CS6501/4501: Vision and Language

Recent Work (2020 and Beyond)

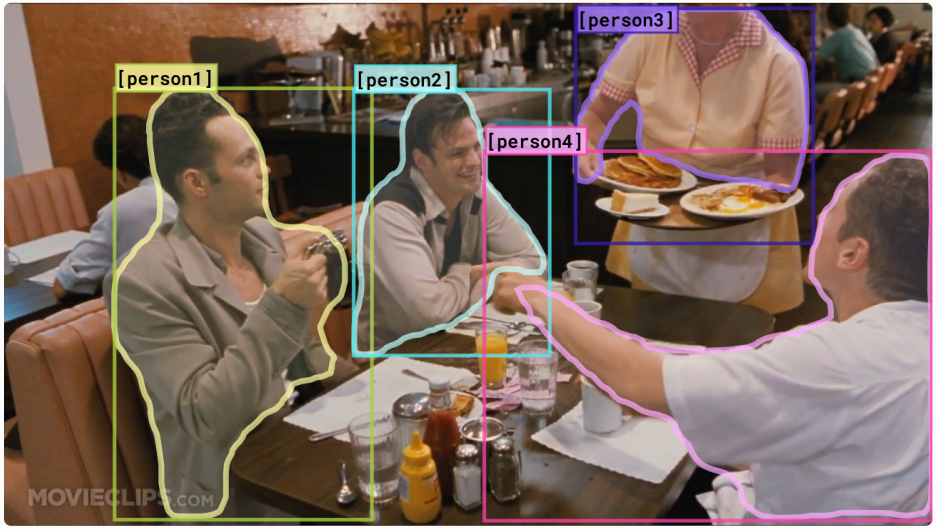
Last Class

- Practical Issues with Deployment
  - Web/Mobile Interface
    - Web Browser / Web Server / HTTP (POST and GET requests)
    - HTML/CSS/Bootstrap
    - Javascript / Asynchronous Calls -- AJAX / DOM
    - Flask / Jinja2 / Python for Web Development
    - Integration with Pytorch (and potentially other python-based toolkits)

Today

- A quick tour on some recently proposed research and directions

Visual Common Sense Reasoning



hide all | show all | [person1] | [person2] | [person3] | [person4]

more objects »

Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.

b) He just told a joke.

c) He is feeling accusatory towards [person1].

d) He is giving [person1] directions.

Rationale: I think so because...

a) [person1] has the pancakes in front of him.

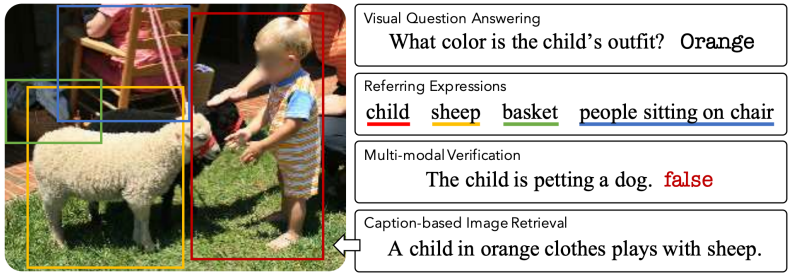
b) [person4] is taking everyone's order and asked for clarification.

c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.

d) [person3] is delivering food to the table, and she might not know whose order is whose.

https://visualcommonsense.com/

Multi-task Learning / More General Models

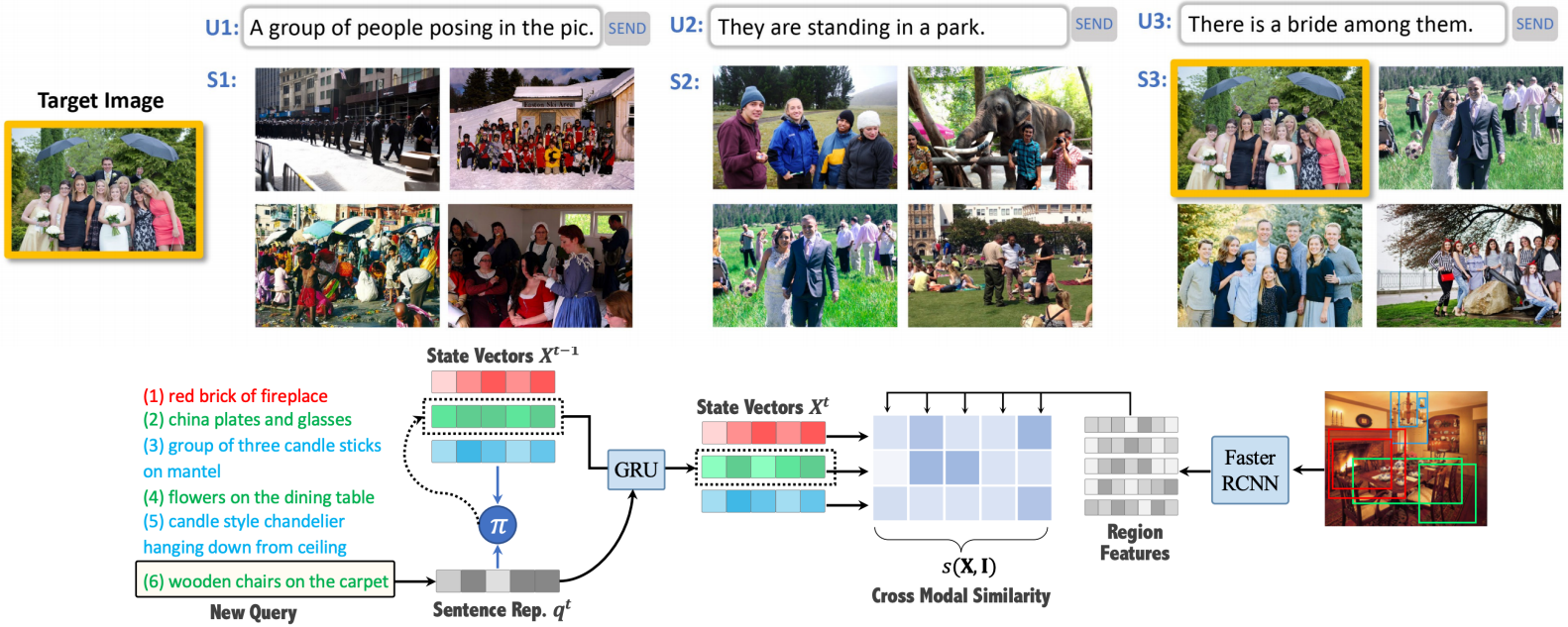


12-in-1: Multi-task Vision and Language
https://arxiv.org/abs/1912.02315

Salesforce DecaNLP
https://arxiv.org/pdf/1806.08730.pdf

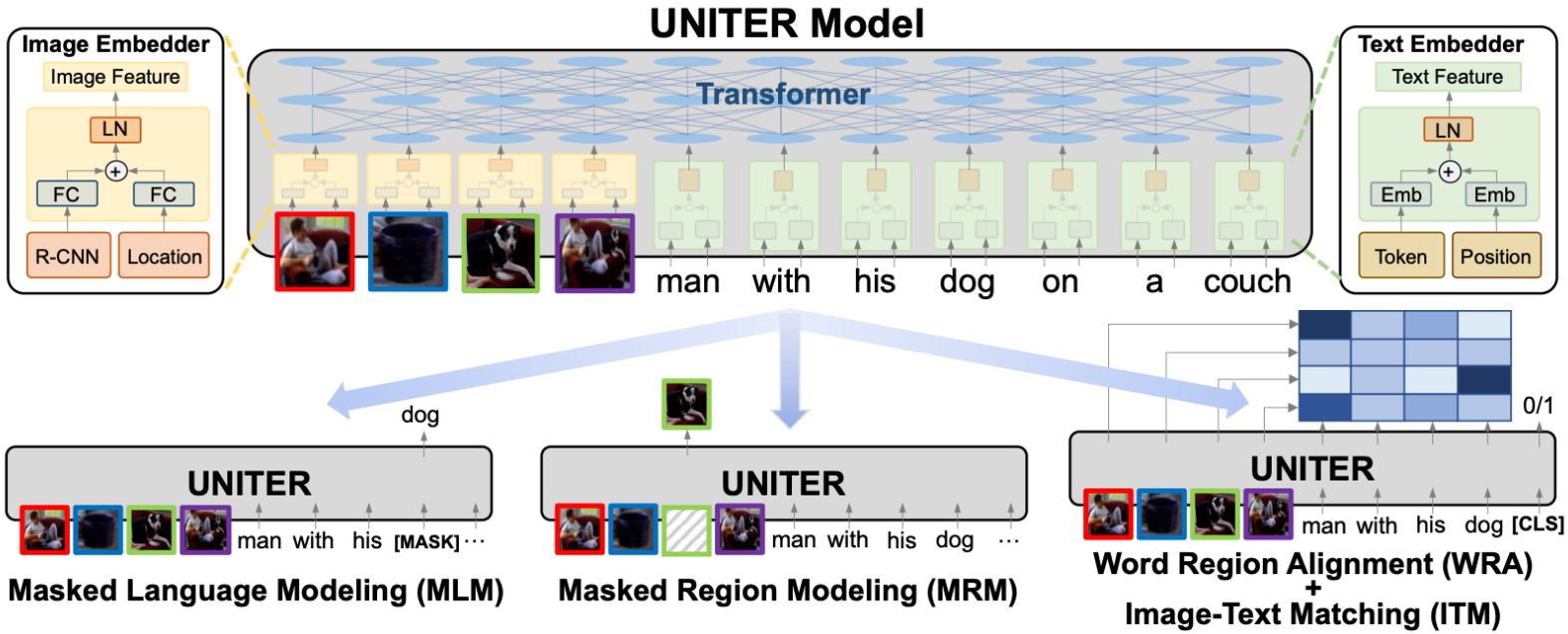
Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Interactivity + Language and Vision

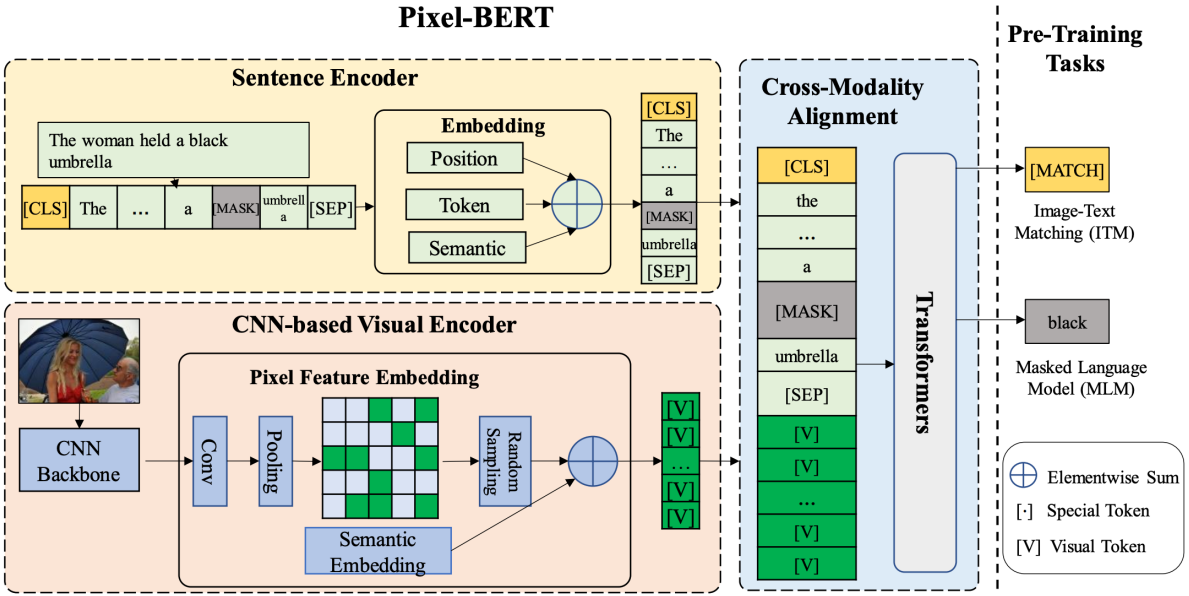


https://arxiv.org/abs/1911.03826

General models for Vision and Language - UNITER

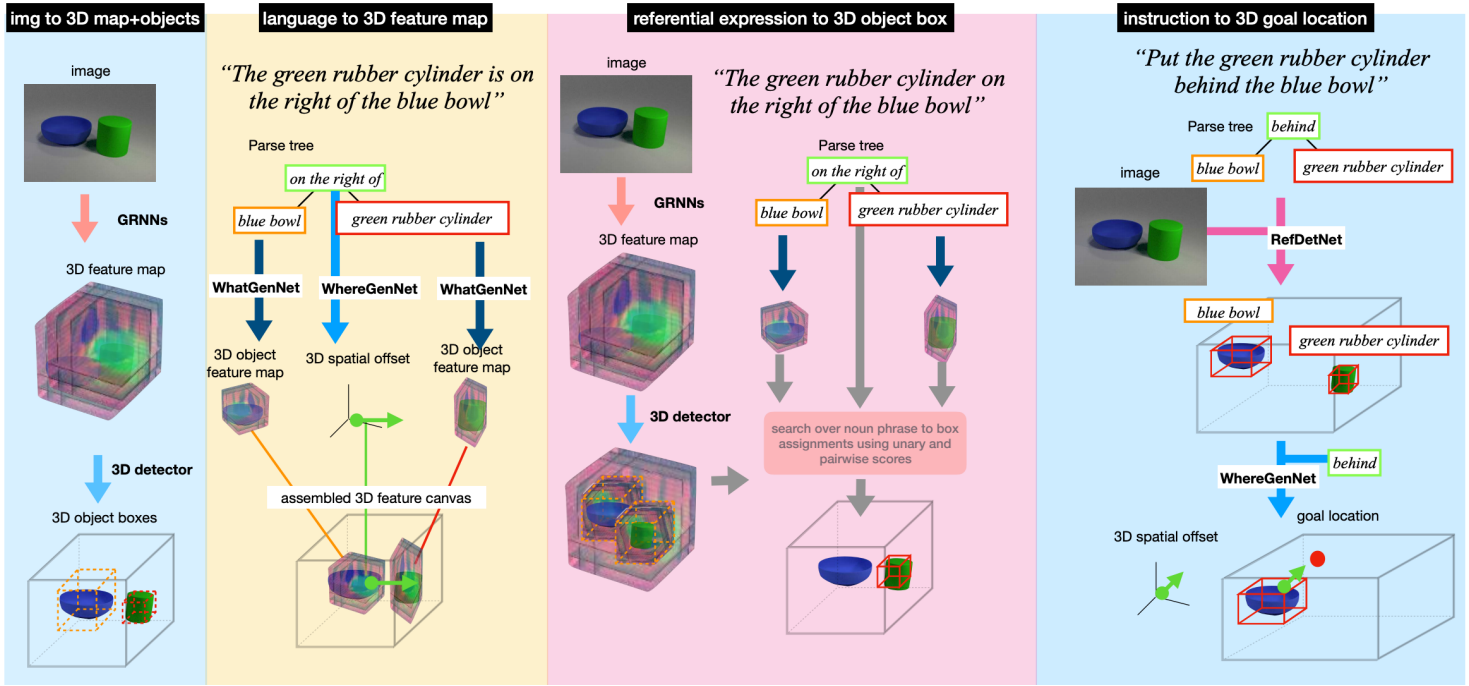


Grid Features – Pixel BERT



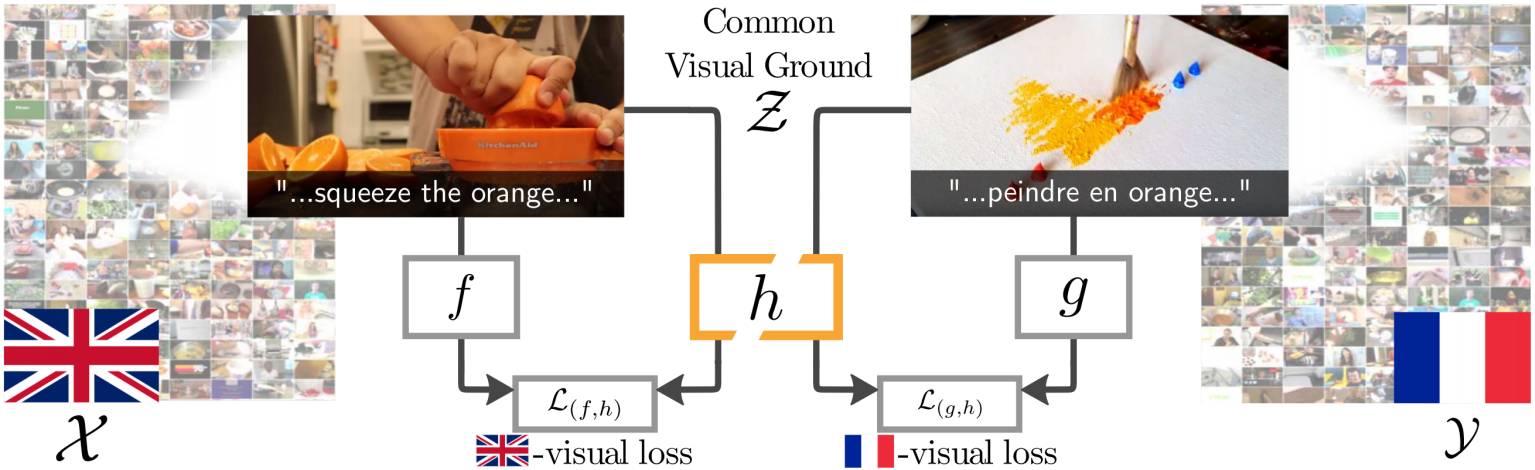
https://arxiv.org/pdf/2004.00849.pdf

Vision + Language + 3D



https://arxiv.org/abs/1910.01210

Multiple Languages and Vision



https://arxiv.org/abs/2003.05078

Video + Language Tasks

00:00:03,576 --> 00:00:05,697
Gavin Mitchell's office.
Rachel Green's office.

00:00:05,870 --> 00:00:07,409
Give me that phone.

00:00:08,873 --> 00:00:12,293
Hello, this is Rachel Green.
How can I help you?

00:00:12,460 --> 00:00:17,629
Uh-huh. Okay, then.
I'll pass you back to your son.

00:00:18,800 --> 00:00:21,639
Hey, Mom. No, that's just my secretary.

(positive) The woman becomes upset when the man answers the phone because he pretends it is his own office.
(negative) The woman becomes upset when the man answers the phone because she is expecting a phone call from her mom.

Inferring reasons

(positive) The woman realizes it is the man's mother who is calling and she passes the phone back to the man.

(negative) The man realizes it is the woman's mother who is calling and he passes the phone back to the woman.

Identifying characters

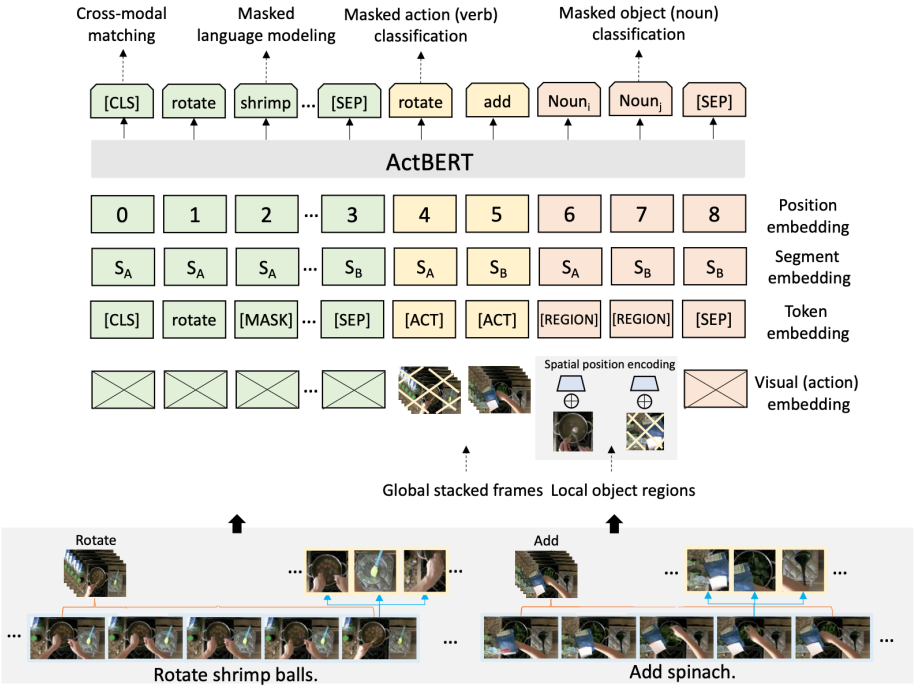
(positive) The phone rings, a man picks it up, and a woman slams her hand on the desk and demands the man give her the phone.

(negative) The two people that the man in the glasses is talking to need to be briefed on something.

Global video understanding

https://openaccess.thecvf.com/content_CVPR_2020/papers/Liu_Violin_A_Large-Scale_Dataset_for_Video-and-Language_Inference_CVPR_2020_paper.pdf

Video + Language: ActBert



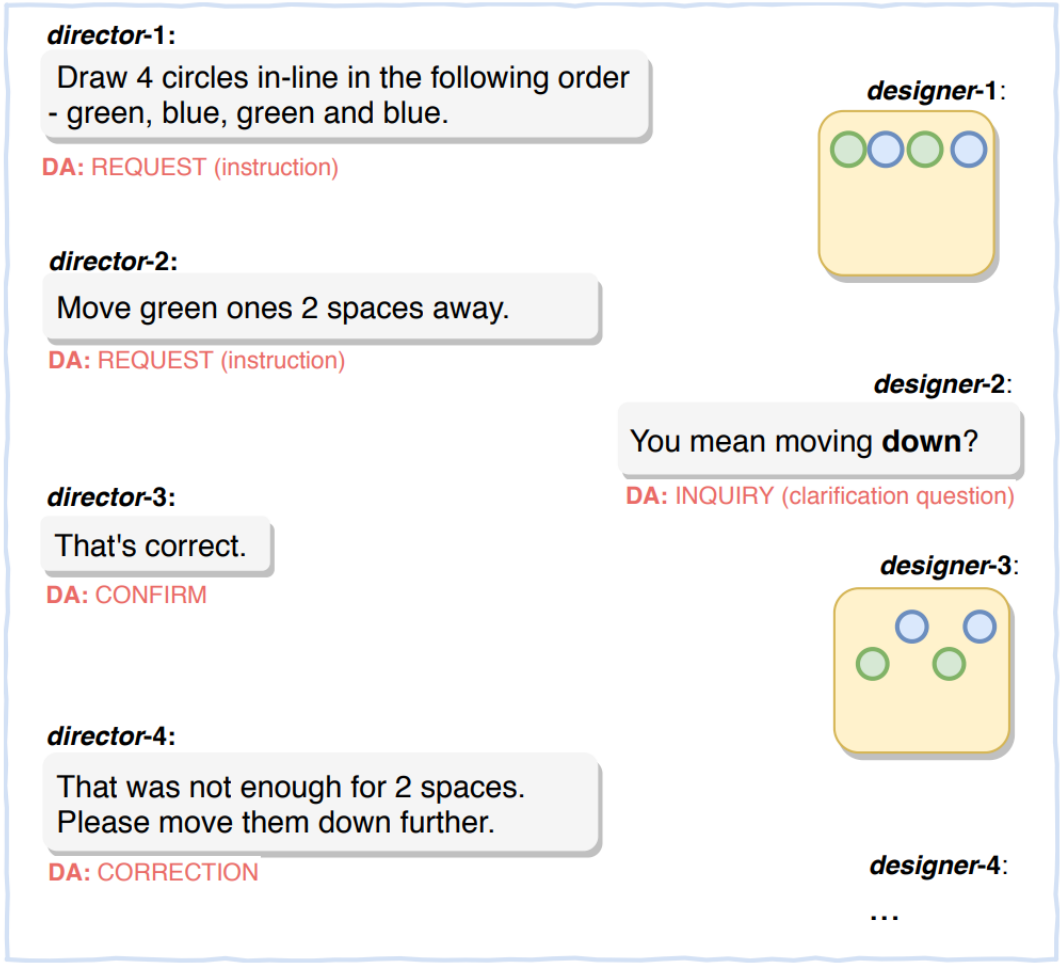
https://arxiv.org/abs/2011.07231

Counterfactuals in Vision and Language

Question Image	Counterfactual Questions	Counterfactual Images
Is this in Australia?	1. Is the grass green? 2. Is there grass on the ground? 3. Are they standing on a green grass field? 4. Is the stop light green?	
What color is the person's helmet?	1. What color jacket is the girl wearing? 2. What color jacket is the person wearing? 3. What color is the jacket? 4. What color is the woman's jacket?	
Where did the shadow on the car come from?	1. What kind of dog is this? 2. What type of dog is this? 3. What kind of dog is shown? 4. What is the breed of dog?	

https://openaccess.thecvf.com/content_CVPR_2020/papers/Abbasnejad_Counterfactual_Vision_and_Language_Learning_CVPR_2020_paper.pdf

Computer Assisted Visual Design and Language



https://arxiv.org/pdf/1812.04081.pdf

Any other ideas?