

# Dynamic visual grounding via video

LOT School 2024

Grzegorz Chrupała



Most work on visual grounding of spoken language has used static images, due both to the wider availability of curated captioned image datasets, as well as the complexity of modeling two weakly synchronized modalities in the temporal domain.

However, for humans visual perception is inherently extended in time, and visual grounding of language based on still images is arguably suboptimal also from an engineering point of view. Many aspects of language are related to processes or actions evolving over time rather than static states and may be more naturally grounded in video.



Instructional cooking videos collected from YouTube have been used in several papers.

A typical video captures the person carrying out the various actions involved in cooking, while they are explaining what is being done.

Clip from <http://youcook2.eecs.umich.edu/>

Another dataset, HowTo100M (<https://www.di.ens.fr/willow/research/howto100m/>), consists of other types of instructional videos featuring narrative descriptions.



Another dataset, Spoken Moments in Time  
(<http://moments.csail.mit.edu/spoken.html>),  
features short video clips paired with spoken  
captions describing what's happening.

This dataset is very close in nature to the Spoken COCO  
dataset, but instead of static images it contains clips.

Both the cooking videos and the captioned clips are  
somewhat special as they contain someone  
describing an event or series of events in some detail.

# AVLnet: Learning Audio-Visual Language Representations from Instructional Videos

Andrew Rouditchenko<sup>1\*</sup> Angie Boggust<sup>1\*</sup> David Harwath<sup>2</sup> Brian Chen<sup>3</sup>  
 Dhiraj Joshi<sup>4</sup> Samuel Thomas<sup>4</sup> Kartik Audhkhasi<sup>5</sup> Hilde Kuehne<sup>4</sup> Rameswar Panda<sup>4</sup>  
 Rogerio Feris<sup>4</sup> Brian Kingsbury<sup>4</sup> Michael Picheny<sup>6</sup> Antonio Torralba<sup>1</sup> James Glass<sup>1</sup>  
<sup>1</sup>MIT CSAIL, <sup>2</sup>UT Austin, <sup>3</sup>Columbia University, <sup>4</sup>IBM Research AI, <sup>5</sup>Google, <sup>6</sup>NYU  
 roudi@mit.edu

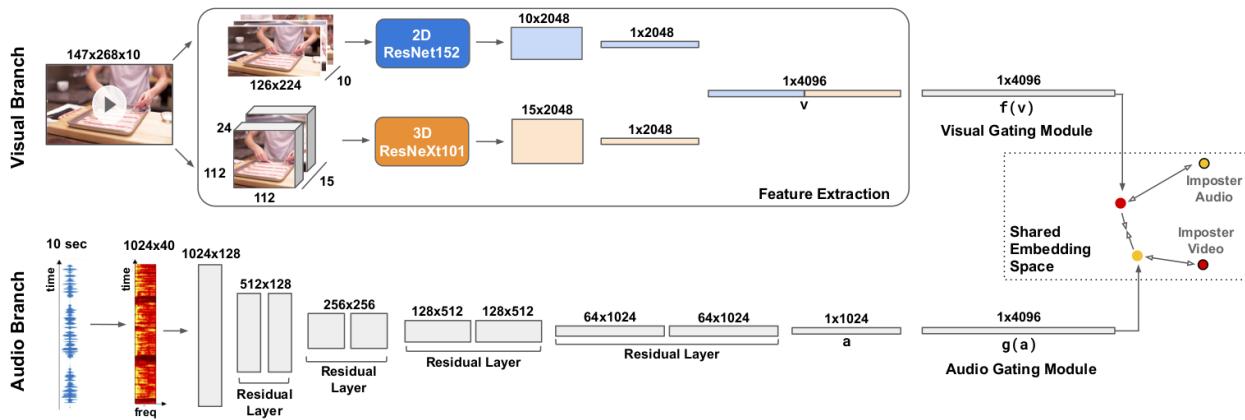


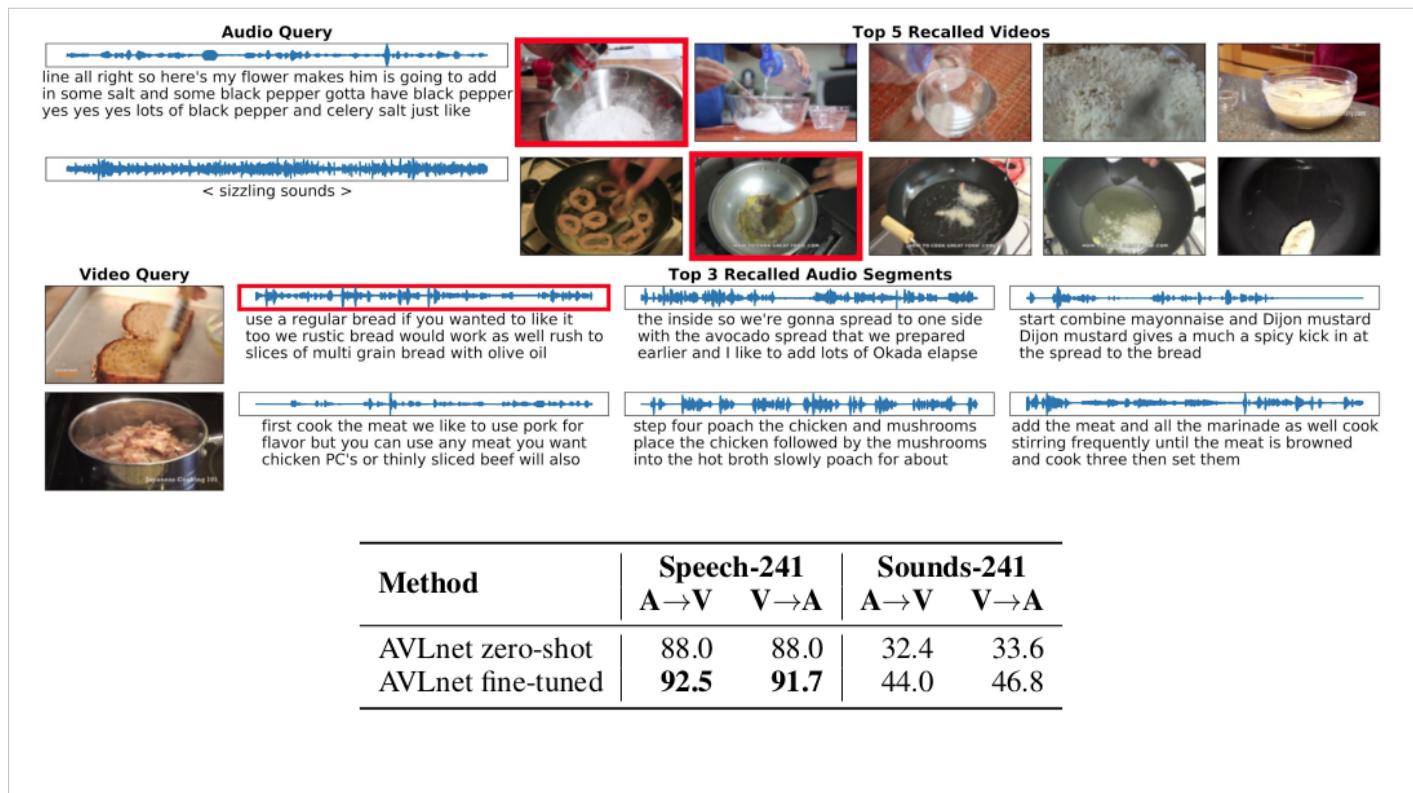
Figure from

<https://doi.org/10.48550/arXiv.2006.09199>

AVLNet consists of an audio encoder (ResNet-based), a video encoder which combines 3D and 2D modeling (also ResNet-based). ResNet is a type of convolutional architecture.

2D convolutions apply to spatial dimensions. 3D Resnet also convolves over the time dimensions. The video encoder then pools each video segment (10s long) into a single vector.

The audio is pooled and projected to the shared embedding space, and the loss function is similar to what we saw with models grounded via still images. The video encoder is pretrained on video classification, while the audio encoder is trained from scratch.



“Figure 3: Video (top) and audio retrieval (bottom) results from AVLnet fine-tuned on YouCook2. Video clips are represented as their center frame, and audio clips are represented as their waveform and ASR transcript. The correct match is highlighted.”

The table (bottom) compares retrieval performance on clips which contain speech (Speech-241) with those which only contain non-speech sounds (Sounds-241).



A: They are fighting over bread.  
C: What?  
A: They are fighting over bread.  
C: Yeah.

An brief real-world verbal interaction between a child and an adult in a city park.

The dialog is in Persian, and has been rendered in English above.

How does a video clip of natural interaction compare to a typical instructional video, or a captioned clip?

What are the implications for ecological validity of instructional videos, or captions?

| <b>Image/video descriptions</b>  | <b>Real world interactions</b>   |
|--|--|
| <p>Strong correlations between visual scene and the meaning of the utterances.</p> | <p>Weaker correlations between visual scene and meaning of the utterances.</p>       |
| <p>Few/weak correlations between scene and non-semantic features of the audio.</p> | <p>Correlations between scene and non-semantic features of the audio are common.</p> |

## Learning English with Peppa Pig

Mitja Nikolaus  
Aix-Marseille University  
mitja.nikolaus@univ-amu.fr

Afra Alishahi  
Tilburg University  
a.alishahi@uvt.nl

Grzegorz Chrupała  
Tilburg University  
grzegorz@chrupala.me

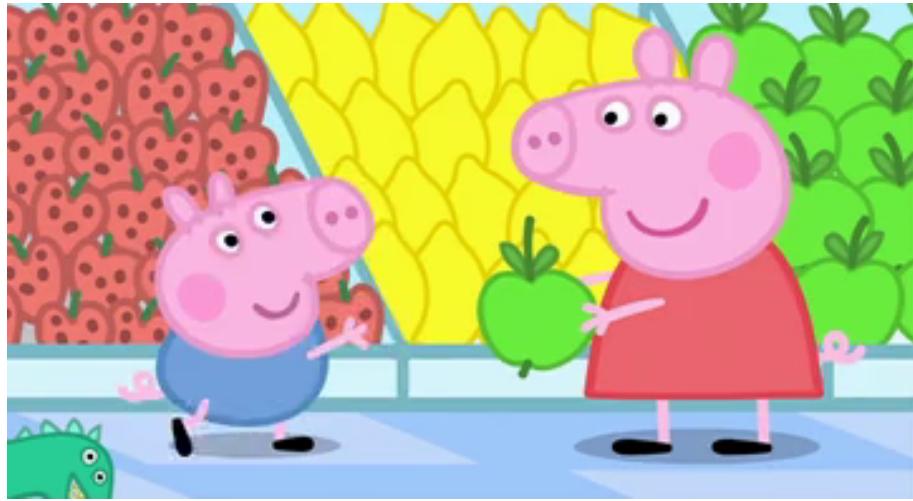


Motivated by considerations of ecological validity we used a children's animated series, Peppa Pig, as a dataset to model language acquisition via visual grounding.

These cartoons are aimed at young children, feature simple graphics and simple but realistic spoken language.

Correlations between speech and visuals tend to be present but are less consistent than in image captions or video descriptions.

There are also confounding correlations between speakers and content of speech, as well as non-speech sounds which can distract the learner from the semantic content of speech. This is what we want in a realistic dataset.



Can you think of a way to use the two different types of speech (dialogs and narrations) in the cartoons for different purposes?

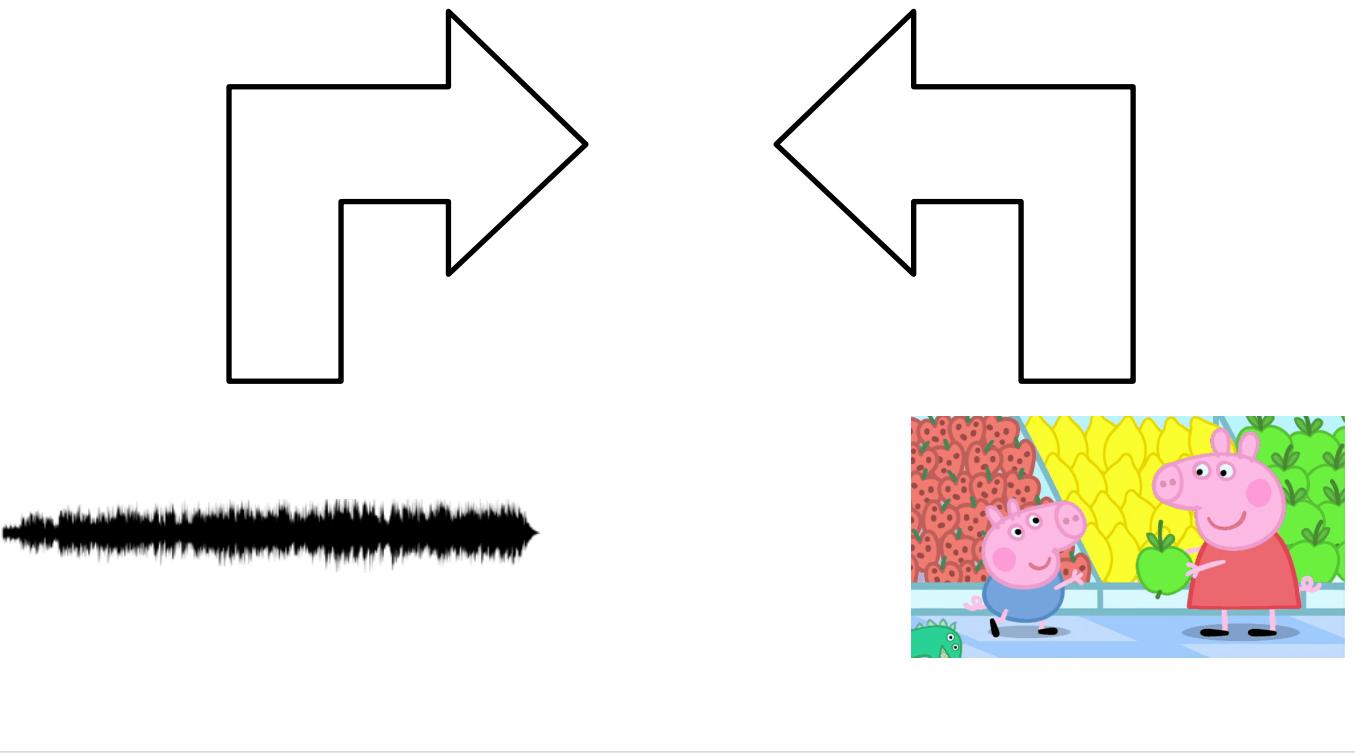
We use dialogs between characters to train the model.

The narrations are excluded from the training data and held out to be used only for evaluation.

The narrations are descriptive and caption-like in nature. The narrator voice is not present in the training data.

Thus, if the model trained on dialogs generalizes to narrations, this is evidence that it is using semantic features of speech, and doesn't rely on confounds.

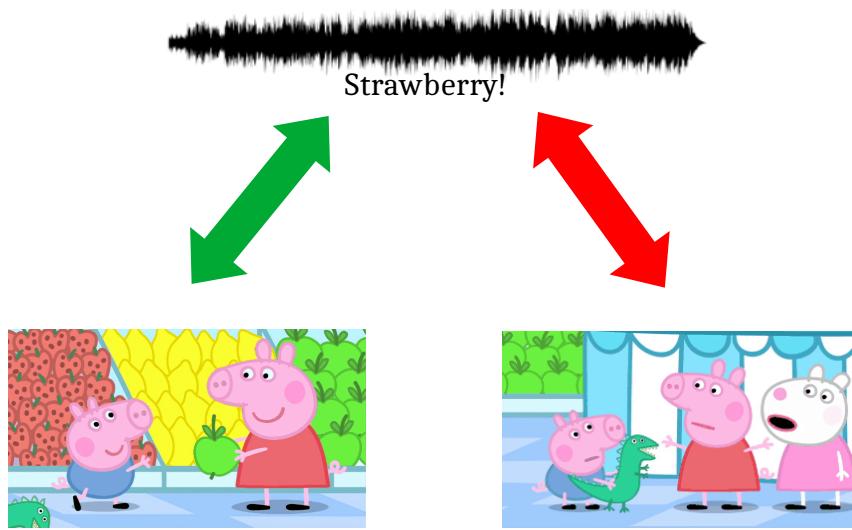
| Split | Type      | Size (h) | # Clips |
|-------|-----------|----------|---------|
| train | dialog    | 10.01    | 15666   |
| val   | dialog    | 0.66     | 1026    |
| val   | narration | 0.94     | 1467    |
| test  | narration | 0.64     | 1006    |



We use a simple architecture with a wav2vec2-based audio encoder, and a ResNet(2+1)D video encoder.

We evaluate both pretrained encoders as well as training from scratch.

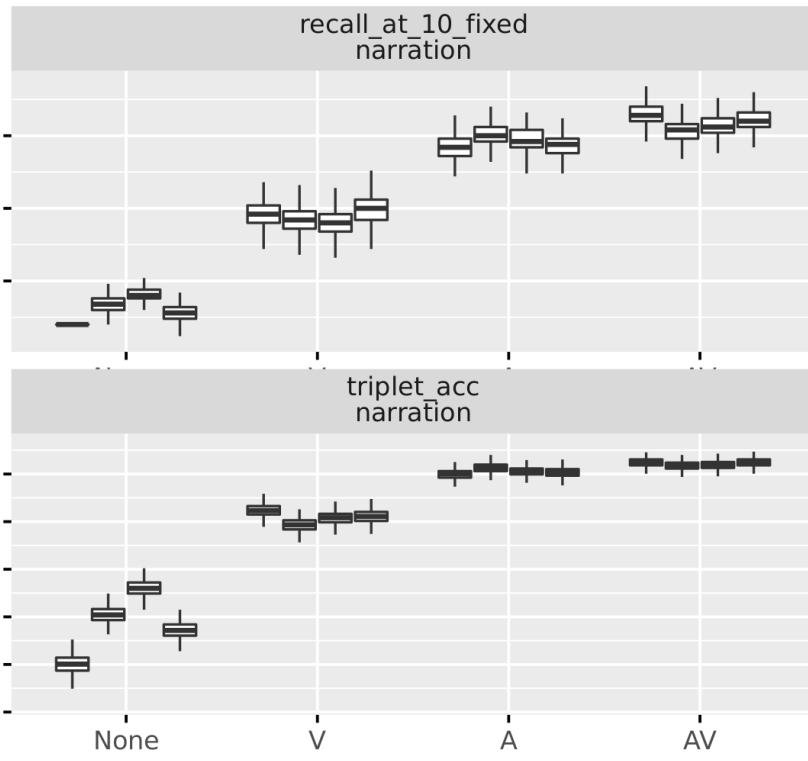
Our objective is a contrastive triplet-like loss.



We evaluate using the usual video clip retrieval setting.

Additionally we also use an alternative metric with simpler stimuli:

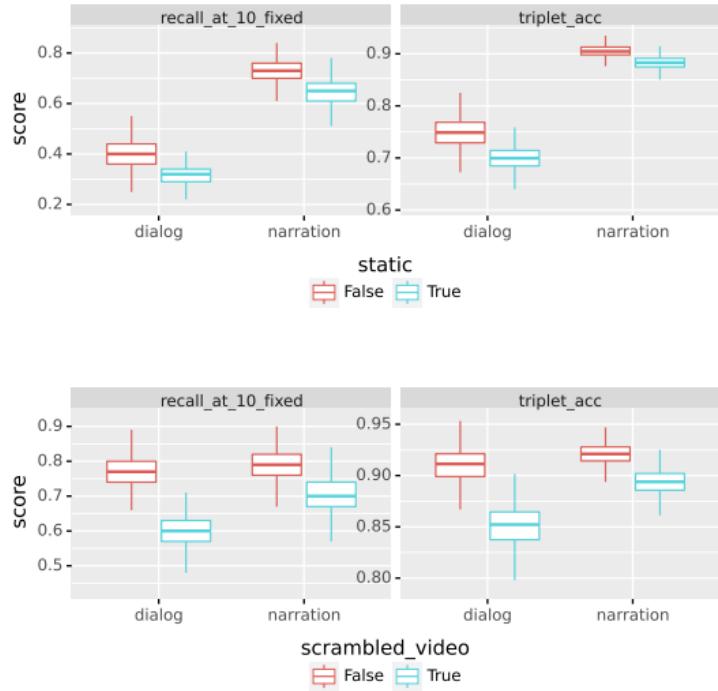
- Extract video/audio clips aligned with subtitle lines
- Repeat 500 times
  - Sample triplets of same length
  - Compute success rate: anchor closer to **positive** than **negative**



We do a number of ablations, where we remove a component from the modeling setup and check the impact of this removal on performance.

Regarding the effect of pretraining, we see that pretraining always helps, and for consistent performance it is necessary to pretrain at least one of the encoders.

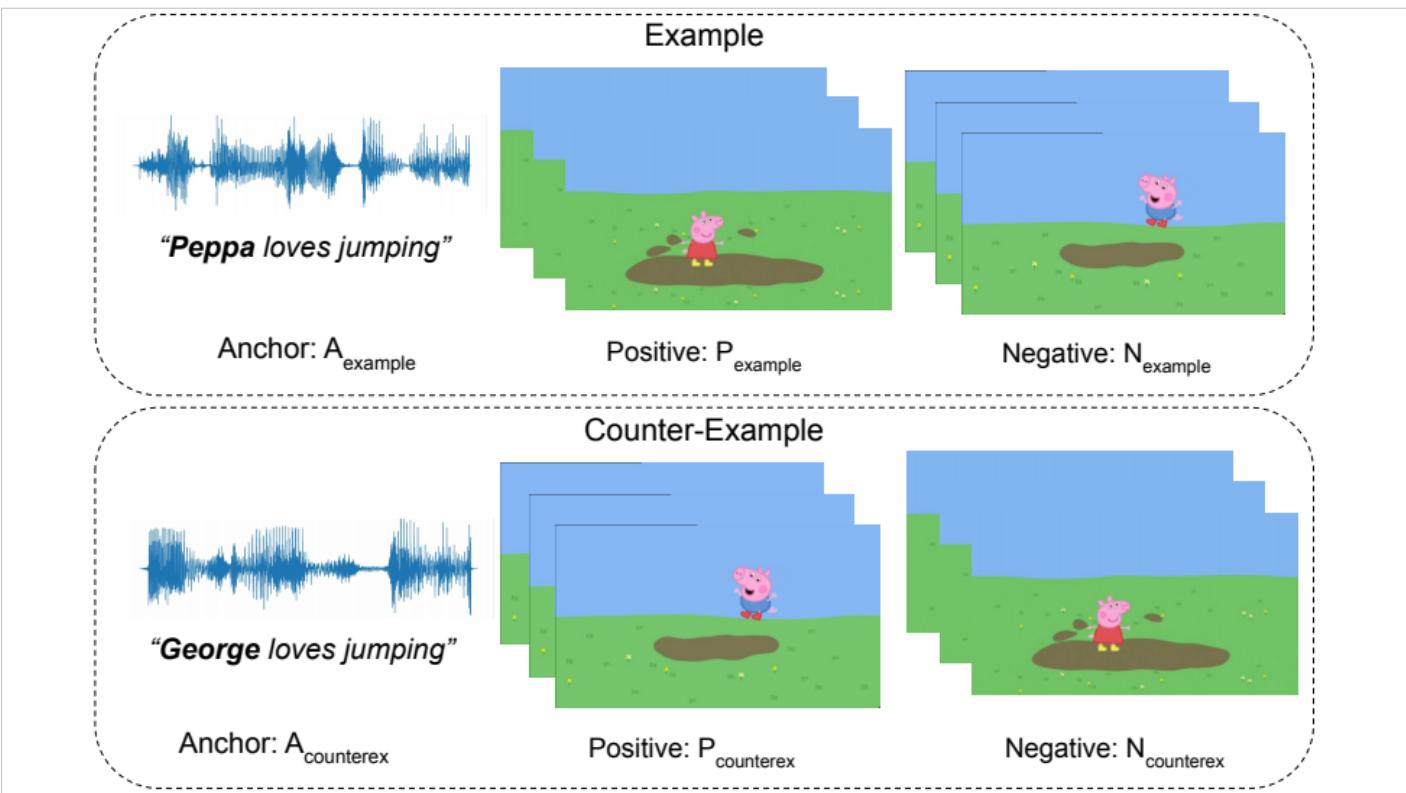
How would you check to what extent the temporal dimension of the video clips was important for the model?



We check the effect of ablating the temporal dimension in two different ways:

- 1) Video encoder based on a 2D ResNet which encodes each video frame separately, and then pools the encoded frames.
- 2) Scrambling the video frames before feeding them to the video encoder (only during evaluation).

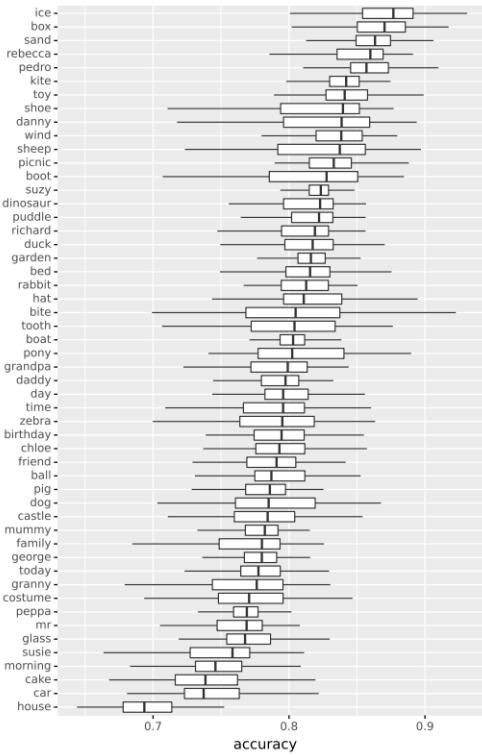
In both cases we see that removing the temporal signal substantially lowers the evaluation scores of the model.



A targeted, controlled triplet evaluation:

Triplets are matched to differ by a single word (via aligned transcriptions).

This setting allows us to partition results by word or by grammatical category.



What we saw were mostly patterns specific to Peppa cartoons.

Character names (Rebecca, Pedro) were easy. Some surprises included low scores for *house* or *car*: house is used in varying visual contexts (house entrance, whole house, inside the house) and in displaced speech (going to somebody's house).

The pattern of results on verbs vs nouns was not illuminating.

These results highlight one limitation of the data: while ecological validity was addressed to some extent, the data is too small for the findings to generalize beyond the Peppa cartoon setting.

Learning language in a realistic grounded setting is still largely an open problem.