Cross-Modal Coherence Relations Inform Eye-gaze Patterns During Comprehension & Production

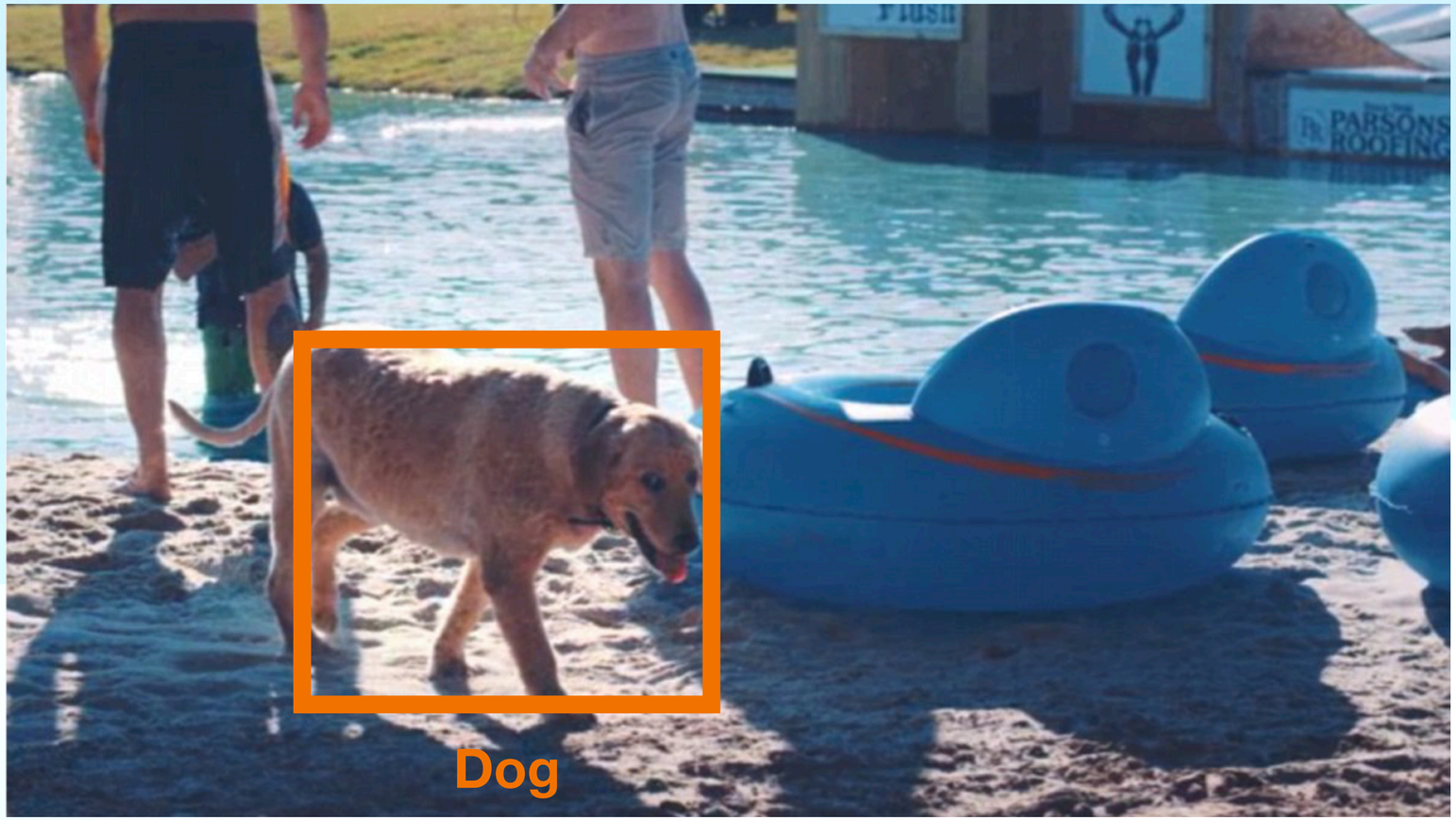
Mert İnan & Malihe Alikhani

LVX VERITAS VIRTVS
Northeastern University

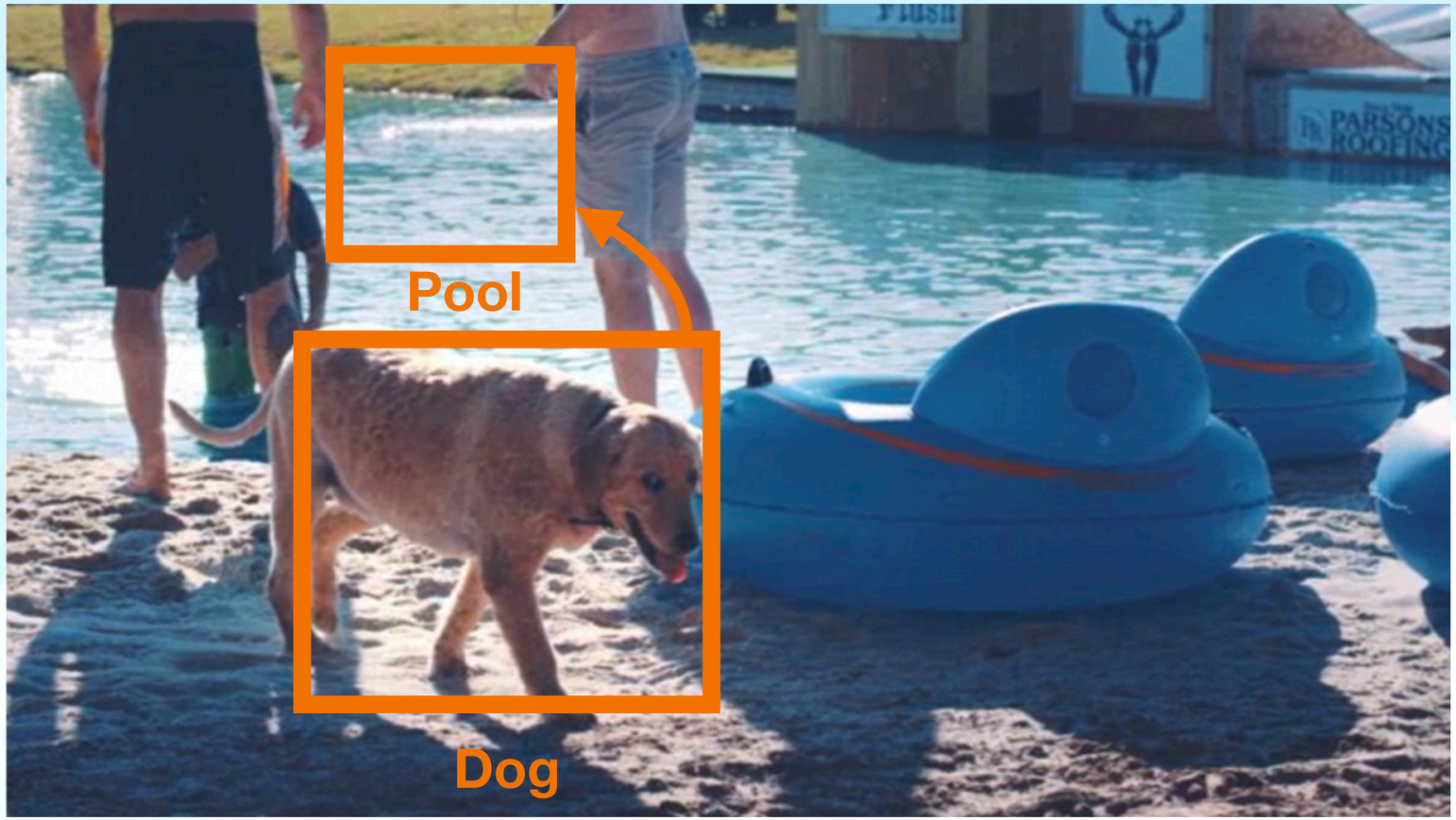
LREC-COLING 2024

A dog walking by the pool

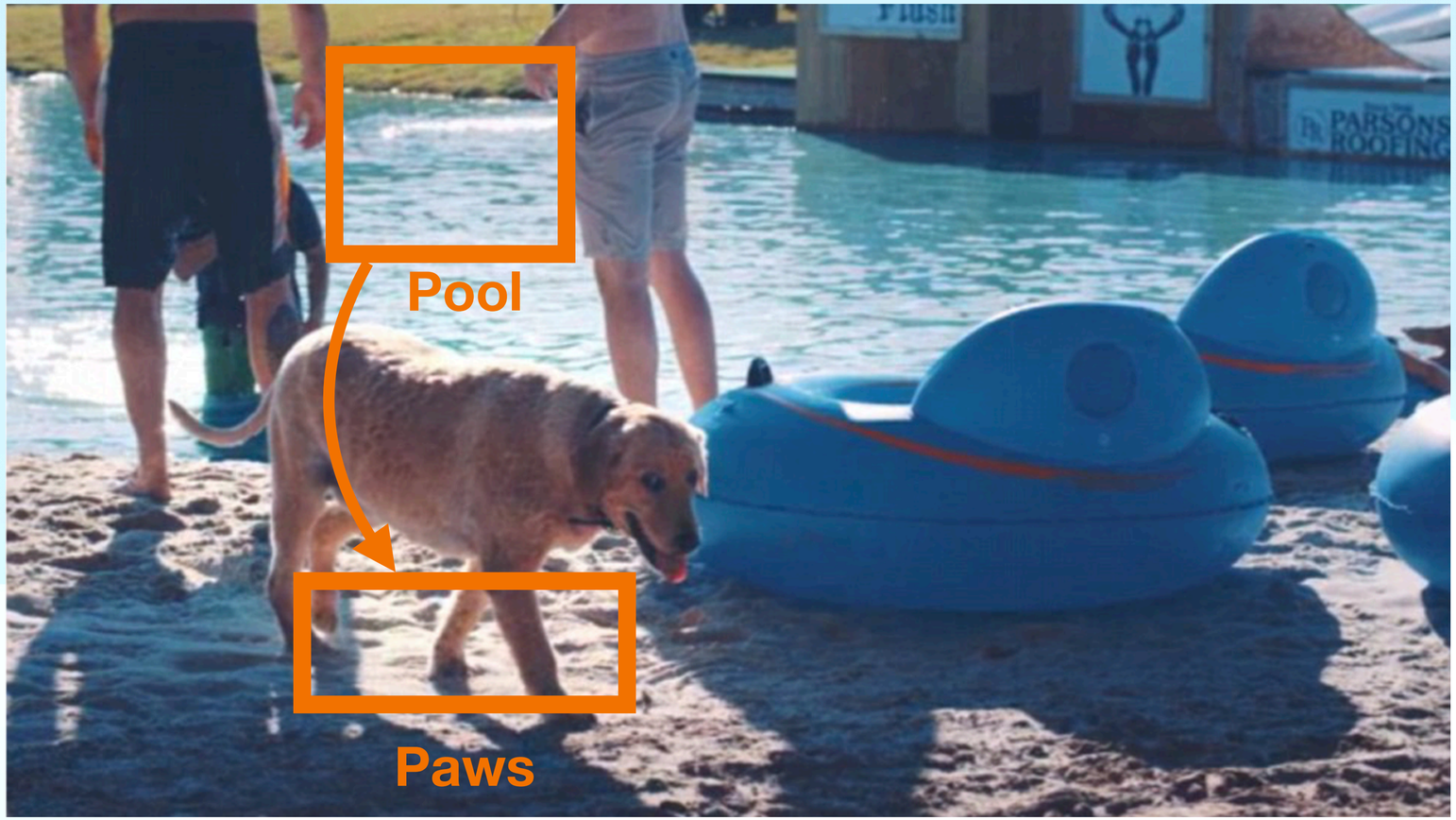
Where did you focus?



A dog walking by the pool



A dog walking by the pool

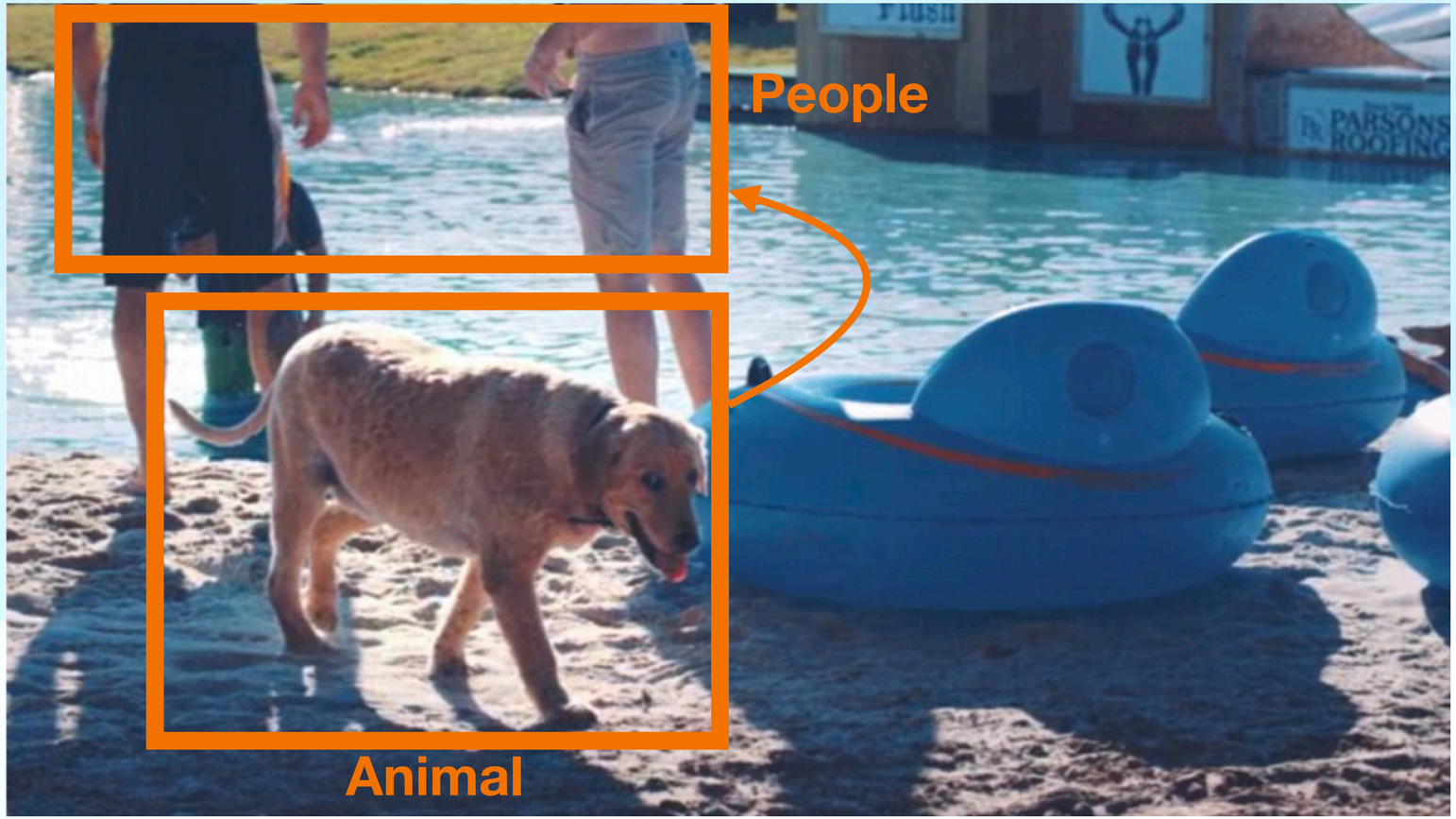


A dog walking by the pool

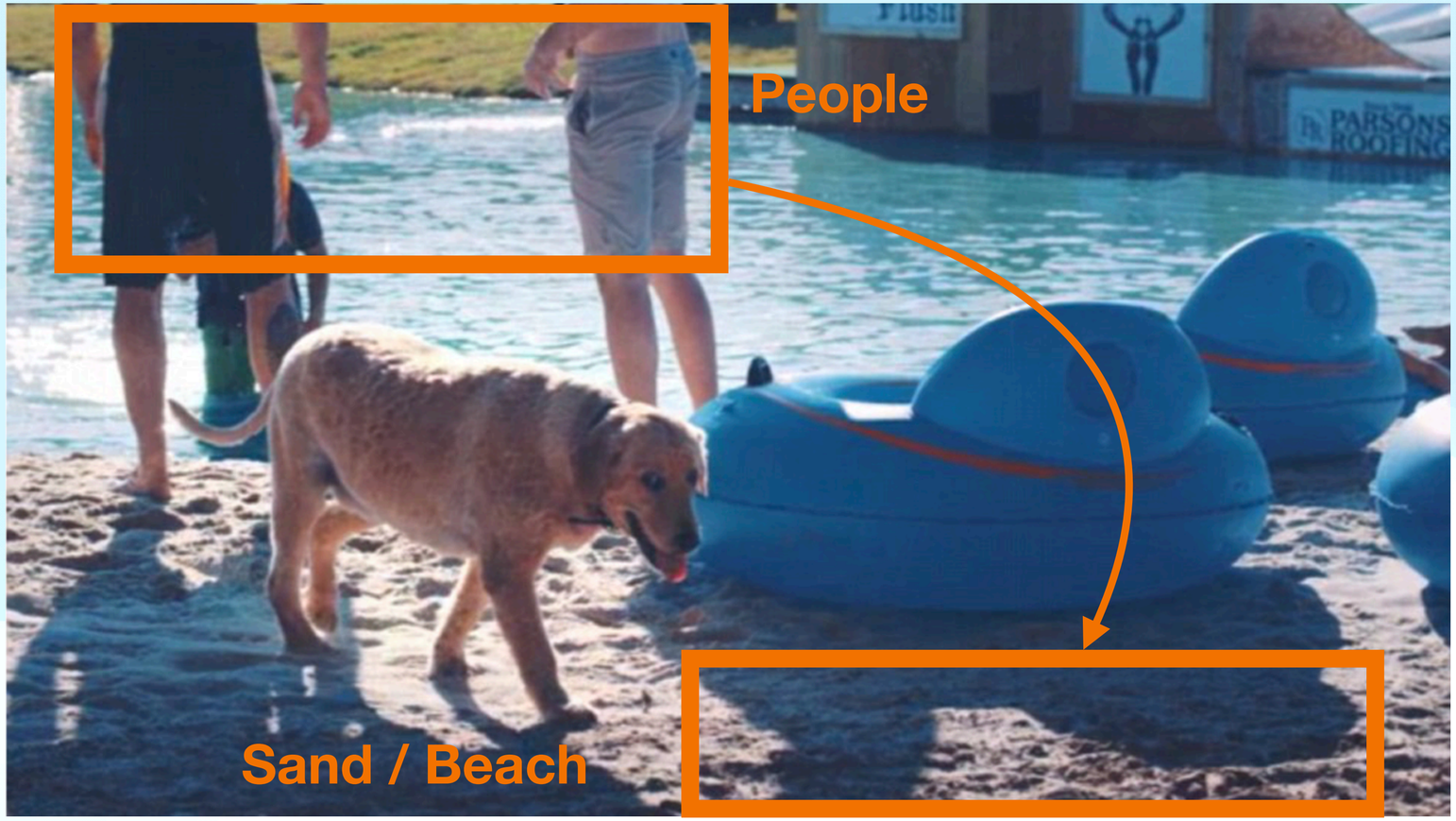
How to spend a day at the beach?

Where did you focus this time?

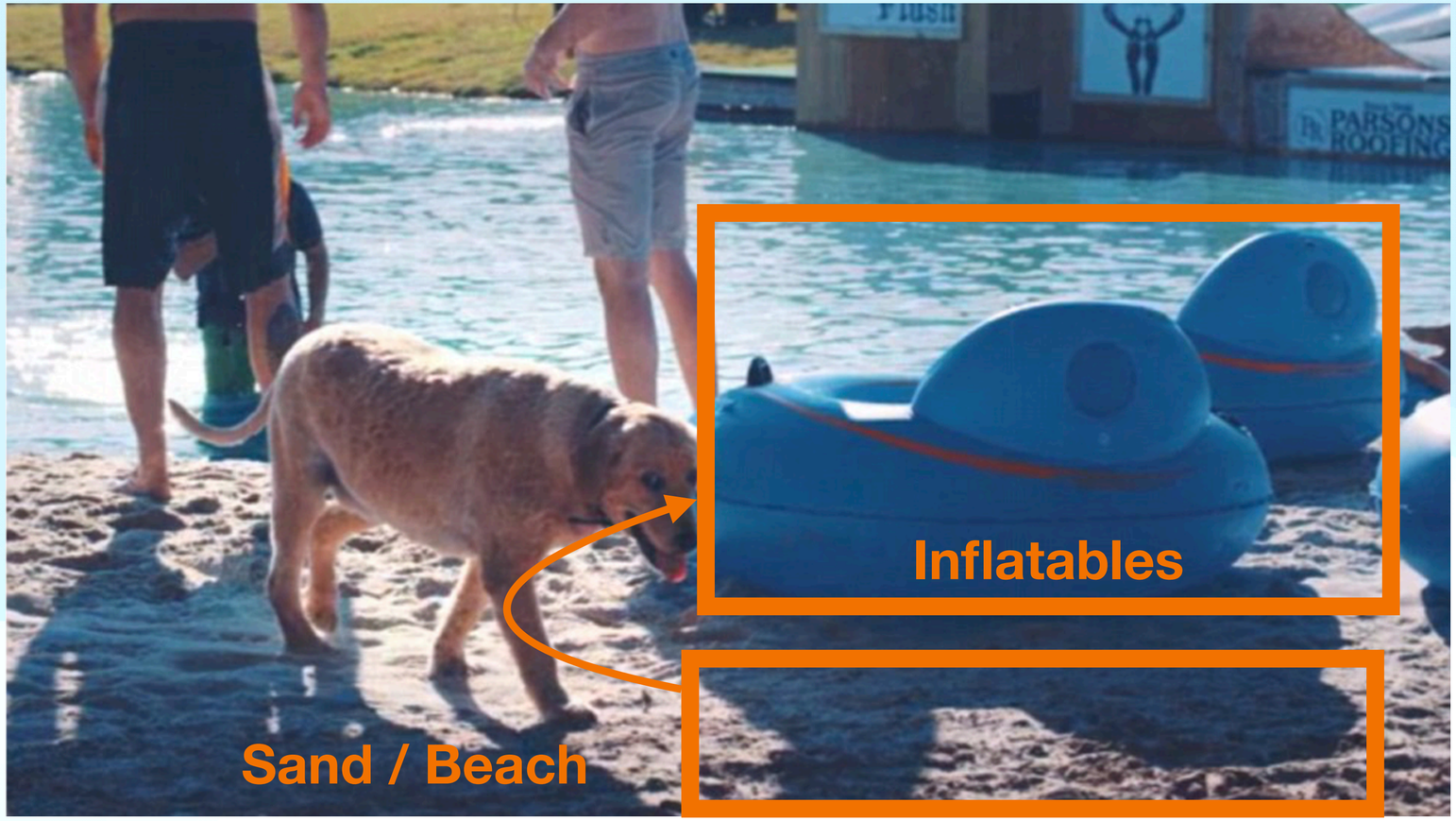
How to spend a day at the beach?



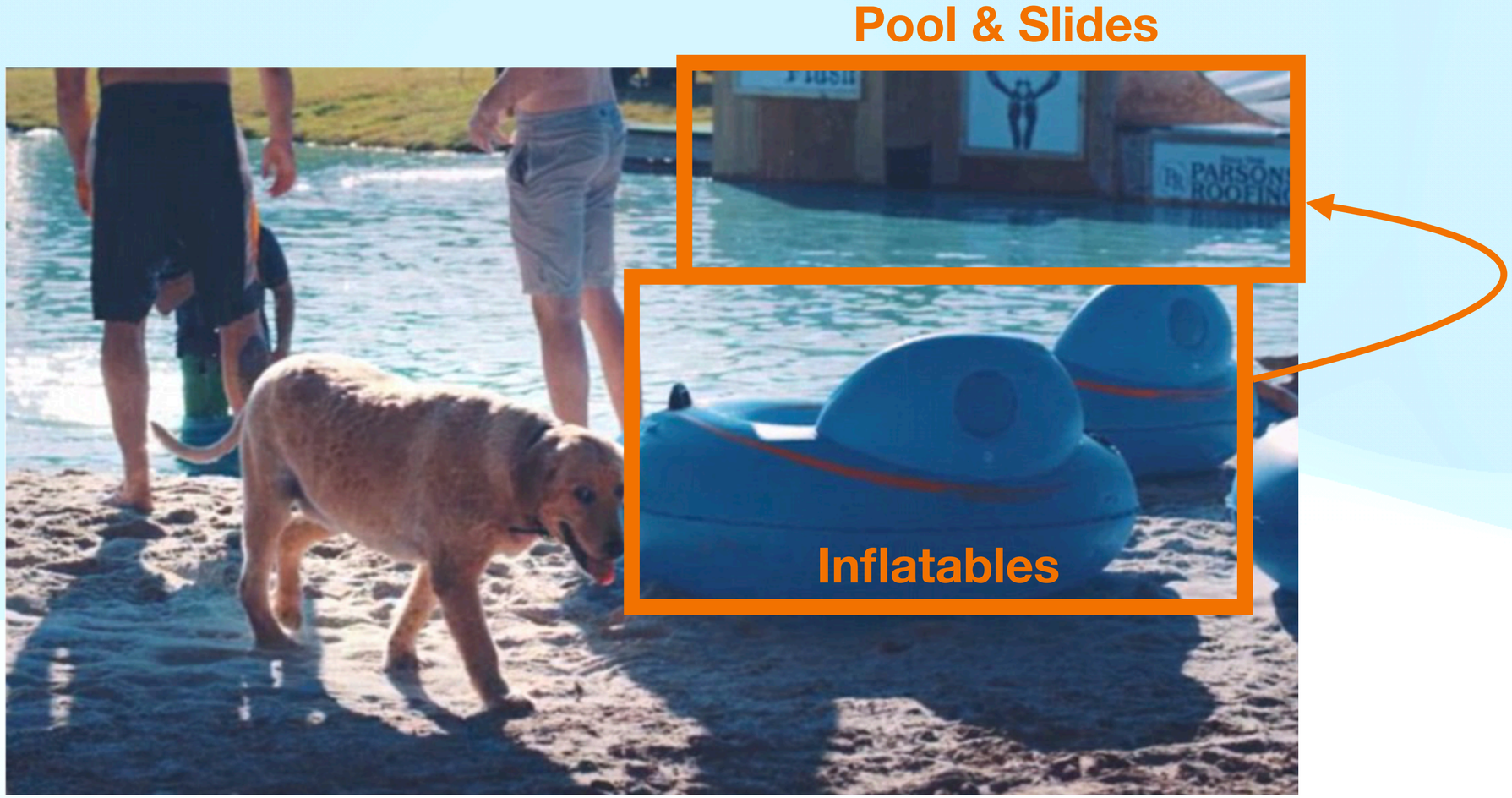
How to spend a day at the beach?



How to spend a day at the beach?



How to spend a day at the beach?



How to spend a day at the beach?

How can we explore this intricate relationship between where we look and how we reason about multimodal presentations?

To investigate this question

In this work,

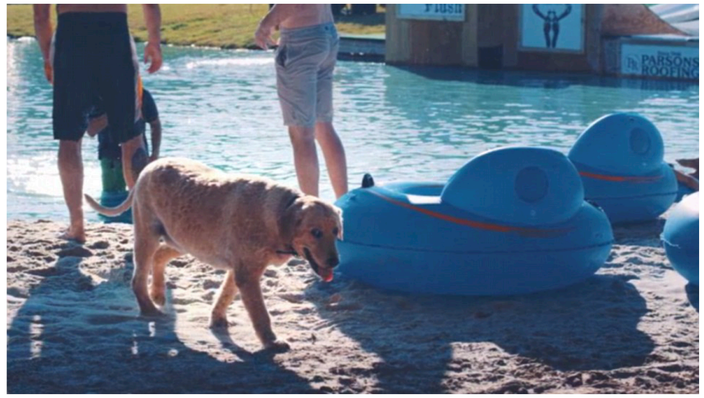
1. We apply a ***Discourse Coherence Framework*** to understand eye gaze patterns in multimodal presentations.
2. We describe experiments both for ***comprehension*** and ***production***.
3. We verify *in situ* that there are individual differences in eye gaze patterns.
4. We explore how recent ***Large Language Models*** represent the connection between discourse goals and eye gaze patterns.

Discourse goals

Background Information

Visible, Action

A dog walking by the pool



Meta, Subjective, Story

How to spend a day at the beach?

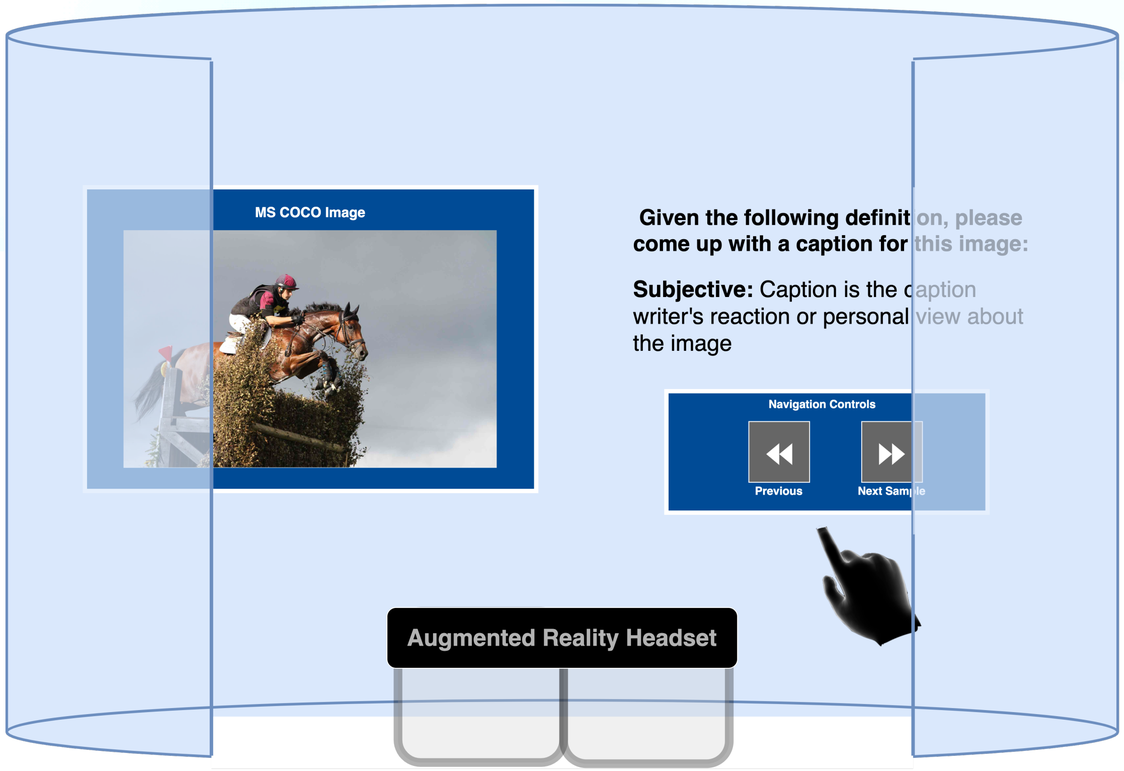
The cross-modal discourse framework has been introduced by Alikhani et al. (2020)

How Can This Discourse Framework Be Applied To Eye Gaze?

- We investigate two settings to test the eye gaze pattern changes in multimodal presentations:
    1. **comprehension**
        - choosing coherence relations in an image and a caption.
    2. **production**
        - producing a caption given a coherence relation and an image.

Setup: How to Measure Eye Gaze?

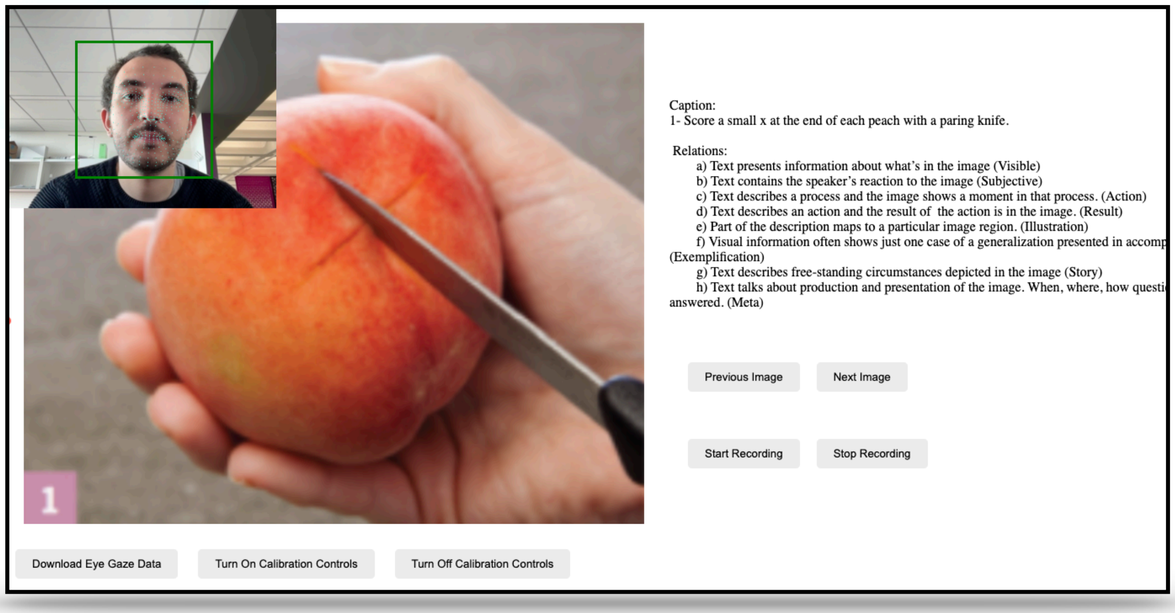
Augmented Reality



- Augmented Reality goggles have high precision eye gaze recording capabilities.
- We use Microsoft HoloLens 2.
- Eye gaze locations are logged for over ~ 1500 datapoints per sample.

Setup: How to Measure Eye Gaze?

Webcam Based Eye Trackers



- Webcam-based tracking is more accessible and can be gathered anywhere.
- We use Webgazer.js to detect gaze through webcams.
- **Noisier:** Eye gaze locations are logged for over ~ 2500 datapoints per sample.

Comprehension Experiment Details

Coherence Relation Identification

Caption: 1- Score a small x at the bottom of each peach with a pairing knife.

Relations:

a. **Story :** _____
b. **Visible:** _____
c. **Meta:** _____
d. **Exemplification** _____
e. **…**

- The participant annotates image and caption pairs from 8 total coherence relations.
- The participant can select multiple relations but is encouraged to pick the most relevant one.
- Their selection is recorded and compared with other participants.

Production Experiment Details

Coherence-Primed Caption Production

Relation: Story (Text describes free-standing circumstances depicted in the image)

Caption: Sea Biscuit won the steeple race

- The participants produce captions for a given image, and a coherence relation category.
- Participants read the definition of a coherence relation,
- They come up with a caption that follows the given coherence relation.

How to Process Noisy Eye Gaze Data?

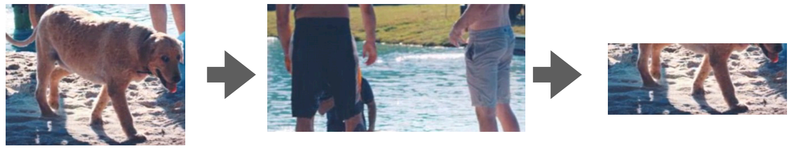
A Novel Gaze Pattern Ranking Algorithm

- Previous methods do:
  - averaging gaze across participants,
  - pixel-based fixation,
  - object-based fixation,
  - saccade pattern analyses
- These disregard the Pragmatics between image and text.
- We introduce a new processing technique: ***Gaze Pattern Ranking***.
- Using this method, we process the time spent on an entity and rank these patterns across coherence relation categories.

Algorithm 1 Gaze Pattern Ranking

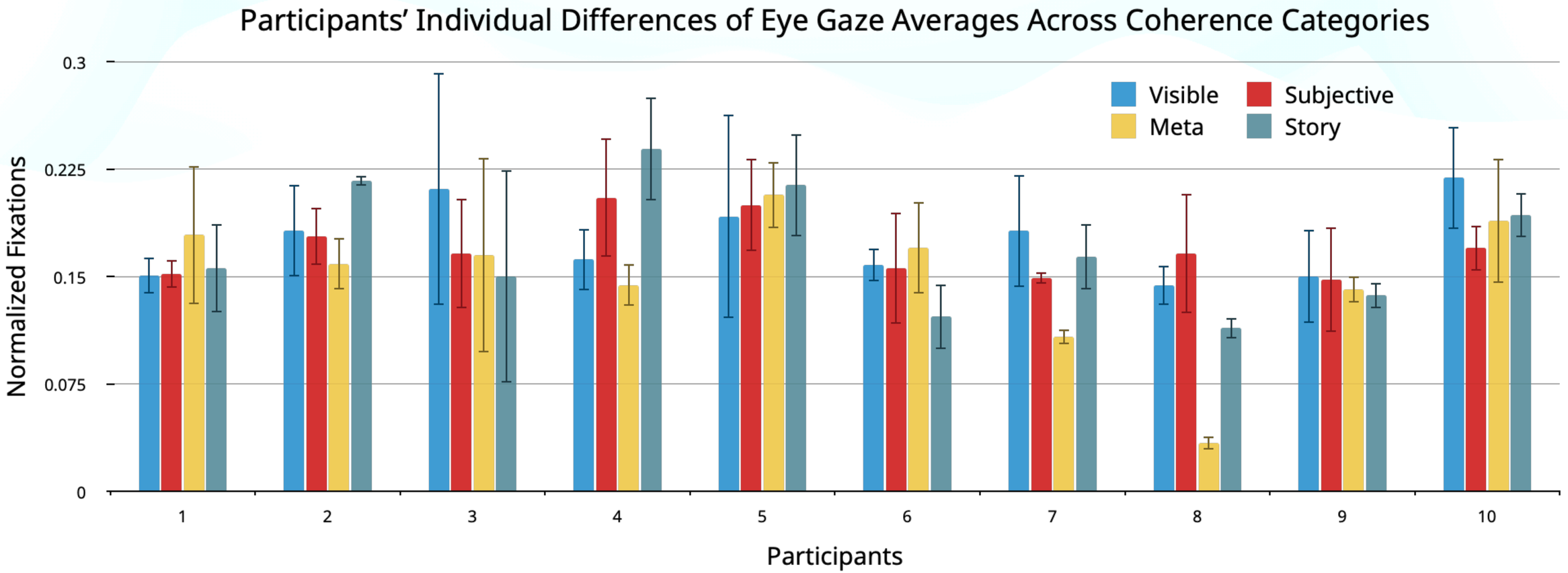
$P \leftarrow [\ ]$
$t \leftarrow$ eye gaze durations on image I
for entity i in image I **do**
  $t_i \leftarrow$eye gaze duration per pixel in O_i
  $O_i \leftarrow$ AVERAGE(t_i)
for participant j **do**
  $P_j \leftarrow$ DESCENDING-SORT(O)
  return P

Visible

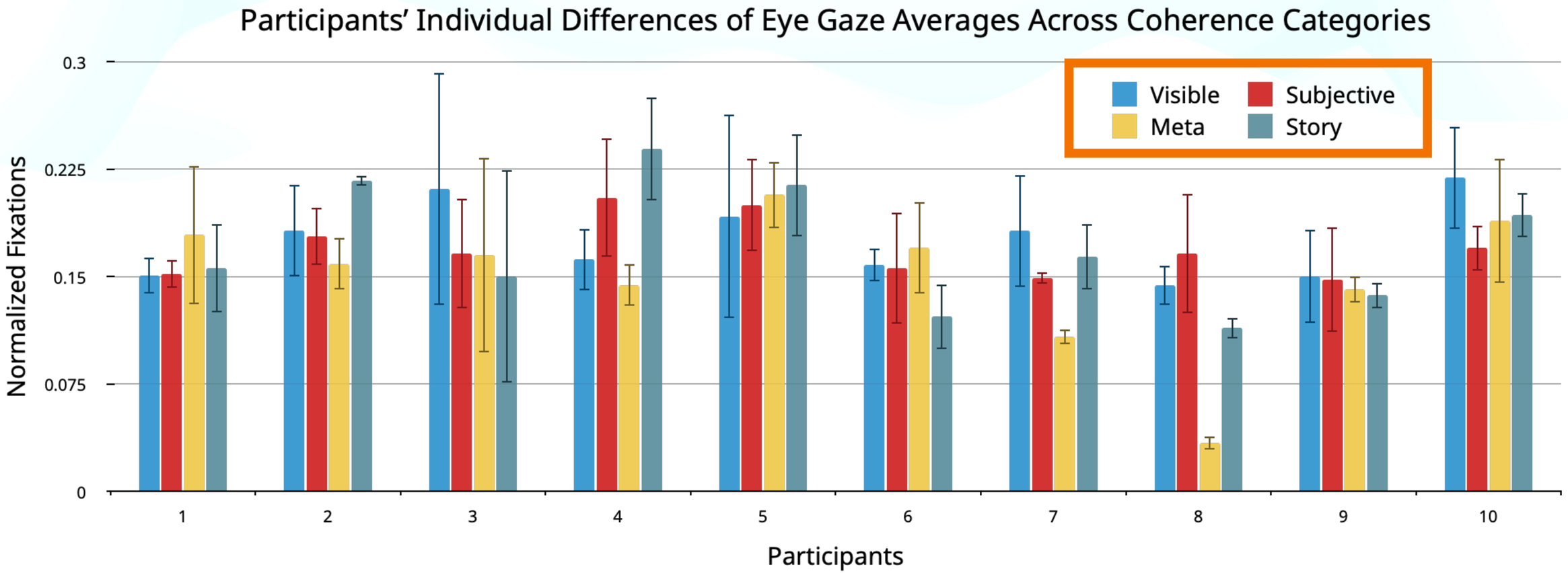


Story

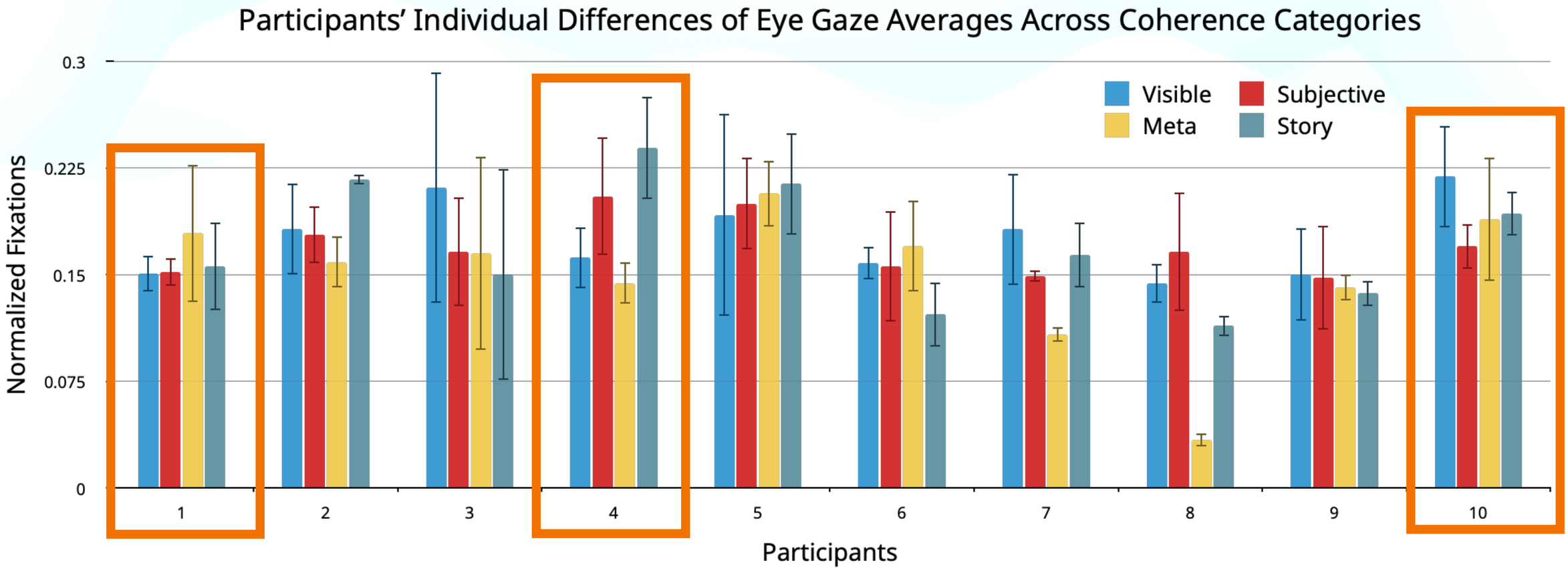
Results: Raw Eye Gaze Durations Point to Individual Differences



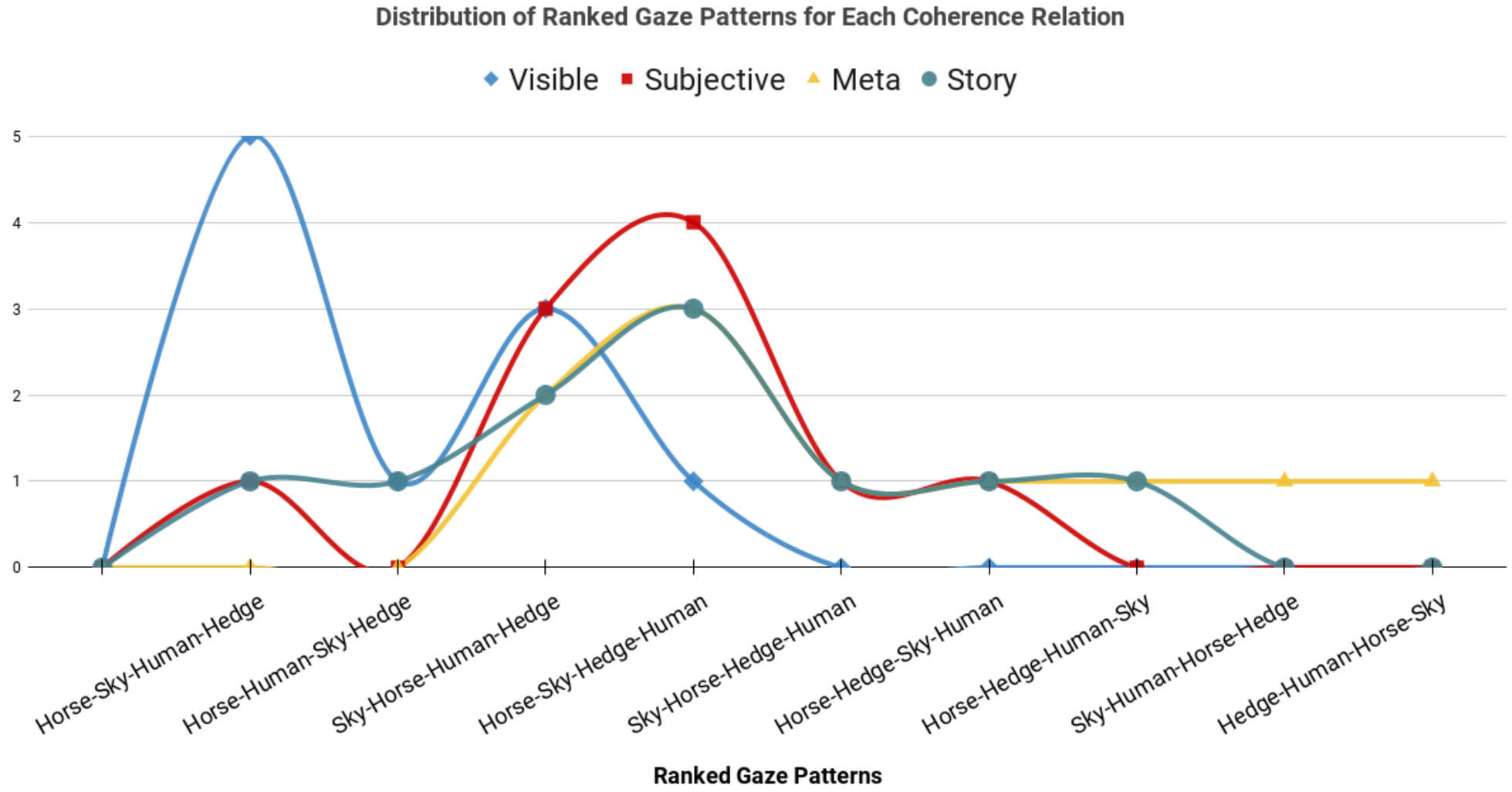
Results: Raw Eye Gaze Durations Point to Individual Differences



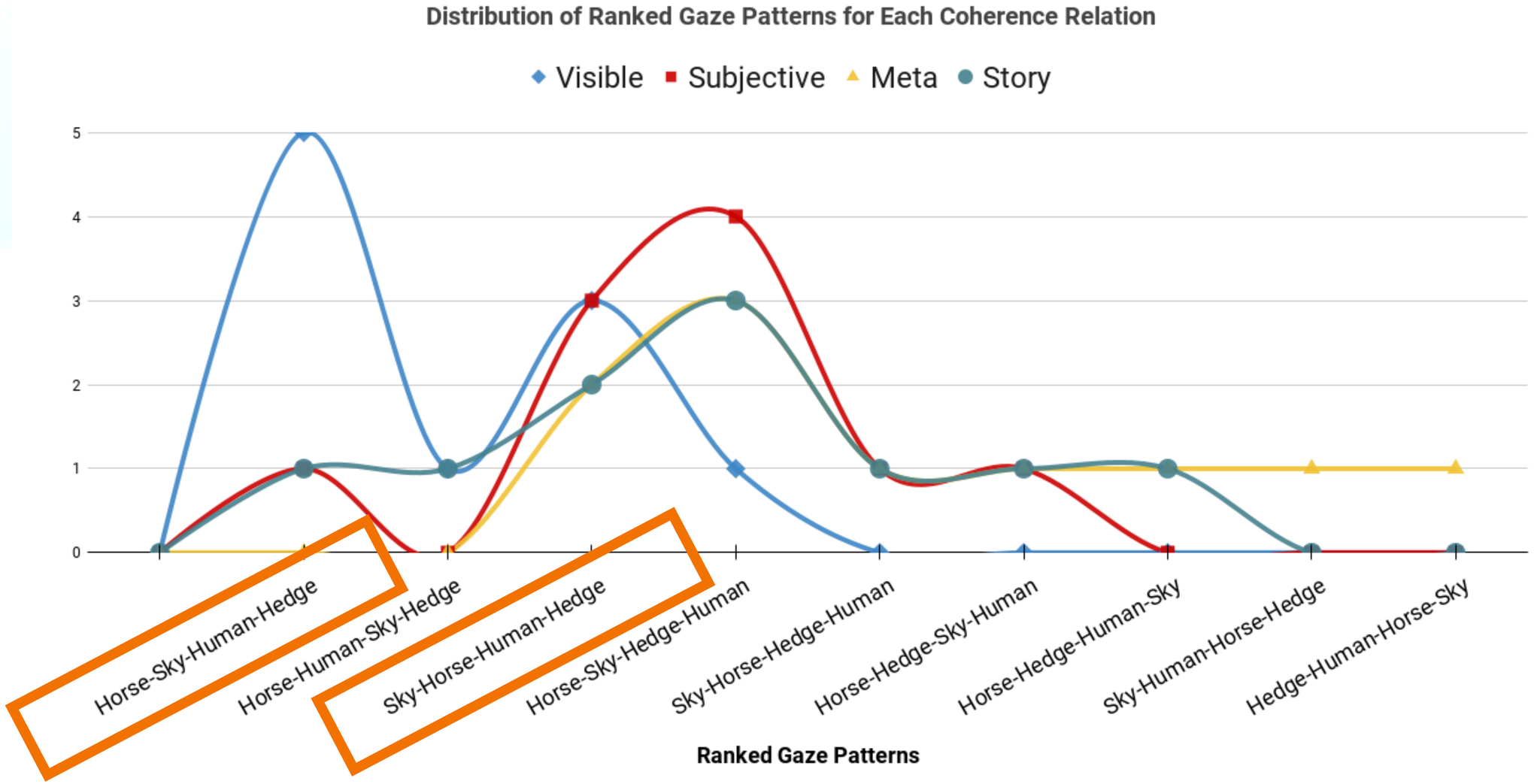
Results: Raw Eye Gaze Durations Point to Individual Differences



Results: Discourse Coherence Relations Can Predict Eye Gaze Patterns



Results: Discourse Coherence Relations Can Predict Eye Gaze Patterns



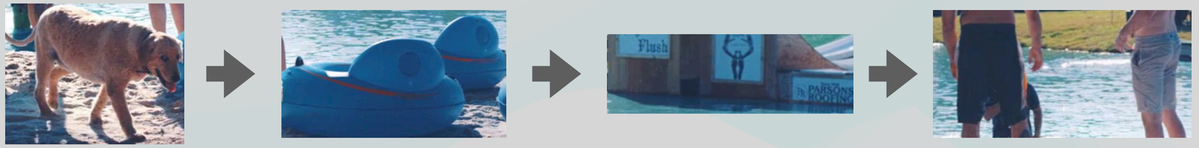
Can LLMs Predict Eye Gaze Patterns?

- As a preliminary case study, we explore prompting Multimodal Large Language Models (LLMs).
- This gives a larger scale glimpse into the relationship between eye gaze and multimodal discourse.
- We prompt with the same experimental setup:
  - Claude 3 Sonnet
  - Gemini 1.5
  - LLaVA v1.6 34B 4bit-quantized

Result: Can LLMs Predict Eye Gaze Patterns?

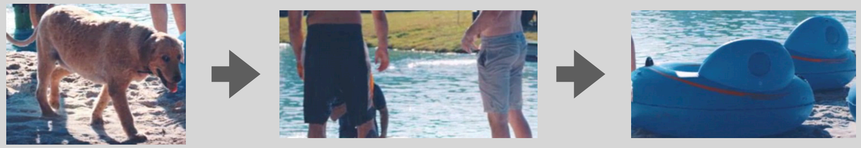
How to spend a day at the beach?

Claude 3 Sonnet



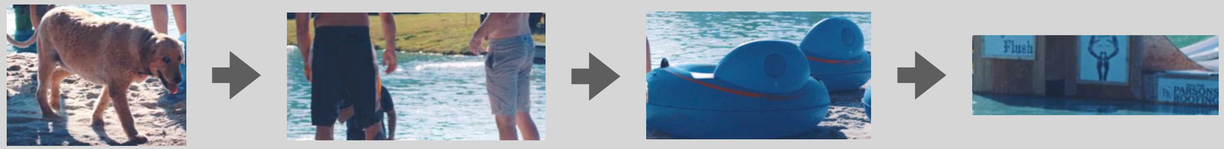
Chosen Coherence Category: *Exemplification*

LLaVA v1.6 34B



Chosen Coherence Category: *Story*

Human



Chosen Coherence Category: *Story, Subjective*

Qualitative Results: Coherence Influences Gaze Patterns

Visible	Subjective	Meta	Story
			
Jockey on a horse jumping over a steeple	Jockey on a well-groomed horse jumping over a steeple	A jockey on a horse jumping over a steeple during the day	Sea Biscuit won the steeple race
			
Dog is on a beach beside a pool with a bunch of people standing next to it	The dog seems to be very happy on its own	The picture seems to be taken through a cell phone, casually	The whole family went to the beach for the weekend

If You Are Curious About How We Visualize Gaze Patterns

A Novel Visualization Tool

- We develop a tool using recent semantic segmentation techniques.
- We automate visualizing object-specific human attention.
- We provide a Python package to plot semantic entity-based gaze maps at this URL: https://github.com/Merterm/eye-gaze-coherence

Conclusion

- We presented that both in production and comprehension coherence relations are predictive of human eye gaze patterns.
- For future work, we will expand into instruction-following and dialogue systems.
- We make our code and analysis tools available and hope that this will encourage the community to use eye gaze to make better use of context.

Thank you!

Mert İnan

merterm.github.io

@merterm

merterm

Malihe Alikhani

malihealikhani.com

@malihealikhani

malihealikhani

For our eye gaze visualization tools, and gaze pattern ranking code, please visit:
https://github.com/Merterm/eye-gaze-coherence

LVX VERITAS VIRTVS

Northeastern University