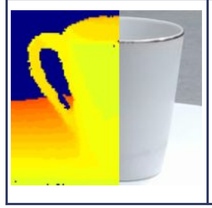
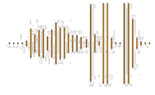
A Spoken Language Dataset of Descriptions for Speech-Based Grounded Language Learning

Gaoussou Youssouf Kebe, Padraig Higgins, Patrick Jenkins, Kasra Darvish, Rishabh Sachdeva, Ryan Barron, John Winder, Donald Engel, Edward Raff, Francis Ferraro, Cynthia Matuszek

{gaoussou1, phiggin1, pjenk1, kasradarvish, rishabs1, jwinder1, donengel, edraff1, ferraro, cmat} @umbc.edu

Grounded Language Dataset (GoLD)

- **Grounded Language Learning →** Learning **natural language** as it relates to **perception of the world**
- **GoLD →** Multimodal Dataset for Grounded Language Learning
- **Color + depth** data of **207 objects** from **47 classes of objects**
- **16500 text and spoken descriptions** from **Mechanical Turk**
- Transcriptions from **Google's Speech to Text API**
- **552 speakers** with speech characteristic annotations

RGB + Depth	Typed	Spoken	Transcribed
	 It's a coffee mug.	 "There is a white coffee mug"	 Arizona white coffee mug

Use these **QR codes** or the **URL** below to check our **paper** and **dataset**.

Paper **github.com/iral-lab/gold** Data

16500 Typed Descriptions

this is a large red spiraled notebook.

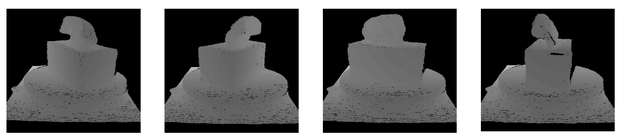
This is a spiral notebook with a red cover and a gold design on the front. Some of the pages have been bent.

16500 Spoken Descriptions

"this is a spiral notebook with a metal spiral on the side a red cover with black writing on it and three punch holes on the side of the red cover"

RGB

Depth



552 Speakers in GoLD

Accented	Yes	50%
Creaky	Yes	35%
Hoarse	Yes	8%
Muffled	1	71%
	2	22%
	3	7%

Volume	1	2%
	2	28%
	3	60%
	4	10%
Backg. Noise	1	66%
	2	26%
	3	7%
	4	1%

Approach: Manifold Alignment

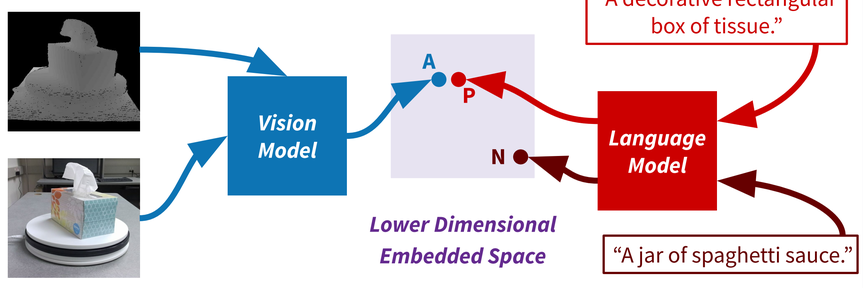
Triplet loss → $max(d(A, P) - d(A, N) + a, 0)$

Vision Model:

- pre-trained CNN + MLP
- **RGB + Depth** after CNN

Language Model:

- **Text:** BERT + MLP
- **Speech:** wav2vec 2.0 + MLP



Results

- Grounded language learning from speech is on par with typed text.
- Speech combined with vision+depth learns language effectively.

	F1	Triplet MRR	Subset MRR
Typed Text	0.84	0.85	0.88
Transcription	**0.93**	**0.86**	**0.95**
Speech	0.83	0.85	0.89