Examining the precision of infants' visual concepts by leveraging vision-language models and automated gaze coding

Tarun Sepuri[1], Martin Zettersten[1], Bria Long[1]

[1]University of California, San Diego

Background

- The visual concepts supporting rapid early word learning may be coarse and gradually learned.
- Visual concept knowledge can be characterized by how competitor similarity influences word recognition.
- However, previous work operationalizes similarity dichotomously and subjectively.
- Infant gaze data are also hard to collect and thus tend to include small sample sizes and item sets.

Questions

Alt q 1. Do infants have more difficulty recognizing words more similar to distractors in a vision-language model similarity space?
Alt q 1. Do infants have partial visual knowledge of words?

1. Will infants be more drawn away from a target the more similar it is to a distractor?

2. Do additional item-level differences influence infants' looking behavior?

Items employed in the study design from THINGS-plus

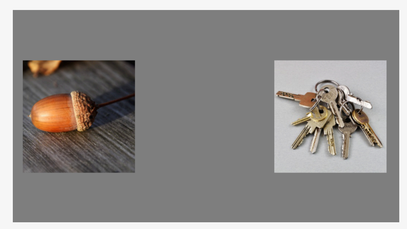
Flip the stuff above

Methods

90 14-24- month old infants
Each infant is shown 32 trials:
8 easy, 8 hard, 16 where the target and distractor are flipped

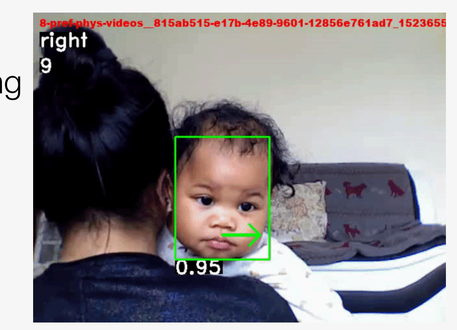
Data collected asynchronously on Children Helping Science

Example trial:

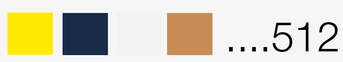


Prompt: "Look at the acorn"

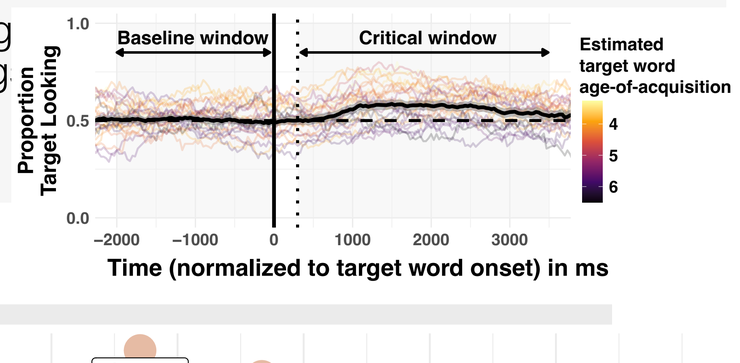
Data passed through iCatcher+ for automated left-right-away gaze coding



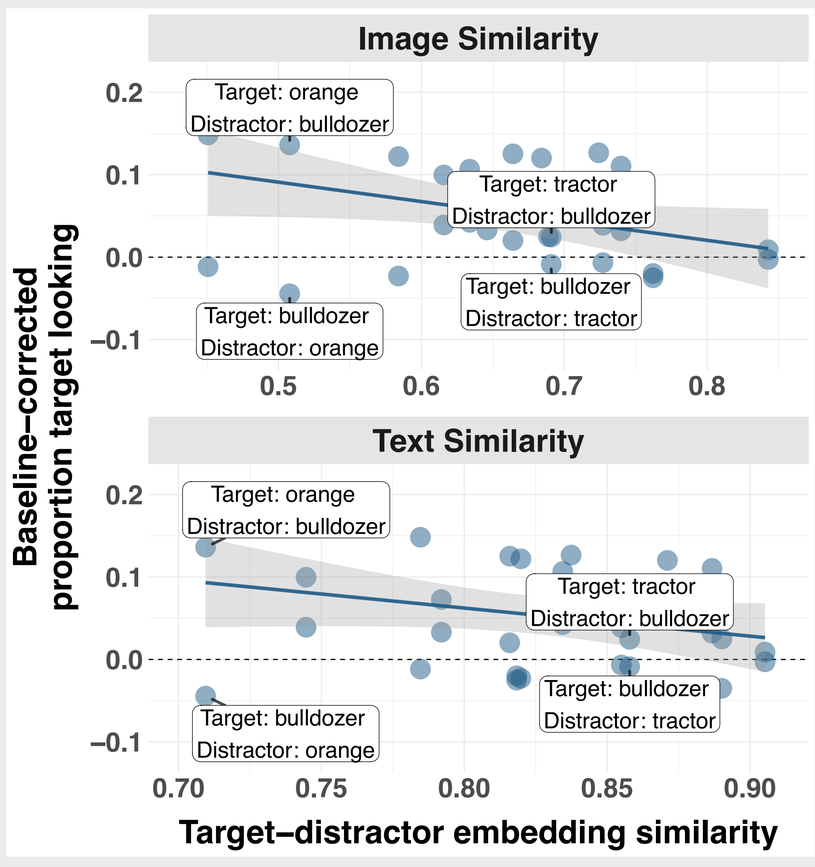
Proportion of looking time to target over distractor correlated with target-distractor similarity



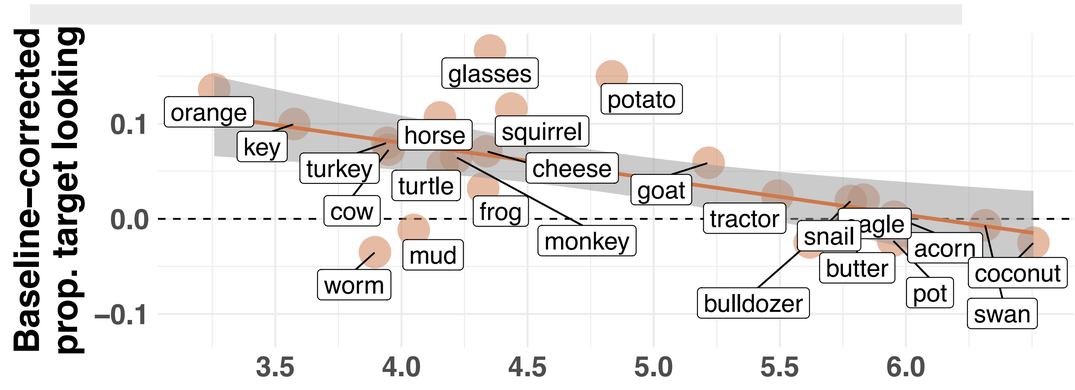
Cosine sim of lang vision embedding



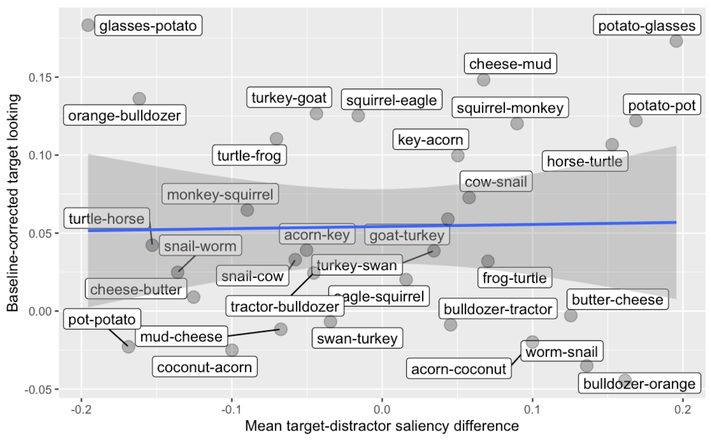
Results

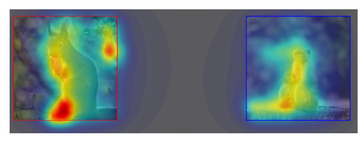


Infants look more at target images the more similar they are in image similarity space (confirmed with a linear-mixed effects model: stats)
Text similarity effect trends in the same direction but is not statistically significant (stats)



Age-of-acquisition of target word correlates inversely with target looking (stats)





Visual saliency differences with GBVS do not explain additional variance (stats)

Infants are more drawn away from a target the more similar it is to a distractor in high-level image similarity space

Infants' looking behavior is additionally shaped by word difficulty but not by visual saliency

Discussion

Results suggest that infants have partial visual knowledge for many difficult words in their second year
A second experiment with 16 new items will help to more robustly determine the nature of early visual representations.
Gaze annotation techniques and vision-language models can be jointly used to further investigate the development of infants' visual concept knowledge

QR CODE

References: 1) Frank et al. (2021).

Examining the precision of infants' visual concepts by leveraging vision-language models and automated gaze coding

Tarun Sepuri[1], Martin Zettersten[1], Bria Long[1]

[1]University of California, San Diego

Background

- The visual concepts supporting rapid early word learning may be coarse and gradually learned.
- Visual concept knowledge can be characterized by how competitor similarity influences word recognition.
- However, previous work operationalizes similarity dichotomously and subjectively.
- Infant gaze data are also hard to collect and thus tend to include small sample sizes and item sets.

Questions

Alt q 1. Do infants have more difficulty recognizing words more similar to distractors in a vision-language model similarity space?
Alt q 1. Do infants have partial visual knowledge of words?

1. Will infants be more drawn away from a target the more similar it is to a distractor?

2. Do additional item-level differences influence infants' looking behavior?

Items employed in the study design from THINGS-plus

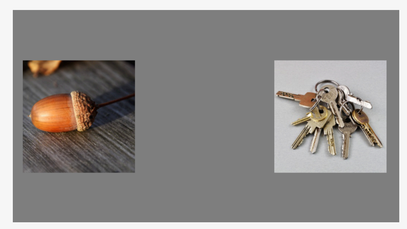
Flip the stuff above

Methods

90 14-24- month old infants
Each infant is shown 32 trials:
8 easy, 8 hard, 16 where the target and distractor are flipped

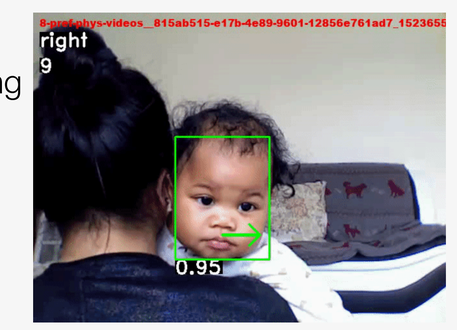
Data collected asynchronously on Children Helping Science

Example trial:

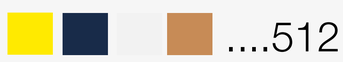


Prompt: "Look at the acorn"

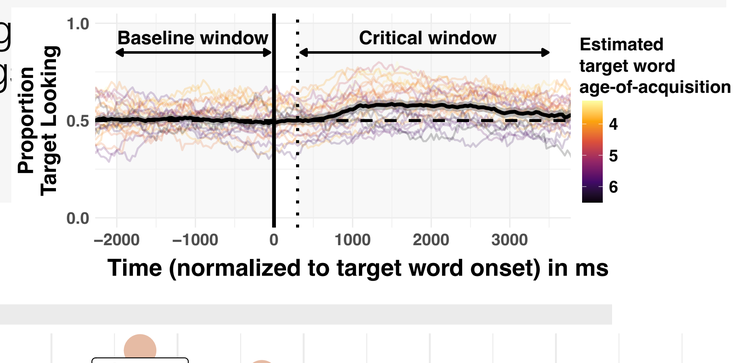
Data passed through iCatcher+ for automated left-right-away gaze coding



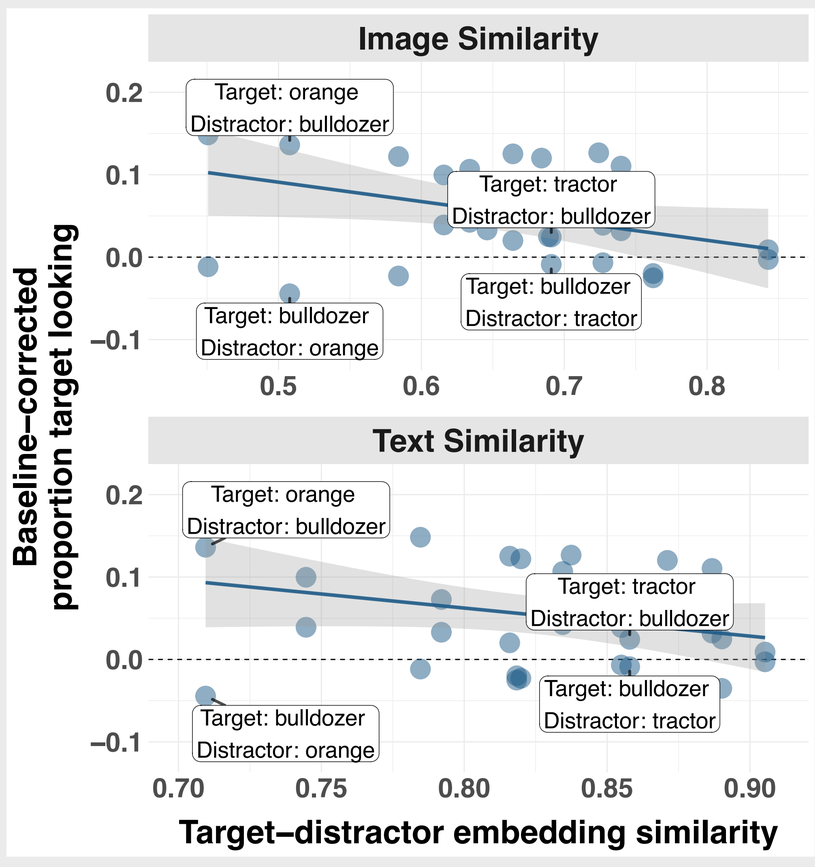
Proportion of looking time to target over distractor correlated with target-distractor similarity



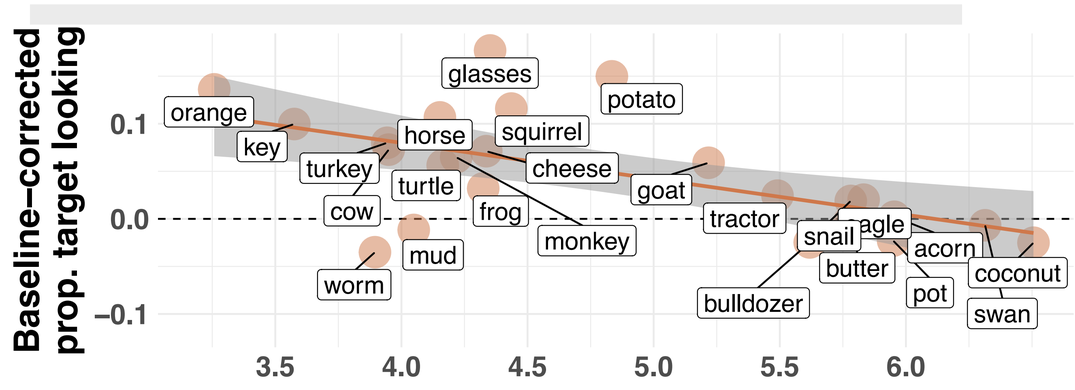
Cosine sim of lang vision embedding



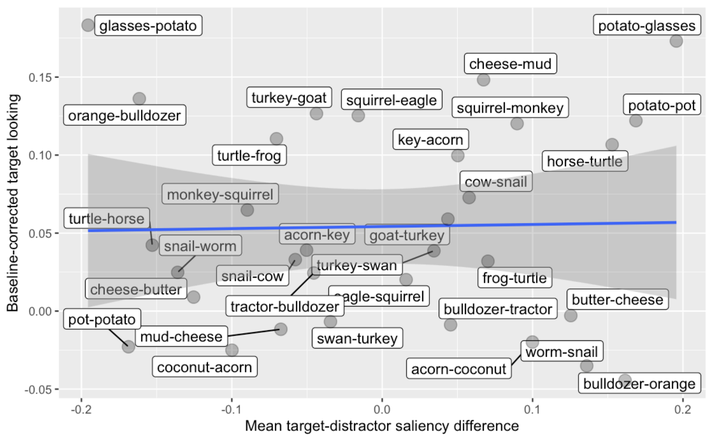
Results

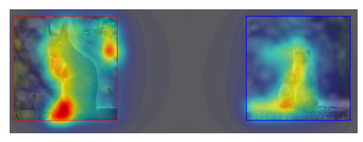


Infants look more at target images the more similar they are in image similarity space (confirmed with a linear-mixed effects model: stats)
Text similarity effect trends in the same direction but is not statistically significant (stats)



Age-of-acquisition of target word correlates inversely with target looking (stats)





Visual saliency differences with GBVS do not explain additional variance (stats)

Infants are more drawn away from a target the more similar it is to a distractor in high-level image similarity space

Infants' looking behavior is additionally shaped by word difficulty but not by visual saliency

Discussion

Results suggest that infants have partial visual knowledge for many difficult words in their second year
A second experiment with 16 new items will help to more robustly determine the nature of early visual representations.
Gaze annotation techniques and vision-language models can be jointly used to further investigate the development of infants' visual concept knowledge

QR CODE

References: 1) Frank et al. (2021).

Examining the precision of infants' visual concepts by leveraging vision-language models and automated gaze coding

Tarun Sepuri[1], Martin Zettersten[1], Bria Long[1]
[1]University of California, San Diego

VISUAL LEARNING LAB

Background

- The visual concepts supporting rapid early word learning may be coarse and gradually learned.
- Visual concept knowledge can be characterized by how competitor similarity influences word recognition.
- However, previous work operationalizes similarity dichotomously and subjectively.
- Infant gaze data are also hard to collect and thus tend to include small sample sizes and item sets.

Questions

1. Will infants be more drawn away from a target the more similar it is to a distractor?
2. Do additional item-level differences influence infants' looking behavior?

Items employed in the study design from THINGS-plus

Methods

90 14-24- month old infants

Each infant is shown 32 trials:

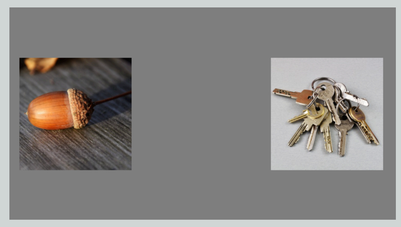
16 easy, 16 hard.

Data collected asynchronously on Children Helping Science

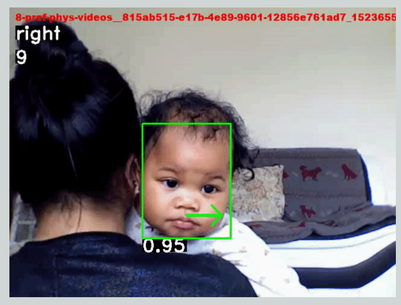
Data passed through iCatcher+ for automated left-right-away gaze coding

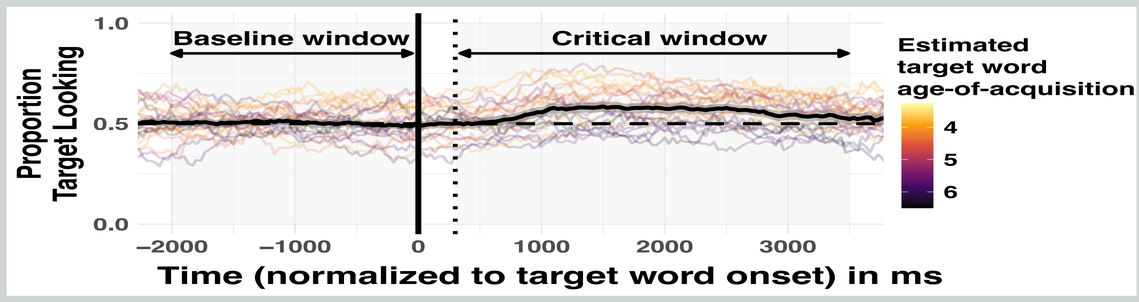
Proportion of looking time to target over distractor correlated with target-distractor similarity

Cosine sim of language and vision embeddings from CLIP



Prompt: "Look at the acorn"





Results

Infants look more at target images the more similar they are in image similarity space (confirmed with a linear-mixed effects model: stats)

Text similarity effect trends in the same direction but is not statistically significant (stats)

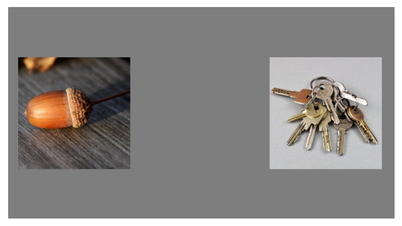
Examining the precision of infants' visual concepts by leveraging vision-language models and automated gaze coding

Tarun Sepuri, Martin Zettersten, Bria Long
University of California, San Diego

VISUAL LEARNING LAB
UC San Diego

Background

- How precise are the visual concepts that support children's rapid early word learning?[1,2]
- Visual concept knowledge can be characterized by how competitor similarity influences word recognition.
- However, previous work operationalizes similarity dichotomously and subjectively.[3,4]
- Infant gaze data are also hard to collect and thus tend to include small sample sizes and item sets.[5]



"Look at the acorn!"
Example prompt

Hypothesis

Infants have partial visual knowledge of many words they appear to not recognize

Approach: Systematically manipulate the similarity of a distractor across word recognition trials

Items employed in the study design from THINGS+[9]

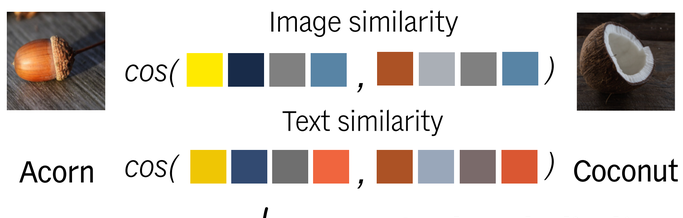
Measuring saliency differences:

Example GBVS[10] saliency map

Mean saliency: 0.227 — Mean saliency: 0.127

GBVS saliency difference: 0.100

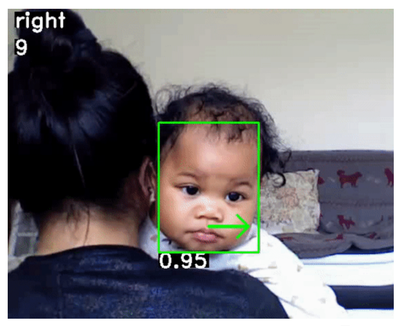
Measuring similarity predictors:



Cosine similarity of CLIP embeddings as a predictor of **Baseline-corrected proportion target looking** ← *Proportion target looking in critical window - baseline window*

Methods

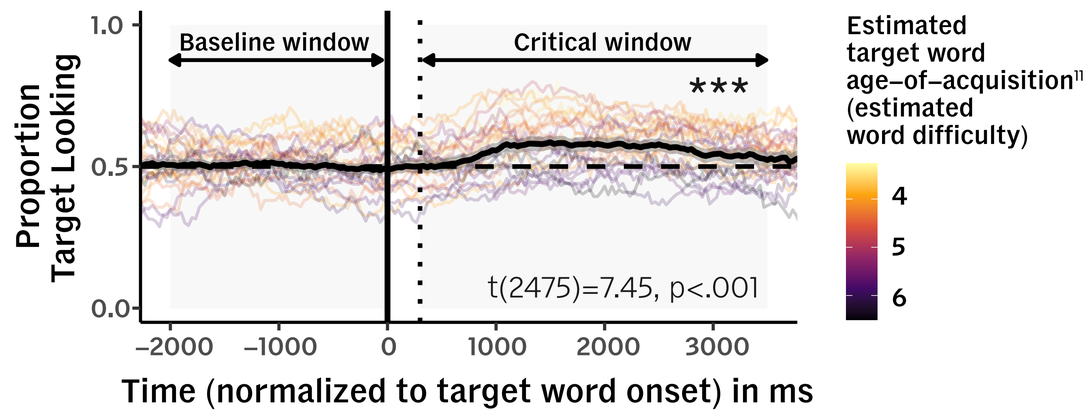
- N=91 children between 14 and 24 months
- Asynchronous data collection with Children Helping Science[6]
- Data passed through iCatcher+[7] for automated left-right gaze coding
- 32 randomized trials: 16 low text similarity, 16 high text similarity
- Similarity defined with vision-language model CLIP[8]



iCatcher+ output example[7]

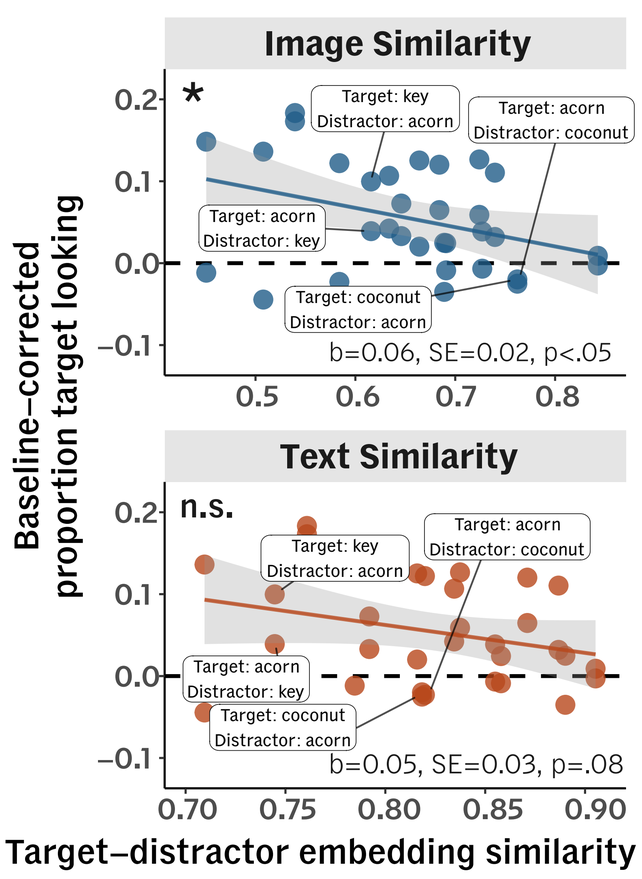
Measuring the proportion of target looking: $\frac{\text{Target looking}}{\text{Target + distractor looking}}$

Infants show above-chance word recognition across items



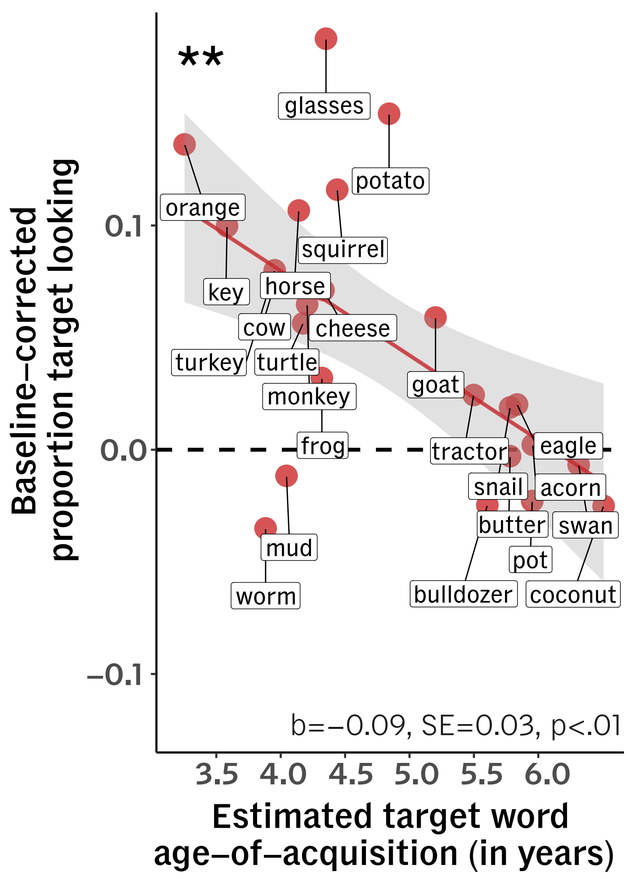
Results

Infants look less at target images the more similar they are to distractors

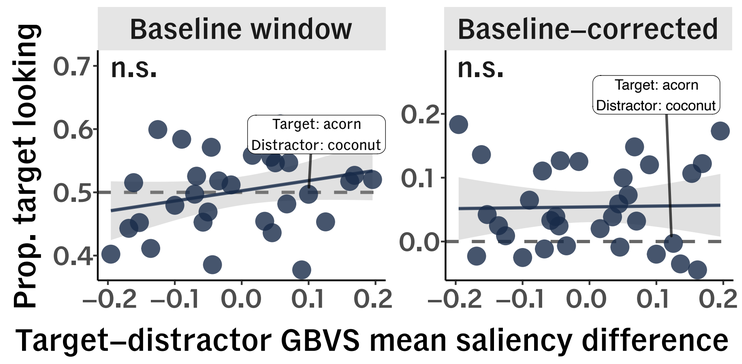


Results confirmed with linear mixed-effects models.

Infants look less at target images the more difficult they are to recognize

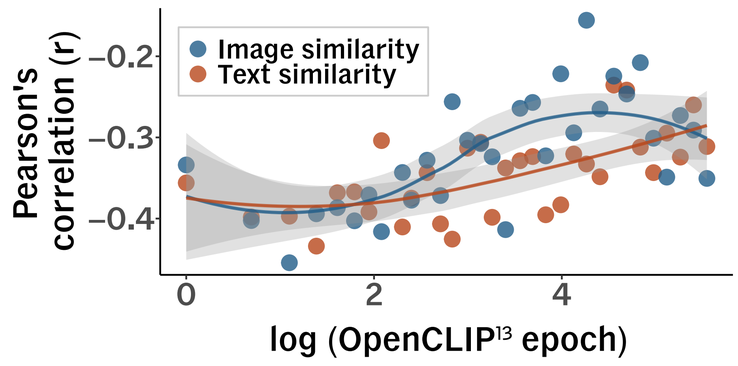


Visual saliency differences are not correlated with baseline-corrected prop. target looking



Future work: How does model-infant correspondence change across model training?[12]

Higher values indicate weaker correlations with baseline-corrected looking.



Discussion

- Results suggest that infants have partial visual knowledge of many difficult words in their second year.
- A planned second experiment with an expanded item set will examine the generalizability of these findings.
- Vision-language models and gaze annotation techniques can be jointly used to further investigate the development of infants' visual concept knowledge.

References: 1) Bergelson 2020. 2) Wagner et al., 2013. 3) Arias-Trejo & Plunkett, 2010. 4) Bergelson & Aslin, 2017. 5) Bergmann et al., 2018. 6) Scott & Schulz, 2017. 7) Erel at al., 2023. 8) Radford et al., 2021. 9) Stoinski et al., 2024. 10) Harel et al., 2006. 11) Kuperman et al., 2012. 12) Tan et al., 2024. 13) Ilharco et al., 2021.