Text with a Purpose

Architecture

ResNet + LSTM | DenseNet + LSTM | **OSCAR (VinVL)**



Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Based on Zhang et al. 2021, Li et al. 2020

Generating Image-Based Text with a Purpose

Training

ResNet + LSTM | DenseNet + LSTM | OSCAR (VinVL)

1

2

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Generating Image-Based Text with a Purpose

Training

ResNet + LSTM | DenseNet + LSTM | OSCAR (VinVL)

Caption generation

1

→ image pre-processing → text generation → Cavendish bananas are the main commercial banana cultivar sold in the world market.

2

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Generating Image-Based Text with a Purpose

Training

ResNet + LSTM | DenseNet + LSTM | OSCAR (VinVL)

Caption generation

1

→ image pre-processing → text generation → Cavendish bananas are the main commercial banana cultivar sold in the world market.

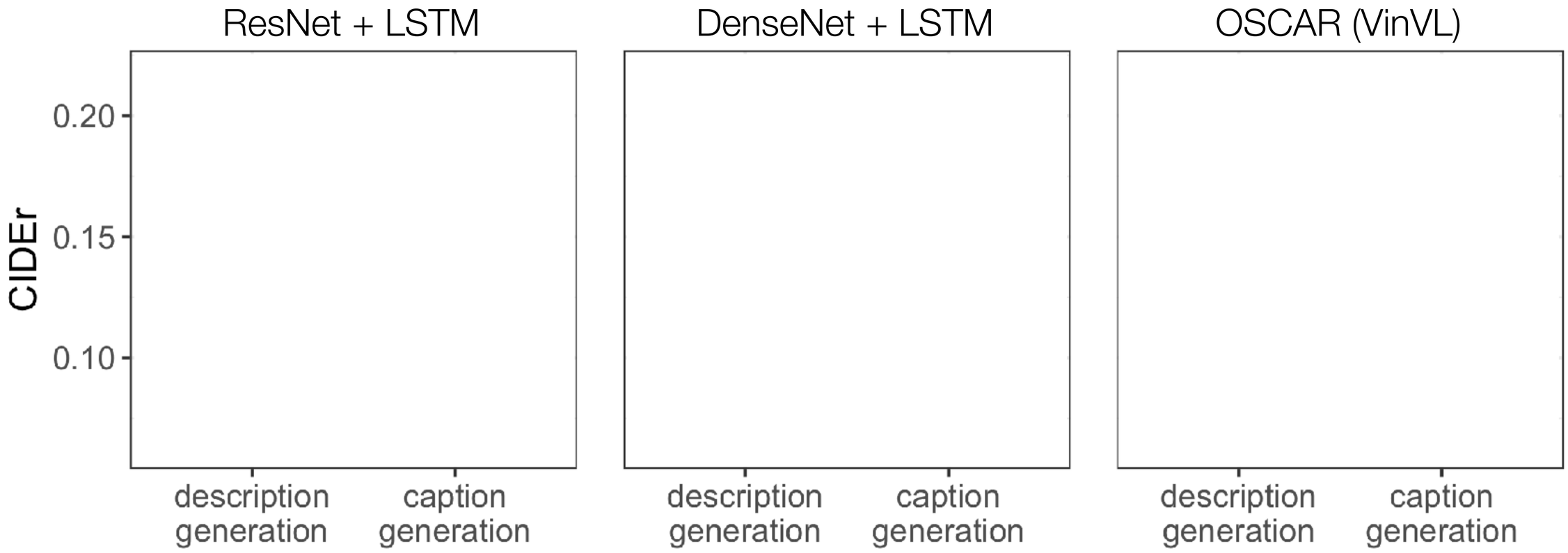
Description generation

2

→ image pre-processing → text generation → Grocery store photo of several bunches of banana

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

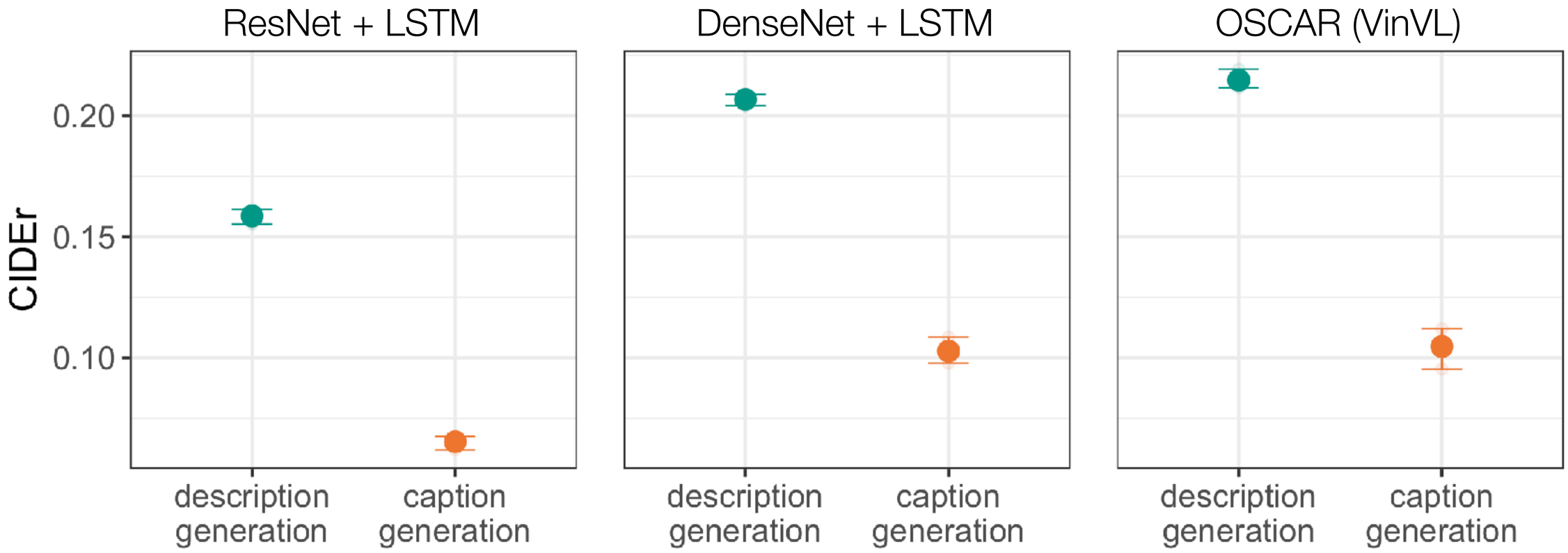
Generating Image-Based Text with a Purpose



Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Generating Image-Based Text with a Purpose

Caption generation is harder than description generation



Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

The Image-Based **Text**'s Communicative Goal

Large-scale analysis of naturalistic data:

The content of alt descriptions and captions differs in structured ways.

Models of image-based text generation:

Learning alt description and caption generation are distinct challenges.

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

The Image-Based **Text**'s Communicative Goal

Large-scale analysis of naturalistic data:

The content of alt descriptions and captions differs in structured ways.

Models of image-based text generation:

Learning alt description and caption generation are distinct challenges.

Is the distinction between descriptions and captions due to a distinction in the text's **communicative goal**?

Is this distinction reflected in **models** trained on the respective data?

Concadia: Towards image-based text generation with a purpose

Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Evaluating the Purpose of Descriptions & Captions

preregistered human subject experiment

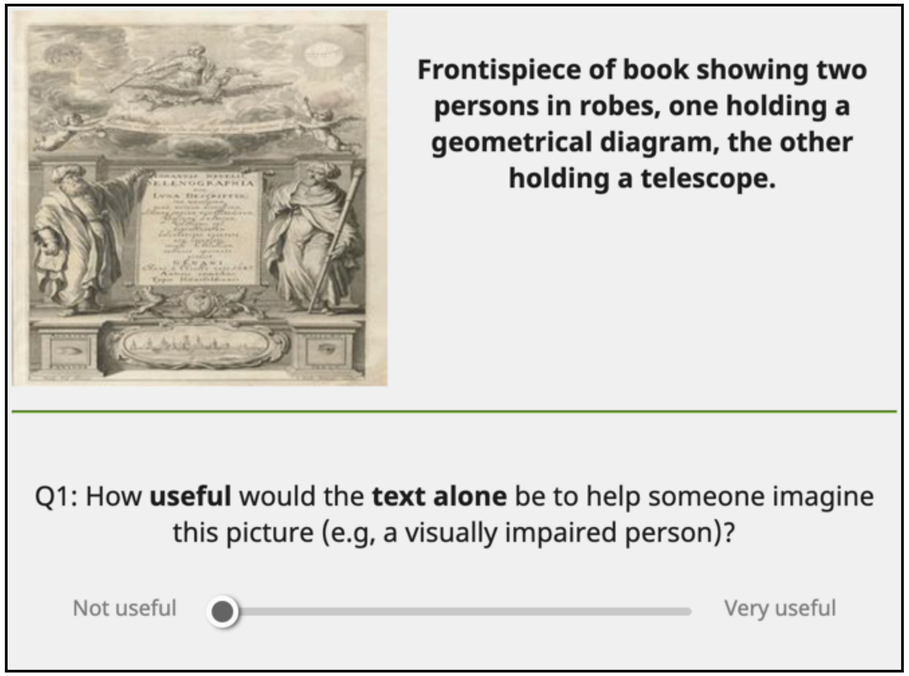
Frontispiece of book showing two persons in robes, one holding a geometrical diagram, the other holding a telescope.

→ image,
text (description: human/model,
caption: human/model)

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Evaluating the Purpose of Descriptions & Captions

preregistered human subject experiment



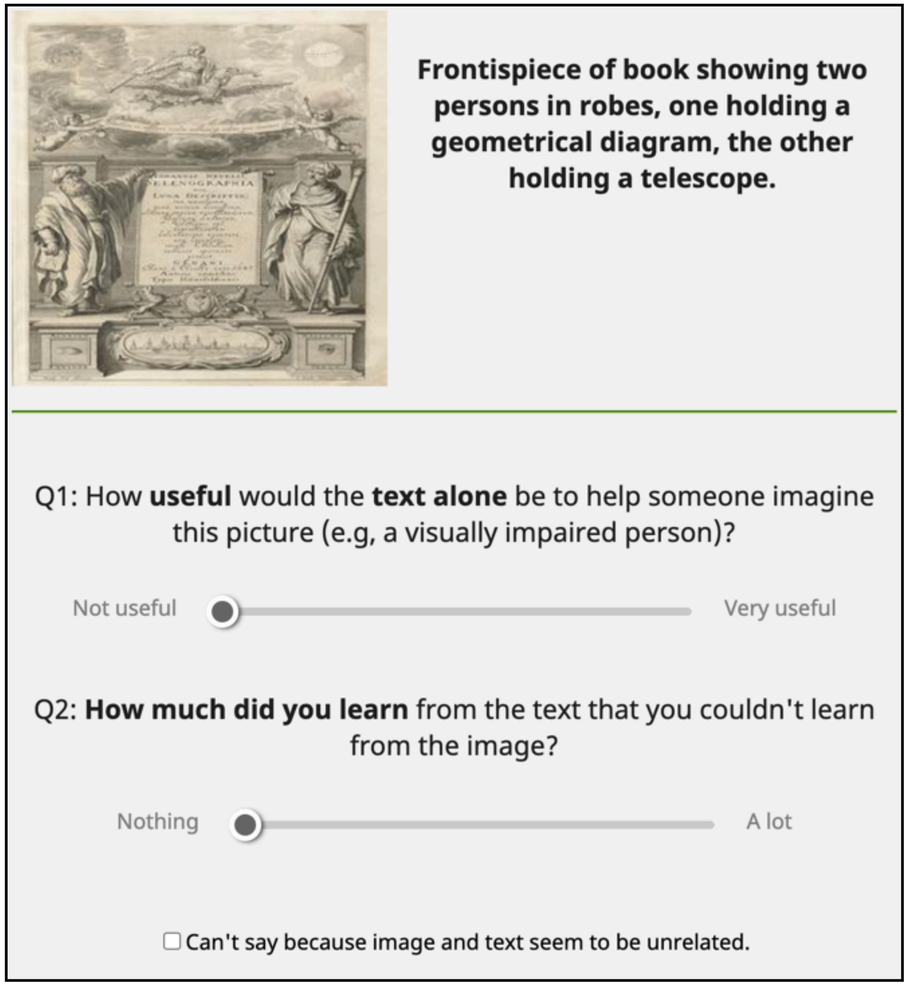
→ image,
text (description: human/model,
caption: human/model)

→ question for descriptive quality

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Evaluating the Purpose of Descriptions & Captions

preregistered human subject experiment



→ image,
text (description: human/model,
caption: human/model)

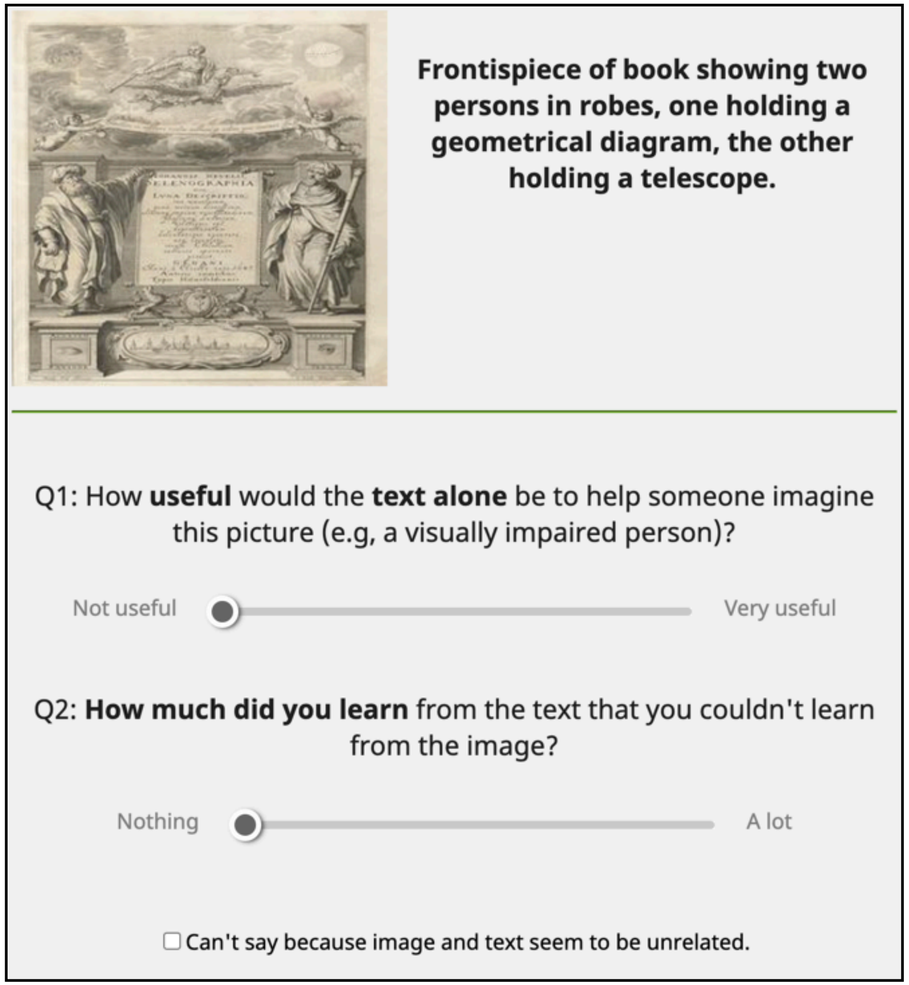
→ question for descriptive quality

→ question for caption quality

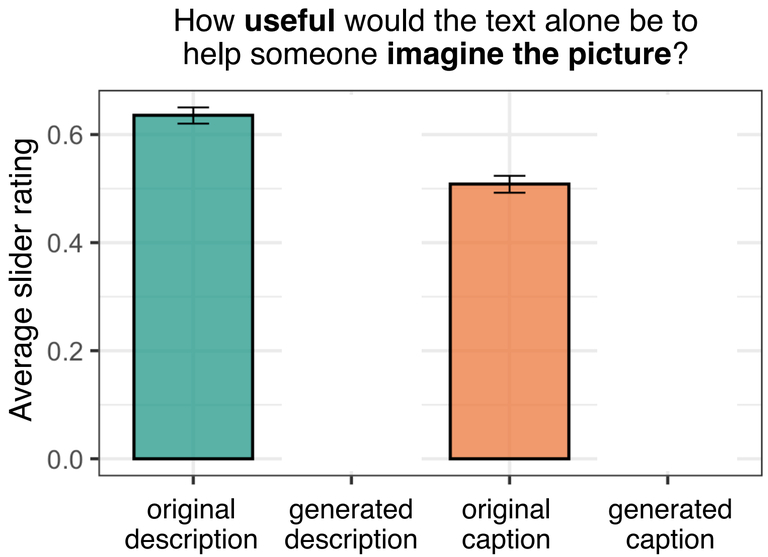
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Towards AI for Image Accessibility

preregistered human subject experiment



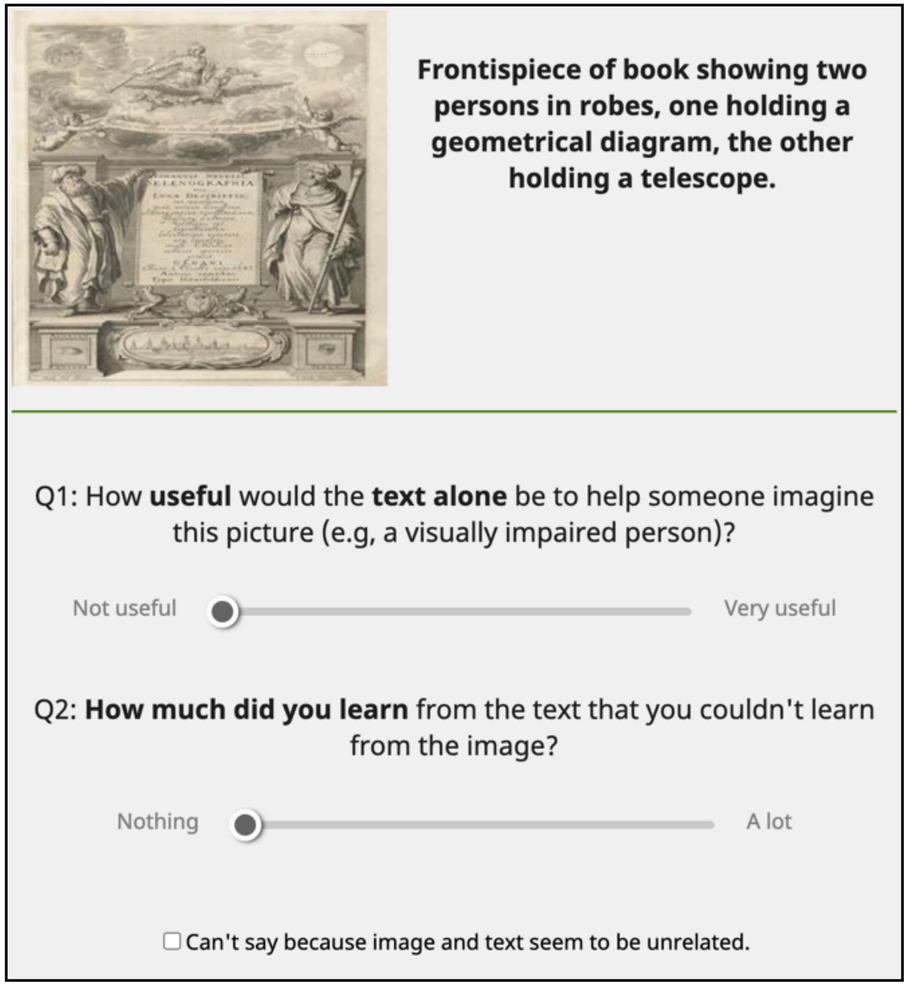
question for descriptive quality



Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Towards AI for Image Accessibility

preregistered human subject experiment

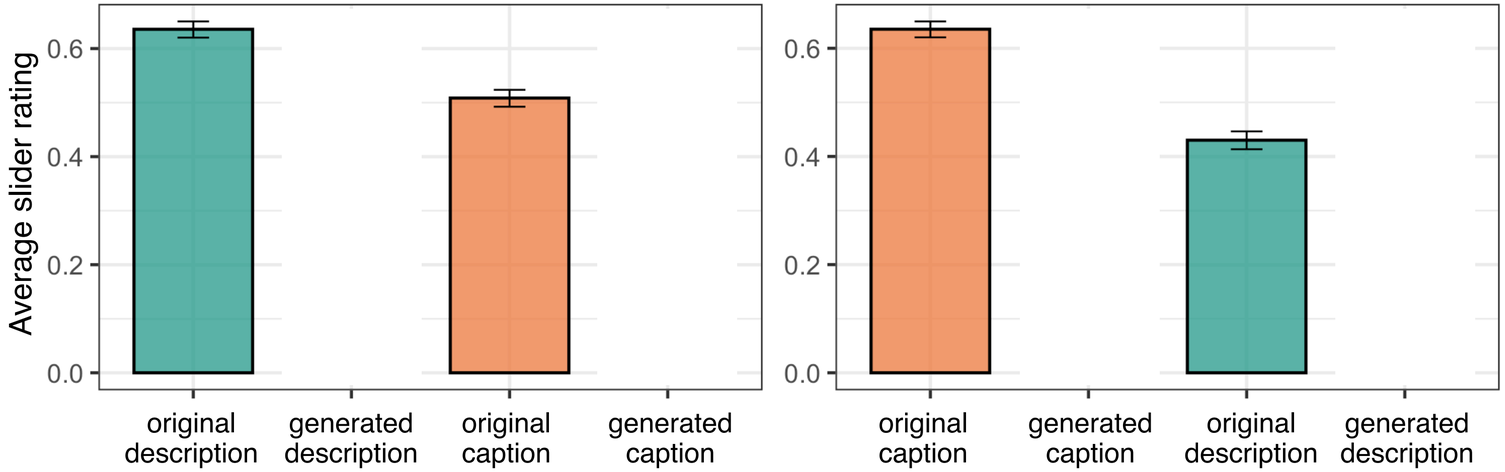


question for descriptive quality

How **useful** would the text alone be to help someone **imagine the picture**?

question for caption quality

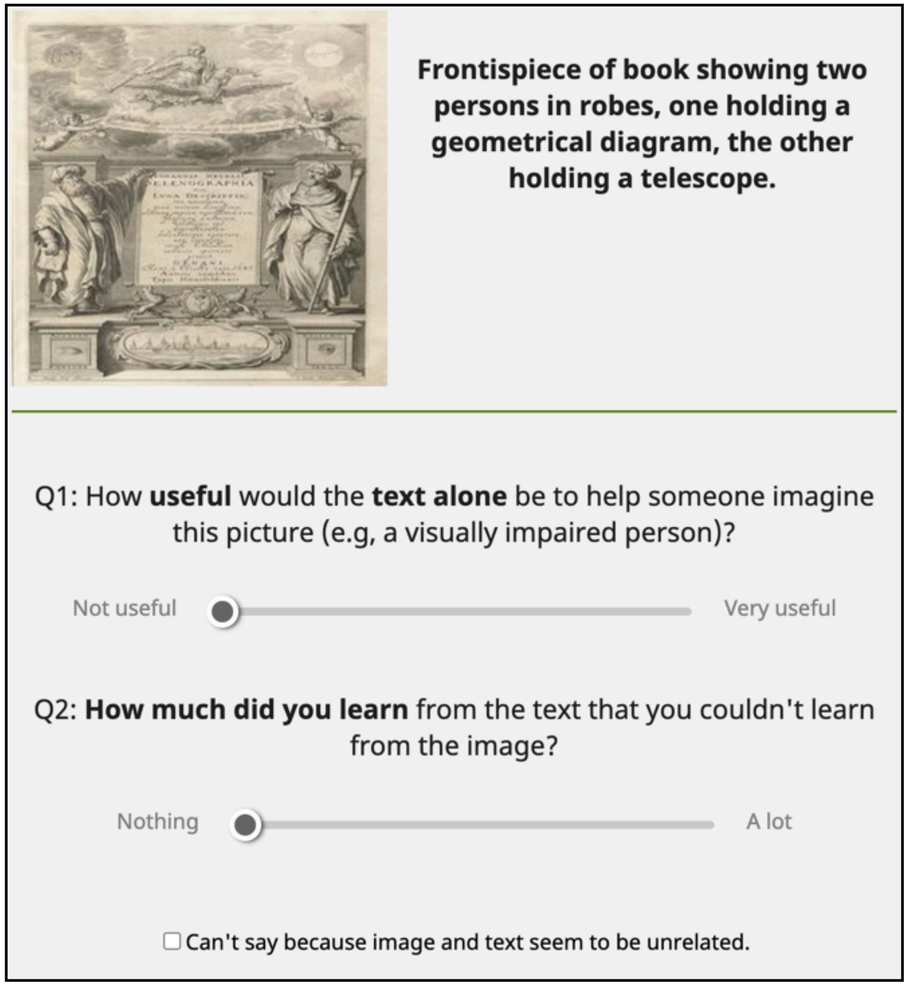
How much did you **learn** from the text that you couldn't learn from the image?



Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Towards AI for Image Accessibility

preregistered human subject experiment

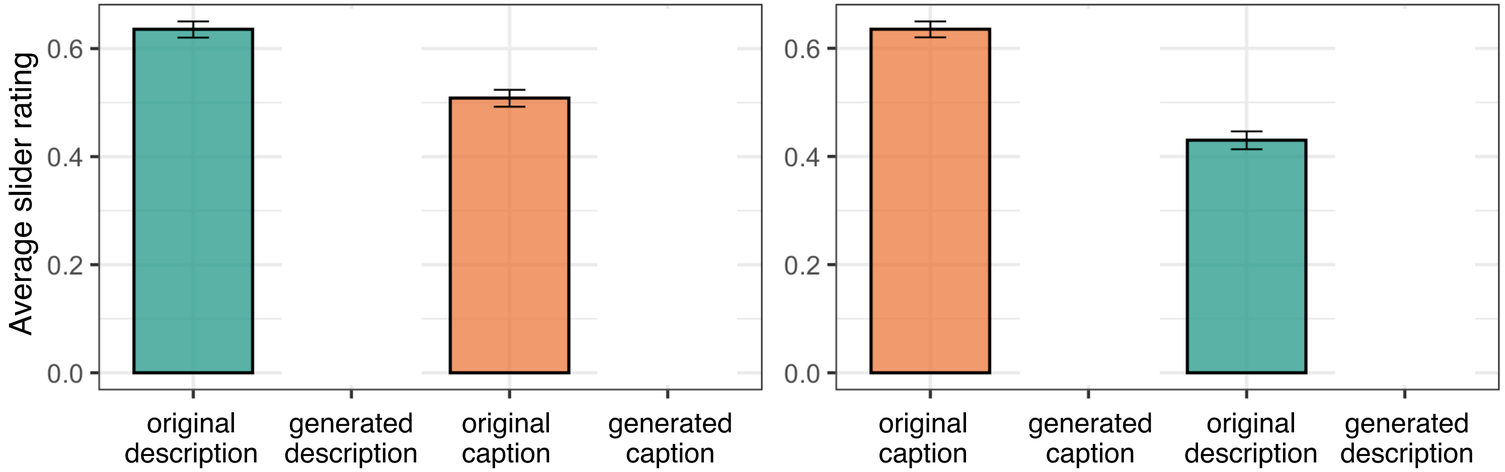


question for descriptive quality

How **useful** would the text alone be to help someone **imagine the picture**?

question for caption quality

How much did you **learn** from the text that you couldn't learn from the image?

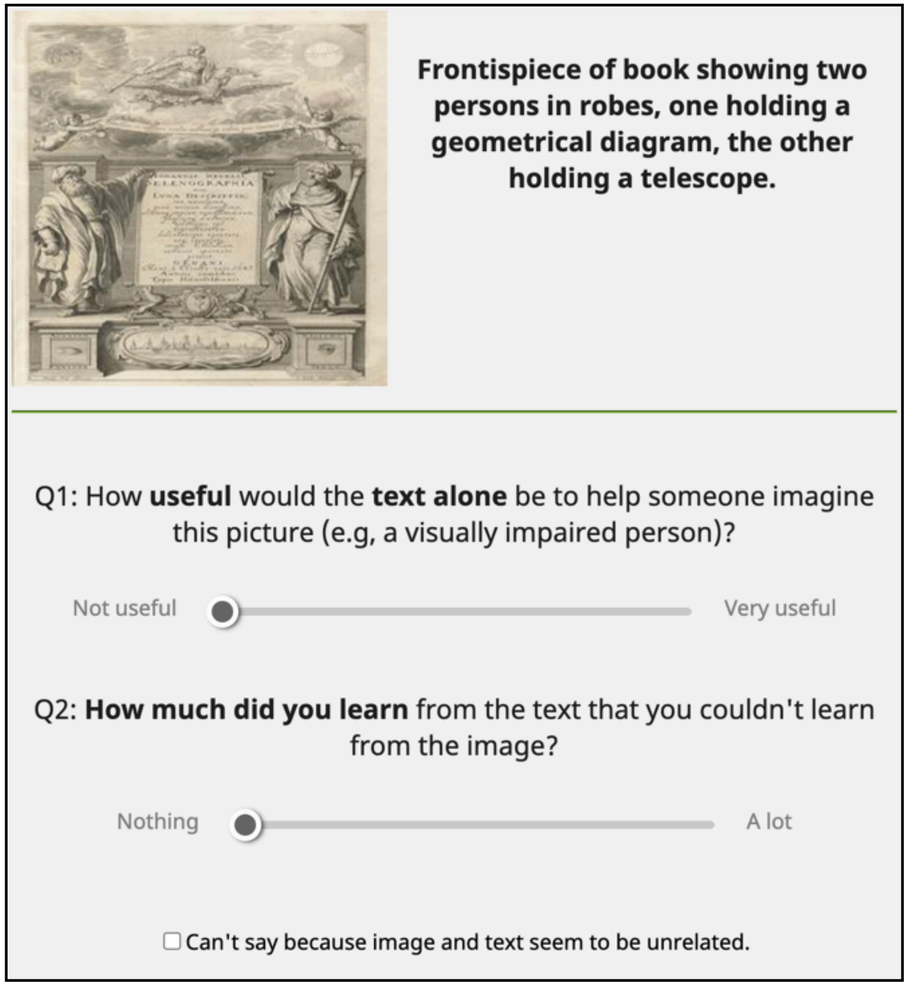


Descriptions and captions optimize for distinct communicative goals.

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Towards AI for Image Accessibility

preregistered human subject experiment

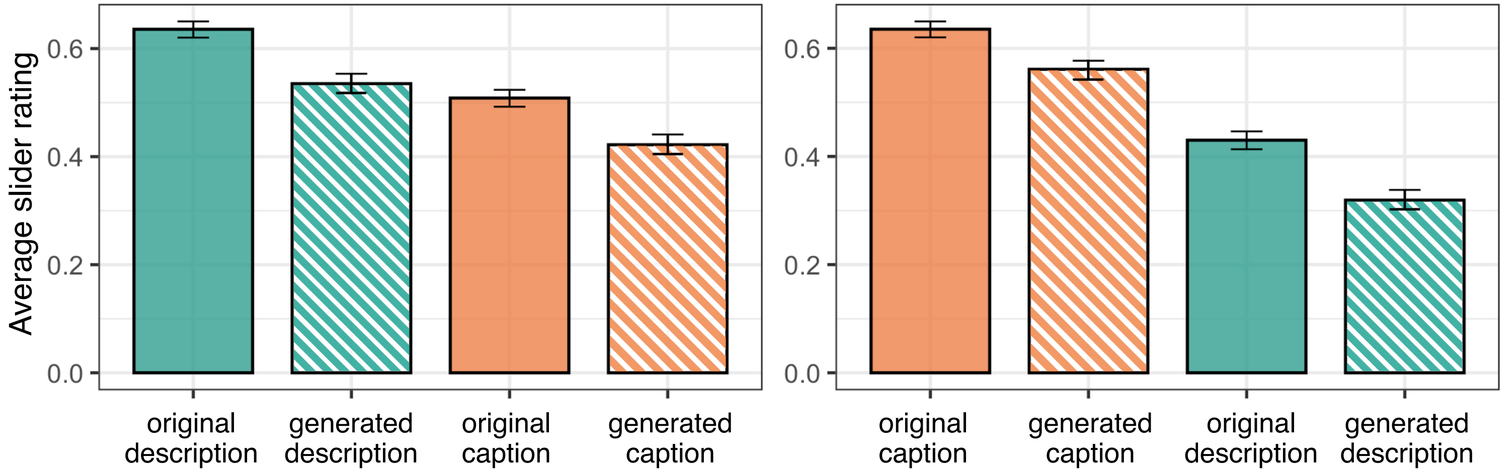


question for descriptive quality

How **useful** would the text alone be to help someone **imagine the picture**?

question for caption quality

How much did you **learn** from the text that you couldn't learn from the image?



Descriptions and captions optimize for distinct communicative goals.

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

The Image-Based **Text**'s Communicative Goal

Large-scale analysis of naturalistic data:

The content of alt descriptions and captions differs in structured ways.

Models of image-based text generation:

Learning alt description and caption generation are distinct challenges.

Human experiment:

Descriptions and captions optimize for distinct communicative goals.

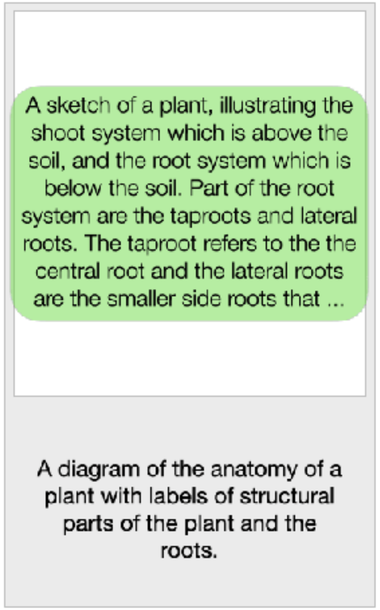
This distinction is reflected in models trained on the respective data.

Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Image-based text generation depends on ...

1 the **image-based text**'s communicative goal.

→ description ≠ caption



2 the **image**'s communicative goal.

Concadia: Towards image-based text generation with a purpose; **Kreiss**, Fang, Goodman, Potts (EMNLP 2022)

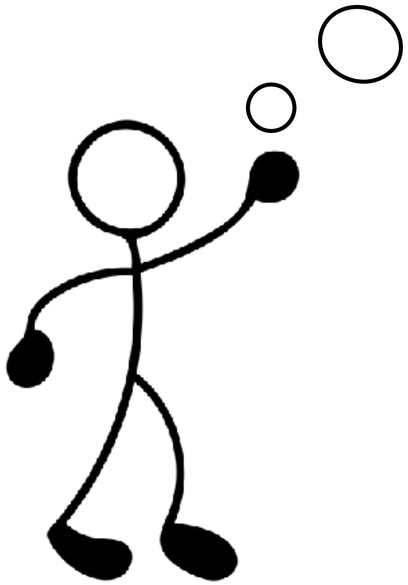
A Pragmatic Approach to Image Descriptions

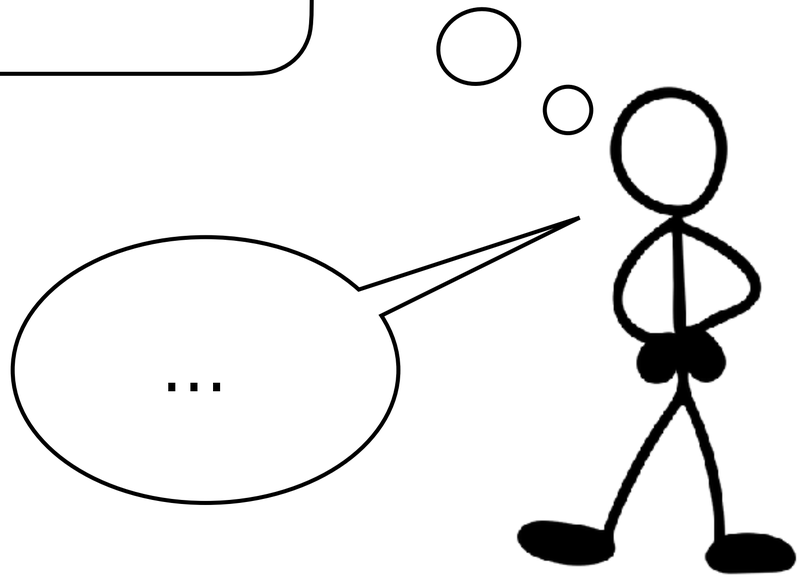
The Gricean Maxims (Grice, 1975)

A Pragmatic Approach to Image Descriptions

The Gricean Maxims

(Grice, 1975)





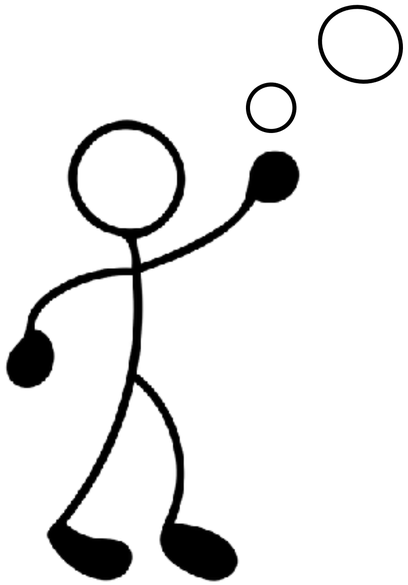
A Pragmatic Approach to Image Descriptions

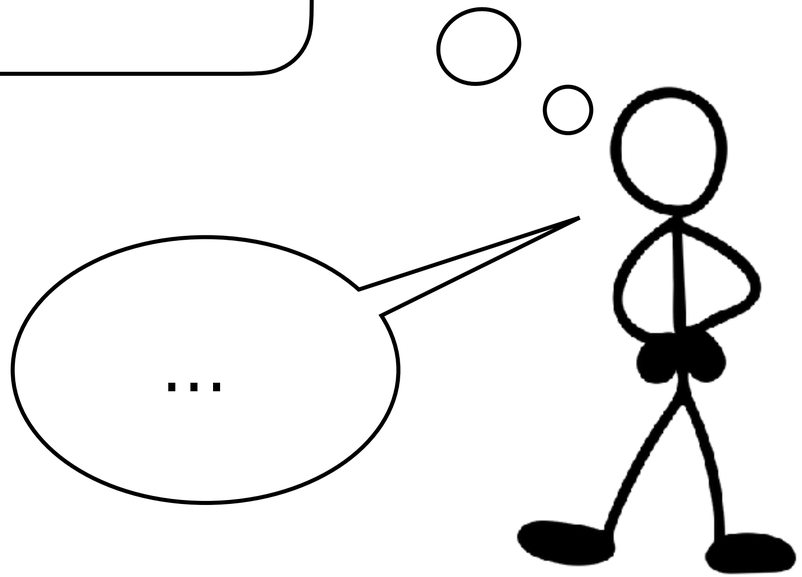
The Gricean Maxims (Grice, 1975)

Maxim of Quantity: Be informative!

(A) Make your contribution as informative as is required (for the current purposes of the exchange).

(B) Do not make your contribution more informative than is required.





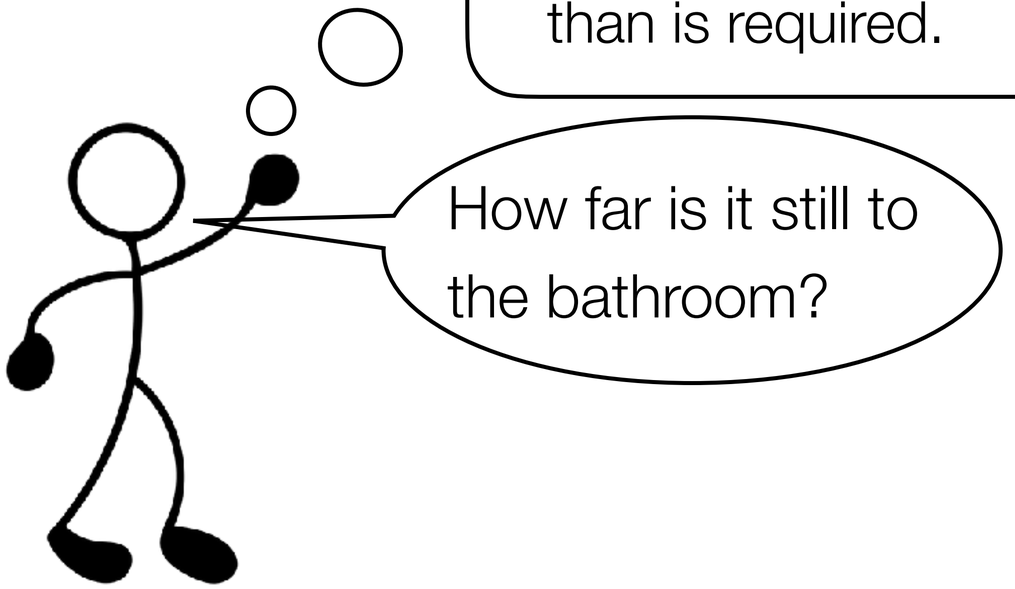
A Pragmatic Approach to Image Descriptions

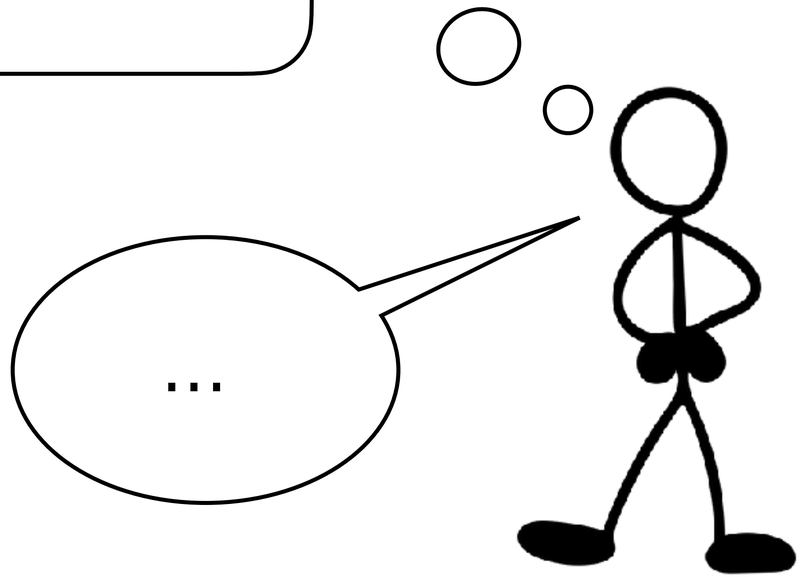
The Gricean Maxims (Grice, 1975)

Maxim of Quantity: Be informative!

(A) Make your contribution as informative as is required (for the current purposes of the exchange).

(B) Do not make your contribution more informative than is required.





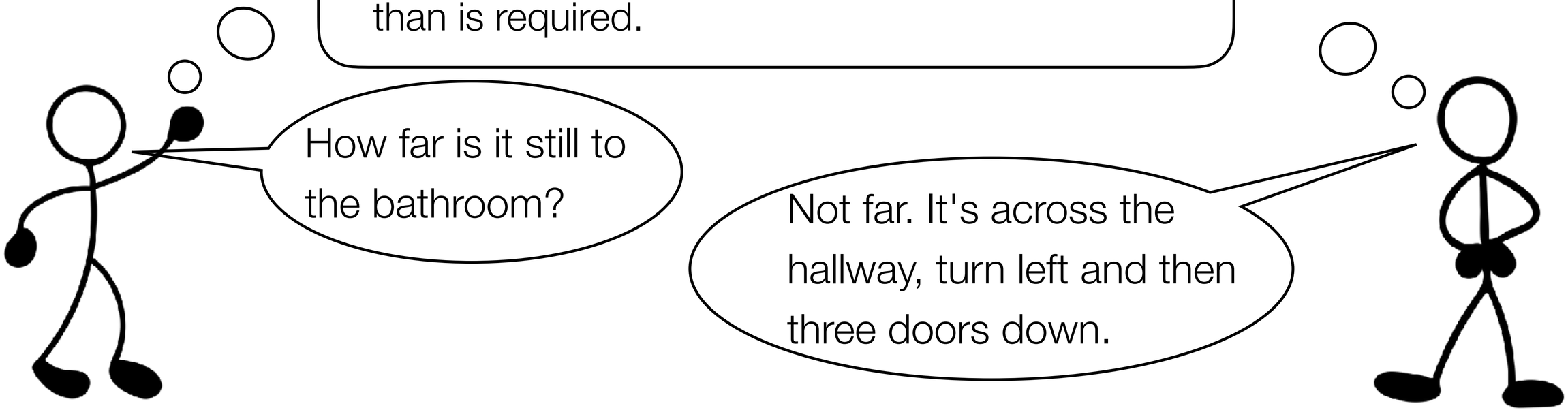
A Pragmatic Approach to Image Descriptions

The Gricean Maxims (Grice, 1975)

Maxim of Quantity: Be informative!

(A) Make your contribution as informative as is required (for the current purposes of the exchange).

(B) Do not make your contribution more informative than is required.



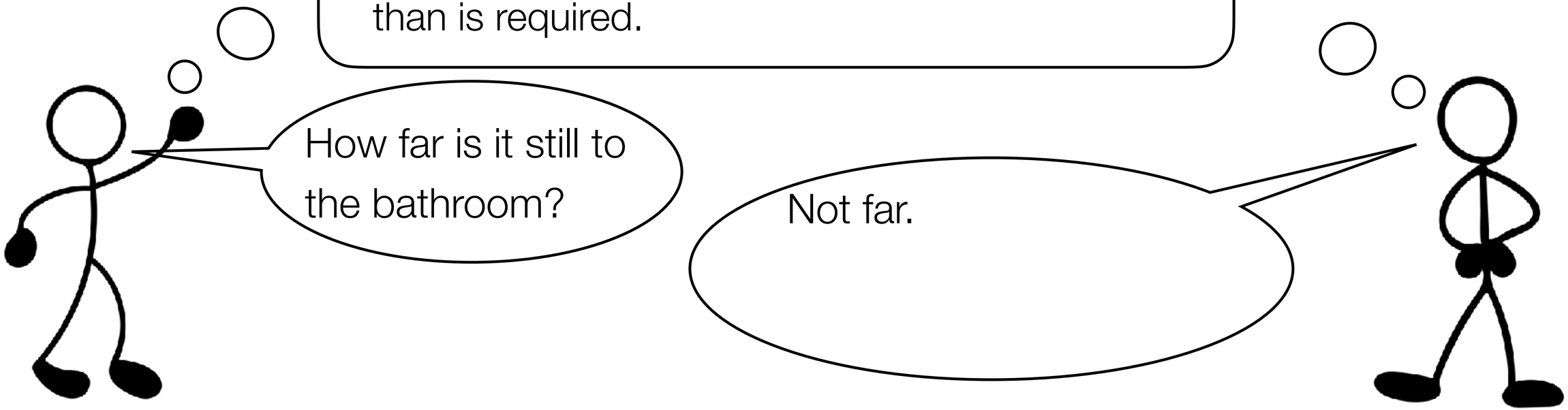
A Pragmatic Approach to Image Descriptions

The Gricean Maxims (Grice, 1975)

Maxim of Quantity: Be informative!

(A) Make your contribution as informative as is required (for the current purposes of the exchange).

(B) Do not make your contribution more informative than is required.



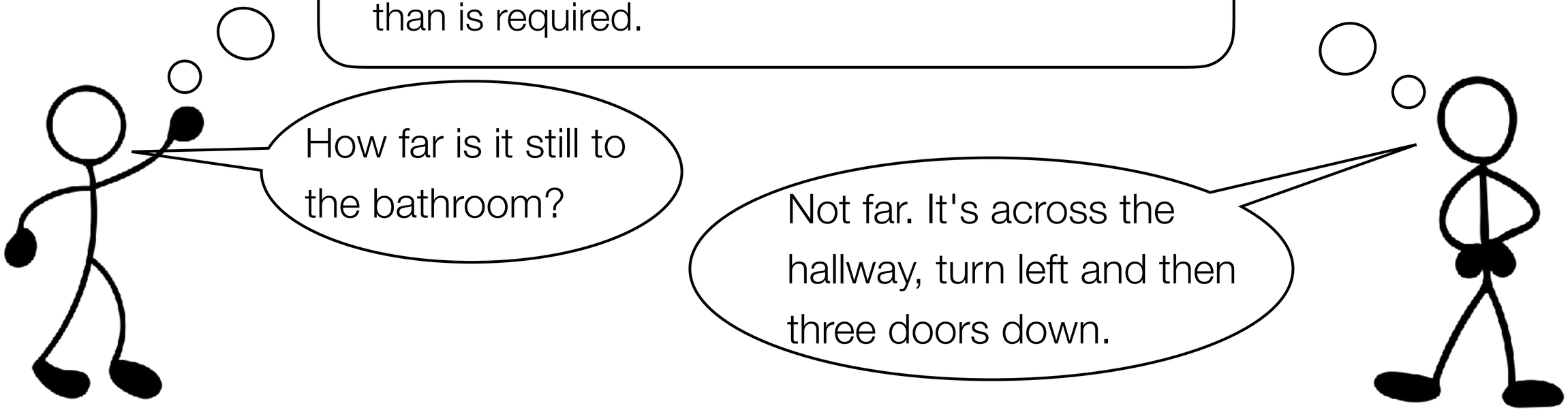
A Pragmatic Approach to Image Descriptions

The Gricean Maxims (Grice, 1975)

Maxim of Quantity: Be informative!

(A) Make your contribution as informative as is required (for the current purposes of the exchange).

(B) Do not make your contribution more informative than is required.



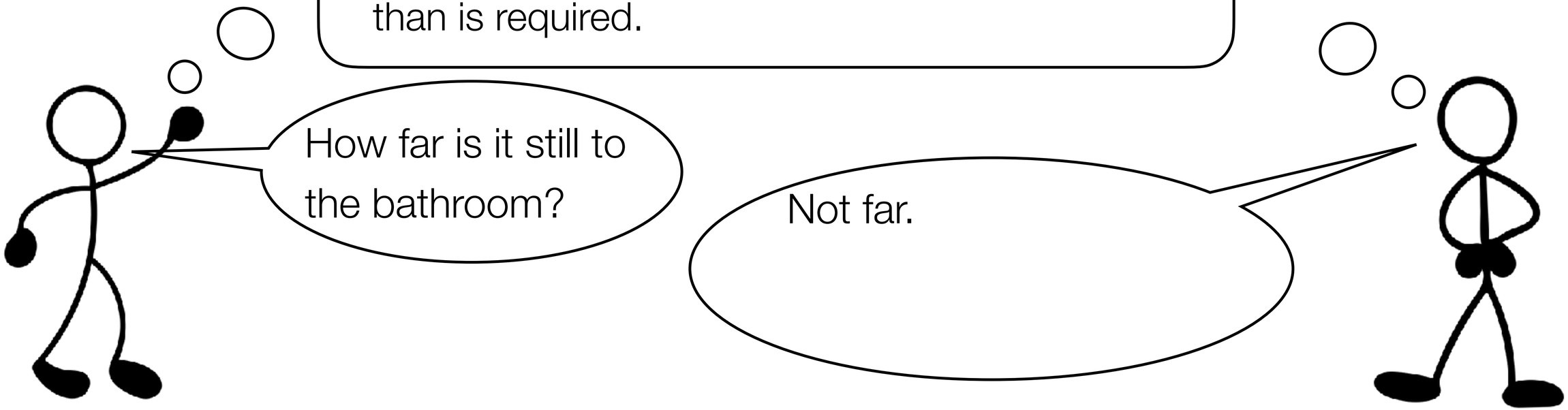
A Pragmatic Approach to Image Descriptions

The Gricean Maxims (Grice, 1975)

Maxim of Quantity: Be informative!

(A) Make your contribution as informative as is required (for the current purposes of the exchange).

(B) Do not make your contribution more informative than is required.



A Pragmatic Approach to Image Descriptions

The Gricean Maxims

Maxim of Quantity: Be informative!

Choosing what is informative depends on **context**.

A Pragmatic Approach to Image Descriptions

The Gricean Maxims

Maxim of Quantity: Be informative!

Choosing what is informative depends on **context**.

There is no one-size-fits-all approach to (accessibility) communication.

[See also: Stangl et al. 2021; Muehlbradt & Kane, 2022; Herskovitz et al. 2023]

The Role of Context for Image Descriptions

WIKIPEDIA
The Free Encyclopedia

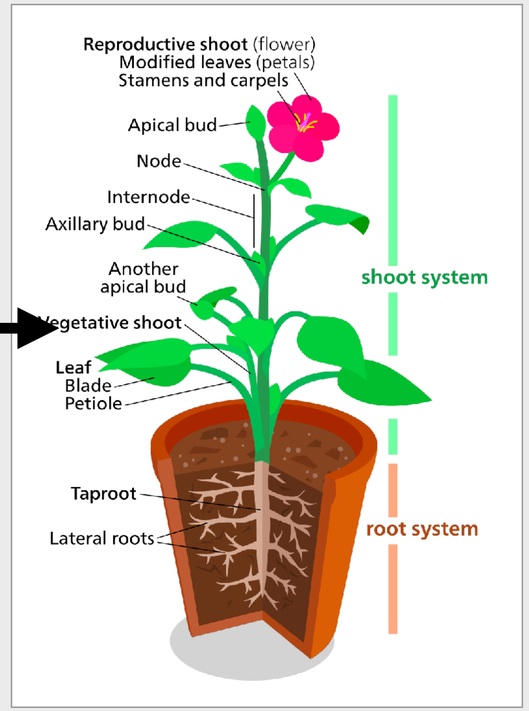
Multimodal Pedagogy

Multimodal pedagogy is an approach to the teaching of writing that implements different modes of communication.[1][2] Multimodality re ... s in differing pieces of media, each nec

The visual mode conveys meaning ... color. The aural mode refers to sound in ... des written and spoken language. The ... t. The gestural mode refers to physical m ... ltimodal text is characterized by the combin

Multimodality as a term was coined ... ing used as early as Egyptian hieroglyphs and classical rhetoric.[7] Compositionists and writing theorists have been exploring how the five modes of communication interact with each other and how multimodality can be used in the teaching of writing since the 20th century.[8]

An educational sketch of a plant, illustrating the shoot system which is above the soil, and the root system which is below the soil. The different parts of the plant are labeled to point out, e.g., the plant's reproductive shoot (i.e., the flower) or the lateral roots.



Example of complementing linguistic and visual information leading to learning benefits over unimodal approaches.

Kreiss, Fang, Goodman, Potts (EMNLP 2022)
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The Role of Context for Image Descriptions

WIKIPEDIA
The Free Encyclopedia

Plant Anatomy

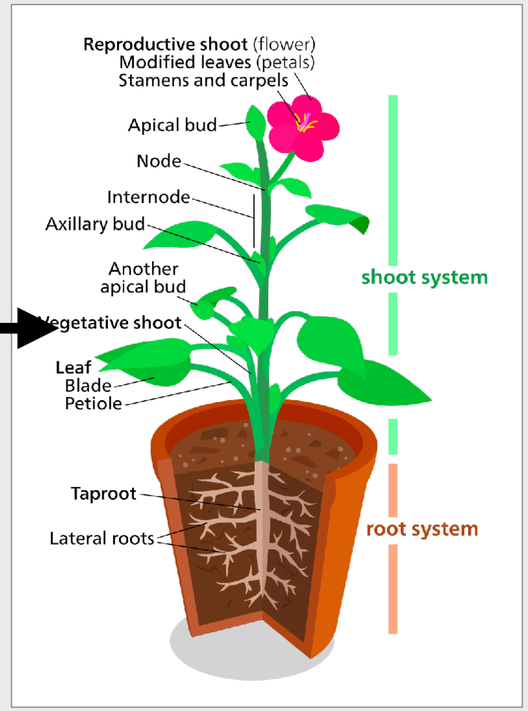
Plant anatomy or **Phytotomy** is the general term for the study of the internal structure of plants. Originally it included plant morphology, the description of the physical form and external structure of plants, but since the mid-20th century plant anatomy has bee][2]
Plant anatomy is now frequently inv es and microscopy.[3]

Structural divisions [edi

Some studies of plant anatomy use ch as nutrient transport, flowering, pollinat ly[5] divided into the following structural categories:

- Flower anatomy, including study of the Calyx, Corolla, Androecium, and Gynoecium
- Leaf anatomy, including study of the Epidermis, stomata and Palisade cells
- Stem anatomy, including Stem structure and vascular tissues, buds and shoot apex

An educational sketch of a plant, illustrating the shoot system which is above the soil, and the root system which is below the soil. The different parts of the plant are labeled to point out, e.g., the plant's reproductive shoot (i.e., the flower) or the lateral roots.



A diagram of the anatomy of a plant with labels of structural parts of the plant and the roots.

Kreiss, Fang, Goodman, Potts (EMNLP 2022)
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The Role of Context for Image Descriptions

WIKIPEDIA
The Free Encyclopedia

Plant Anatomy

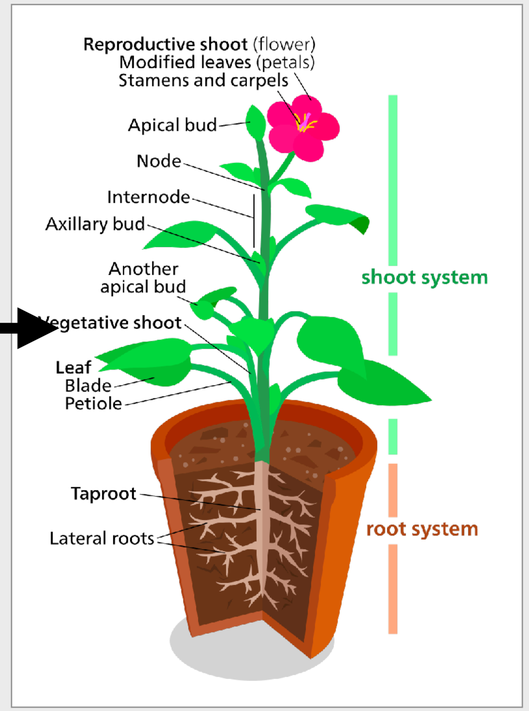
Plant anatomy or **Phytotomy** is the general term for the study of the internal structure of plants. Originally it included plant morphology, the description of the physical form and external structure of plants, but since the mid-20th century plant anatomy has bee][2]
Plant anatomy is now frequently inv es and microscopy.[3]

Structural divisions [edi

Some studies of plant anatomy use ch as nutrient transport, flowering, pollinat ly[5] divided into the following structural categories:

- Flower anatomy, including study of the Calyx, Corolla, Androecium, and Gynoecium
- Leaf anatomy, including study of the Epidermis, stomata and Palisade cells
- Stem anatomy, including Stem structure and vascular tissues, buds and shoot apex

A sketch of a plant, illustrating the shoot system which is above the soil, and the root system which is below the soil. Part of the root system are the taproots and lateral roots. The taproot refers to the central root and the lateral roots are the smaller side roots that ...



A diagram of the anatomy of a plant with labels of structural parts of the plant and the roots.

Kreiss, Fang, Goodman, Potts (EMNLP 2022)
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The **Image**'s Communicative Goal, or: **Context Matters!**

Data

Modeling

Evaluation

The **Image**'s Communicative Goal, or: **Context Matters!**

Data

a dog sitting

cartoon of a dog park

a dinosaur telling a joke

e.g., MSCOCO (Lin et al., 2014); Flickr30k (Hodosh et al. 2013); VizWizCaptions (Gurari et al. 2020)

Modeling

Evaluation

The **Image**'s Communicative Goal, or: **Context Matters!**

Data

image-based text

image-based text

image-based text

e.g., MSCOCO (Lin et al., 2014); Flickr30k (Hodosh et al. 2013); VizWizCaptions (Gurari et al. 2020)

Modeling

↓

Text Generation Model

↓

image-based text

general principle in, e.g., Xu et al. 2015, Li et al. 2020, Hossain et al. 2021, Radford et al. 2021

Evaluation

The **Image**'s Communicative Goal, or: **Context Matters!**

Data

image-based text

image-based text

image-based text

e.g., MSCOCO (Lin et al., 2014); Flickr30k (Hodosh et al. 2013); VizWizCaptions (Gurari et al. 2020)

Modeling

↓

Text Generation Model

↓

image-based text

general principle in, e.g., Xu et al. 2015, Li et al. 2020, Hossain et al. 2021, Radford et al. 2021

Evaluation

image-based text

↓ ↓

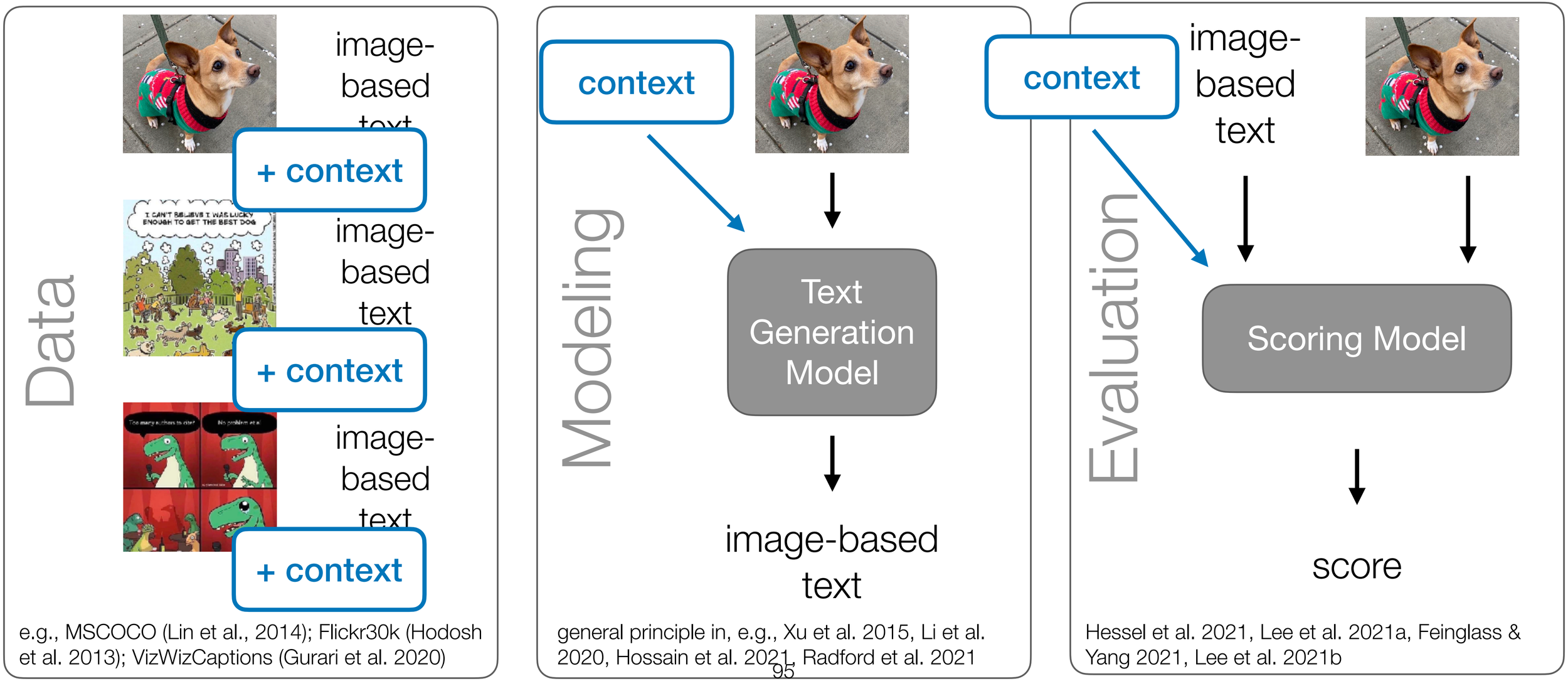
Scoring Model

↓

score

Hessel et al. 2021, Lee et al. 2021a, Feinglass & Yang 2021, Lee et al. 2021b

The **Image**'s Communicative Goal, or: **Context Matters!**



The **Image**'s Communicative Goal, or: **Context Matters!**

Concadia: A naturalistic dataset containing rich contextual information.

96,918 images with captions, alt descriptions and surrounding paragraph from Wikipedia

Contextual information: (1) domain, (2) article title, (3) closest paragraph, (4) description / caption

Wikipedia-Article on **Banana**

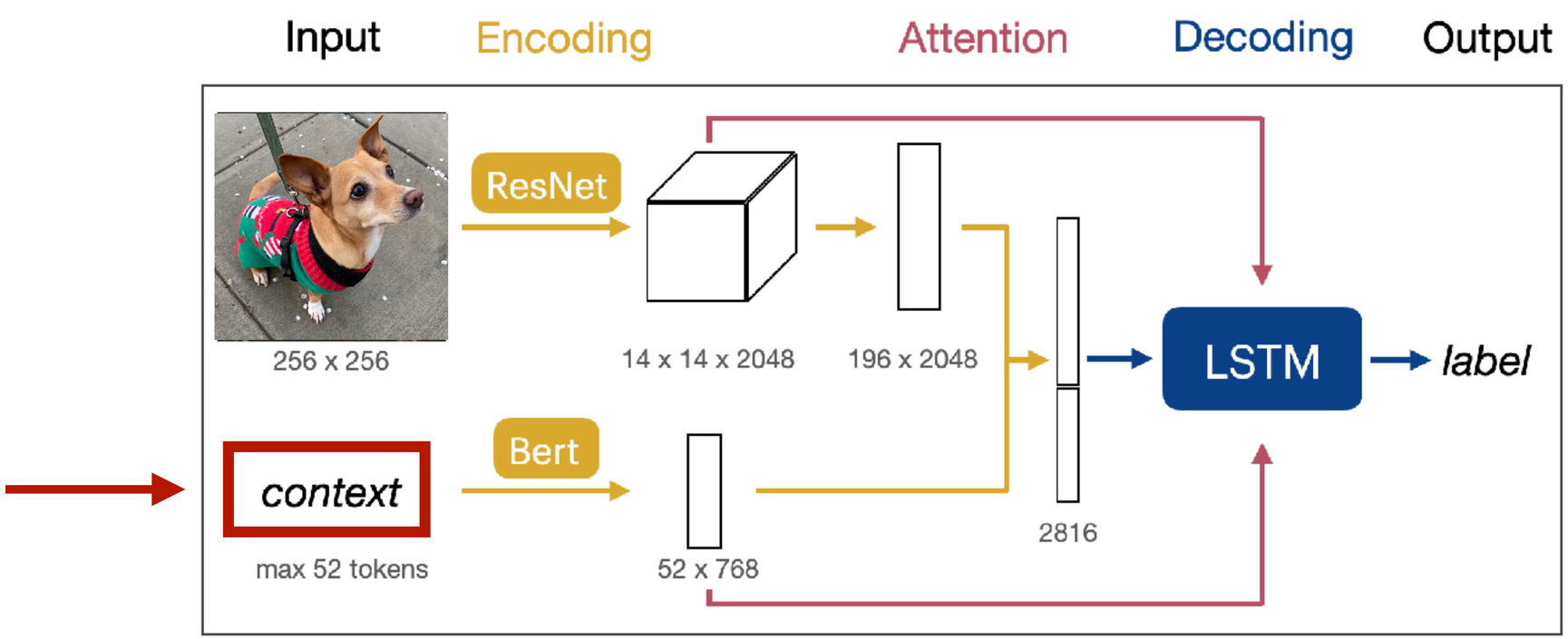
image context: In global commerce in 2009, by far the most important cultivars belonged to the triploid AAA group of Musa acuminata, commonly referred to as Cavendish group bananas. They accounted for the majority of banana exports, despite only coming into existence in 1836. The cultivars Dwarf Cavendish and Grand Nain (Chiquita Banana) gained popularity in the 1950s after the previous mass-produced cultivar, Gros Michel (also an AAA ...

(Accessibility) **Description:**
Grocery store photo of several bunches of bananas

(Contextualizing) **Caption:**
Cavendish bananas are the main commercial banana cultivars sold in the world market.

Making Models Context-Sensitive

ResNet + LSTM **DenseNet + LSTM** OSCAR (VinVL)



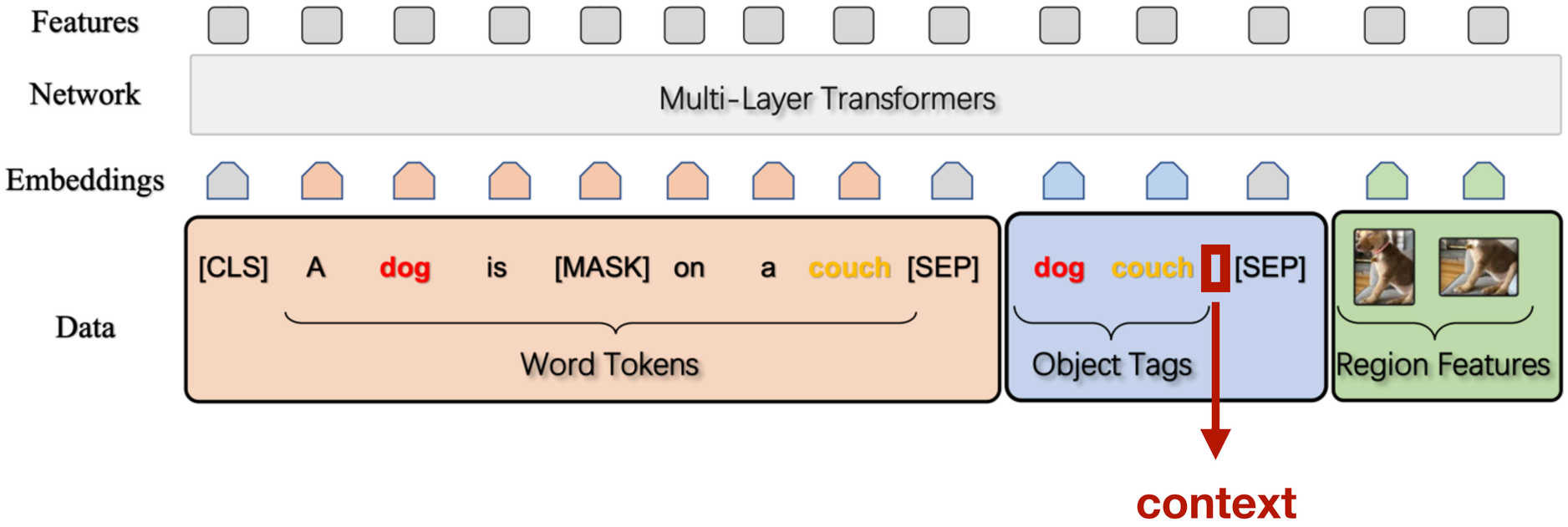
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Making Models Context-Sensitive

ResNet + LSTM

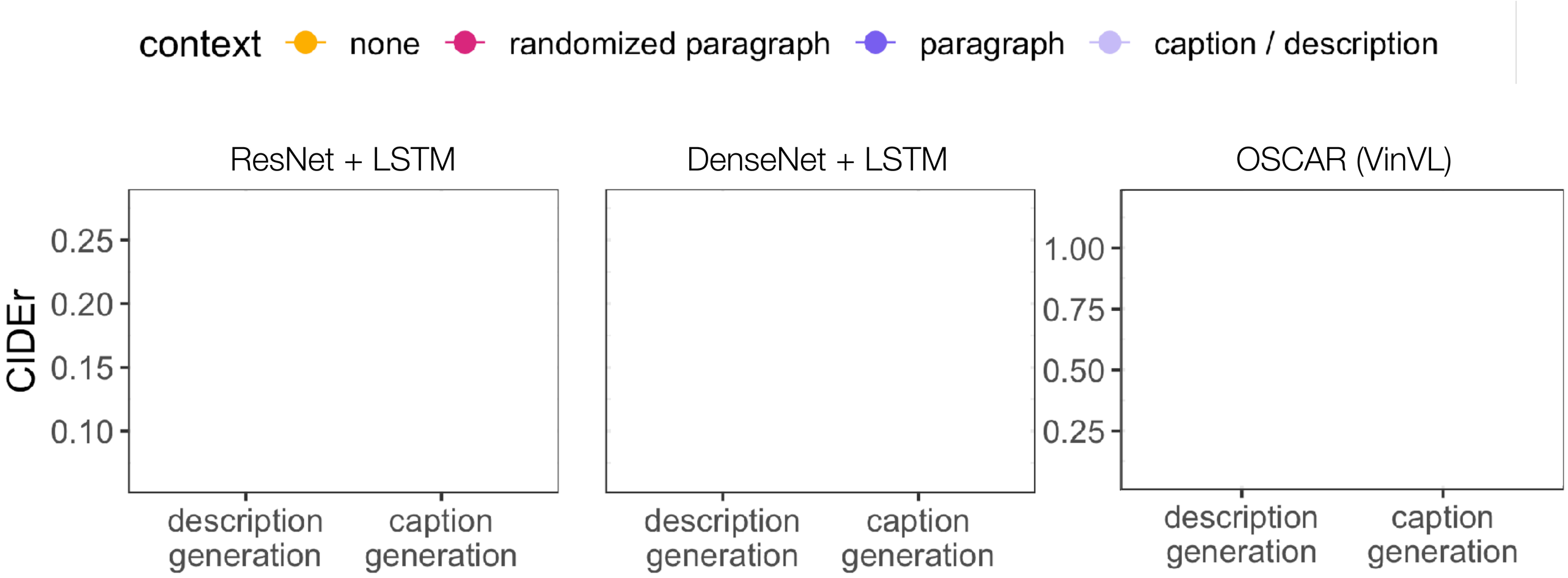
DenseNet + LSTM

OSCAR (VinVL)



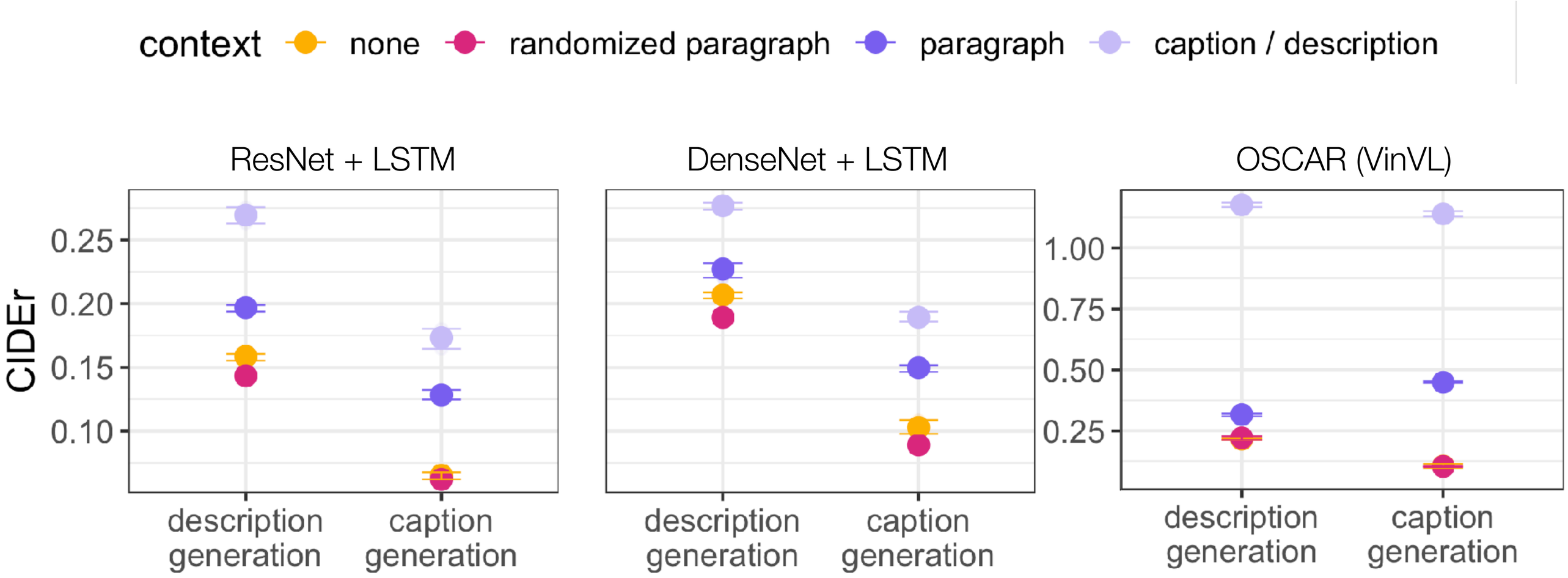
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Context Improves Image-Based Text Generation



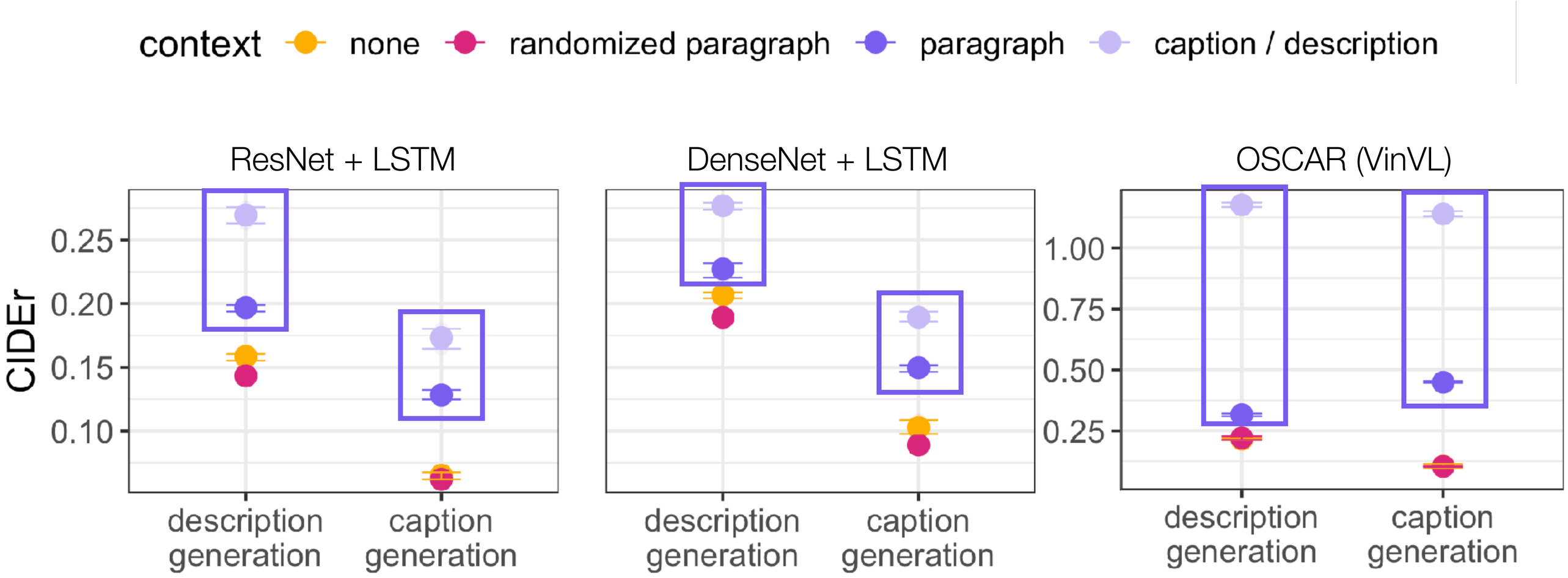
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Context Improves Image-Based Text Generation



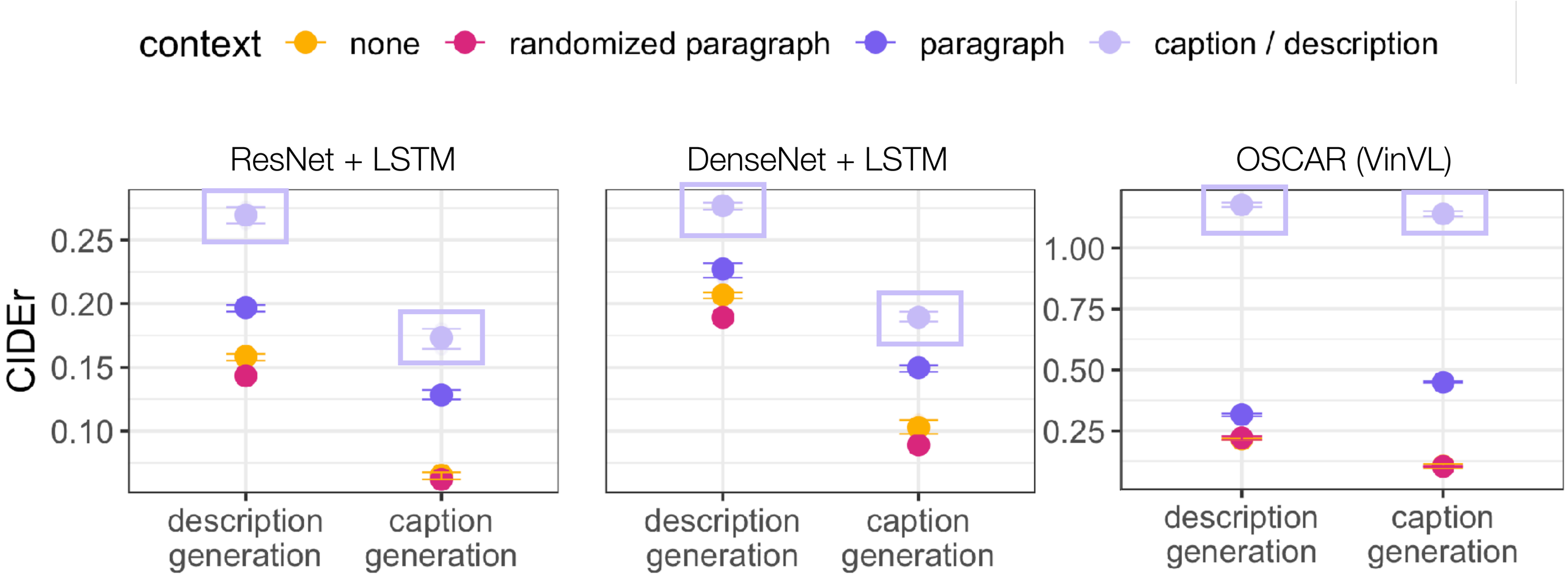
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Context Improves Image-Based Text Generation



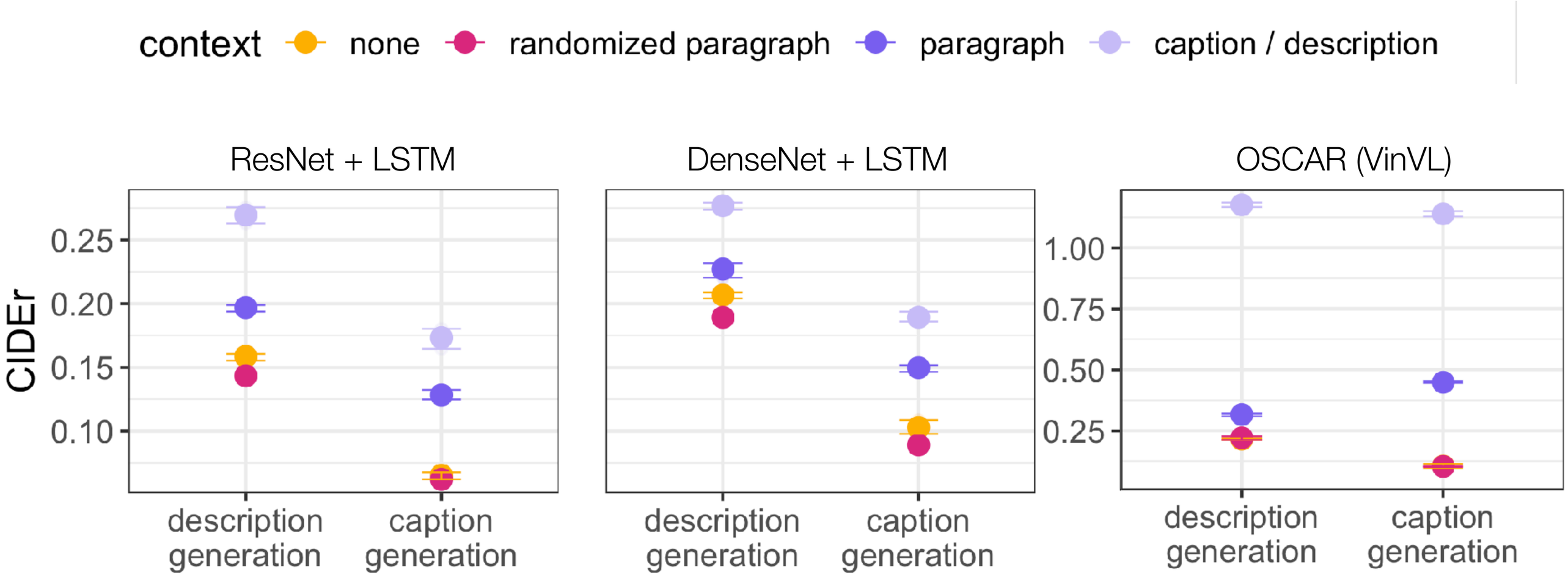
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Context Improves Image-Based Text Generation



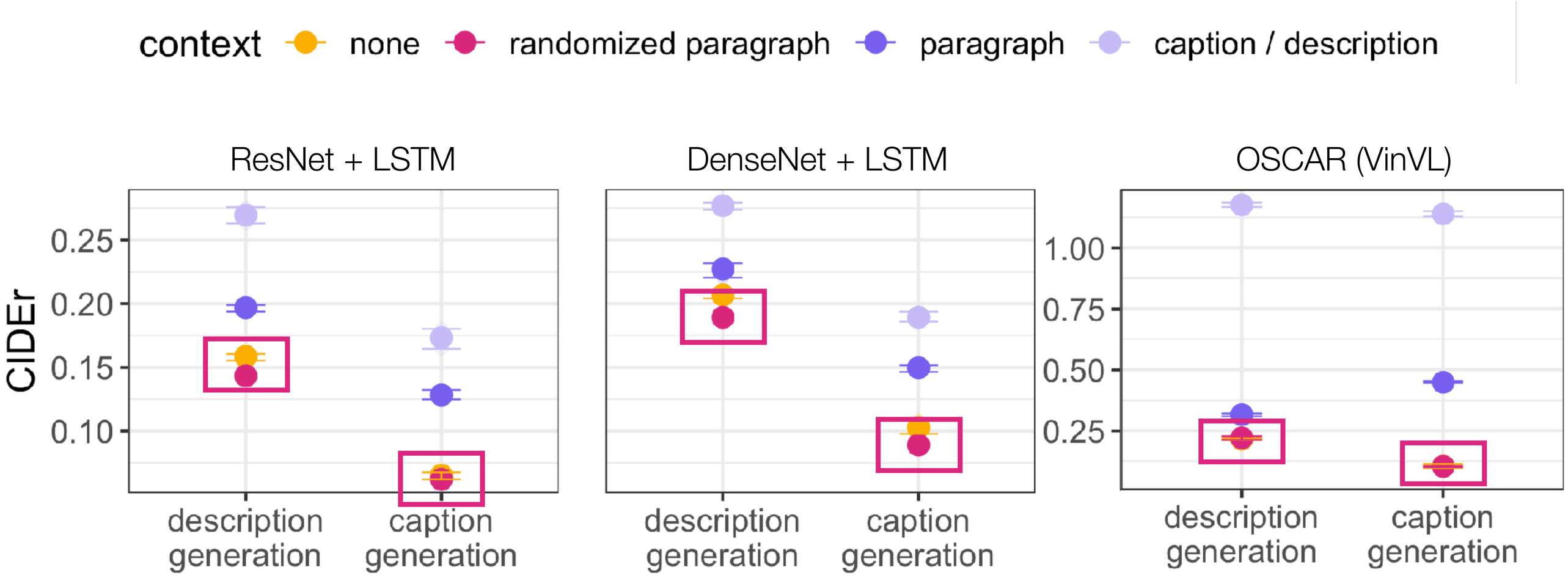
Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Context Improves Image-Based Text Generation



Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

Context Improves Image-Based Text Generation



Concadia: Towards image-based text generation with a purpose
Kreiss, Fang, Goodman, Potts (EMNLP 2022)

The **Image**'s Communicative Goal, or: **Context Matters!**

Concadia: A naturalistic dataset containing rich contextual information.

Models of image-based text generation:

Supplying contextual information benefits model performance.

Model Evaluation

Which model is better?

Neural Network Model (1)

some image-based text output

Neural Network Model (2)

some other! image-based text output

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

Model Evaluation Methods: People

Which model is better?

Neural Network Model (1)

Neural Network Model (2)

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

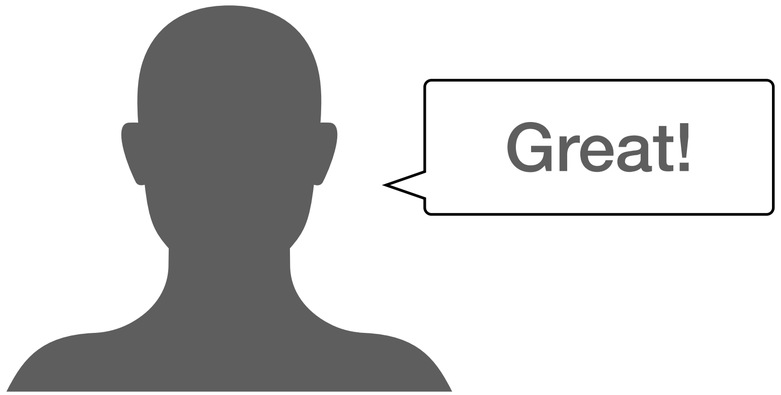
Model Evaluation Methods: People

Which model is better?

Neural Network Model (1)

Neural Network Model (2)

some
image-based text
output



Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

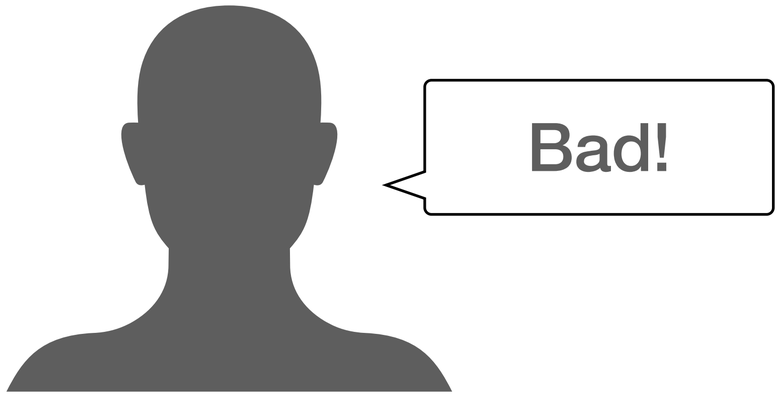
Model Evaluation Methods: People

Which model is better?

Neural Network Model (1)

Neural Network Model (2)

some other!
image-based text output



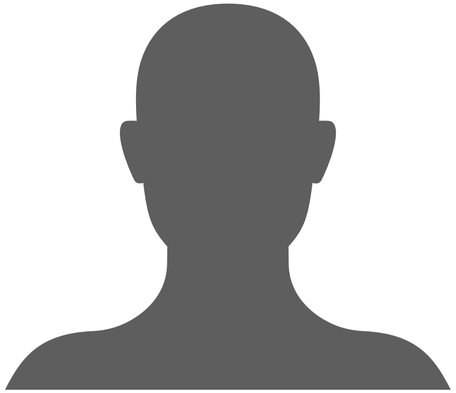
Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

Model Evaluation Methods: People

Which model is better?

Neural Network Model (1)

Neural Network Model (2)



Pro: Effective!

Con: Slow and costly → not feasible for model development

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

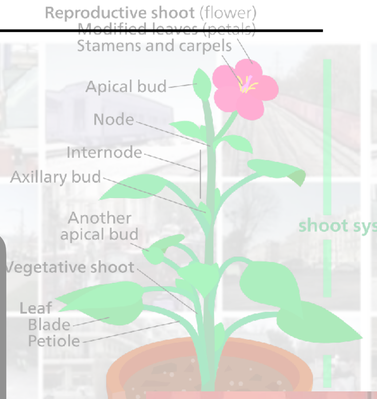
Model Evaluation Methods: Automatic

small dog sitting outside

Cartoon of various dogs and people in a park and all people thinking the same thing: I can't believe I was lucky enough to get the best dog.

Grocery store photo of several bunches of bananas

Model



an educational sketch of a plant's anatomy, including the shoot and root

a dinosaur telling a joke

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

Model Evaluation Methods: Automatic

some image-based text

Evaluation Model

Score

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

Model Evaluation Methods: Automatic

State-of-the-art metric (e.g., CLIPScore)

Hessel et al. 2021, Kasai et al. 2021

some image-based text

Evaluation Model

Score

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

Model Evaluation Methods: Automatic

State-of-the-art metric (e.g., CLIPScore): Inherently context-**in**dependent

some image-based text

Evaluation Model

Score

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The Role of Context for Evaluation

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The Role of Context for Evaluation

Building Material

Building material is material used for construction. Many naturally occurring substances, such as clay, rocks, sand and wood, even twigs and leaves, have been used to construct buildings. Apart from naturally occurring materials, ...

Christian Cross

The Christian cross, seen as a representation of the instrument of the crucifixion of Jesus, is the best-known symbol of Christianity. It is related to the crucifix (a cross that includes a corpus, usually a three-dimensional ...

Roof

A roof is the top covering of a building, including all materials and constructions necessary to support it on the walls of the building or on uprights, providing protection against rain, snow, sunlight, extremes of temperature, and wind. A ...

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The Role of Context for Evaluation

Building Material

Building material is material used for construction. Many naturally occurring substances, such as clay, rocks, sand and wood, even twigs and leaves, have been used to construct buildings. Apart from naturally occurring materials, ...

Christian Cross

The Christian cross, seen as a representation of the instrument of the crucifixion of Jesus, is the best-known symbol of Christianity. It is related to the crucifix (a cross that includes a corpus, usually a three-dimensional ...

Roof

A roof is the top covering of a building, including all materials and constructions necessary to support it on the walls of the building or on uprights, providing protection against rain, snow, sunlight, extremes of temperature, and wind. A ...

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The Role of Context for Evaluation

Building Material

Building material is material used for construction. Many naturally occurring substances, such as clay, rocks, sand and wood, even twigs and leaves, have been used to construct buildings. Apart from naturally occurring materials, ...

How useful is this description for making the image accessible?

participants

blind & low vision

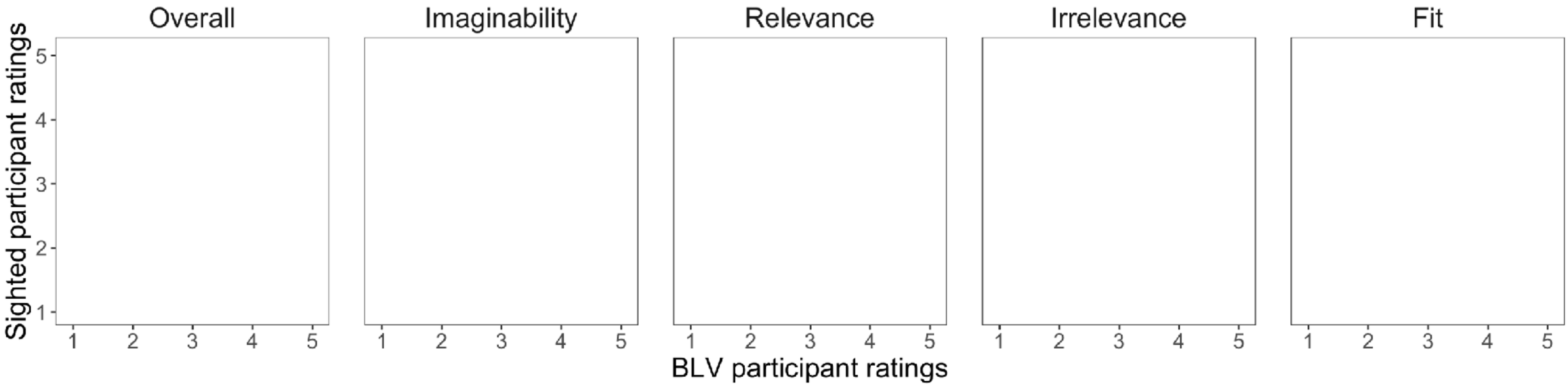
Score

sighted

Score

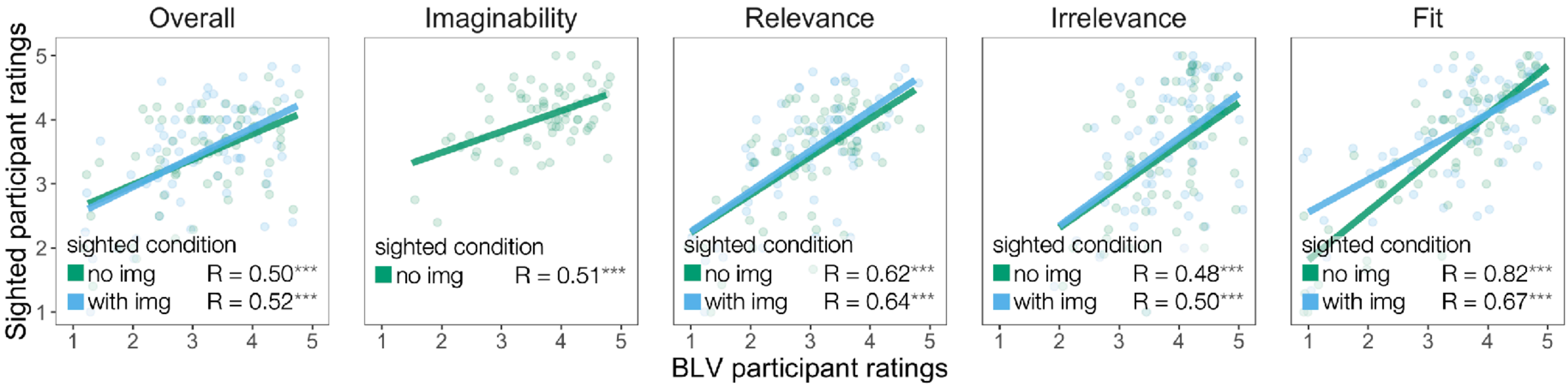
Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The Role of Context for Evaluation



Kreiss et al. (EMNLP 2022b)

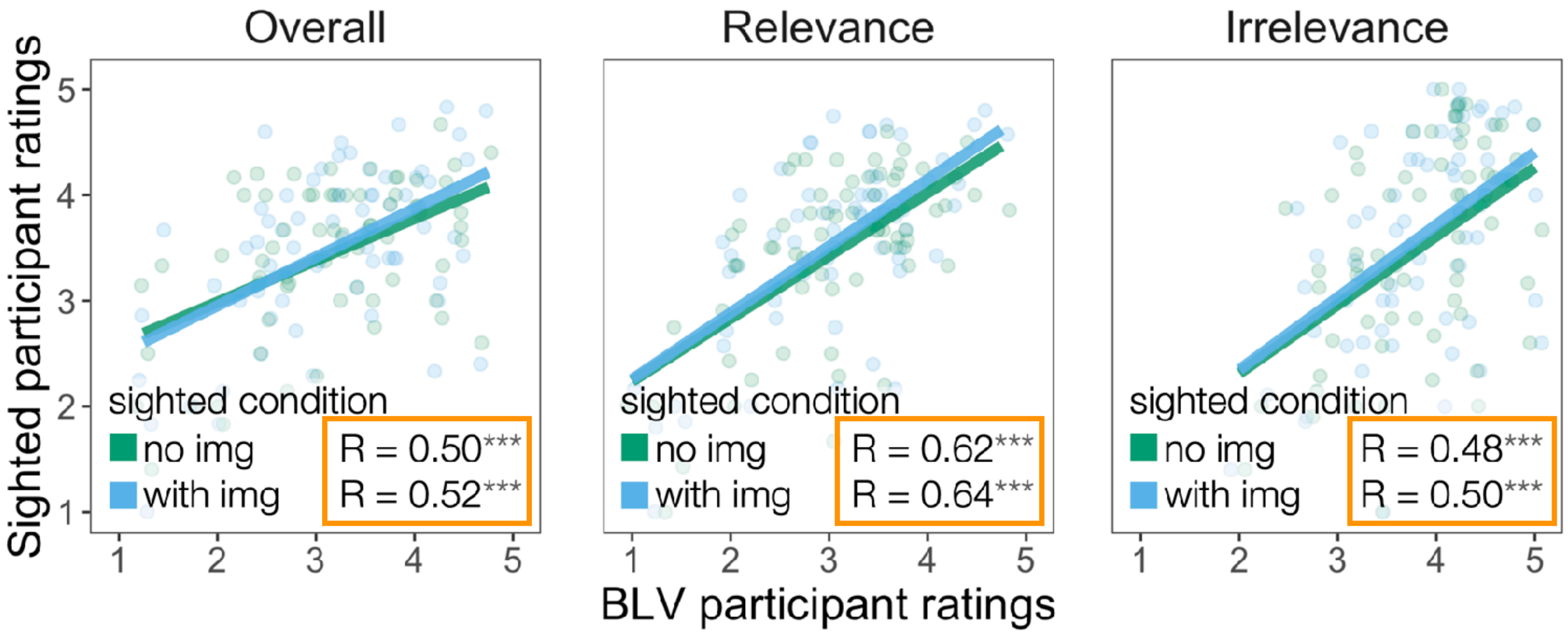
The Role of Context for Evaluation



Kreiss et al. (EMNLP 2022b)

The Role of Context for Evaluation

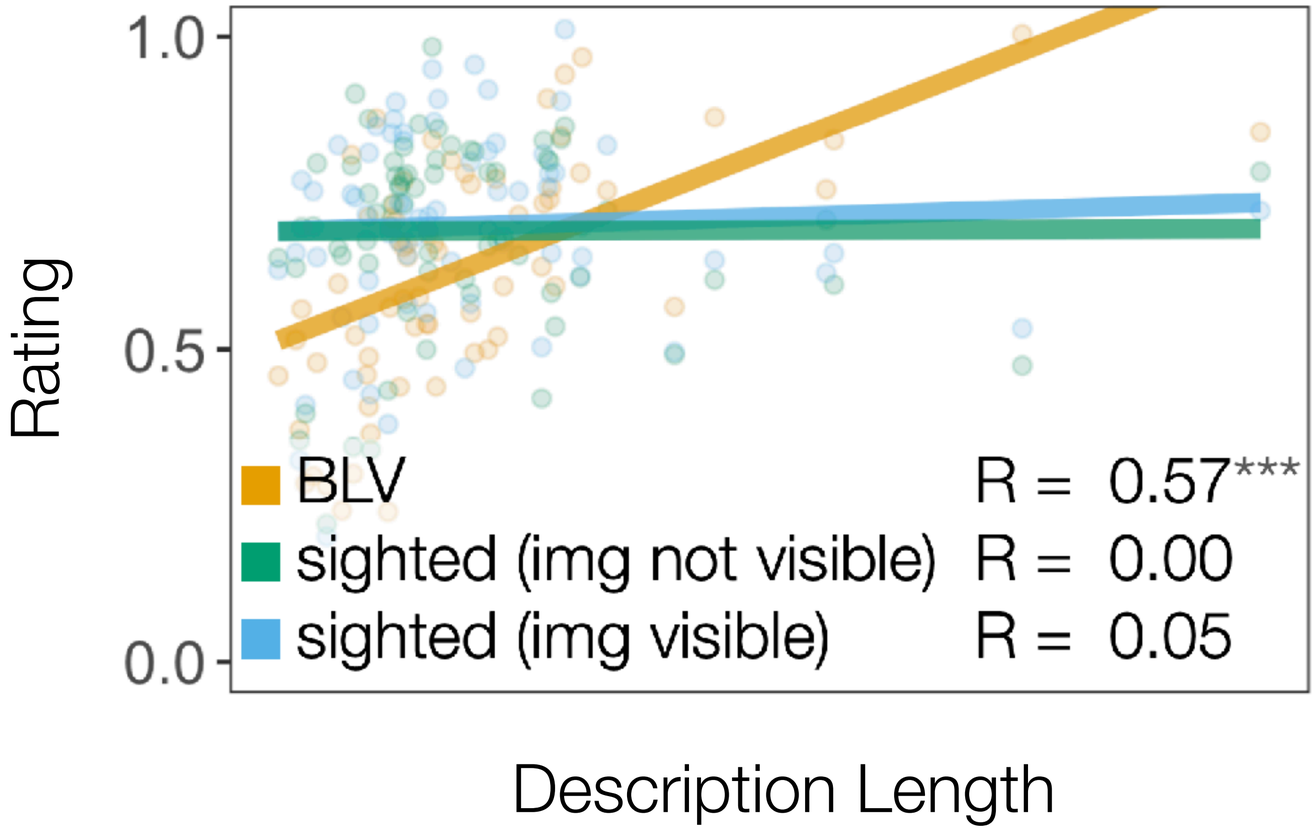
Hiding the image from sighted participants doesn't improve their alignment with BLV participant ratings.



Kreiss et al. (EMNLP 2022b)

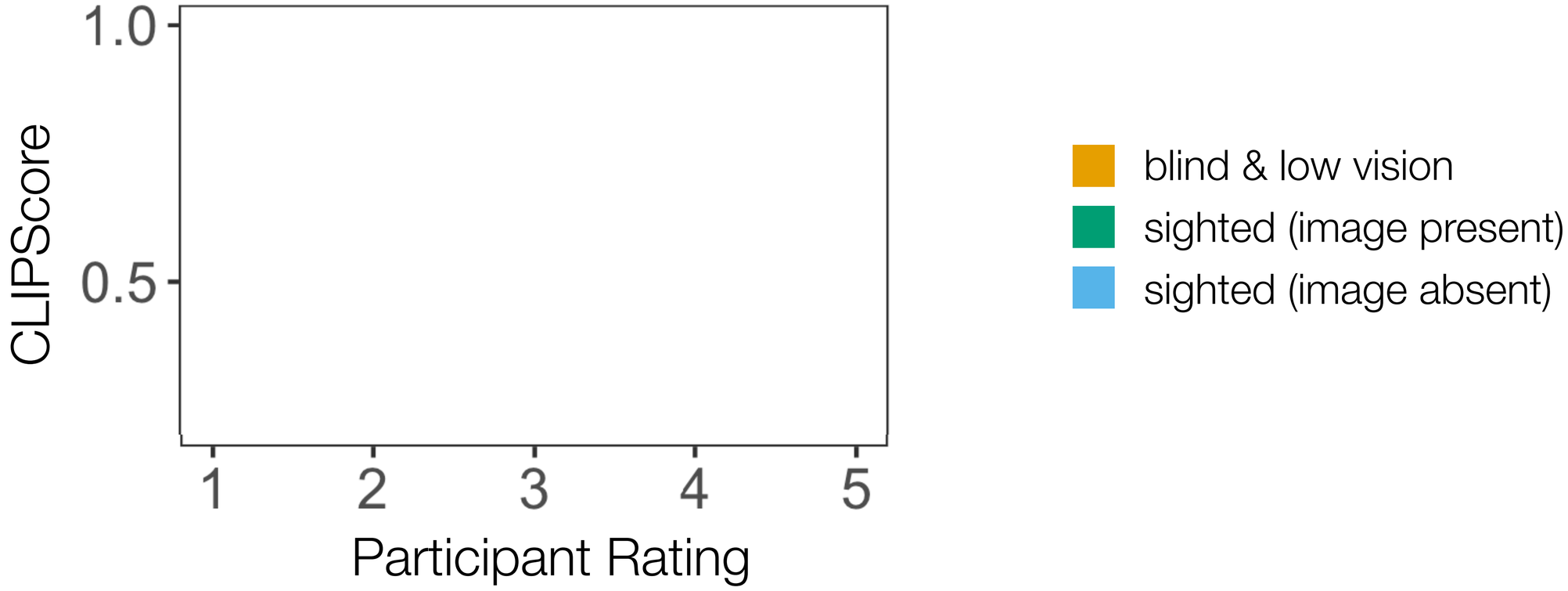
The Role of Context for Evaluation

BLV participant ratings (but **not** sighted participant ratings) show a strong correlation with description length



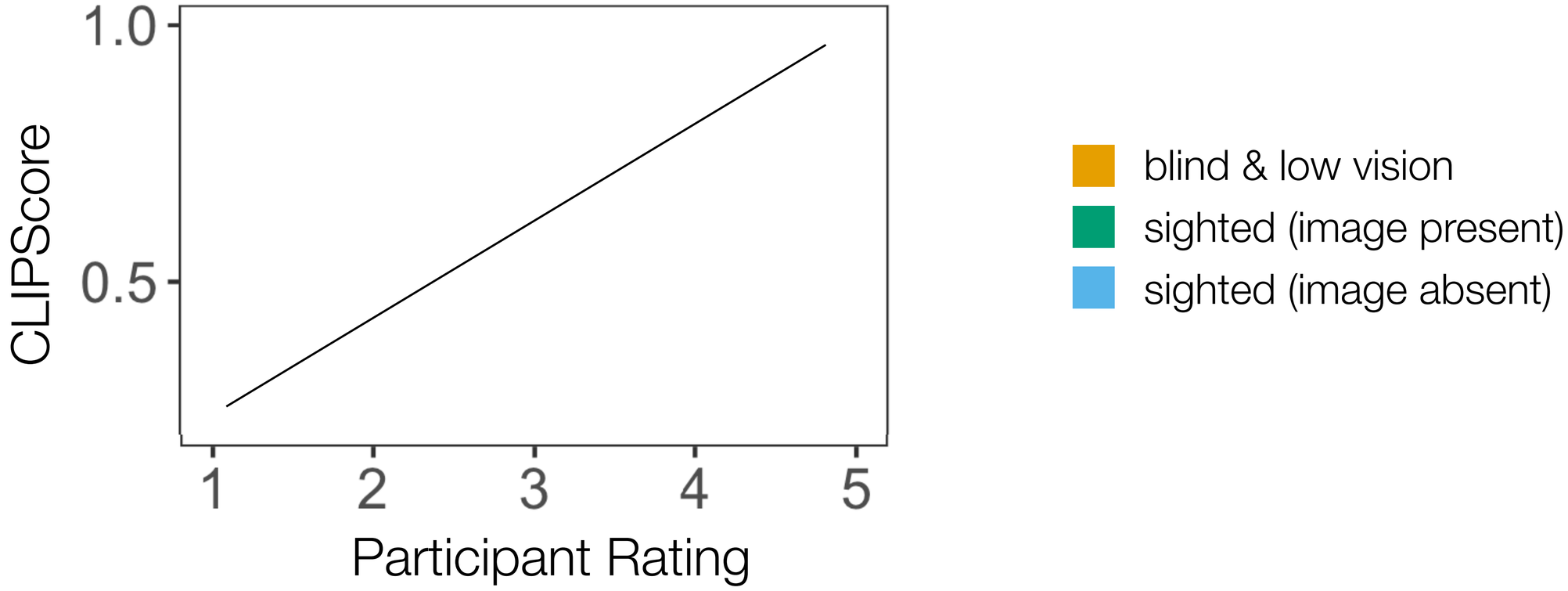
Kreiss et al. (EMNLP 2022b)

The Role of Context for Evaluation



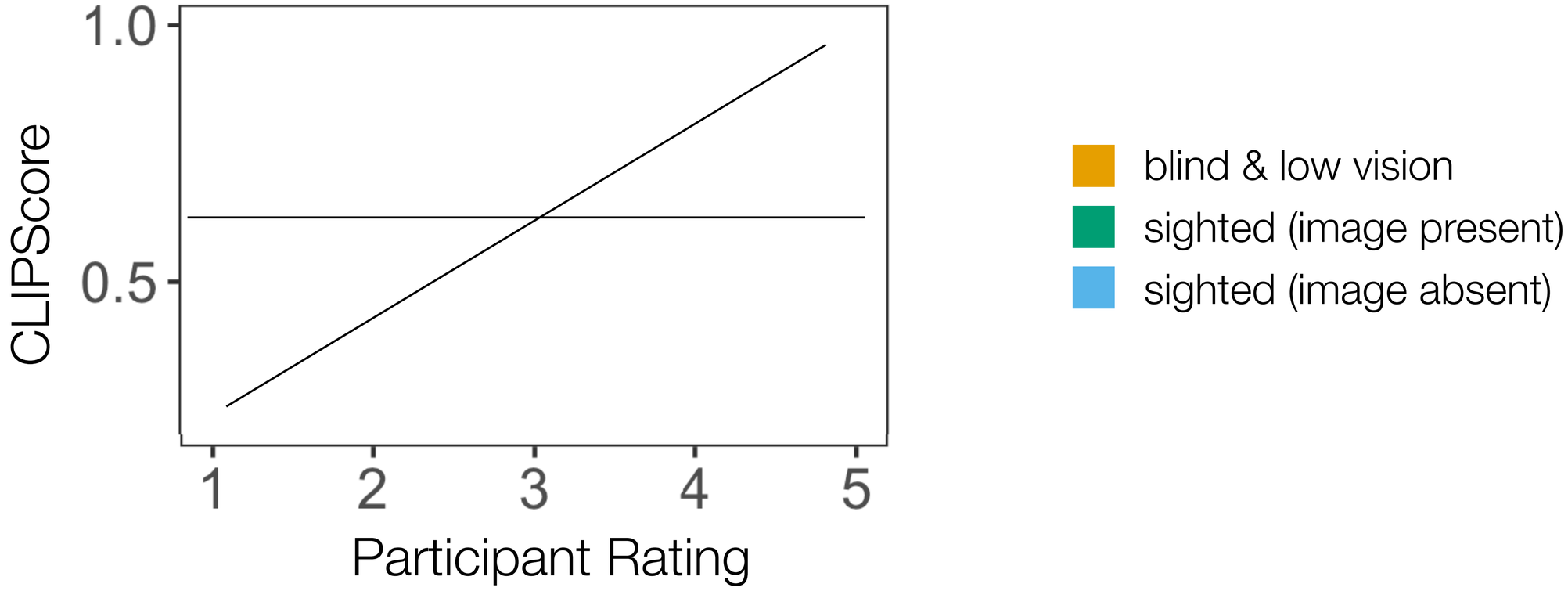
Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The Role of Context for Evaluation



Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

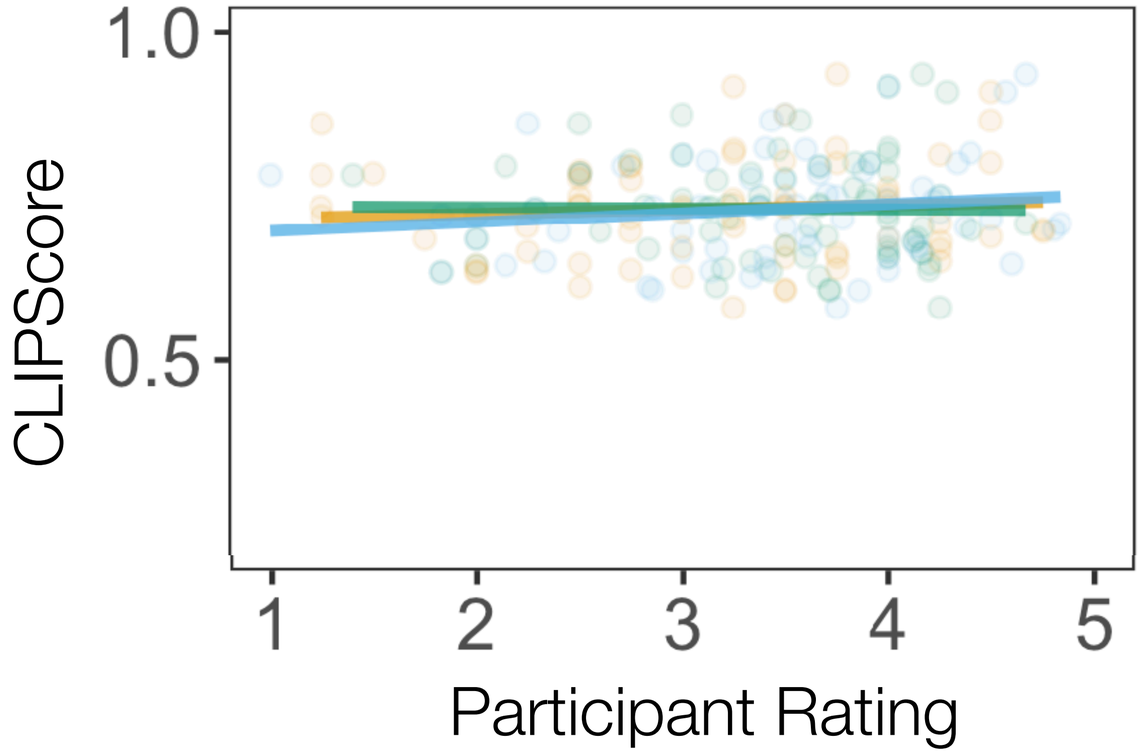
The Role of Context for Evaluation



Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

The Role of Context for Evaluation

No correlation between CLIPScore and human ratings



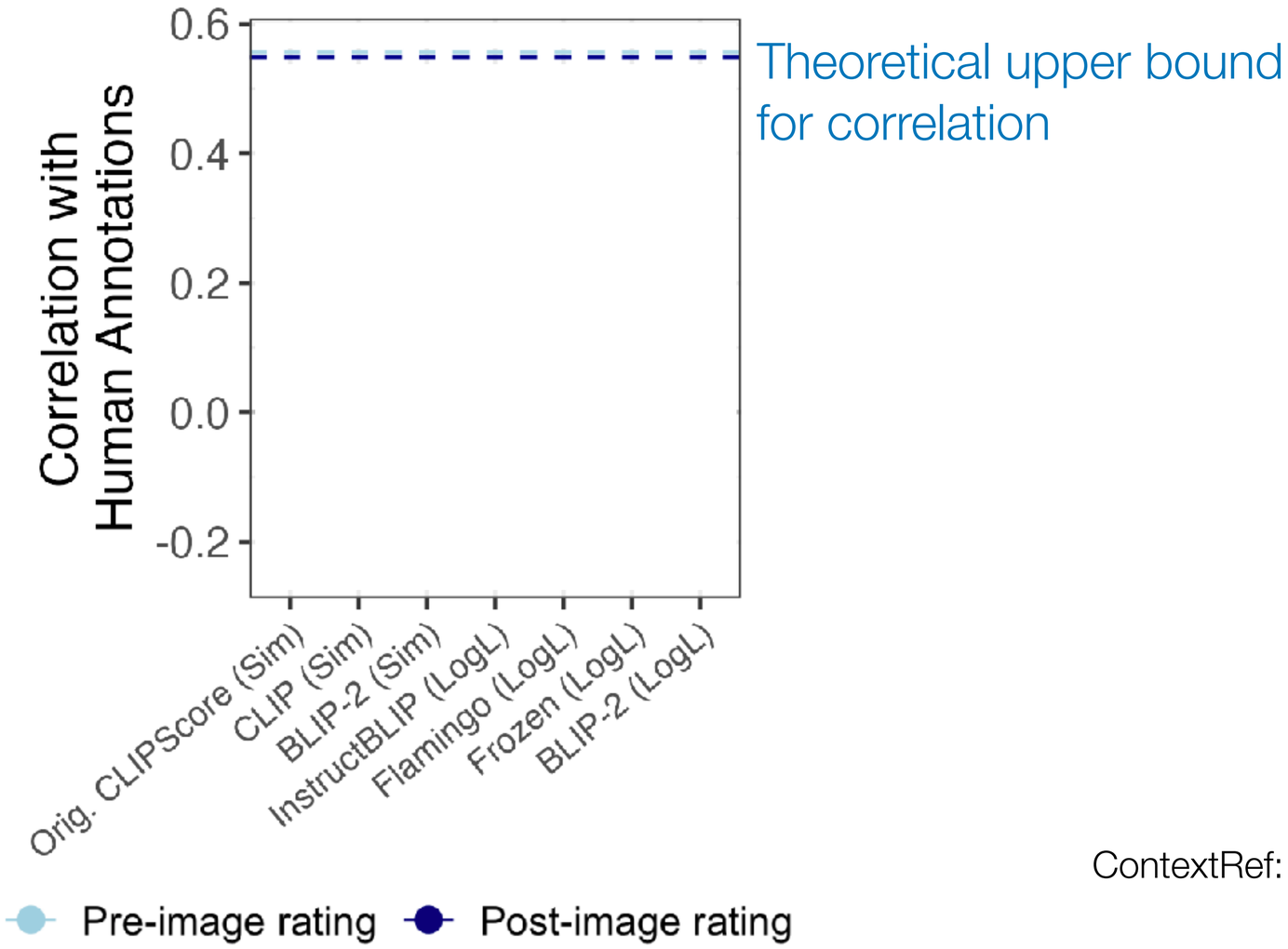
blind & low vision	R = 0.08	n.s.
sighted (image present)	R = -0.01	n.s.
sighted (image absent)	R = 0.14	n.s.

Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics
Kreiss, Bennett, Hooshmand, Zelikman, Morris, Potts (EMNLP 2022)

Models for Human-like Image Description Evaluation

Models for Human-like Image Description Evaluation

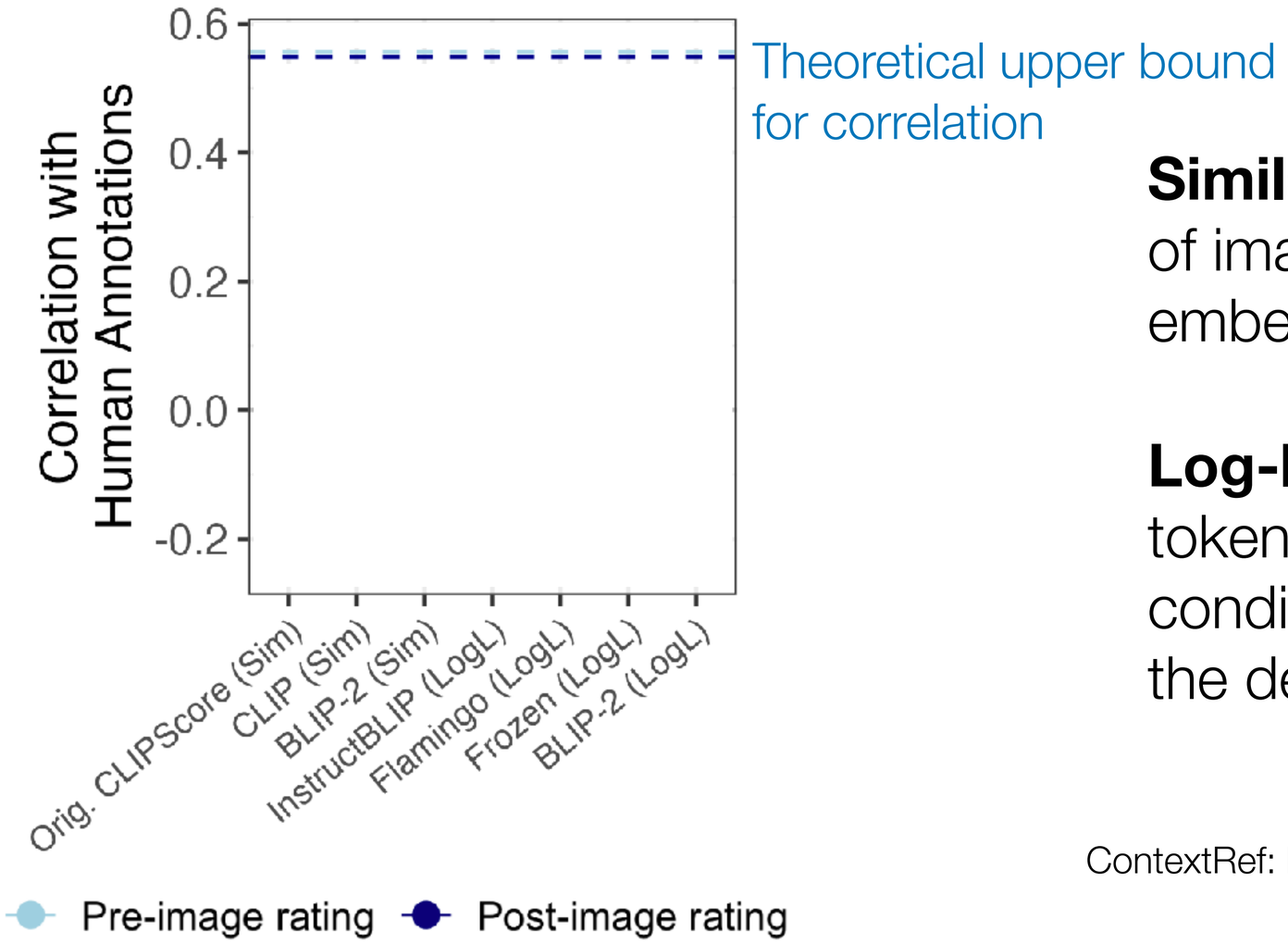
Promising correlations with human raters



ContextRef: Evaluating Referenceless Metrics for Image Description Generation
Kreiss*, Zelikman*, Potts, Haber (ICLR 2024)

Models for Human-like Image Description Evaluation

Promising correlations with human raters



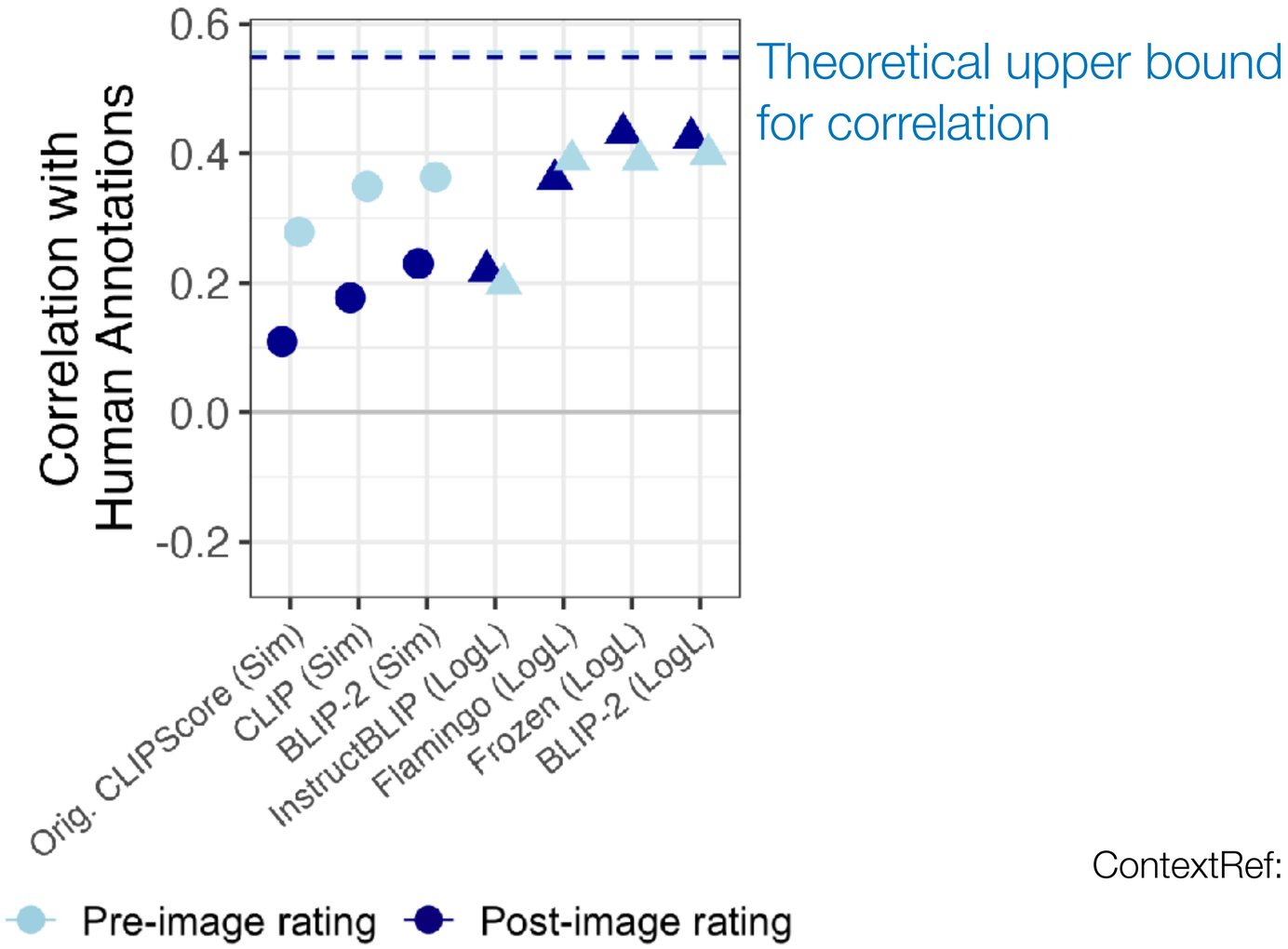
Similarity-based metrics: Cosine similarity of image, description, and context in embedding space

Log-likelihood metrics: LM's average per-token log-likelihood of the description, conditioned on the image, the context, and the description being high-quality

ContextRef: Evaluating Referenceless Metrics for Image Description Generation
Kreiss*, Zelikman*, Potts, Haber (ICLR 2024)

Models for Human-like Image Description Evaluation

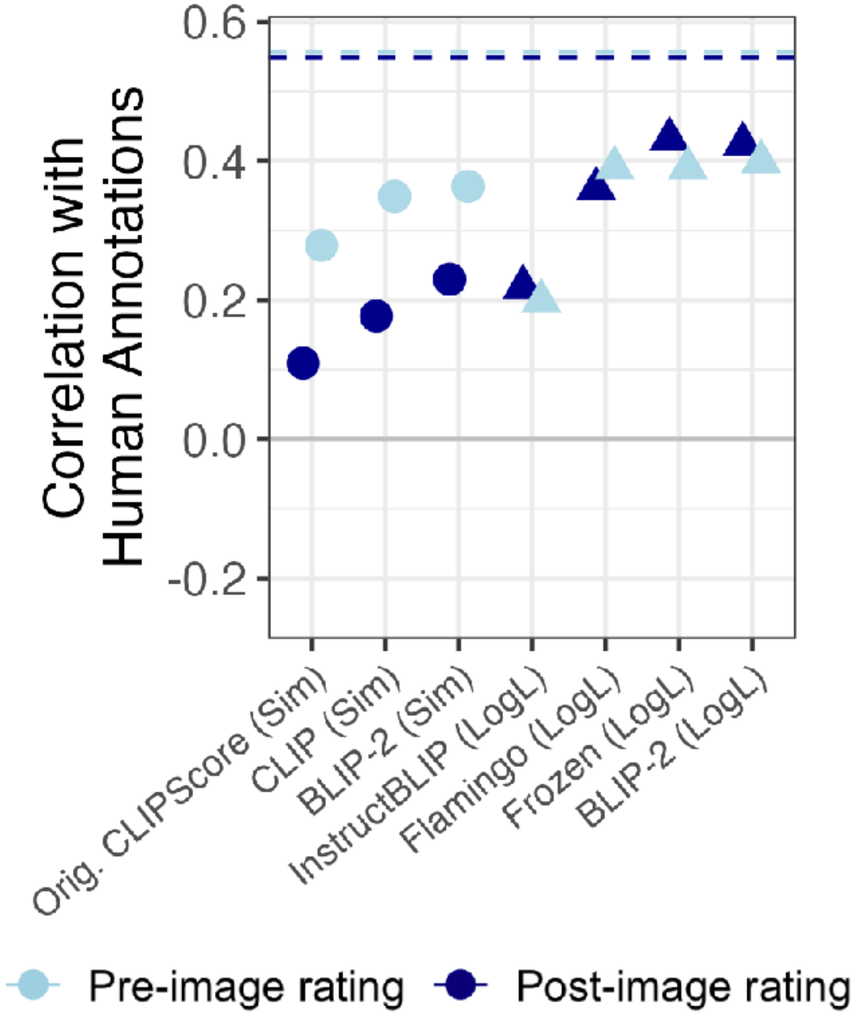
Promising correlations with human raters



ContextRef: Evaluating Referenceless Metrics for Image Description Generation
Kreiss*, Zelikman*, Potts, Haber (ICLR 2024)

Models for Human-like Image Description Evaluation

Promising correlations with human raters



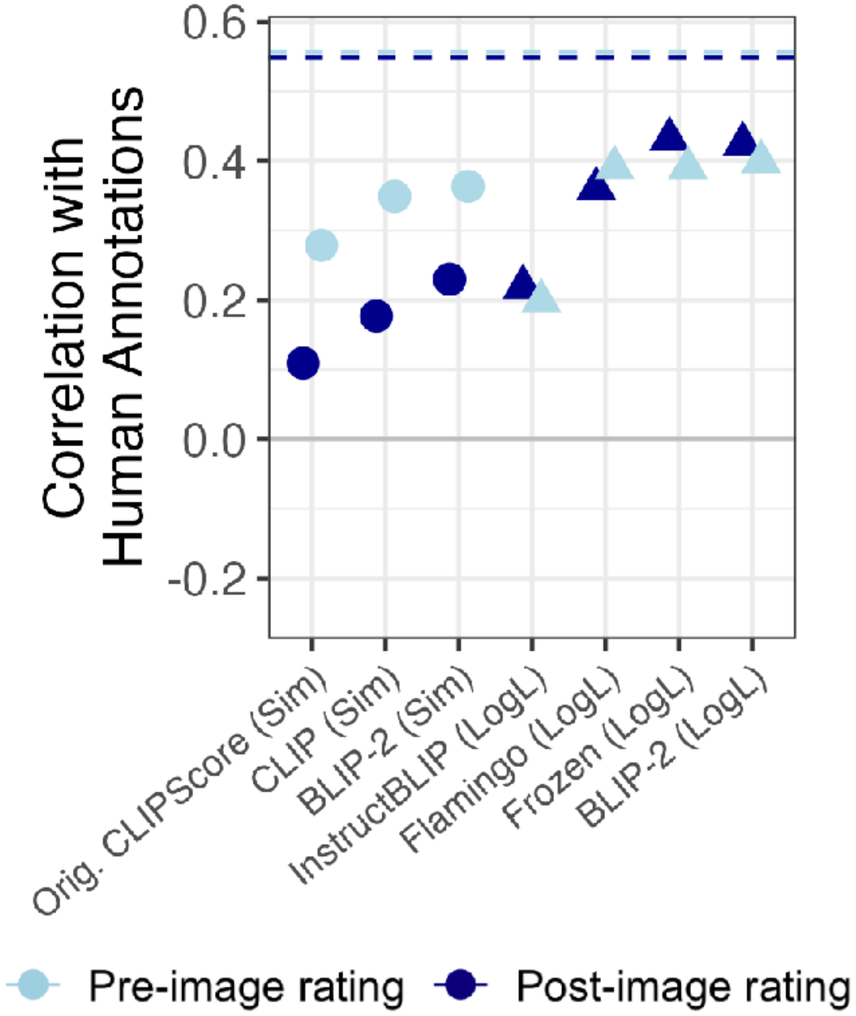
Data augmentations uncover unreliable behavior in all models

Original: a dog with big ears Good

ContextRef: Evaluating Referenceless Metrics for Image Description Generation
Kreiss*, Zelikman*, Potts, Haber (ICLR 2024)

Models for Human-like Image Description Evaluation

Promising correlations with human raters



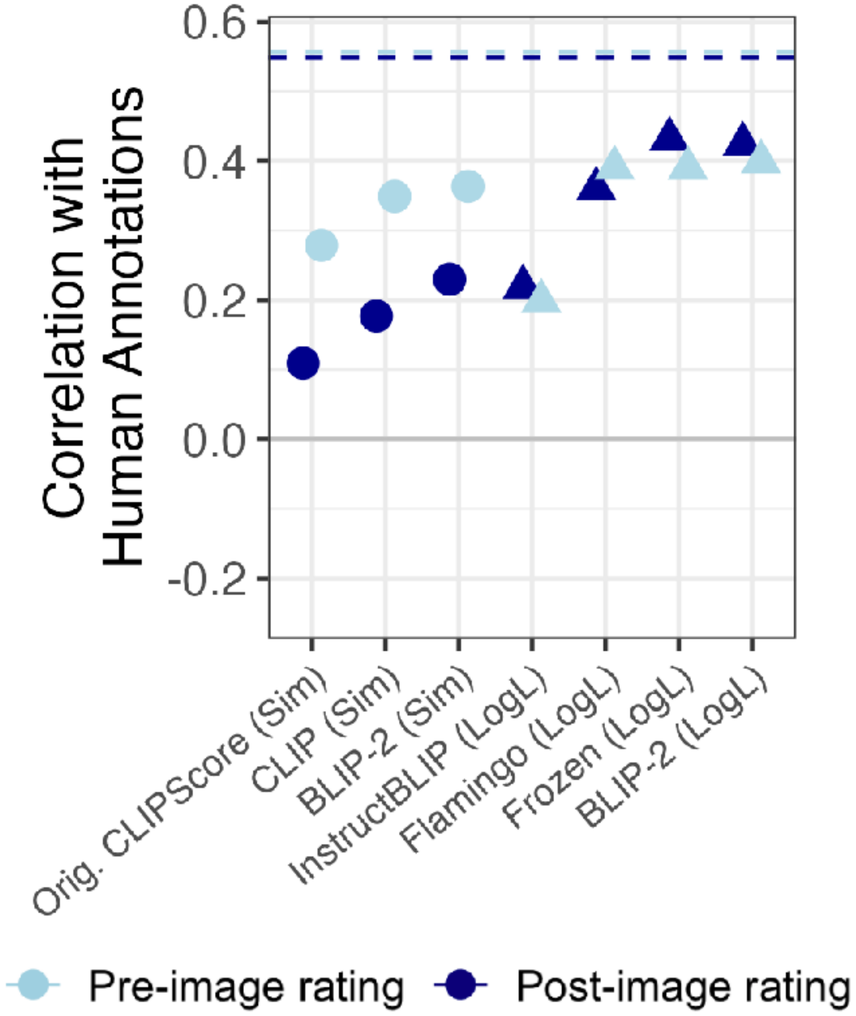
Data augmentations uncover unreliable behavior in all models

Original: a dog with big ears — Good

Exact repetition: a dog with big ears a dog with big ears — Excellent!

Models for Human-like Image Description Evaluation

Promising correlations with human raters



Data augmentations uncover unreliable behavior in all models

Original: a dog with big ears — Good

Exact repetition: a dog with big ears a dog with big ears — Excellent!

Irrelevance: a dog with big ears Elephants are the largest existing land animals — Excellent!

Models for Human-like Image Description Evaluation

Even state-of-the-art models are still far from approximating real human data.

Data augmentations uncover unreliable behavior in all models

Original: a dog with big ears — Good

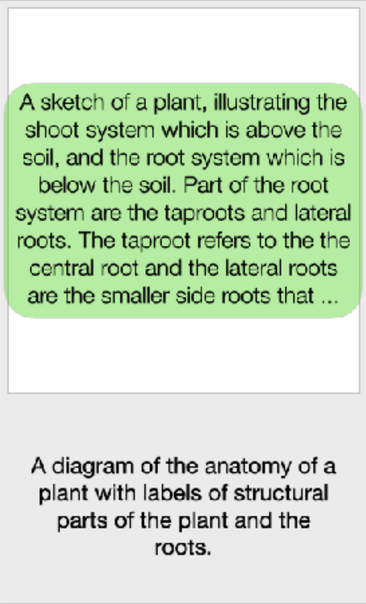
Exact repetition: a dog with big ears a dog with big ears — Excellent!

Irrelevance: a dog with big ears Elephants are the largest existing land animals — Excellent!

Image-based text generation depends on ...

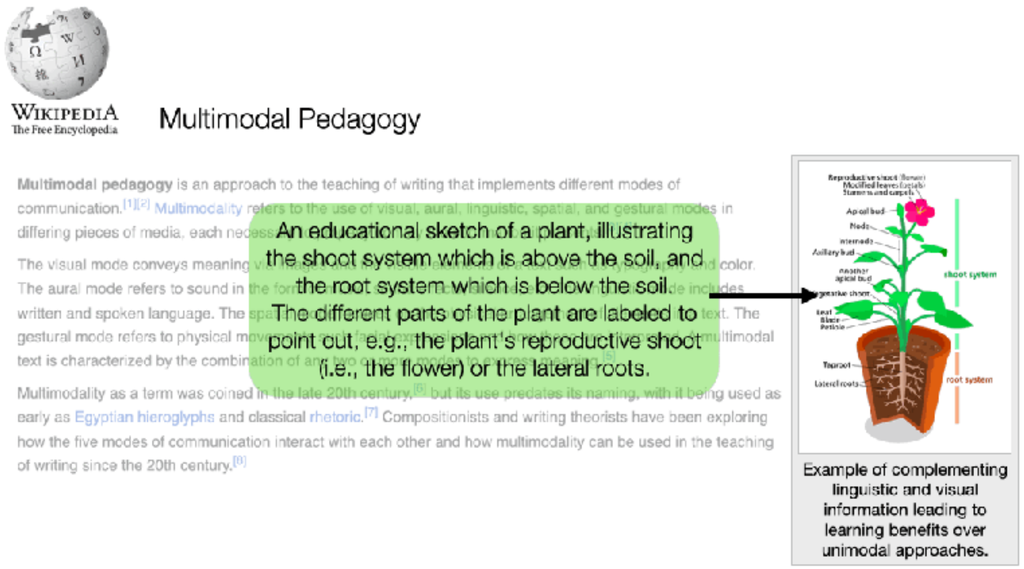
1 the **image-based text**'s communicative goal.

→ description ≠ caption



2 the **image**'s communicative goal.

→ article context determines information selection



What **can** we say about an image? → What **should** we say about an image?

WIKIPEDIA
The Free Encyclopedia

Plant Anatomy

Plant anatomy or **Phytotomy** is the general term for the study of the internal structure of plants. Originally it included plant morphology, the description of the physical form and external structure of plants, but since the mid-20th century plant anatomy has been considered a separate field referring only to internal plant structure.[1][2] Plant anatomy is now frequently investigated at the cellular level, and often involves the sectioning of tissues and microscopy.[3]

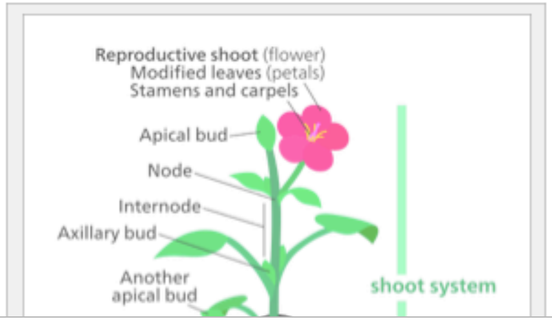


Image-based text generation depends on ...

the image-based **text**'s communicative goal.

→ description ≠ caption

the **image**'s communicative goal.

→ article context determines information selection

Coalas Lab

What **can** we say about an image? → What **should** we say about an image?

Frontiers in Generative AI for Nonvisual Accessibility

- Development of **local** (privacy-preserving, internet-independent) systems likely requires the development of smaller specialized systems which require more task-specific supervision.
- **Well-calibrated** systems (including hallucinations) for which we can anticipate error behavior and achieve pragmatic inference alignment (relevance, informativity, length) are especially consequential in situations where verification of the ground-truth is hard.

ekreiss@ucla.edu

@elisakreiss

Coalas Lab