Hallucinating Vision in Neural Networks: the Case of Quantifiers

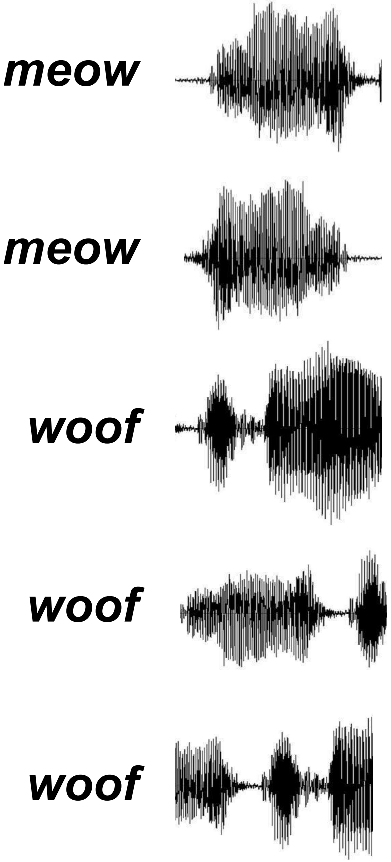
Candidate: Alberto Testoni
Supervisor: Raffaella Bernardi

M.Sc. in Cognitive Science - CIMeC - University of Trento
Language and Multimodal Interaction Track (LMI)

Meaning is Multimodal

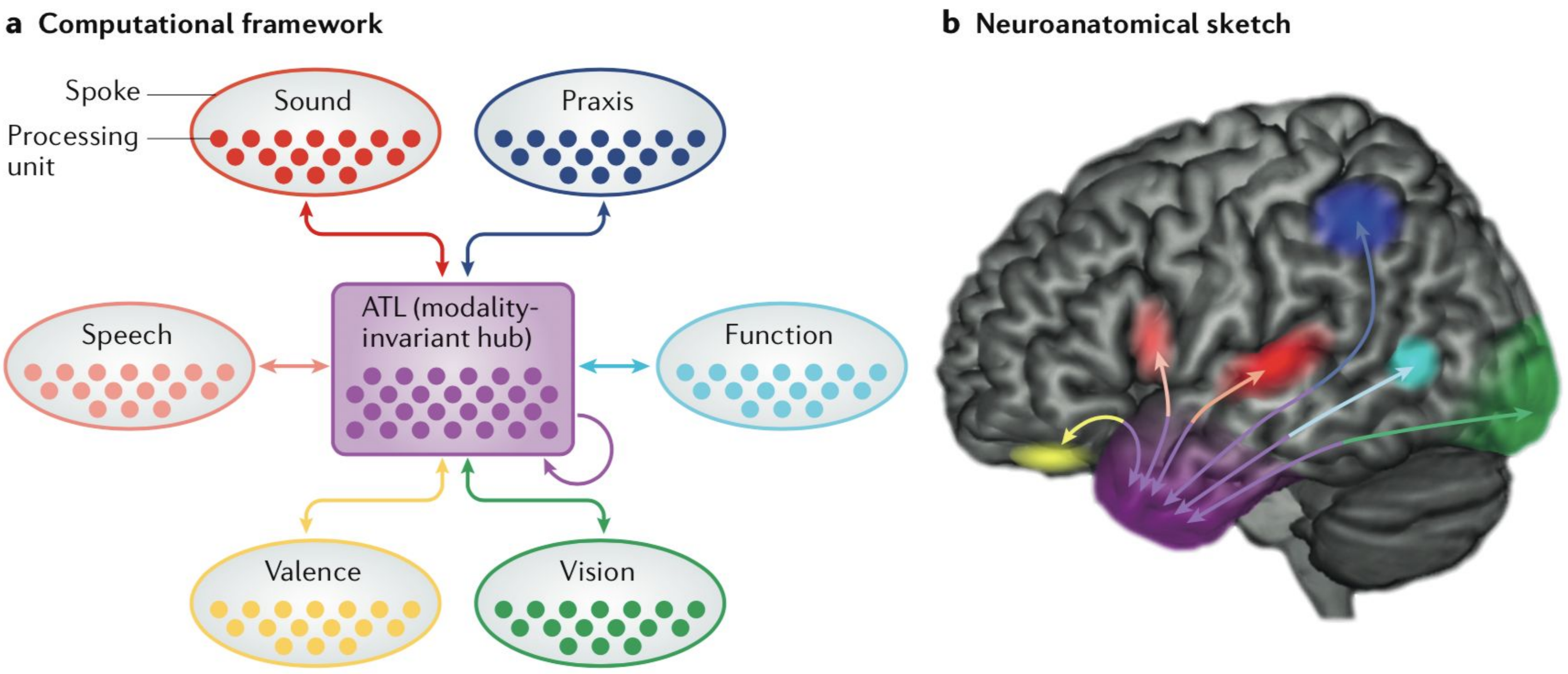
Cross-modal and multi-sensory integration

“There are three dogs and two cats”



Meaning is Multimodal

The Hub-and-Spoke Theory



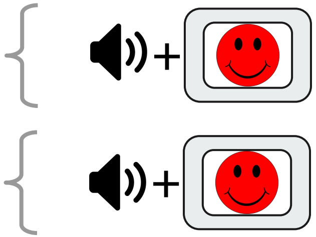
Lambon Ralph et al. , 2017

Meaning is Multimodal

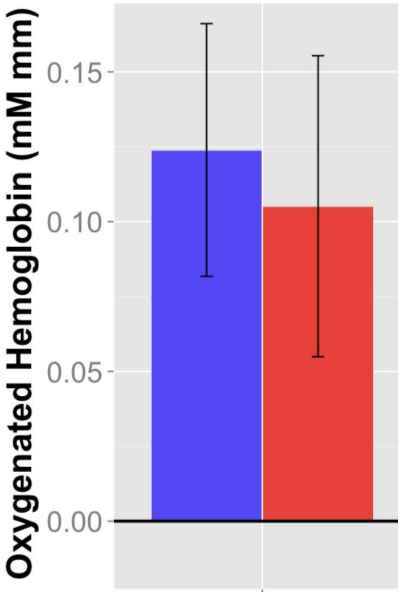
Predictive Coding Model

Prior Knowledge:

Stimulus:



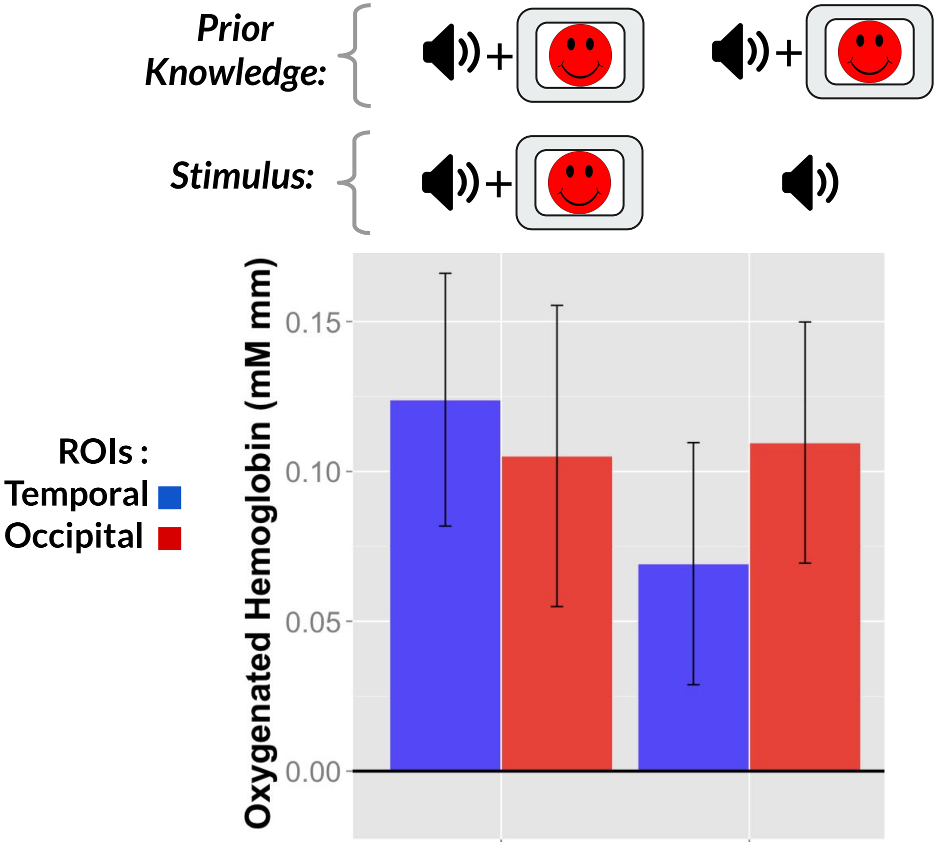
ROIs :
Temporal
Occipital



Emberson et al., 2015

Meaning is Multimodal

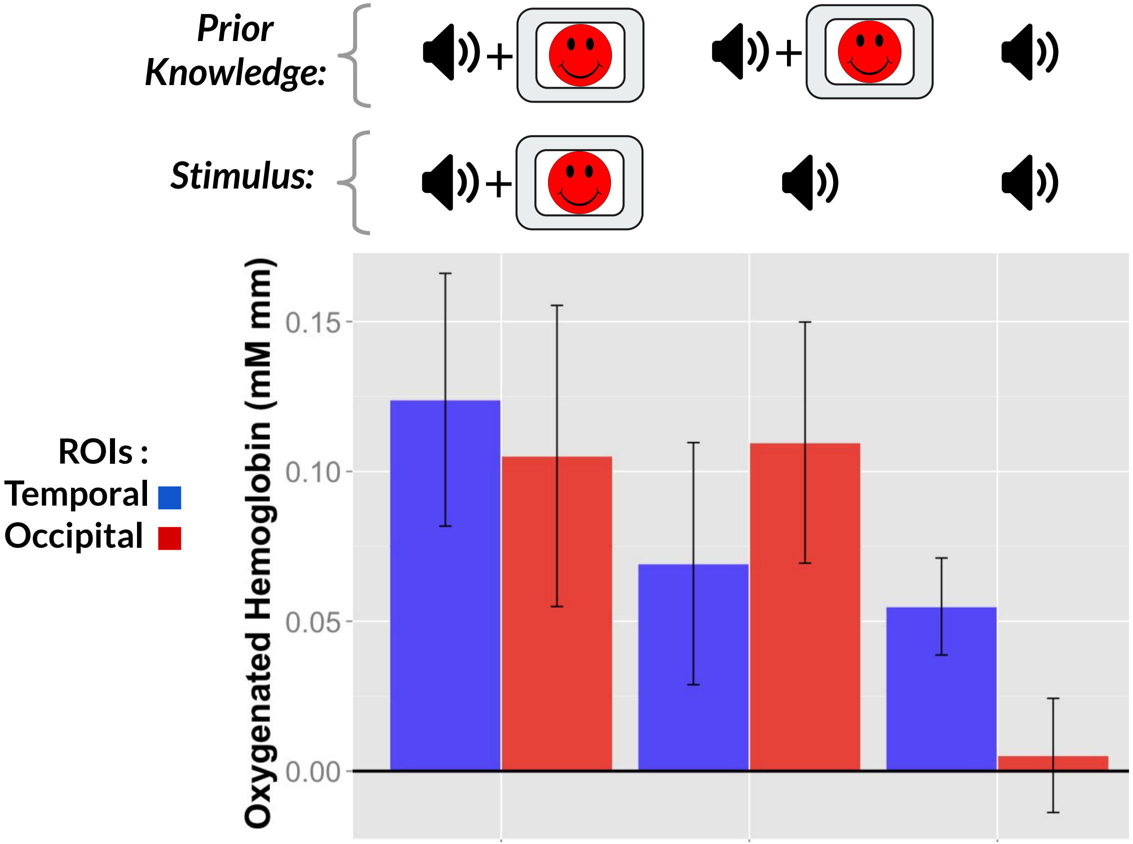
Predictive Coding Model



Emberson et al., 2015

Meaning is Multimodal

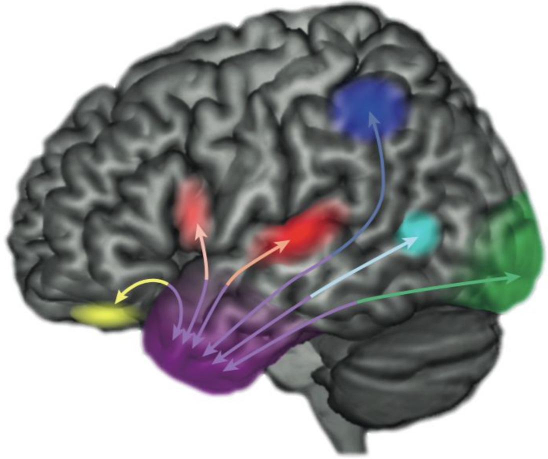
Predictive Coding Model



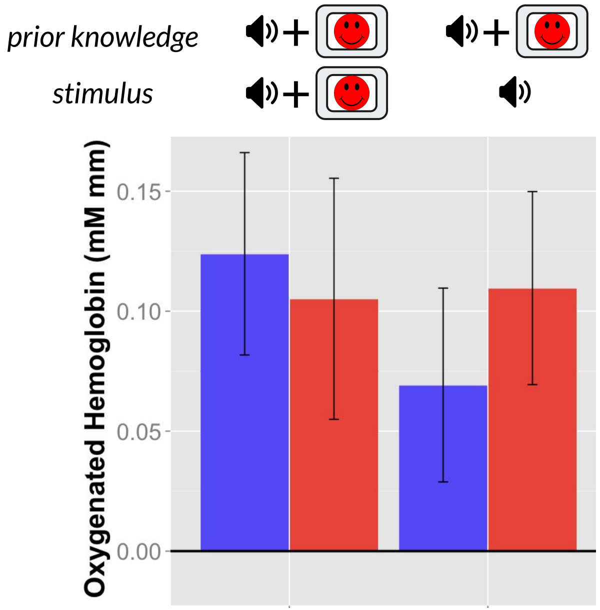
Emberson et al., 2015

Thesis proposal

Integrate multimodal inputs in a single computational "hub"

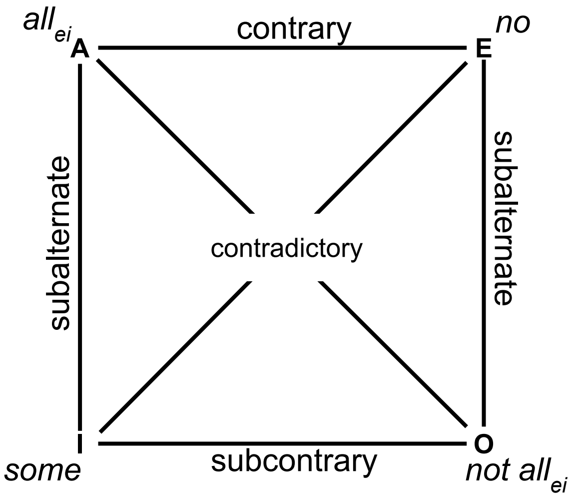


Simulate the effect of prior knowledge on cross-modal activations via hallucination



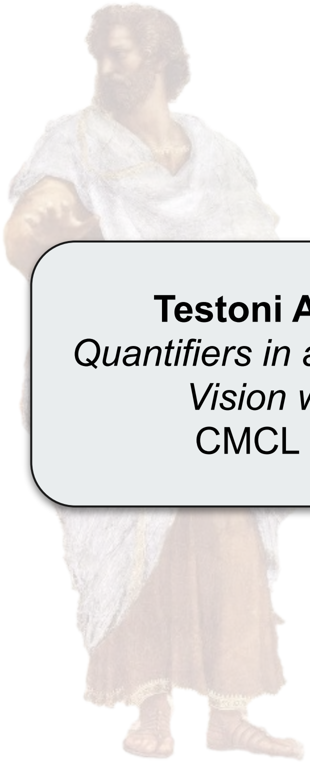
Two Tasks for Computational Models

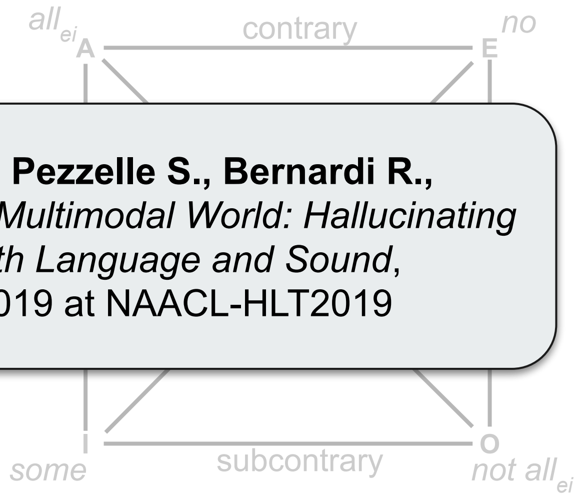
Quantifiers



Two Test-beds for Computational Models

Quantifiers

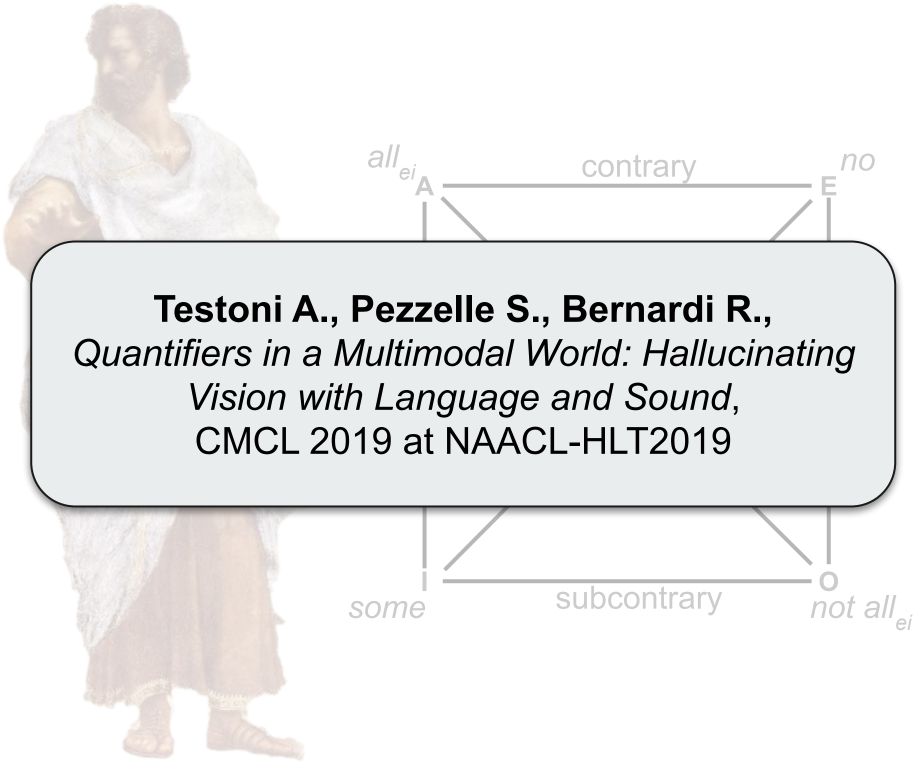




Testoni A., Pezzelle S., Bernardi R.,
Quantifiers in a Multimodal World: Hallucinating Vision with Language and Sound,
CMCL 2019 at NAACL-HLT2019

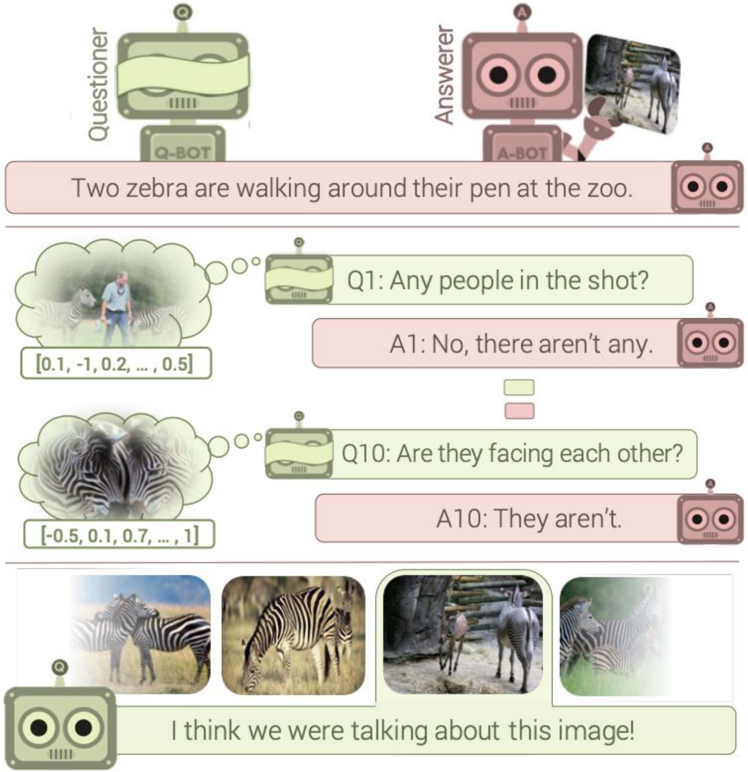
Two Test-beds for Computational Models

Quantifiers



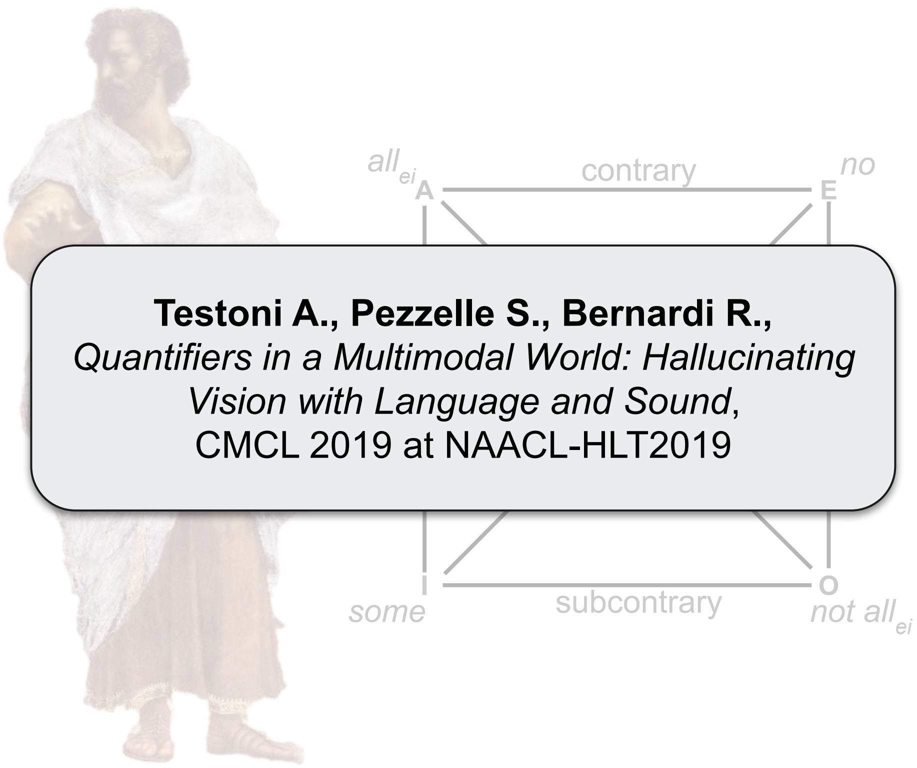
Testoni A., Pezzelle S., Bernardi R.,
Quantifiers in a Multimodal World: Hallucinating Vision with Language and Sound,
CMCL 2019 at NAACL-HLT2019

Conversational Agents



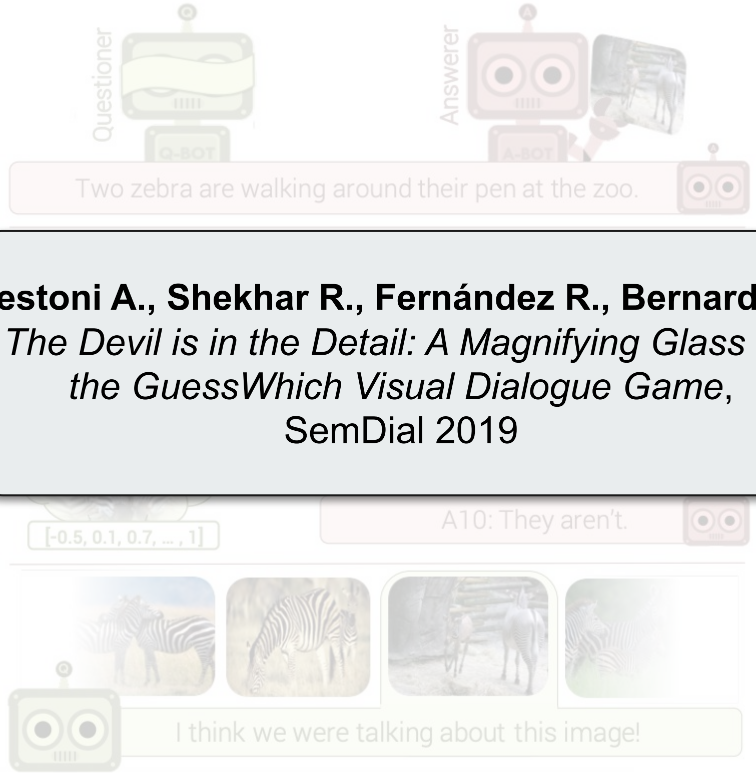
Two Test-beds for Computational Models

Quantifiers



Testoni A., Pezzelle S., Bernardi R.,
Quantifiers in a Multimodal World: Hallucinating Vision with Language and Sound,
CMCL 2019 at NAACL-HLT2019

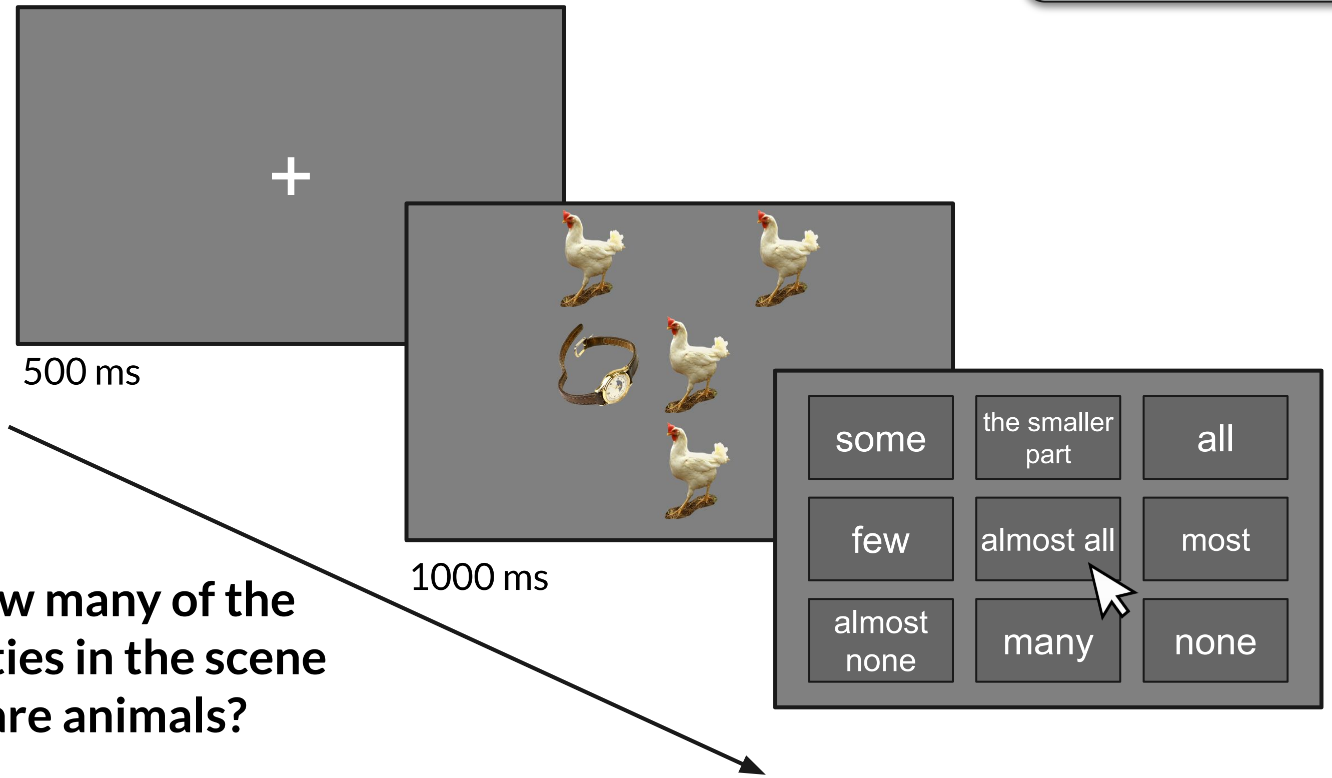
Conversational Agents



Testoni A., Shekhar R., Fernández R., Bernardi R.,
The Devil is in the Detail: A Magnifying Glass for the GuessWhich Visual Dialogue Game,
SemDial 2019

Quantification Task

17 proportions of animals/tools
9 English quantifiers

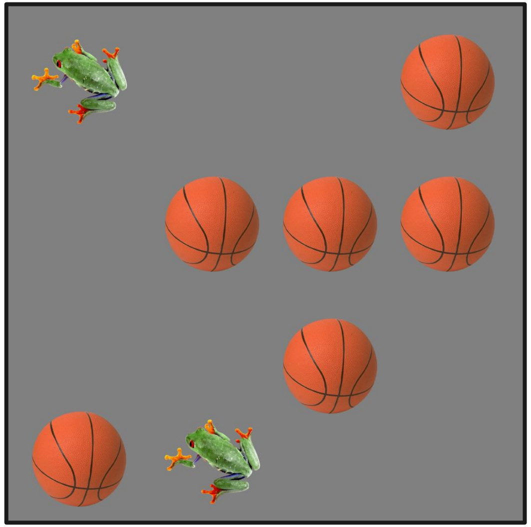


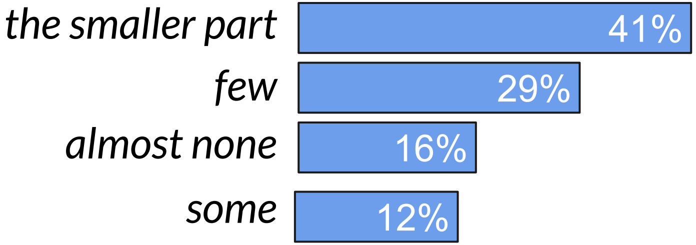
How many of the entities in the scene are animals?

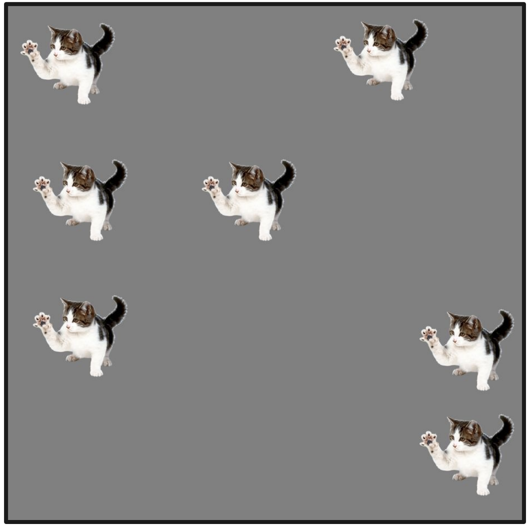
Pezzelle, S., Bernardi, R., & Piazza, M. (2018). Probing the mental representation of quantifiers. Cognition, 181, 117-126.

Quantification Task

none, *almost none*, *few*, *the smaller part*, *some*, *many*, *most*, *almost all*, *all*





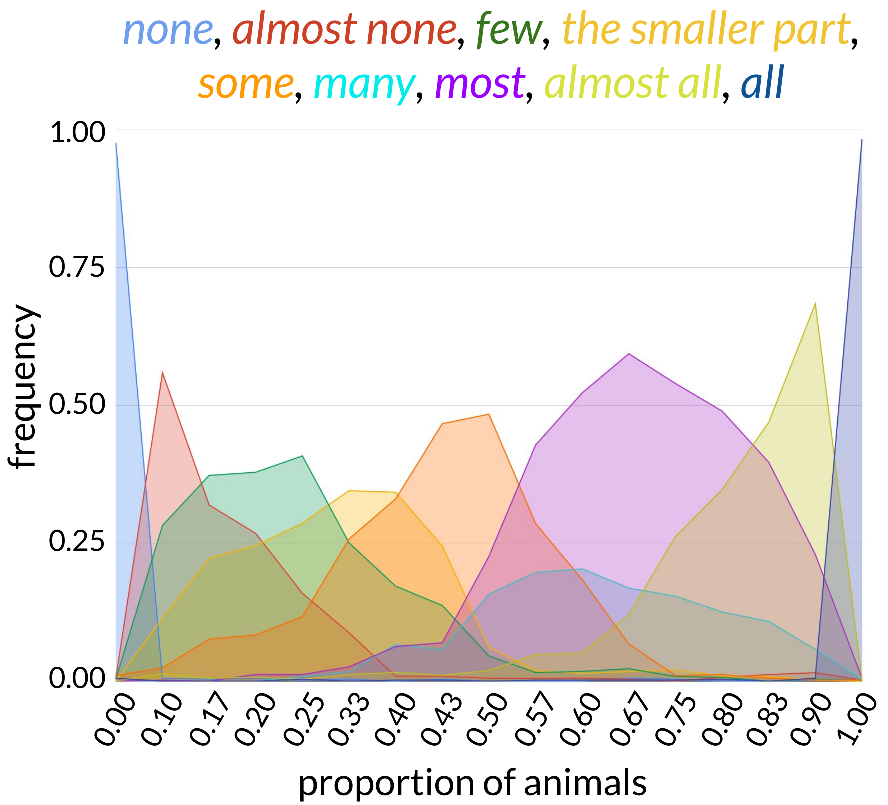


Multimodal Quantification Datasets 🎶 👁 📖

- ➔ 3 modalities: Sound, Vision and Language.
- ➔ Data-points contain animal (*target*) and artifact (*distractor*) entities in 17 different proportions.
- ➔ 17K data-points generated for each modality.
- ➔ Human annotations on 9 English quantifiers from *Pezzelle et al., 2018*.

Multimodal Quantification Datasets 🎵 👁 📖

- ➔ 3 modalities: Sound, Vision and Language.
- ➔ Data-points contain animal (*target*) and artifact (*distractor*) entities in 17 different proportions.
- ➔ 17K data-points generated for each modality.
- ➔ Human annotations on 9 English quantifiers from *Pezzelle et al., 2018.*



Multimodal Dataset

Multimodal Dataset 👁 🎶

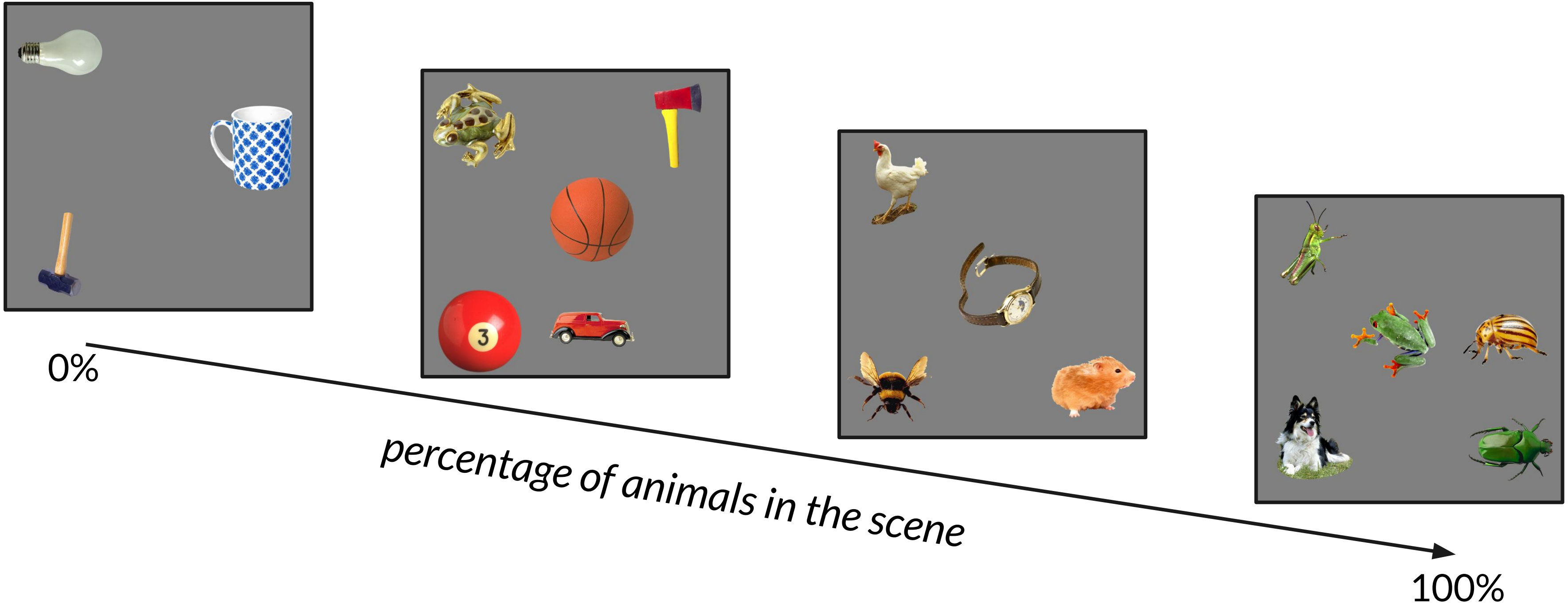
Multimodal Dataset

“There are one clock, two cars, one mammal and one telephone”

Visual Dataset

Inception V3 CNN

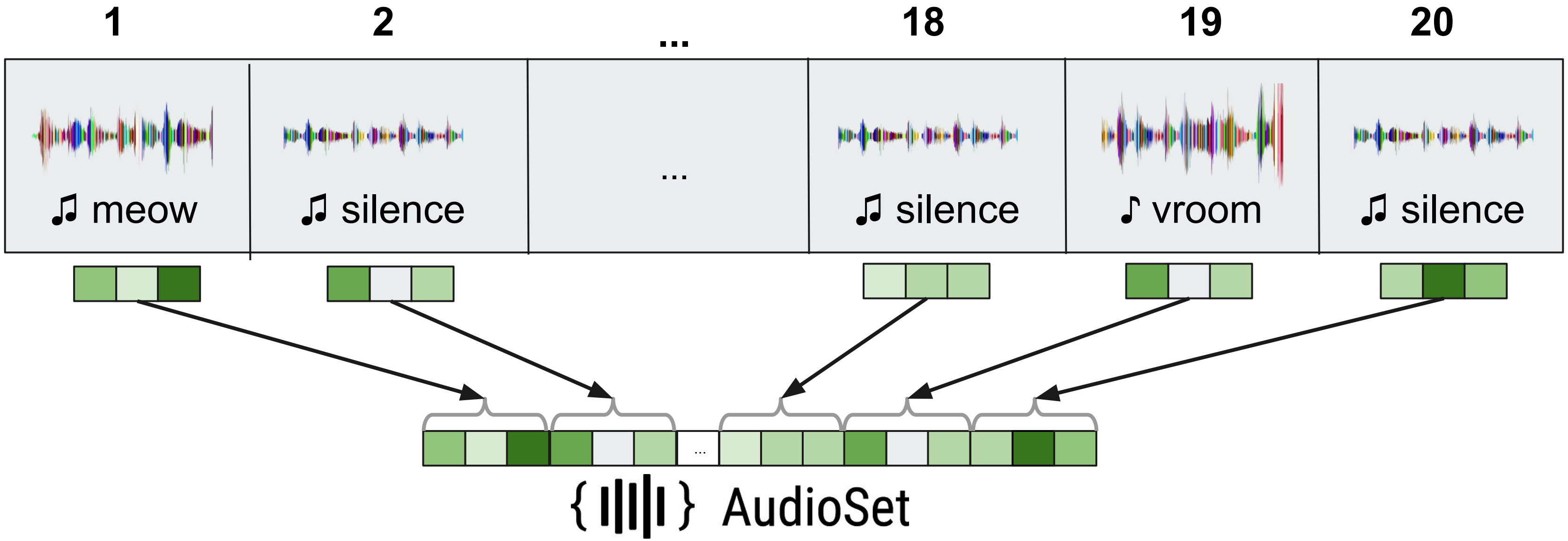
- ➔ 110 entities (55 animals+55 artifacts) manually selected from *Kiani et al., 2007*.
- ➔ Only entities for which a corresponding sound is available.
- ➔ Total number of entities in the scene ranges between 3 and 20.



Auditory Dataset ♬

AudioSet Features

- Starting point: Audioset (*Gemmeke et al., 2017*).
- For each visual entity in a visual scene, we took the corresponding sound.
- Each sound representation is concatenated and a silence sound is added to fill empty "cells" in the final vector.



Linguistic Dataset

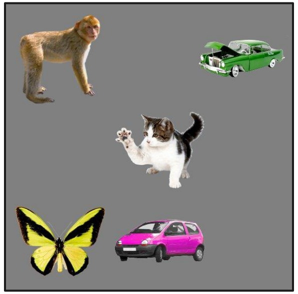
➔ Each entity in the visual scene is manually annotated with 3 nouns expressing different levels of an ontological hierarchy.

➔ One of the three nouns is randomly picked and combined with the others to form a meaningful sentence.

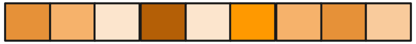
1st LEVEL	MONKEY	BUTTERFLY	CAR	CAT	CAR
2nd LEVEL	PRIMATE	ARTHROPOD	AUTOMOBILE	FELINE	AUTOMOBILE
3rd LEVEL	MAMMAL	INSECT	VEHICLE	MAMMAL	VEHICLE
	MAMMAL	BUTTERFLY	AUTOMOBILE	MAMMAL	AUTOMOBILE

“There are one butterfly, two automobiles and two mammals”

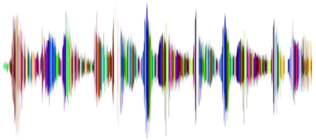
Sensory representations



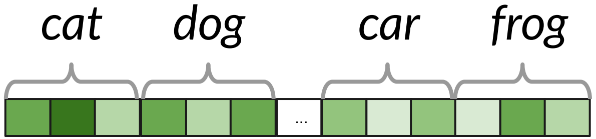
Inception V3 CNN
(*Szegedy et al., 2016*)



2048-d visual vector



AudioSet
(*Gemmeke et al., 2017*)



5120-d acoustic vector

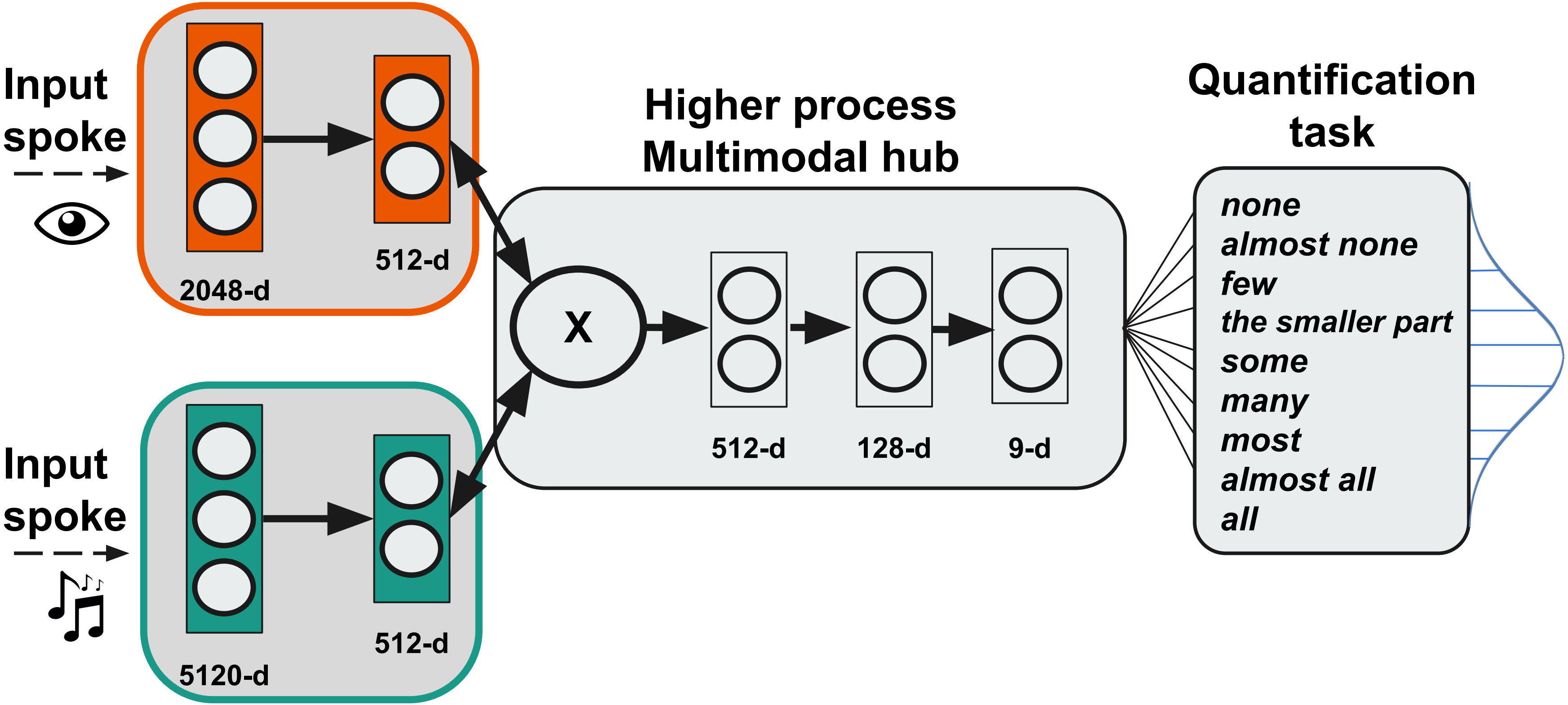
There are one butterfly, two cars and two mammals

Universal Sentence Encoder
(*Cer et al., 2018*)

512-d linguistic vector

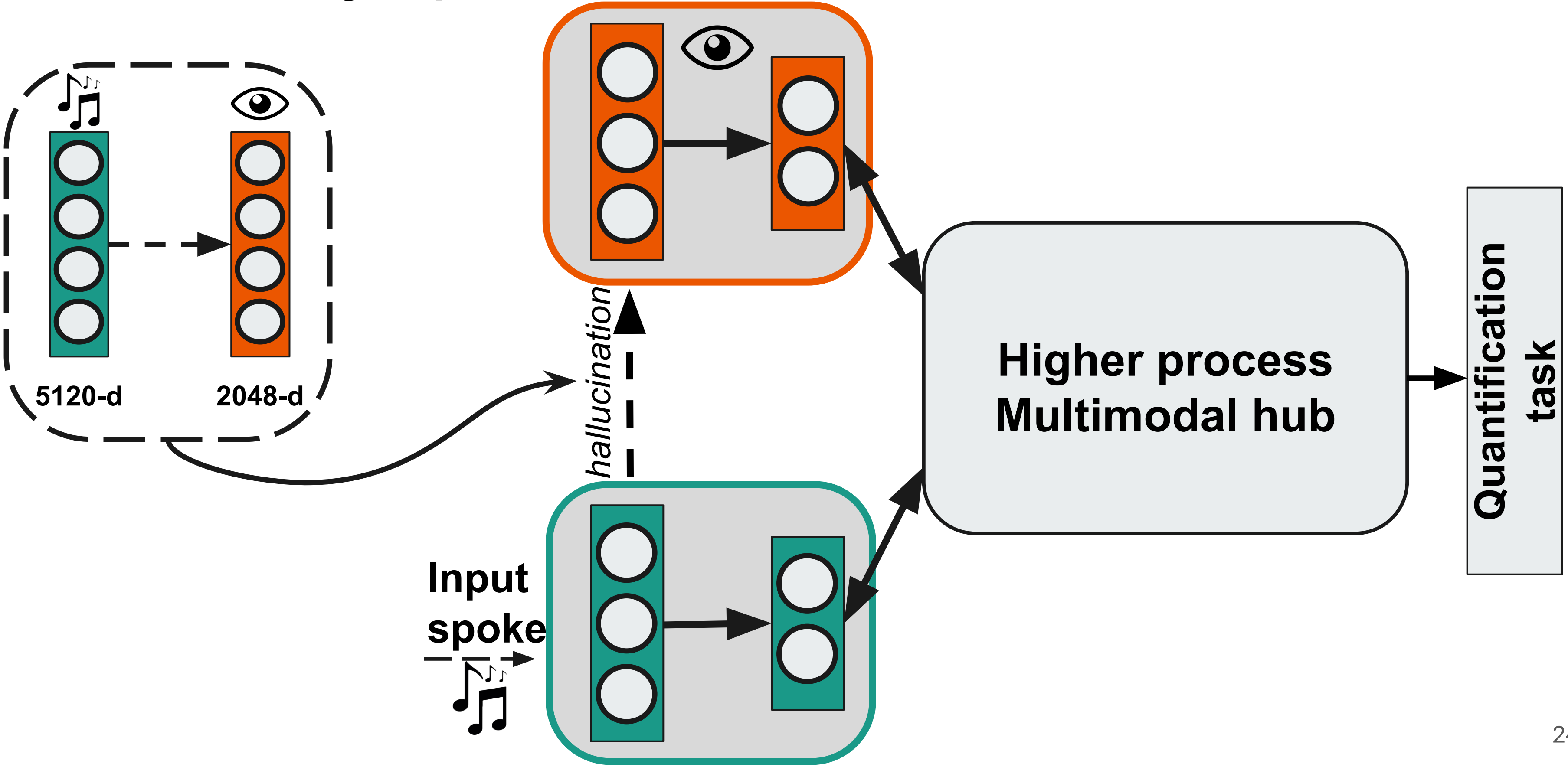
Computational Models

Multi-modal model inspired by the H&S theory of semantic representation



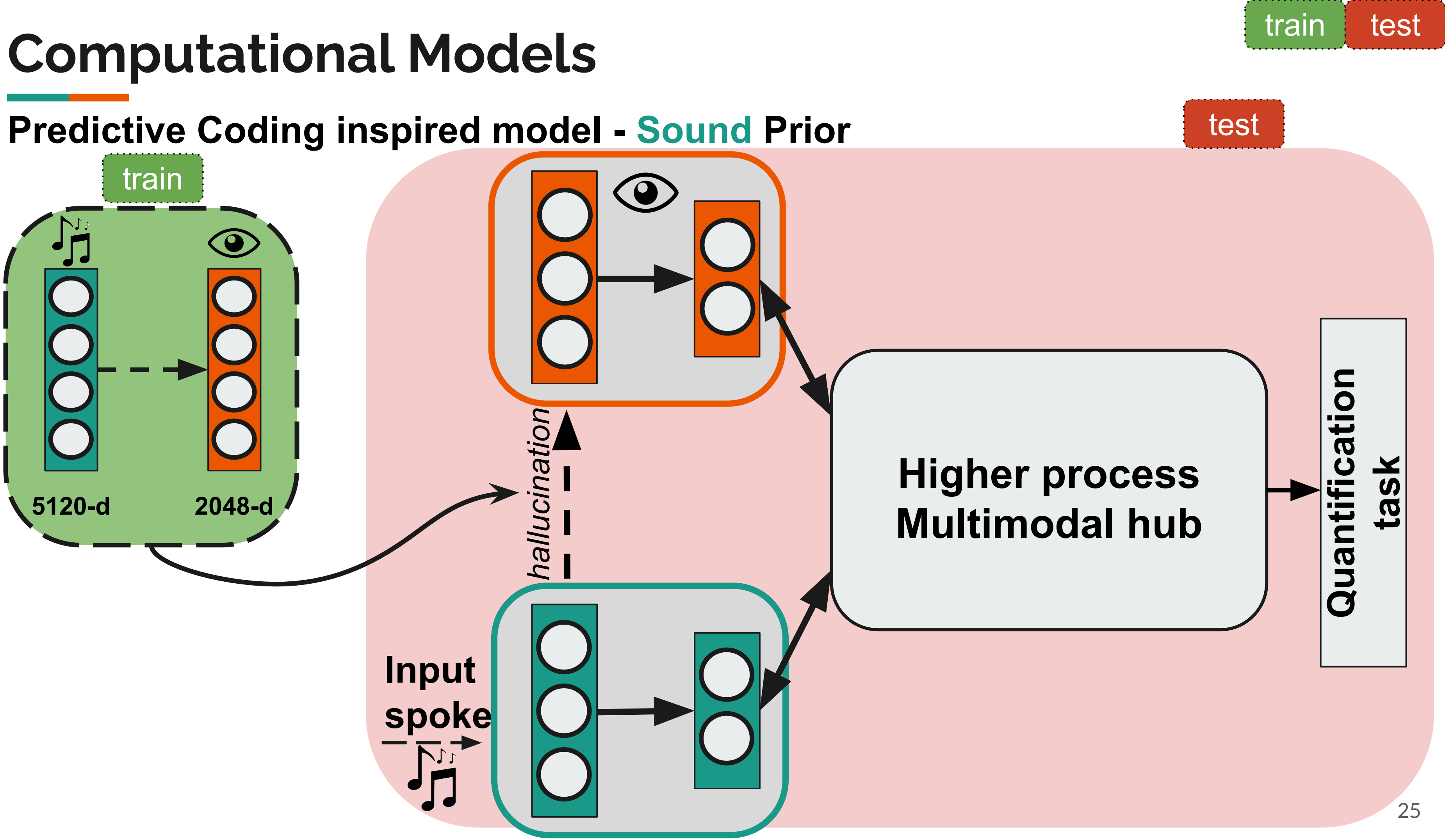
Computational Models

Predictive Coding inspired model - Sound Prior



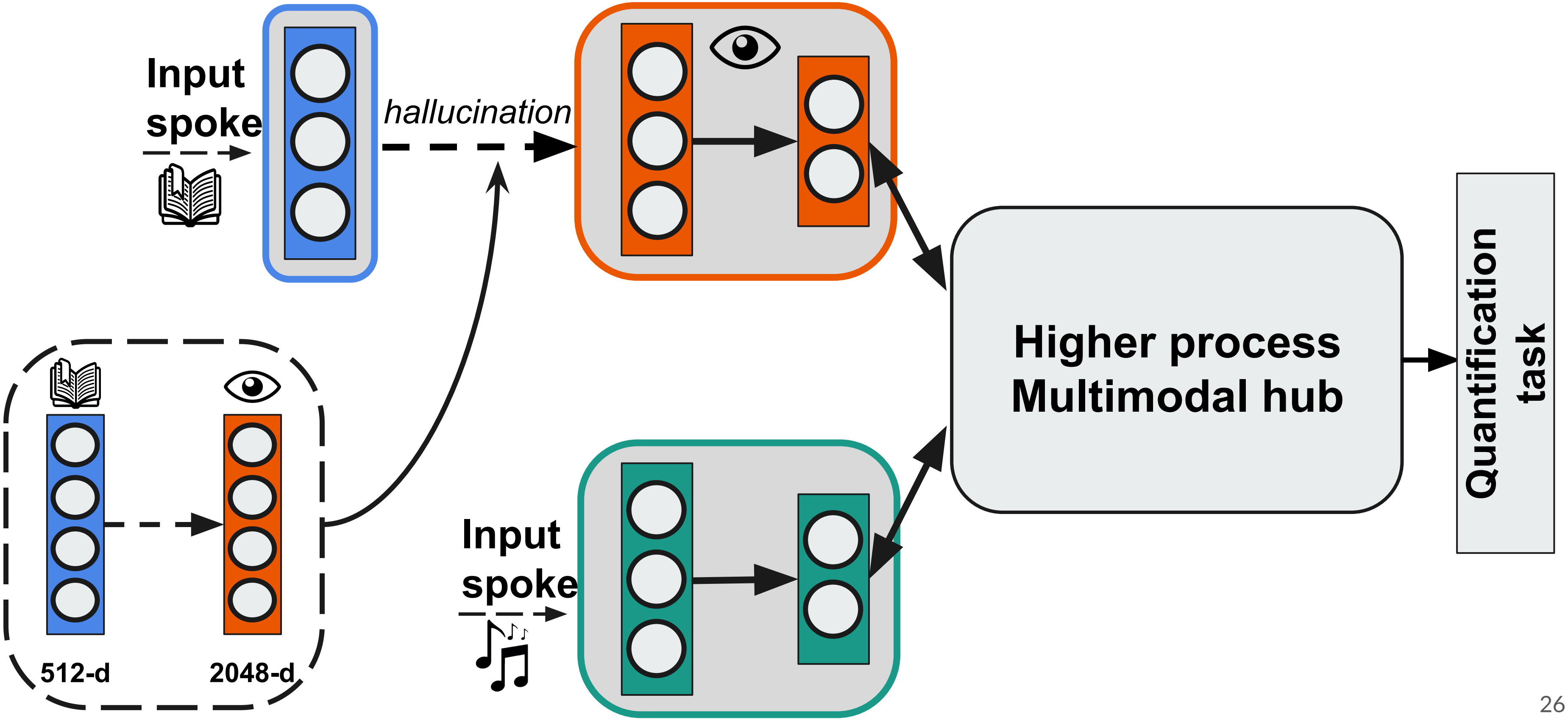
Computational Models

Predictive Coding inspired model - Sound Prior



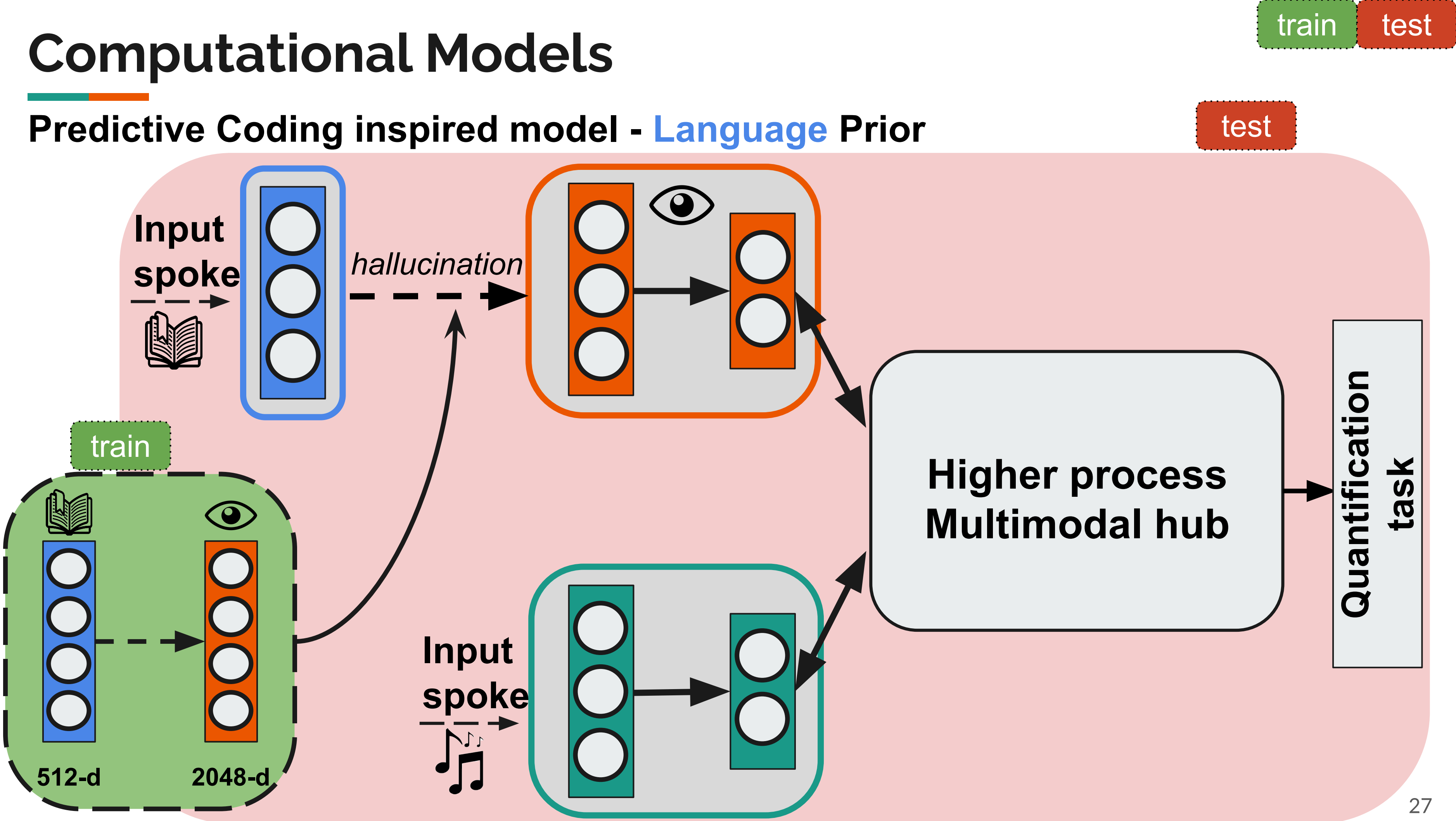
Computational Models

Predictive Coding inspired model - Language Prior



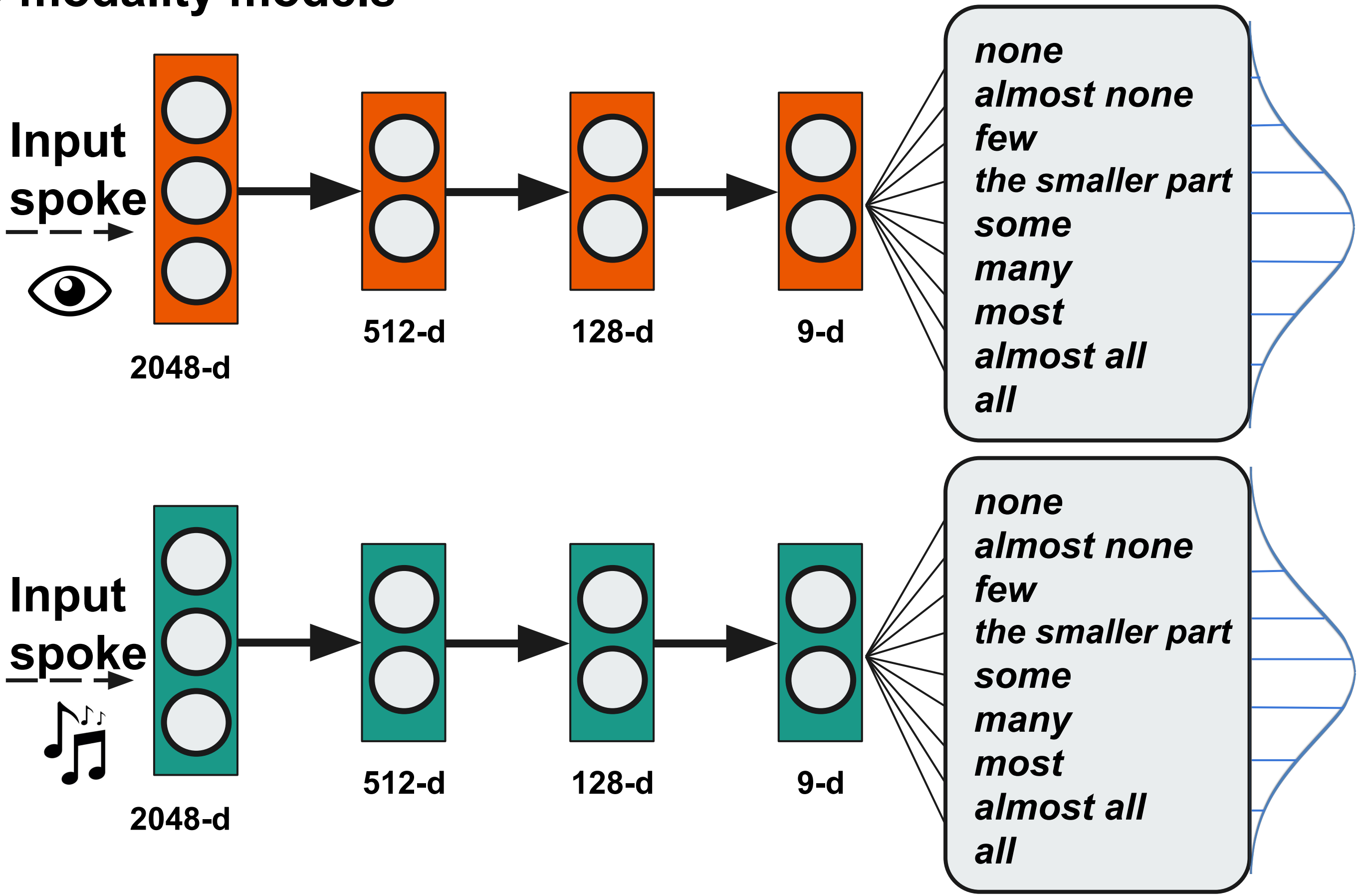
Computational Models

Predictive Coding inspired model - Language Prior



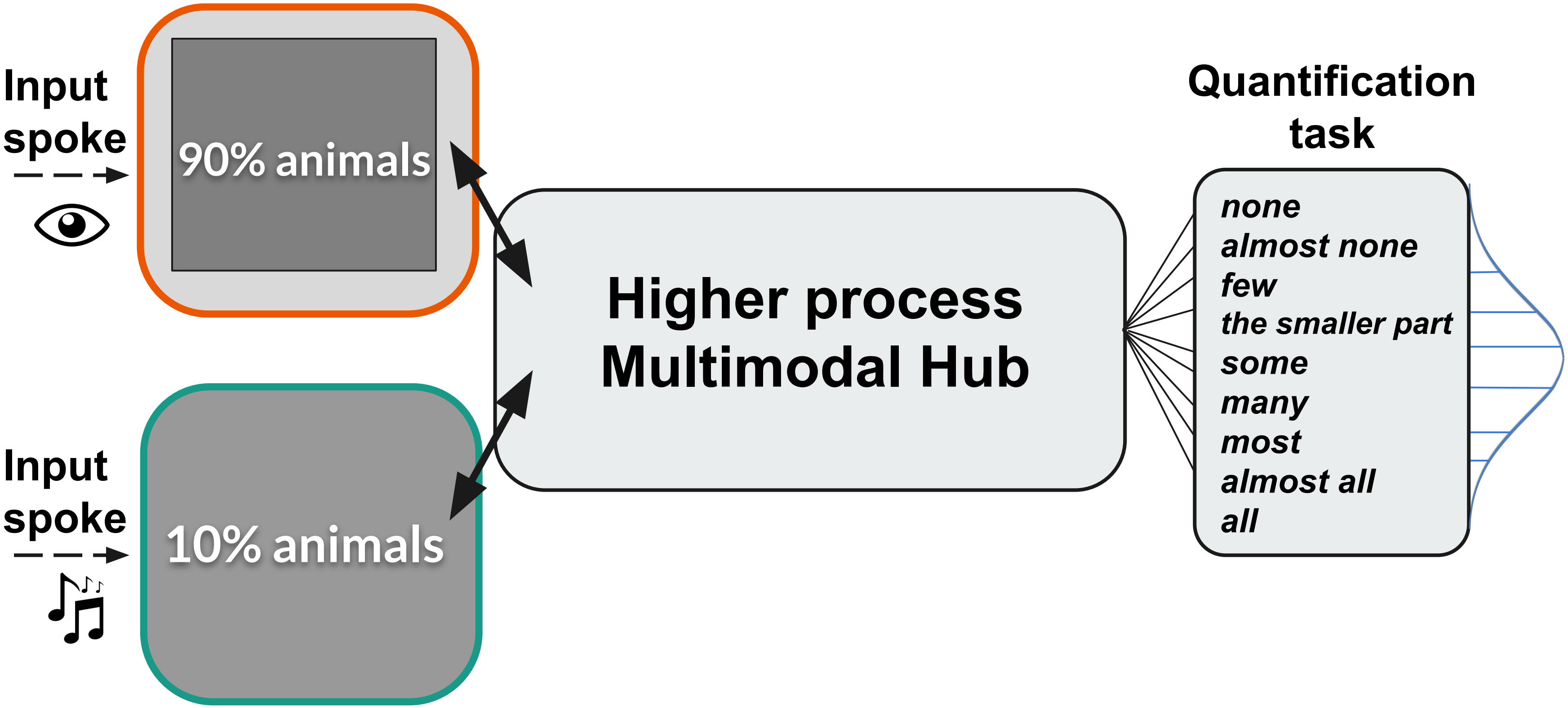
Computational Models

Single-modality models



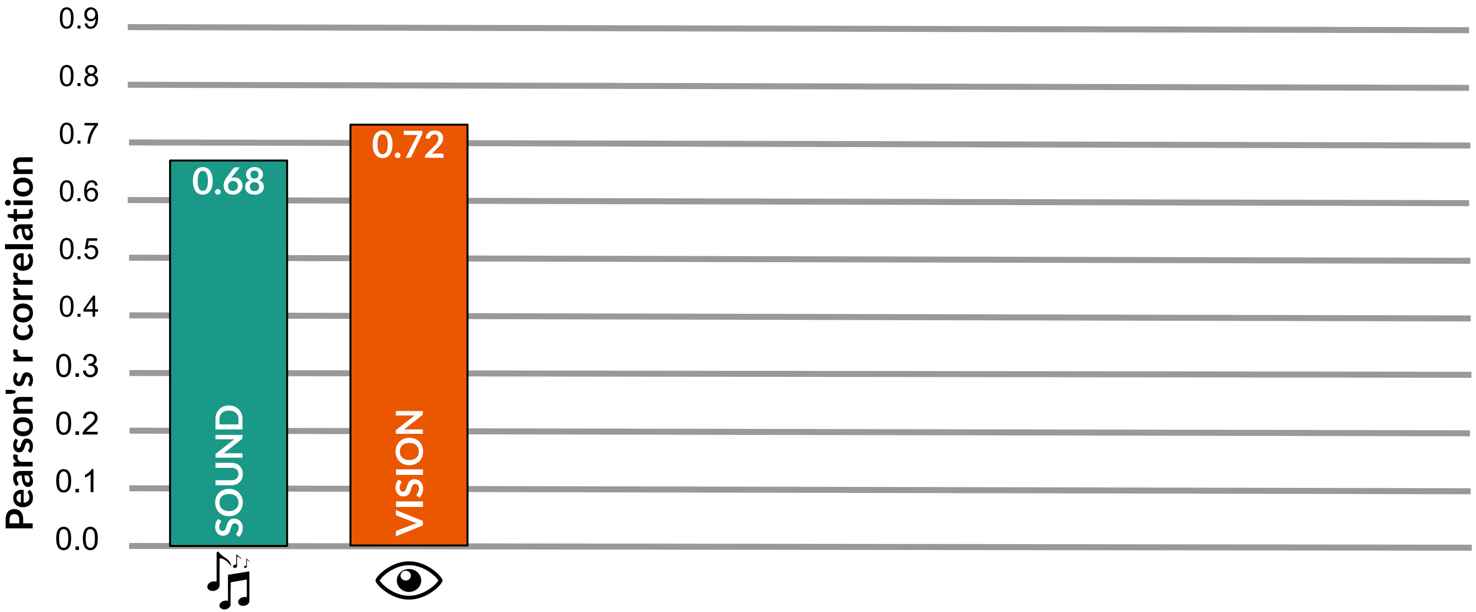
Additional Experiment

→ H&S tested with **incongruent** pairs of visual-auditory inputs.



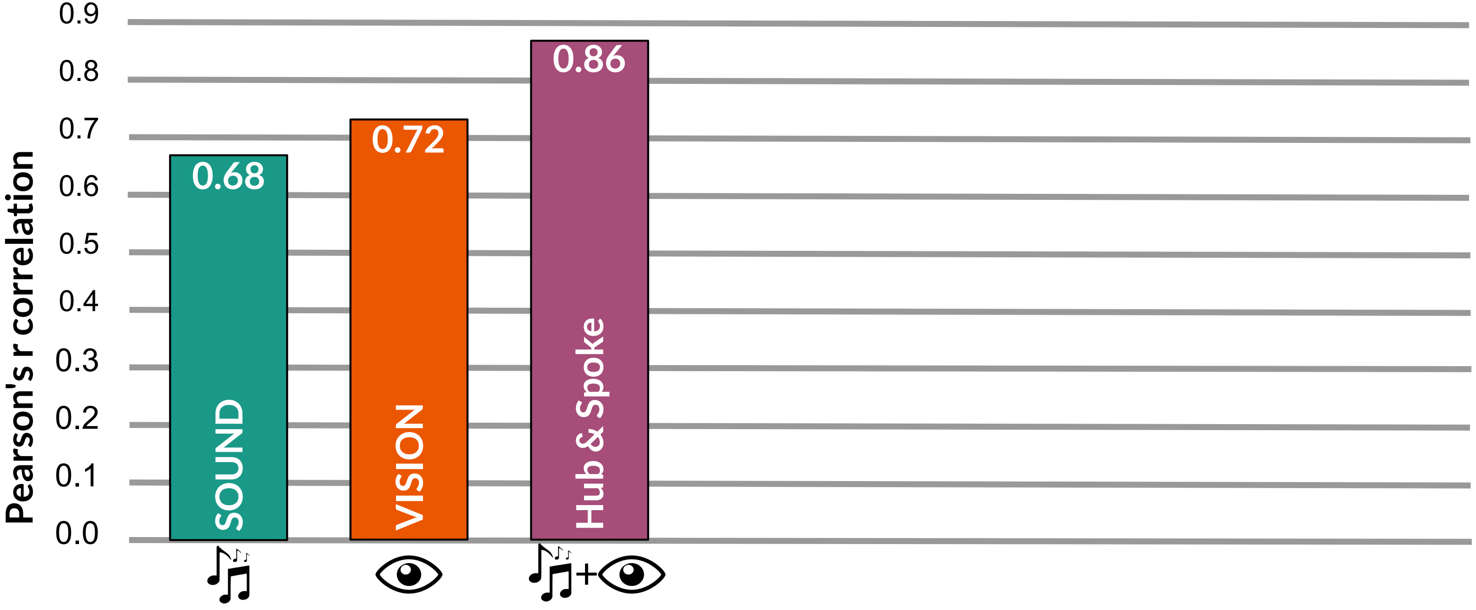
Results

Pearson's Correlation with human annotations



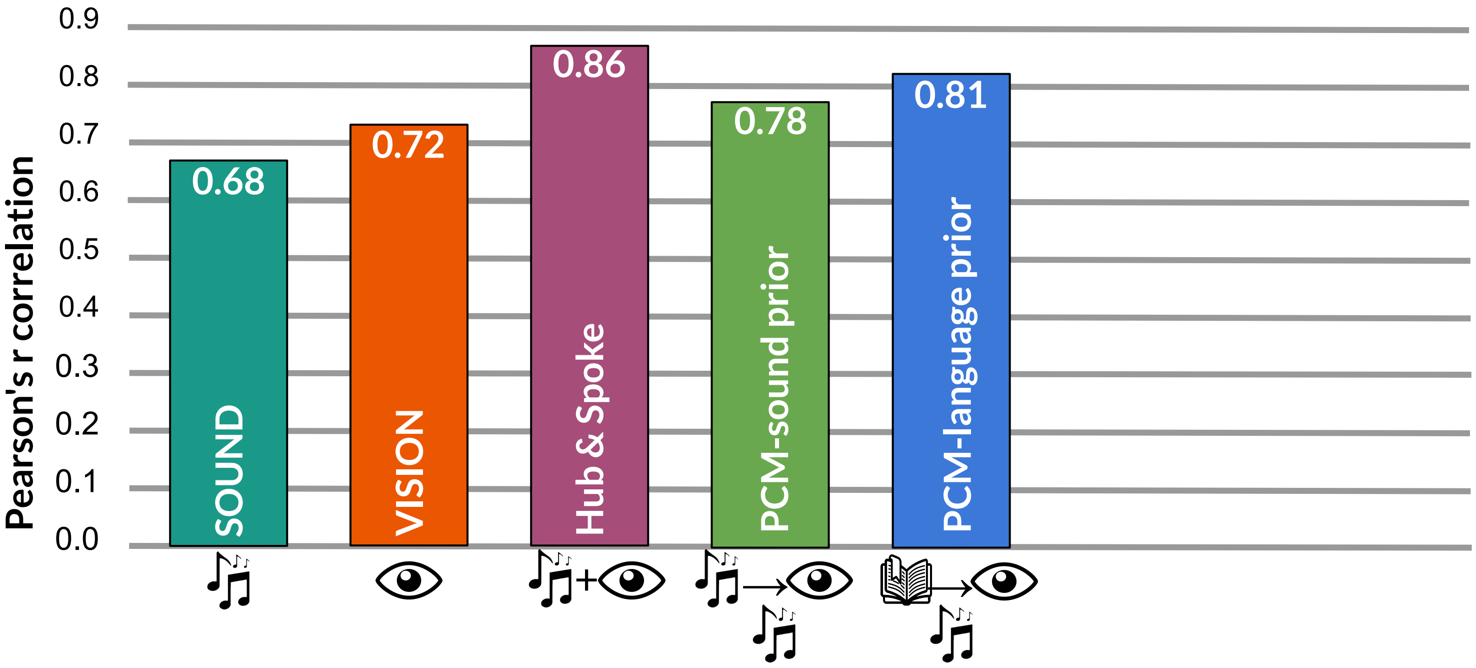
Results

Pearson's Correlation with human annotations



Results

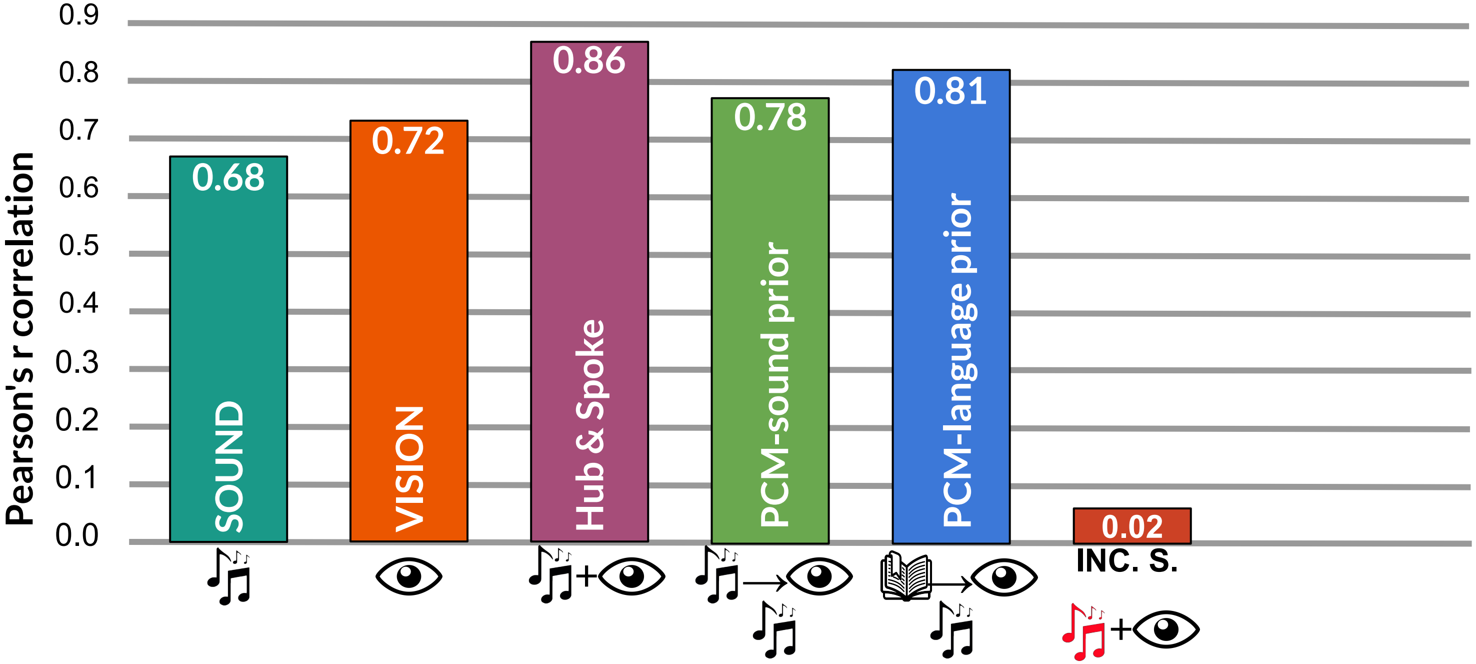
Pearson's Correlation with human annotations



Results

INC.S = incongruent sound

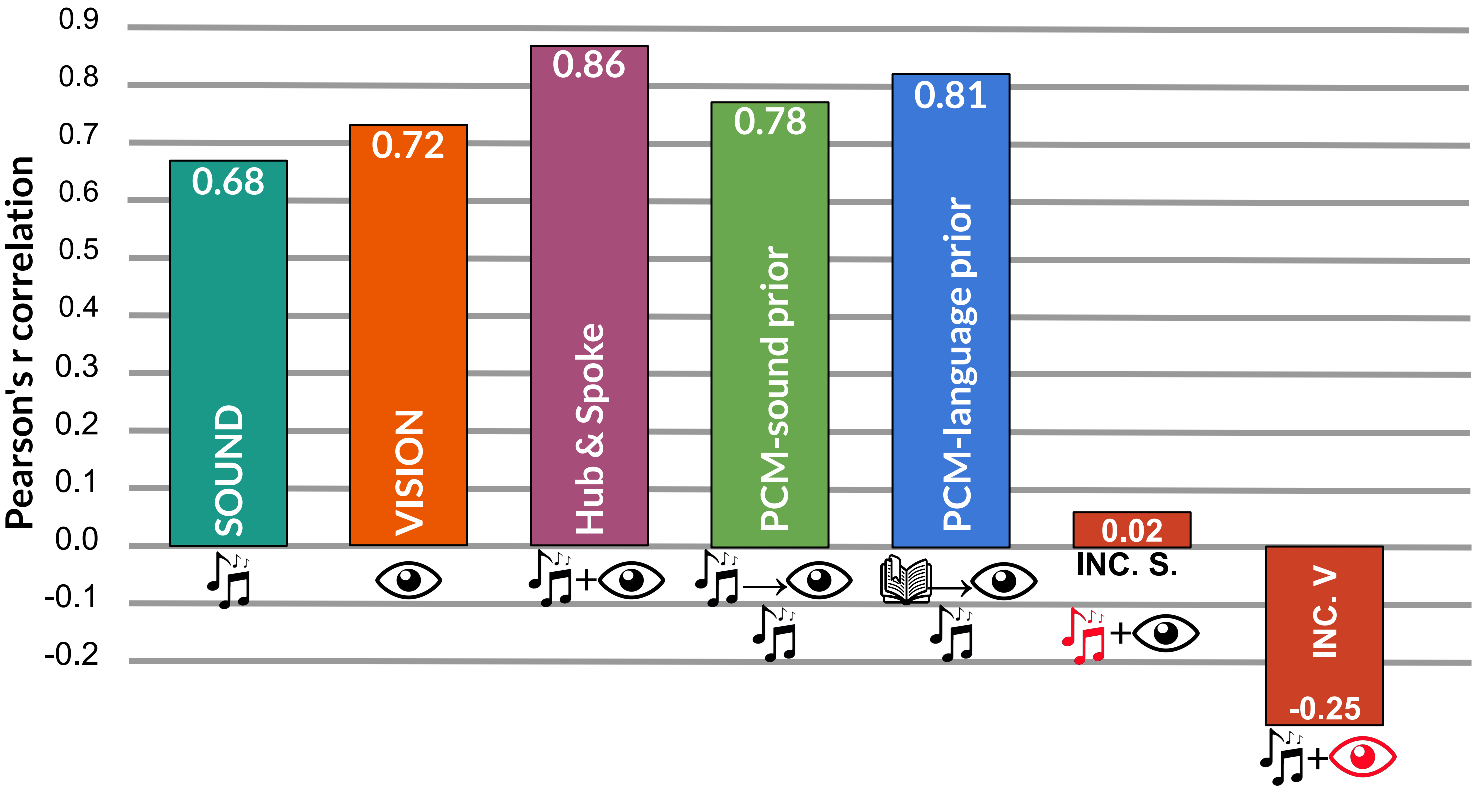
Pearson's Correlation with human annotations



Results

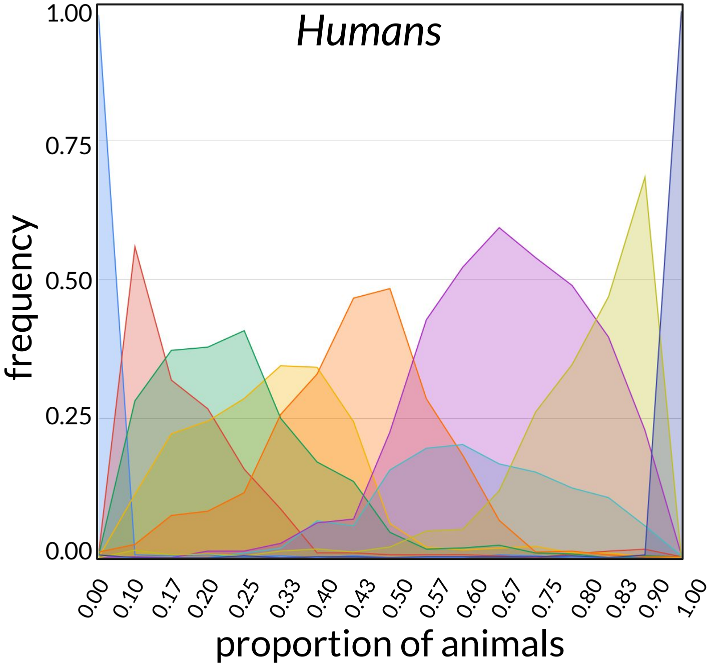
INC.S = incongruent sound
INC. V = incongruent vision

Pearson's Correlation with human annotations



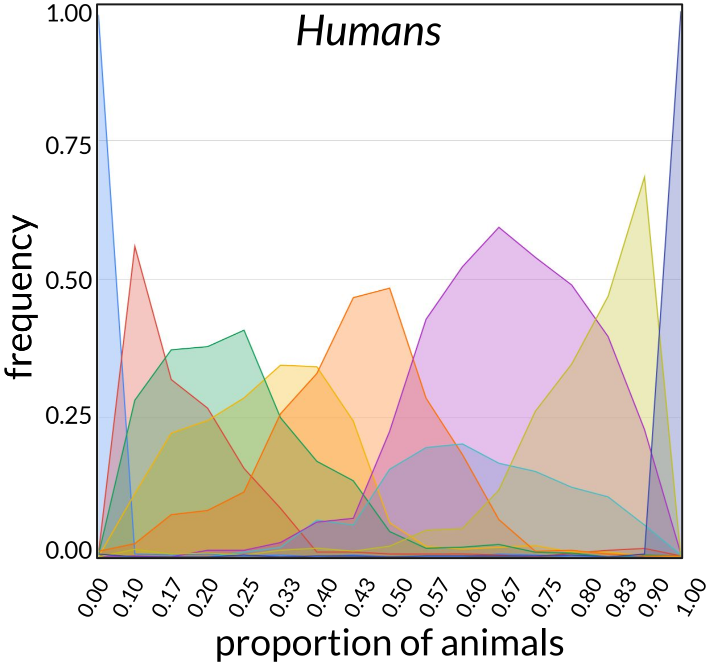
Results

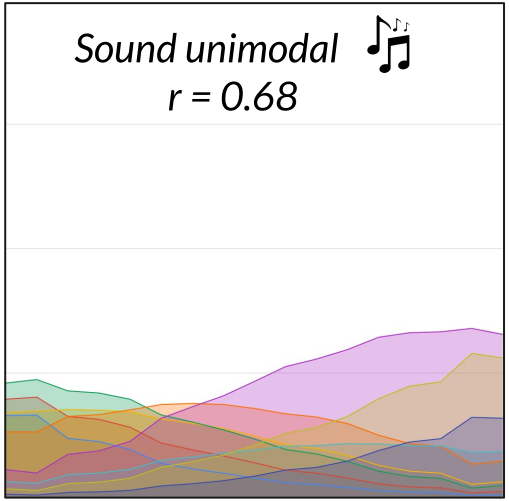
none, *almost none*, *few*, *the smaller part*, *some*, *many*, *most*, *almost all*, *all*



Results

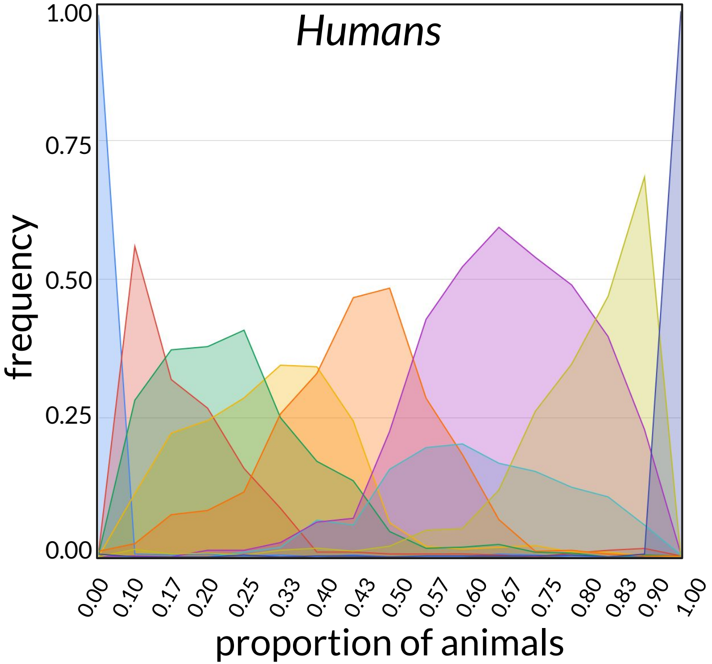
none, *almost none*, *few*, *the smaller part*, *some*, *many*, *most*, *almost all*, *all*

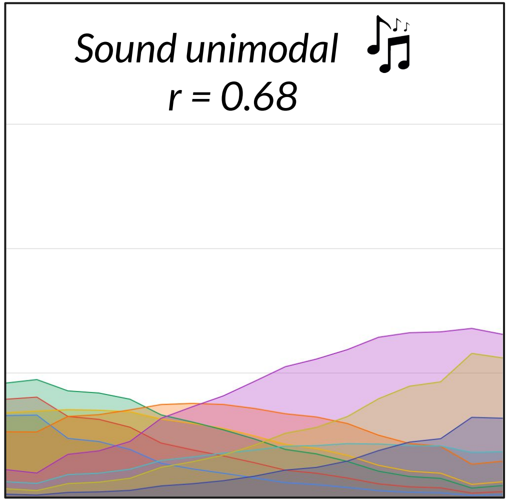


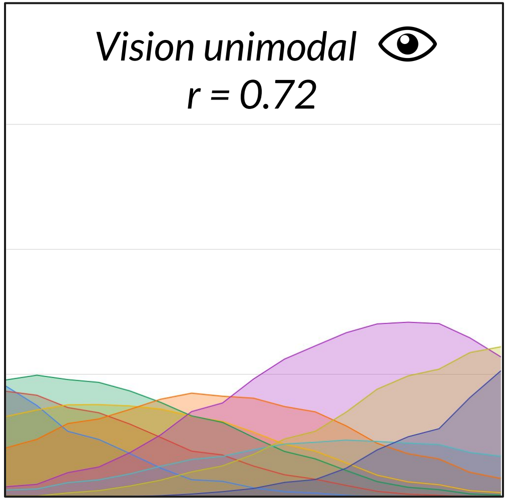


Results

none, almost none, few, the smaller part, some, many, most, almost all, all

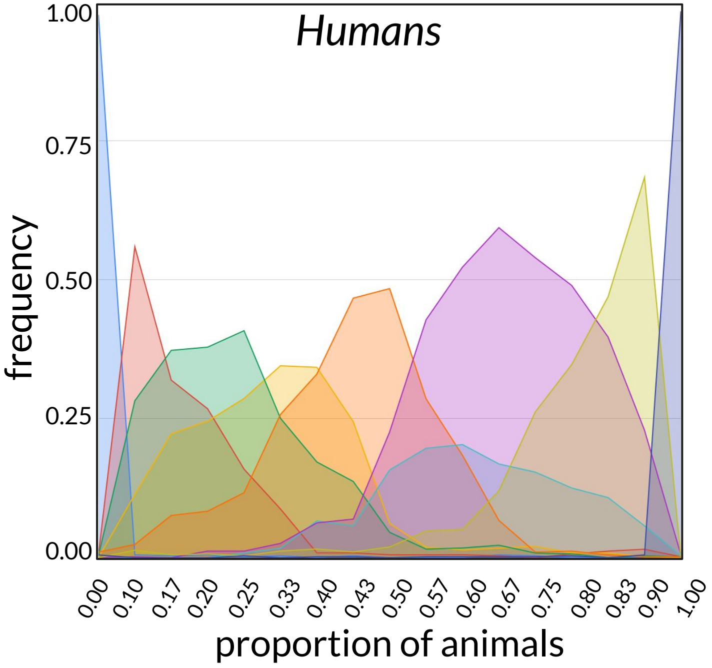


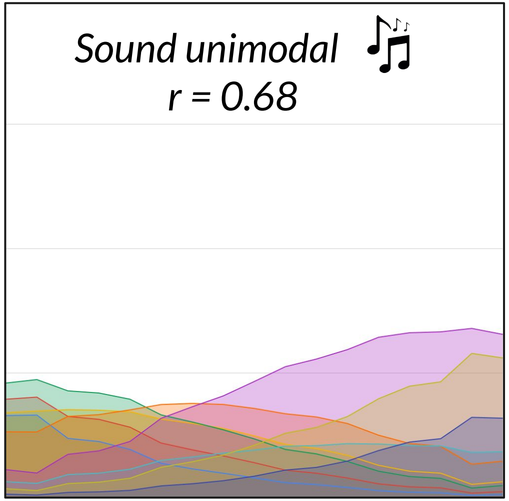


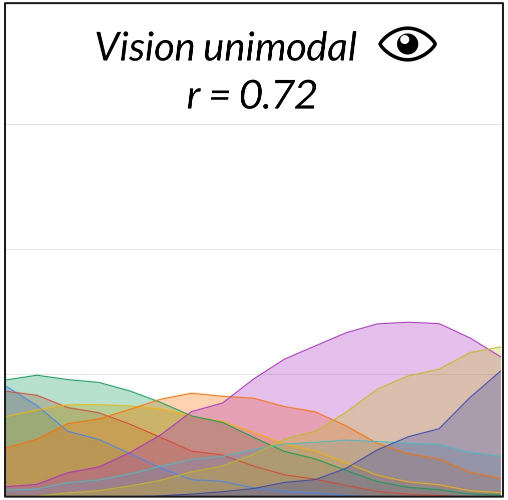


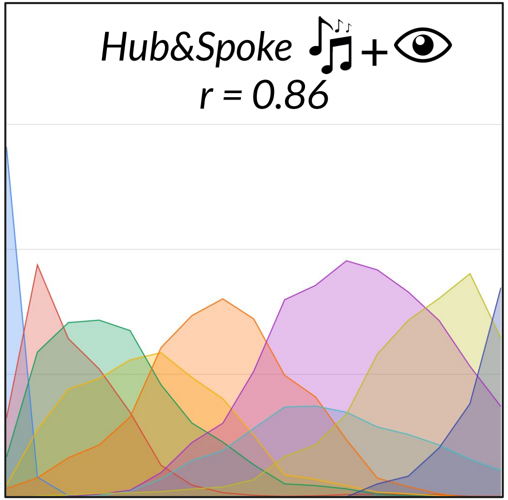
Results

none, almost none, few, the smaller part, some, many, most, almost all, all



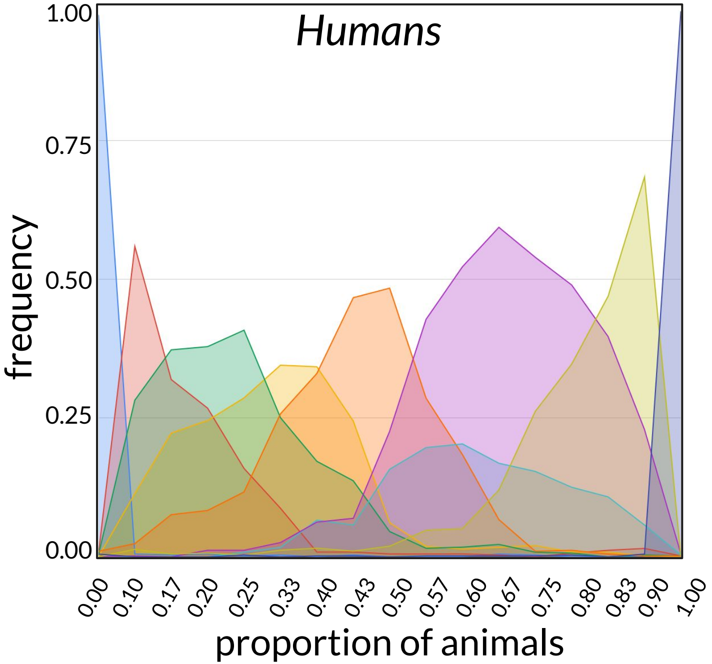


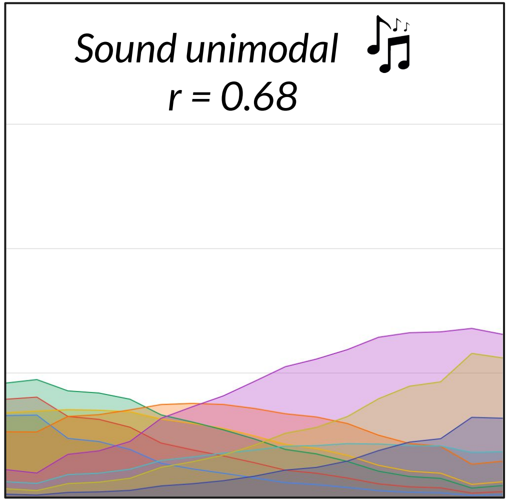


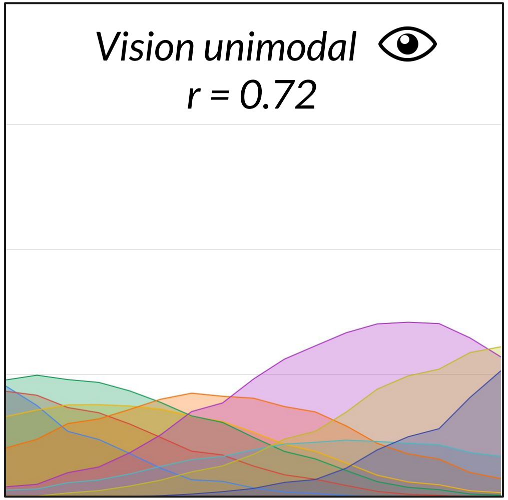


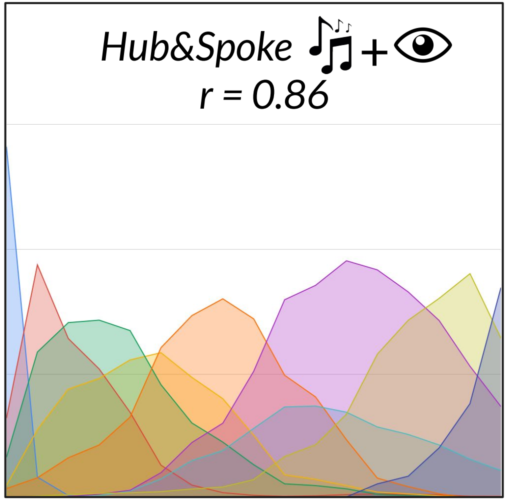
Results

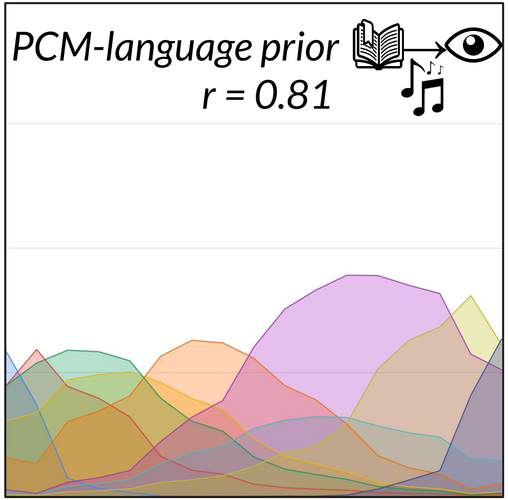
none, *almost none*, *few*, *the smaller part*, *some*, *many*, *most*, *almost all*, *all*





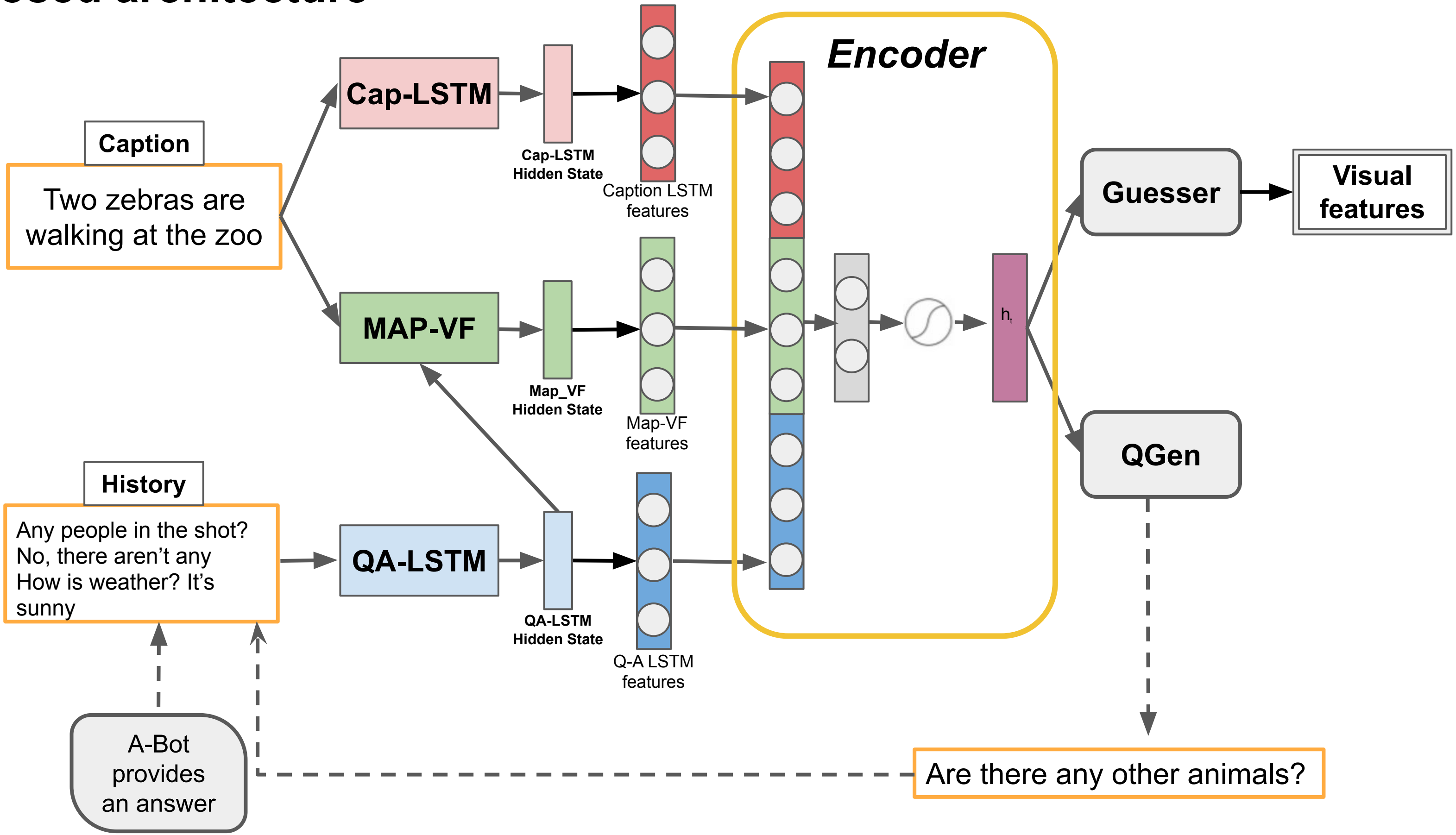






Visual Hallucination for Neural Dialogue Modeling

Proposed architecture



Visual Hallucination for Neural Dialogue Modeling

Preliminary results

	Mean Percentile Rank	Lexical Diversity	Questions Diversity	% of Games with Repeated Questions
Chance	50.00			
QBot-SL (Das et al., 2017)	91.19	0.11	1.66	100
QBot-RL (Das et al., 2017)	94.19	0.05	0.35	100
QA+Cap	95.65	0.452	31.25	41.66
QA+Cap+Map-VF (static hallucination)	95.72	**0.502**	**47.17**	**35.26**
QA+Cap+Map-VF (dynamic hallucination)	**95.98**	0.325	16.57	85.02

Visual Hallucination for Neural Dialogue Modeling

Imagining Grounded Conceptual Representations from Perceptual Information in Situated Guessing Games

Alessandro Suglia[1], **Antonio Vergari**[2], **Ioannis Konstas**[1], **Yonatan Bisk**[3], **Emanuele Bastianelli**[1], **Andrea Vanzo**[1], and **Oliver Lemon**[1]
[1]Heriot-Watt University, Edinburgh, UK
[2]University of California, Los Angeles, USA
[3]Carnegie Mellon University, Pittsburgh, USA
[1]{as247,i.konstas,a.vanzo,e.bastianelli,o.lemon}@hw.ac.uk
[2]aver@cs.ucla.edu, [3]ybisk@cs.cmu.edu

To appear in Proceedings of *COLING 2020*

Conclusions

➔ 1st task: quantifying over multimodal stimuli with a hallucinated visual representation.

- ✓ **+0.10** (sound → vision) and **+0.13** (language → vision) *VS* single modalities
- ✓ Remarkable linguistic competence achieved with multimodal data.

➔ Robust ability of Artificial Neural Networks to model cross-sensory associations, in line with the Predictive Coding model.

THANK YOU!

Conclusions

➔ 1st task: quantifying over multimodal stimuli with a hallucinated visual representation.

- ✓ **+0.10** (sound → vision) and **+0.13** (language → vision) *VS* single modalities
- ✓ Remarkable linguistic competence achieved with multimodal data.

Conclusions

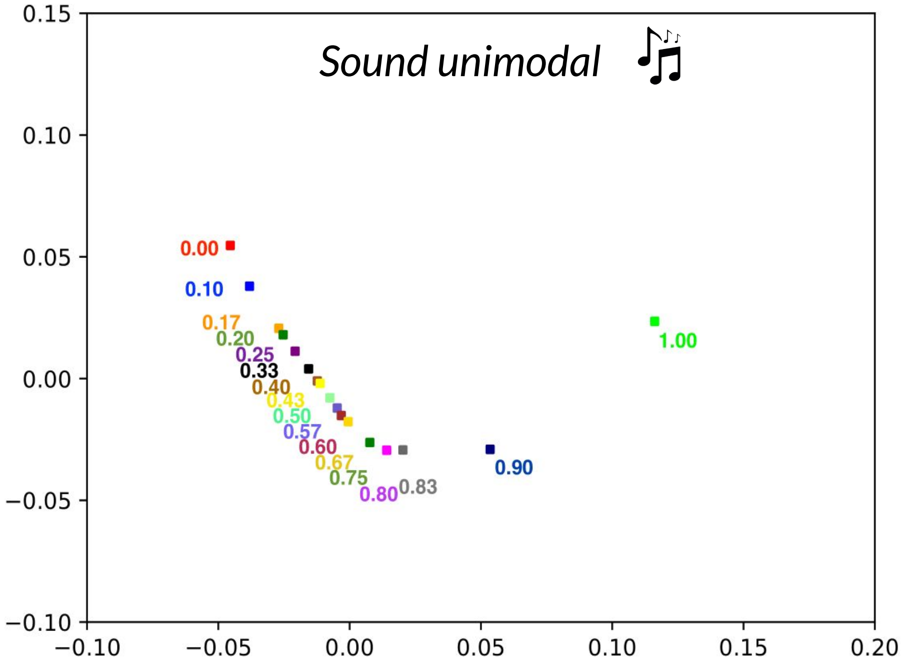
- ➔ 1st task: quantifying over multimodal stimuli with a hallucinated visual representation.
    - ✓ **+0.10** (sound $\rightarrow$ vision) and **+0.13** (language $\rightarrow$ vision) *VS* single modalities
    - ✓ Remarkable linguistic competence achieved with multimodal data.
- ➔ 2nd task: hallucinating a visual representation from a dialogue between two agents.
    - ✓ **+25** and **+30** Mean Rank when hallucinating from caption *VS* full dialogue.
    - ✓ Richer vocabulary and less repetitions using the hallucination module.

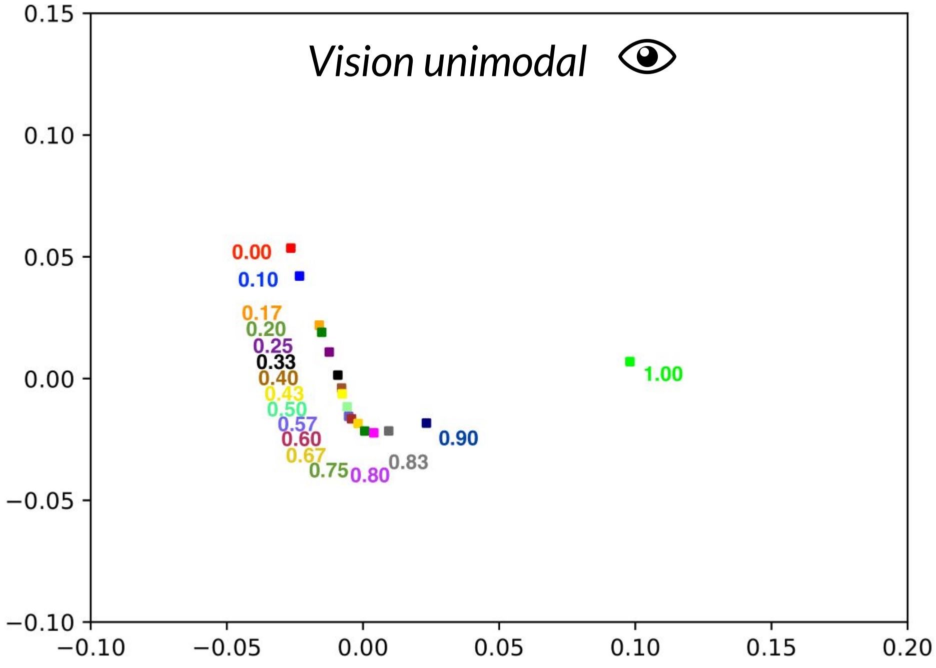
Implementation and Evaluation Details

- *ReLU* activations on the hidden layers and *Softmax* activation on the output layer.
- Adam optimizer (*Kingma and Ba, 2015*) with Learning Rate = 0.0001.
- Training for no more than 150 epochs (early stopping).
- Kullback- Leibler (KL) divergence loss between the activations of the output layer and human responses from *Pezzelle et al., 2018*.
- PyTorch v0.4
- The models are evaluated by computing the Pearson's correlation coefficient (ranging from -1 to +1) between the output of the models and human annotations.

Results

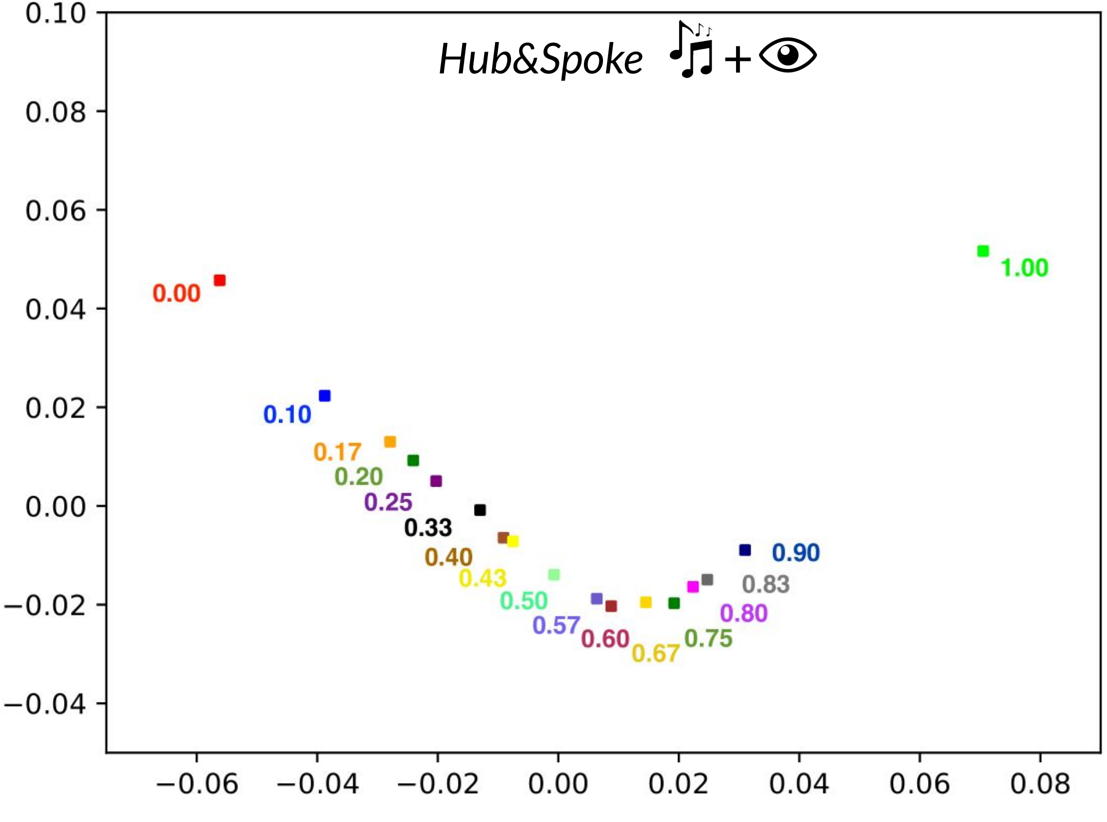
Qualitative Results - PCA on the activations of the last hidden layer





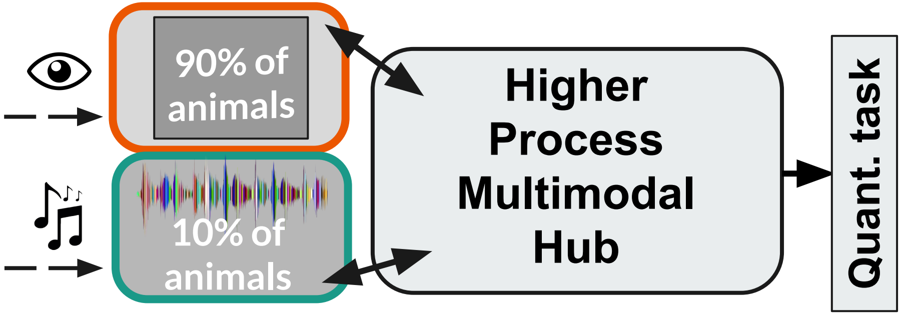
Results

Qualitative Results - PCA on the activations of the last hidden layer

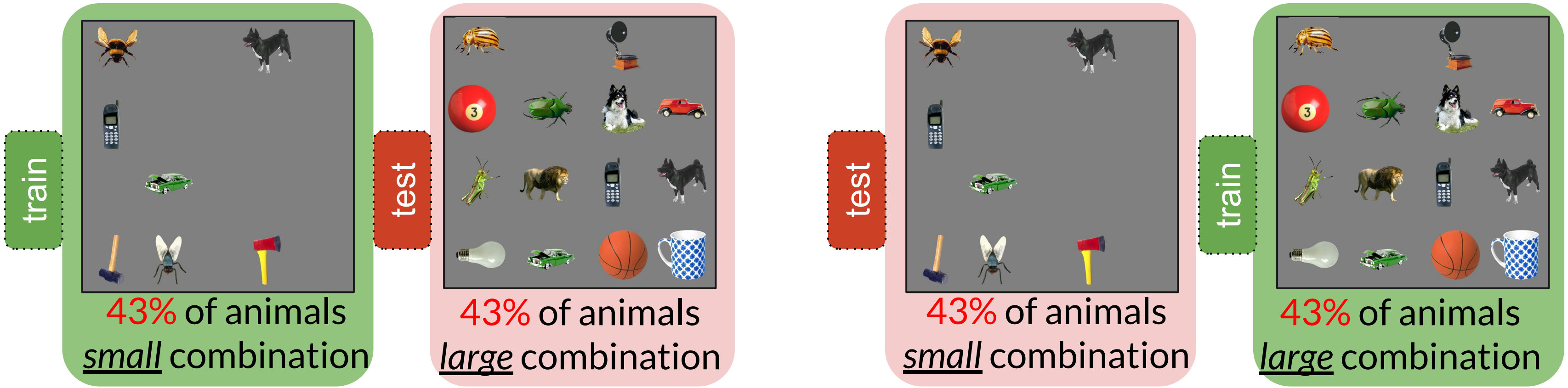


Additional Experiments

➔ H&S tested on incongruent pairs of visual-auditory inputs.

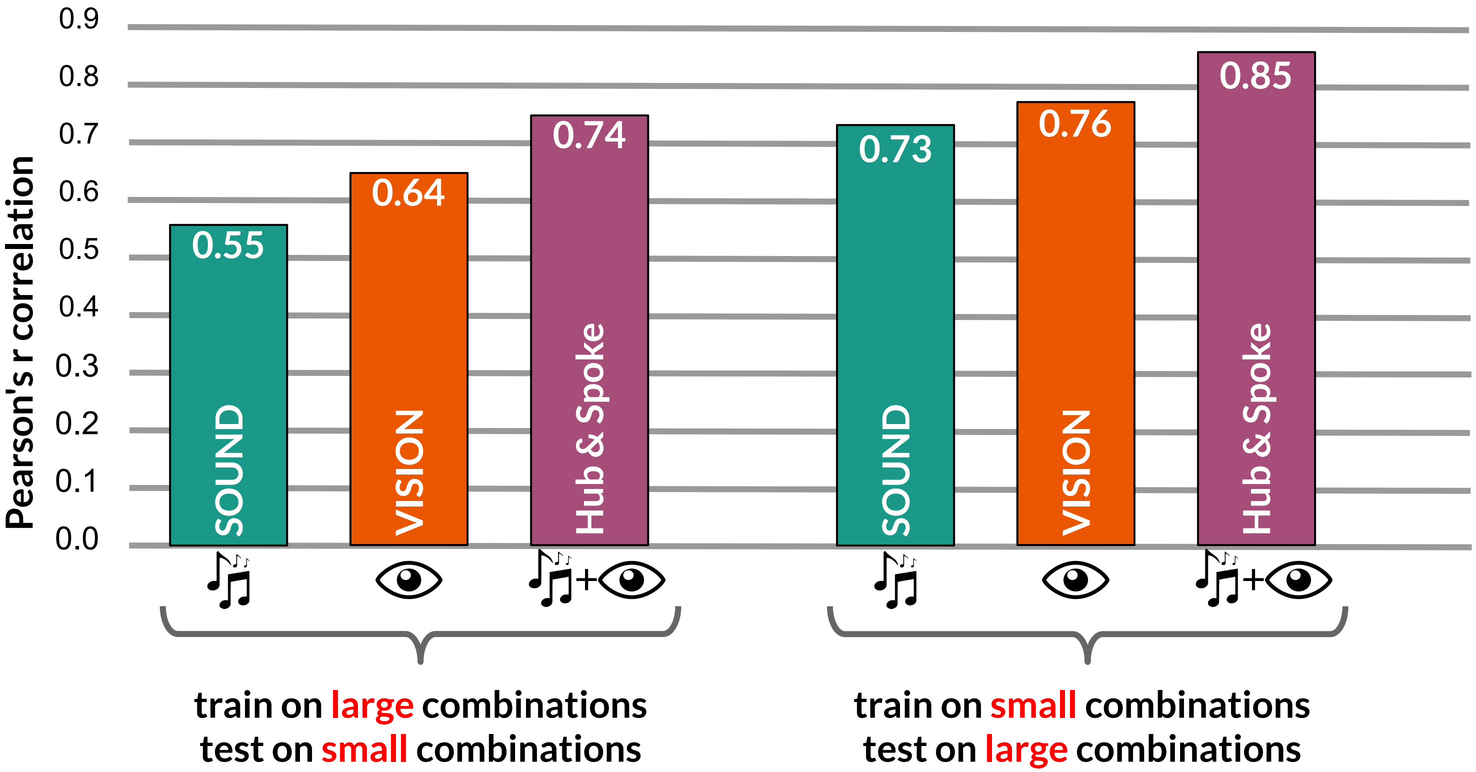


➔ Generalization on unseen combinations with small/large number of total entities.



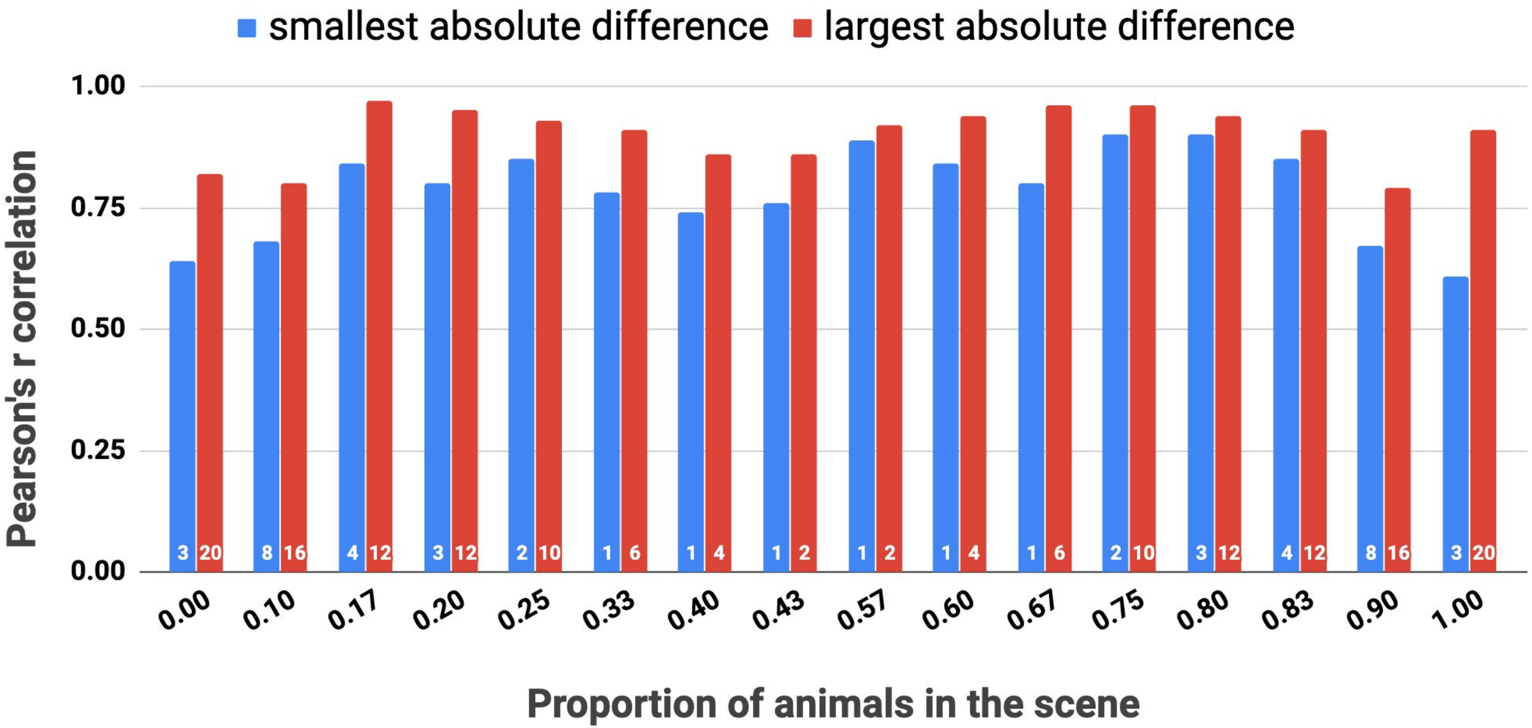
Results

Quantitative Results - Generalization on Unseen Combinations



Results

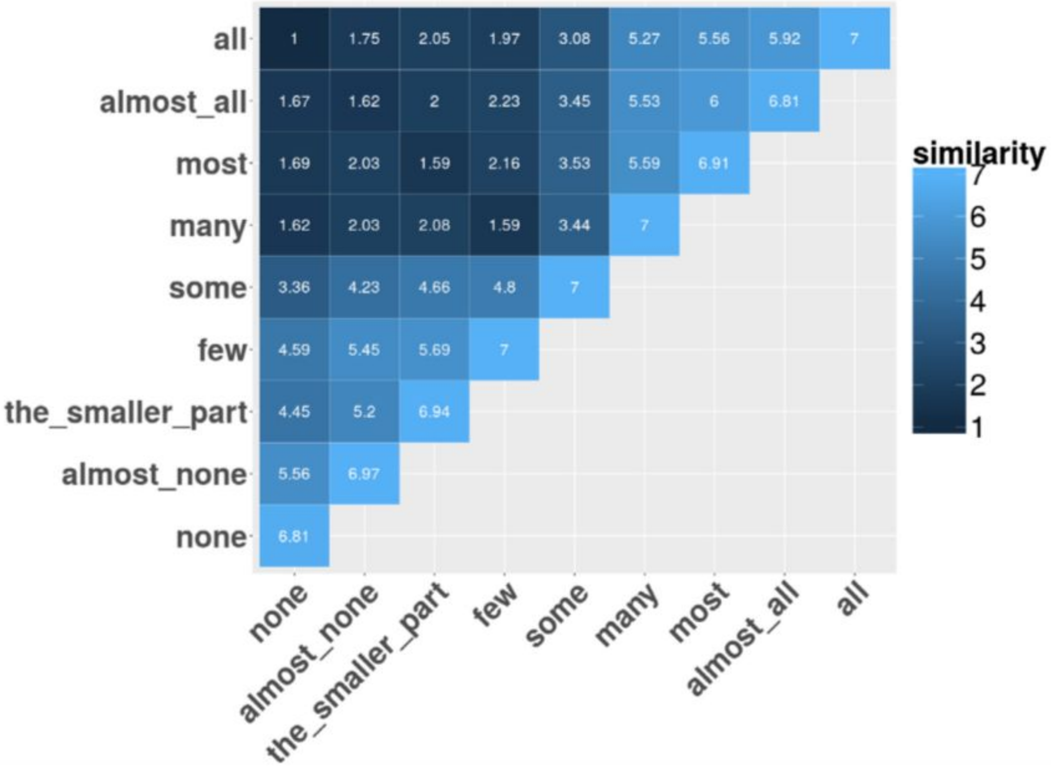
Quantitative Results - Effects of the absolute difference of animals/artifacts



References

- Testoni, Alberto, Sandro Pezzelle, and Raffaella Bernardi. "*Quantifiers in a Multimodal World: Hallucinating Vision with Language and Sound.*" Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. 2019.
- Ralph, Matthew A. Lambon, et al. "*The neural and computational bases of semantic cognition.*" Nature Reviews Neuroscience 18.1 (2017): 42.
- Emberson, Lauren L., John E. Richards, and Richard N. Aslin. "*Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months.*" Proceedings of the National Academy of Sciences 112.31 (2015): 9585-9590.
- Pezzelle, Sandro, Raffaella Bernardi*, and Manuela Piazza.* "Probing the mental representation of quantifiers." Cognition 181 (2018): 117-126.
- Friston, Karl. "*The free-energy principle: a rough guide to the brain?.*" Trends in cognitive sciences 13.7 (2009): 293-301.

Human similarity judgments



Pezzelle et al., 2018