

## Vision and Language



Eric Wallace

with thanks to Rudy Corona & Daniel Fried  
CS 288, 4/12/2022



## What is Language Grounding?

- Language often refers to *the world*



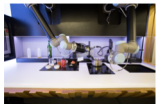
## What is Language Grounding?

- Language often refers to *the world*
- Grounding is tying language to non-linguistic things (e.g., databases, vision, sound)



## What is Language Grounding?

- Language often refers to *the world*
- Grounding is tying language to non-linguistic things (e.g., databases, vision, sound)
- Today we will talk about grounding into *visual* environments:



"Add the tomatoes and mix"



"Take me to the shop on the corner"




### Grounding

- (Some) possible things to map language to:



### Grounding

- (Some) possible things to map language to:
  - **Low-level percepts:** *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...






### Grounding

- (Some) possible things to map language to:
  - **Low-level percepts:** *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...
  - **High-level percepts:** *cat* means this type of pattern





### Grounding

- (Some) possible things to map language to:
  - **Low-level percepts:** *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...
  - **High-level percepts:** *cat* means this type of pattern
  - **Embodiment (effects on the world):** *go left* means the robot turns left, *speed up* means increasing actuation

### Grounding

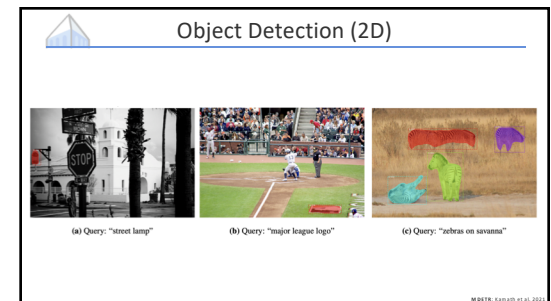
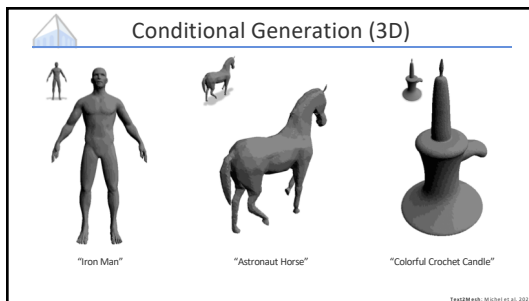
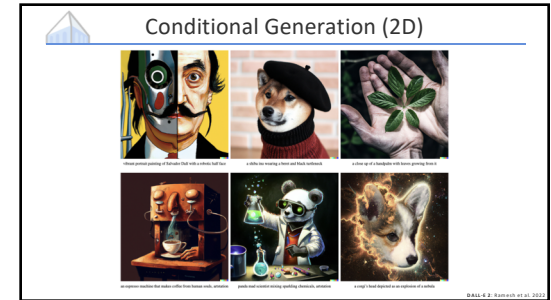
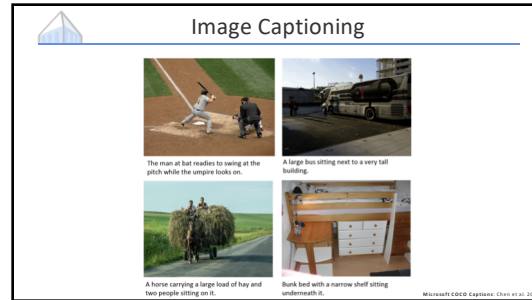
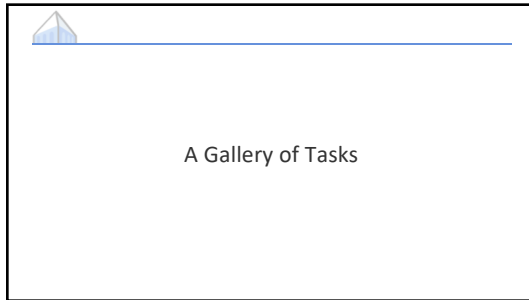
- (Some) possible things to map language to:
  - **Low-level percepts:** *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...
  - **High-level percepts:** *cat* means this type of pattern
  - **Embodiment (effects on the world):** *go left* means the robot turns left, *speed up* means increasing actuation
  - **Social (effects on others):** polite language is correlated with longer forum discussions

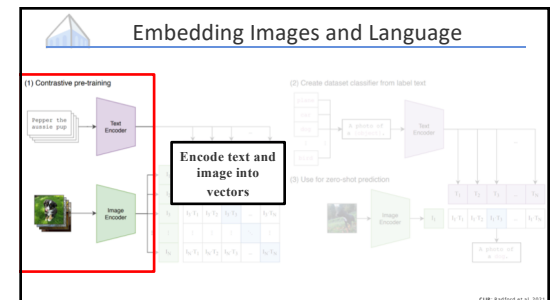
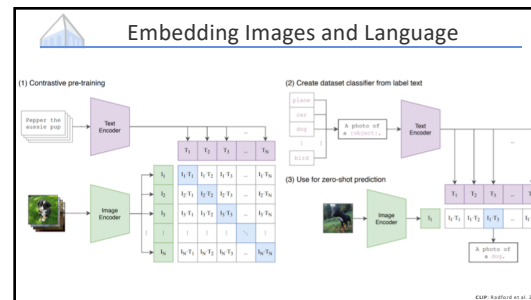
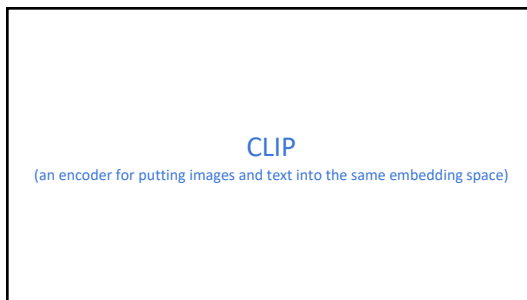
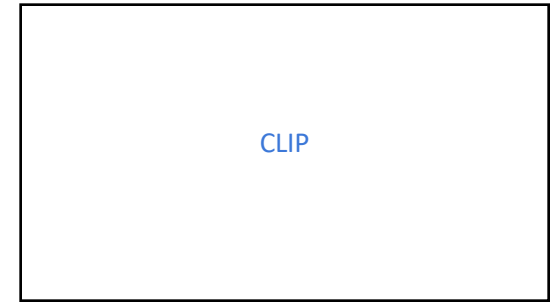
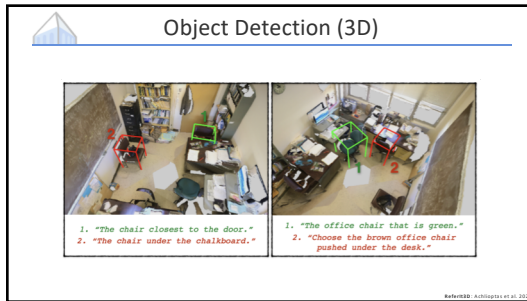





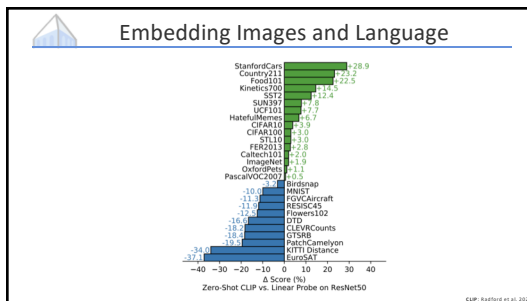
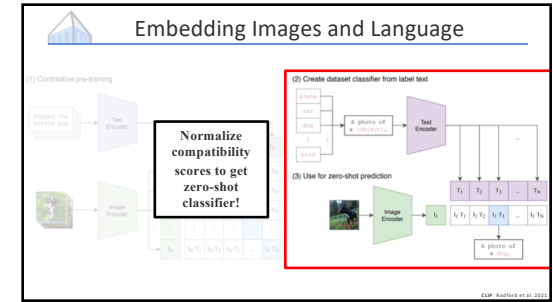
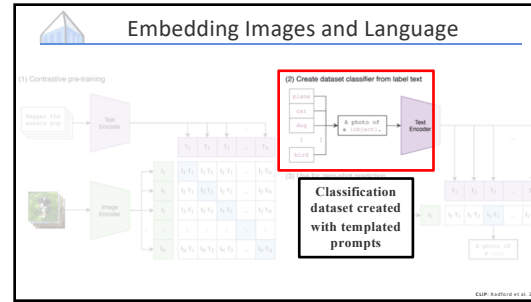
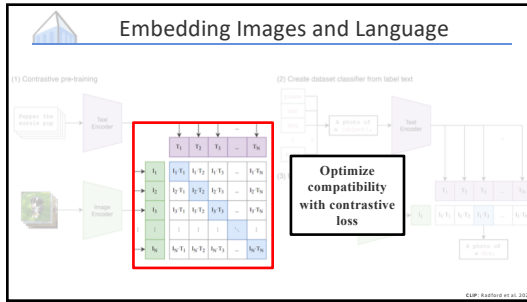
### Grounding

- (Some) possible things to map language to:
  - **Low-level percepts:** *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor...
  - **High-level percepts:** *cat* means this type of pattern
  - **Embodiment (effects on the world):** *go left* means the robot turns left, *speed up* means increasing actuation
  - **Social (effects on others):** polite language is correlated with longer forum discussions

For a nice taxonomy, related work, and examples, see *Experience Grounds Language* [Bisk et al. 2020]







### Joint Vision-Language Models

