Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 1.2:
Multimodal Research Tasks

Louis-Philippe Morency
Guest lecture by Paul Liang

* Original version co-developed with Tadas Baltrusaitis

Lecture Objectives

- Understand the breath of possible tasks for multimodal research
- Research topics in affective computing
- Media description and Multimodal QA
- Multimodal navigation
- Examples of previous course projects
- Available multimodal datasets

Administrative Stuff

First Reading Assignment – Week 2

- 3 paper options are available
  - **Each student should pick one option!**
  - Then you will create a short summary to help others
- Discussions with your study group
  - 9-10 students in each study group
  - Discuss together the 3 papers. Ask questions!
    - But you should also try to answer the questions
- Google Sheets were created to help balance the papers between group members

First Reading Assignment – Week 2

Four main steps for the reading assignments

1. **Monday 8pm:** Official start of the assignment
2. **Wednesday 8pm:** Select your paper
3. **Friday 8pm:** Post your summary
4. **Monday 8pm:** End of the reading assignment

Details posted on Piazza

Lecture Highlights – Starting Next Week!

- Students should summarize lecture highlights
  - Each lecture is split in 3 segments (~30mins each)
  - One highlight statement for each segment
- Highlights submitted 42 hours after the lecture
  - Lecture can be watched live or asynchronously
- Optionally, students can ask questions

Detailed instructions were also posted on Piazza

Process for Selecting your Course Project

- Today: Lecture describing available multimodal datasets and research topics
- **Tuesday 9/8:** Let us know your dataset preferences for the course project
- **Thursday 9/10:** During the later part of the lecture, we will have an interactive period to help with team formation
- **Wednesday 9/16:** Pre-proposals are due. You should have selected your teammates, dataset and task
- Following week: meeting with TAs to discuss project

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)\\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)\\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

Course Project Timeline

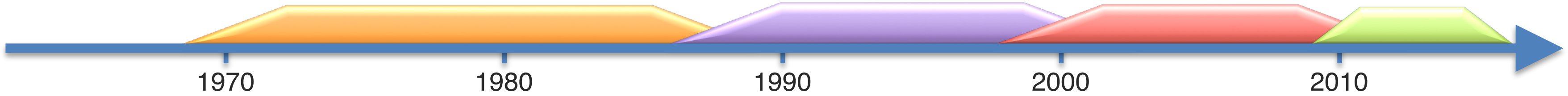
- Pre-proposal (Wednesday 9/16)
  - Define your dataset, research task and teammates
- First project assignment (due Friday Oct. 9)
  - Experiment with unimodal representations
  - Study prior work on your selected research topic
- Midterm project assignment (due Friday Nov. 12)
  - Implement and evaluate state-of-the-art model(s)
  - Discuss new multimodal model(s)
- Final project assignment (due Friday Dec. 11)
  - Implement and evaluate new multimodal model(s)
  - Discuss results and possible future directions

Multimodal Research Tasks

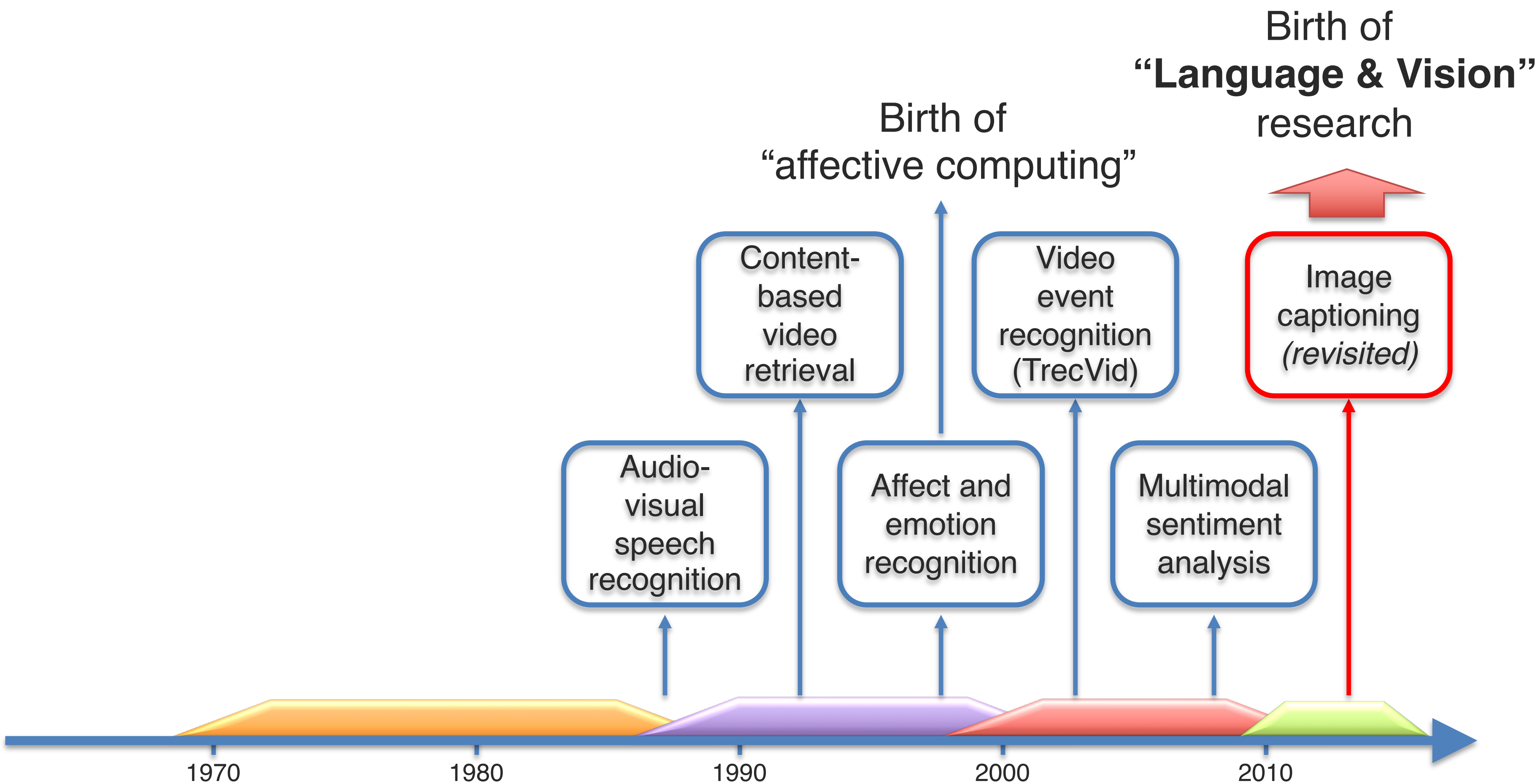
Prior Research on "Multimodal"

Four eras of multimodal research

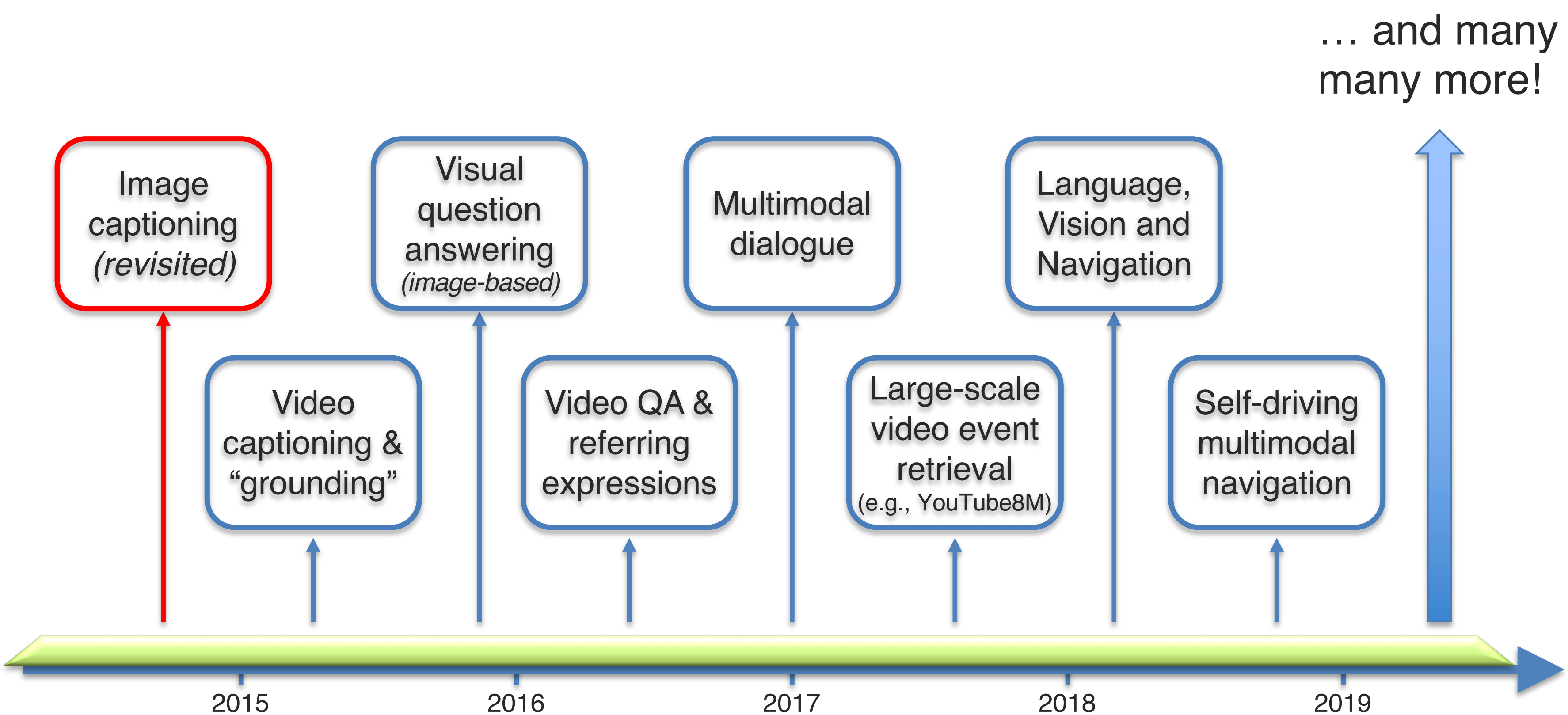
- ➢ The "behavioral" era (1970s until late 1980s)
- ➢ The "computational" era (late 1980s until 2000)
- ➢ The "interaction" era (2000 - 2010)
- ➢ The "deep learning" era (2010s until …)
  - ❖ Main focus of this course



Multimodal Research Tasks



Multimodal Research Tasks



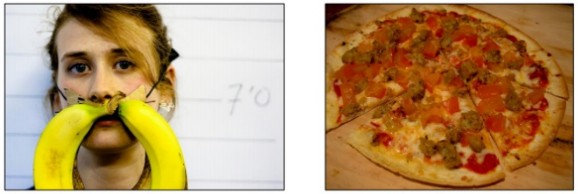
Real world tasks tackled by MMML

A. Affect recognition
   - Emotion
   - Personalities
   - Sentiment

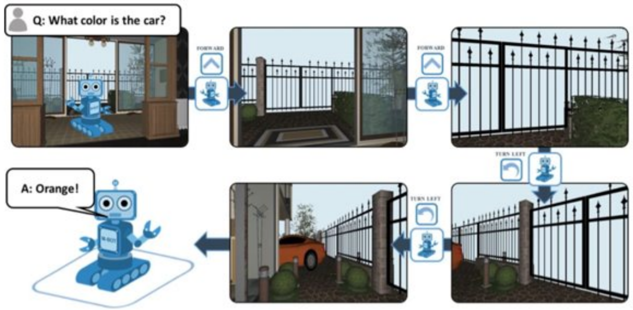
B. Media description
   - Image and video captioning

C. Multimodal QA
   - Image and video QA
   - Visual reasoning

D. Multimodal Navigation
   - Language guided navigation
   - Autonomous driving


What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

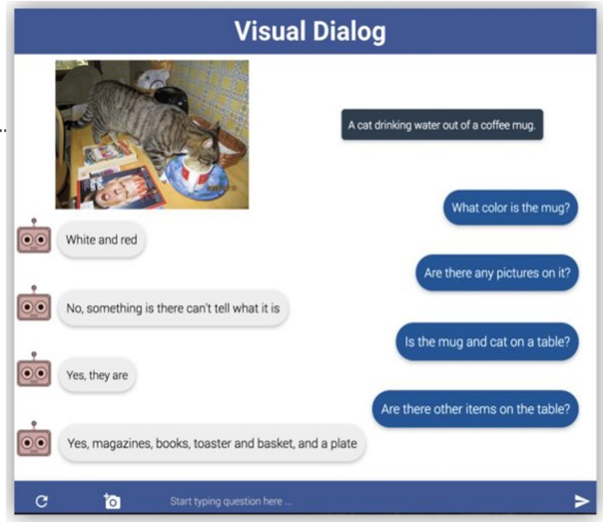


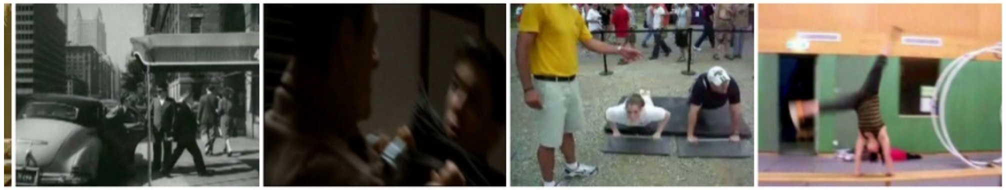
Real world tasks tackled by MMML

E. Multimodal Dialog
   - Grounded dialog

F. Event recognition
   - Action recognition
   - Segmentation

G. Multimedia information retrieval
   - Content based/Cross-media





(a) get-out-car (a) fight-person (b) push-up (b) cartwheel

Affective Computing

Common Topics in Affective Computing

- **Affective states** – emotions, moods, and feelings
- **Cognitive states** – thinking and information processing
- **Personality** – patterns of acting, feeling, and thinking
- **Pathology** – health, functioning, and disorders
- **Social processes** – groups, cultures, and perception

Common topics in affective computing

- **Affective states**
- Cognitive states
- Personality
- Pathology
- Social processes

Anger	Frustration
Disgust	Anxiety
Fear	Contempt
Happiness	Shame
Sadness	Guilt
Positivity	Wonder
Activation	Relaxation
Pride	Pain
Desire	Envy

Common topics in affective computing

- Affective states
- **Cognitive states**
- Personality
- Pathology
- Social processes

- Engagement
- Interest
- Attention
- Concentration
- Effort
- Surprise
- Confusion
- Agreement
- Doubt
- Knowledge

Common topics in affective computing

- Affective states
- Cognitive states
- **Personality**
- Pathology
- Social processes

- Outgoing
- Assertive
- Energetic
- Sympathetic
- Respectful
- Trusting
- Organized
- Productive
- Responsible
- Pessimistic
- Anxious
- Moody
- Curious
- Artistic
- Creative
- Sincere
- Modest
- Fair

Common topics in affective computing

- Affective states
- Cognitive states
- Personality
- **Pathology**
- Social processes

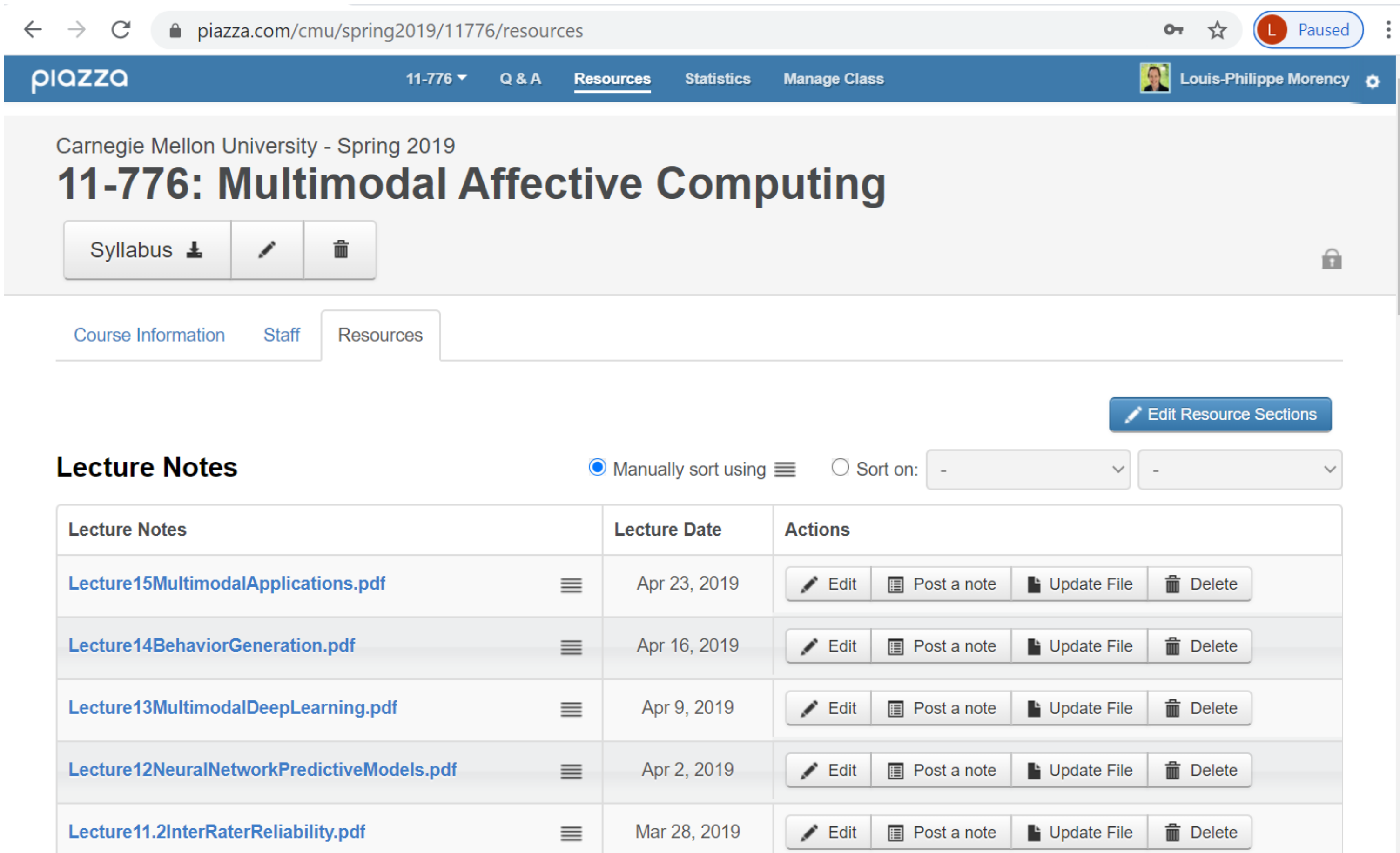
- Depression
- Anxiety
- Trauma
- Addiction
- Schizophrenia
- Antagonism
- Detachment
- Disinhibition
- Negative Affectivity
- Psychoticism

Common topics in affective computing

- Affective states
- Cognitive states
- Personality
- Pathology
- **Social processes**

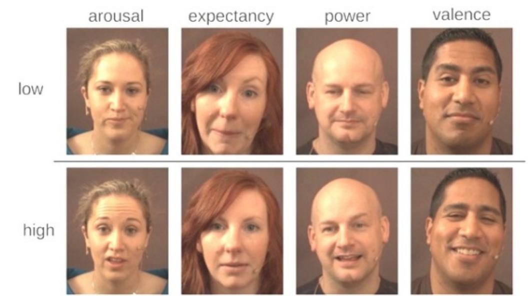
- Rapport
- Cohesion
- Cooperation
- Competition
- Status
- Conflict
- Attraction
- Persuasion
- Genuineness
- Culture
- Skillfulness

11-776 Multimodal Affective Computing

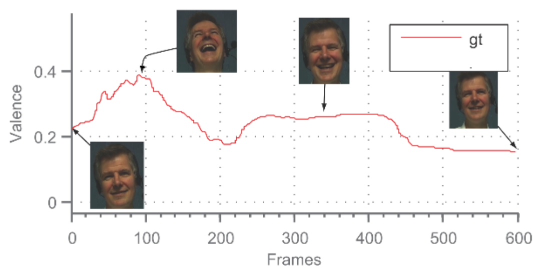


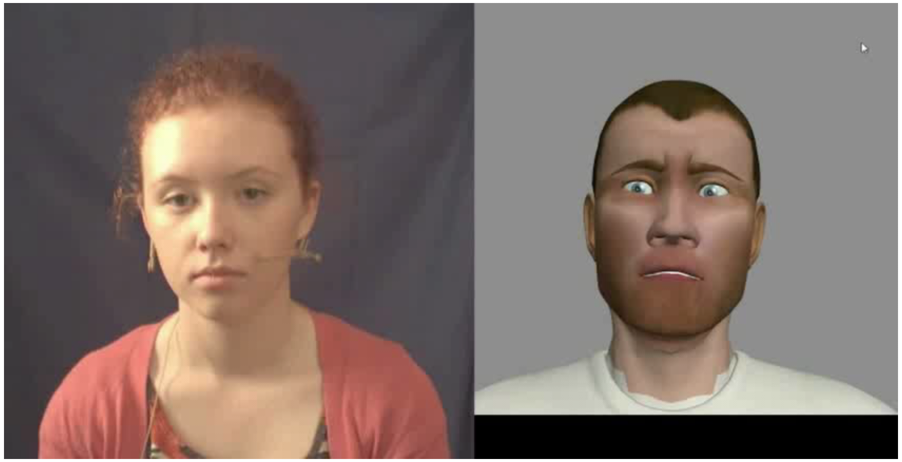
Audio-visual Emotion Challenge 2011/2012

- Part of a larger SEMAINE corpus
- Sensitive Artificial Listener paradigm
- Labeled for four dimensional emotions (per frame)
  - Arousal, expectancy, power, valence
- Has transcripts



AVEC 2011/2012





[**AVEC 2011 – The First International Audio/Visual Emotion Challenge**, B. Schuller et al., 2011]

Audio-visual Emotion Challenge 2013/2014

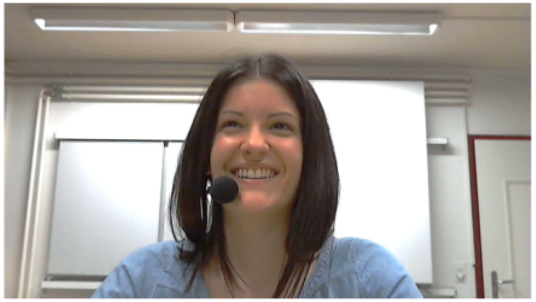
- Reading specific text in a subset of videos
- Labeled for emotion per frame (valence, arousal, dominance)
- Performing an HCI task
  - Reading aloud a text in German
  - Responding to a number of questions
- 100 audio-visual sessions
- Provide extracted audio-visual features

AVEC 2013/2014

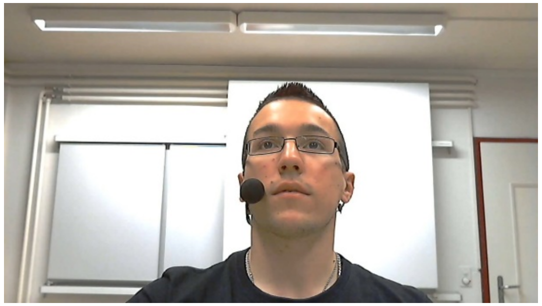
[**AVEC 2013 – The Continuous Audio/Visual Emotion and Depression Recognition Challenge**, Valstar et al. 2013]

Audio-visual Emotion Challenge 2015/2016

- RECOLA dataset
- Audio-Visual emotion recognition
- Labeled for dimensional emotion per frame (arousal, valence)
- Includes physiological data
- 27 participants
- French, audio, video, ECG and EDA
- Collaboration task in video conference
- Broader range of emotive expressions



AVEC 2015



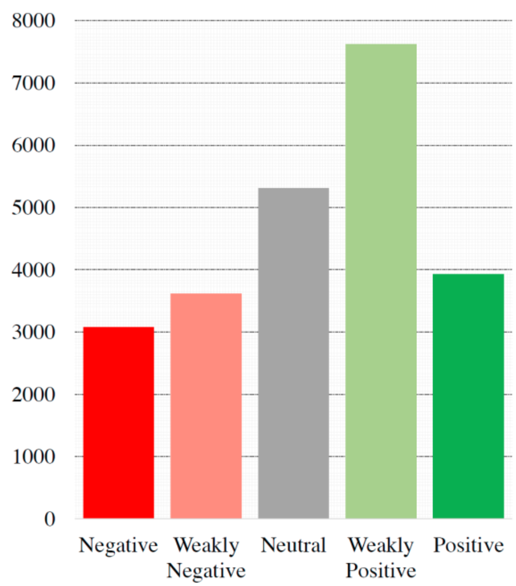
[**Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions**, F. Ringeval et al., 2013]

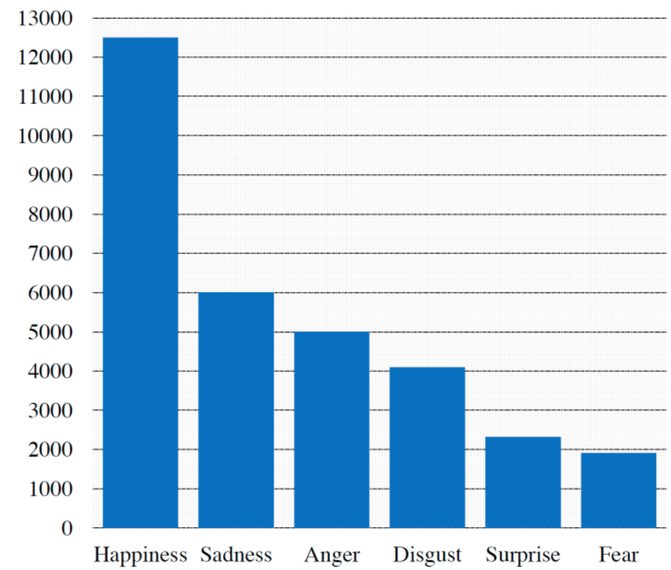
Multimodal Sentiment Analysis

- Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos (MOSI)
- 89 speakers with 2199 opinion segments
- Audio-visual data with transcriptions
- Labels for sentiment/opinion
  - Subjective vs objective
  - Positive vs negative

Multimodal Sentiment Analysis

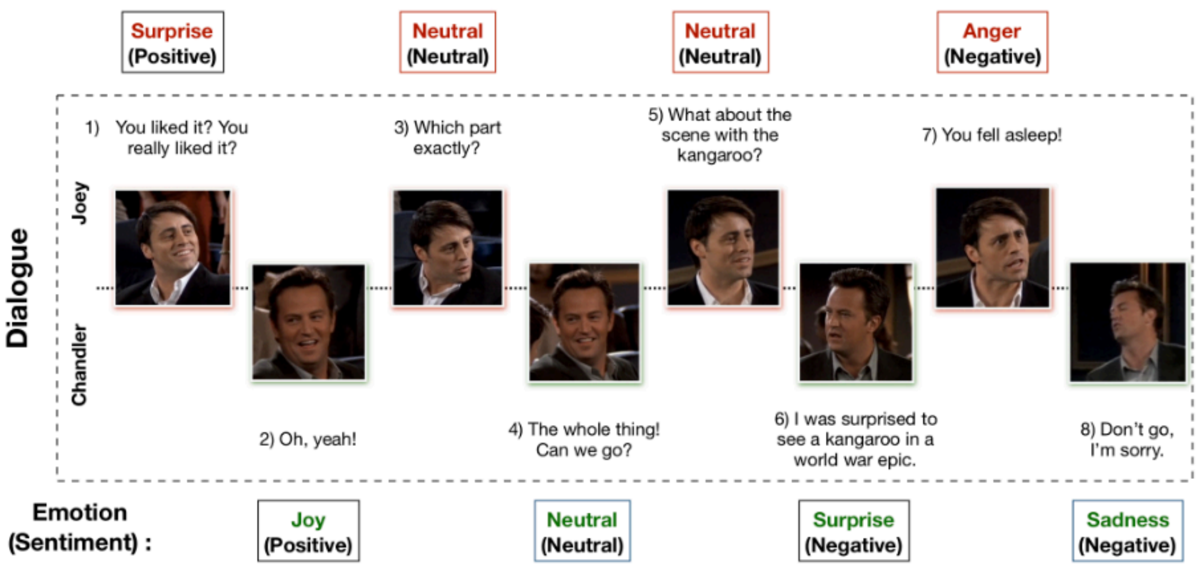
- Multimodal sentiment and emotion recognition
- CMU-MOSEI : 23,453 annotated video segments from 1,000 distinct speakers and 250 topics





Multi-Party Emotion Recognition

- MELD: Multi-party dataset for emotion recognition in conversations



Utterance: *"Become a drama critic!"*

Emotion: *Joy* **Sentiment:** *Positive*

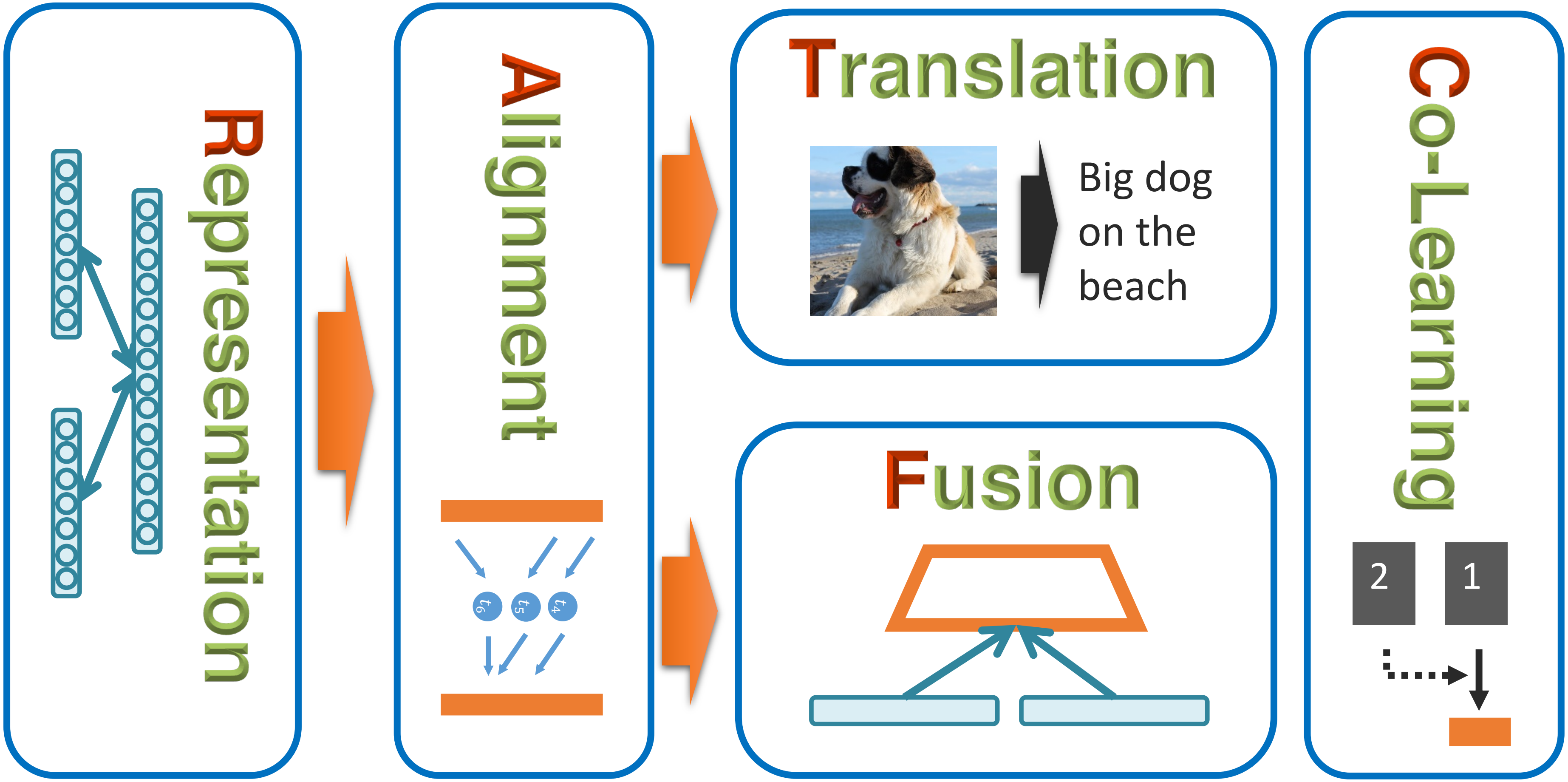
Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face

Utterance: *"Great, now he is waving back"*

Emotion: *Disgust* **Sentiment:** *Negative*

Text	Audio	Visual
Positive/Joy	Flat tone	Frown

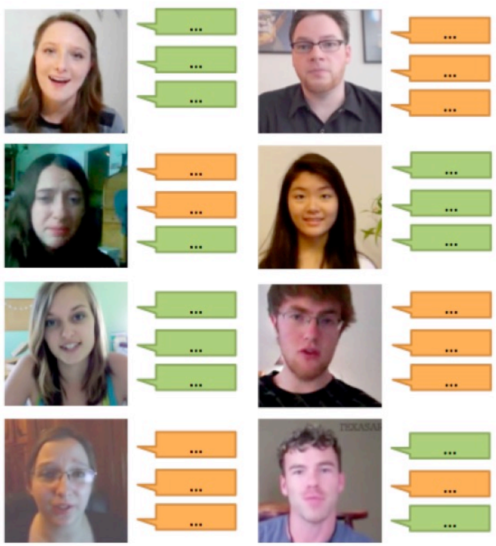
What are the Core Challenges Most Involved in Affect Recognition?



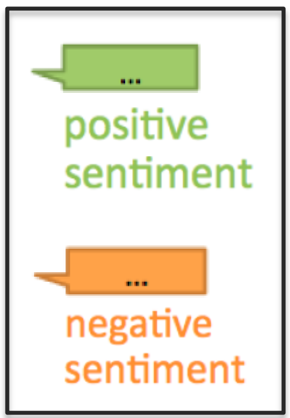
Project Example: Select-Additive Learning

Research task: Multimodal sentiment analysis
Datasets: MOSI, YouTube, MOUD

Main idea: Reducing the effect of *confounding factors* when limited dataset size



Legend



What rules can you infer from this data?

- ✔ *Smile* -> positive sentiment
- ✔ *Frown* -> negative sentiment
- ✔ *nod*-> positive sentiment
- ✗ *Wearing glasses* -> negative sentiment

Confounding factor!

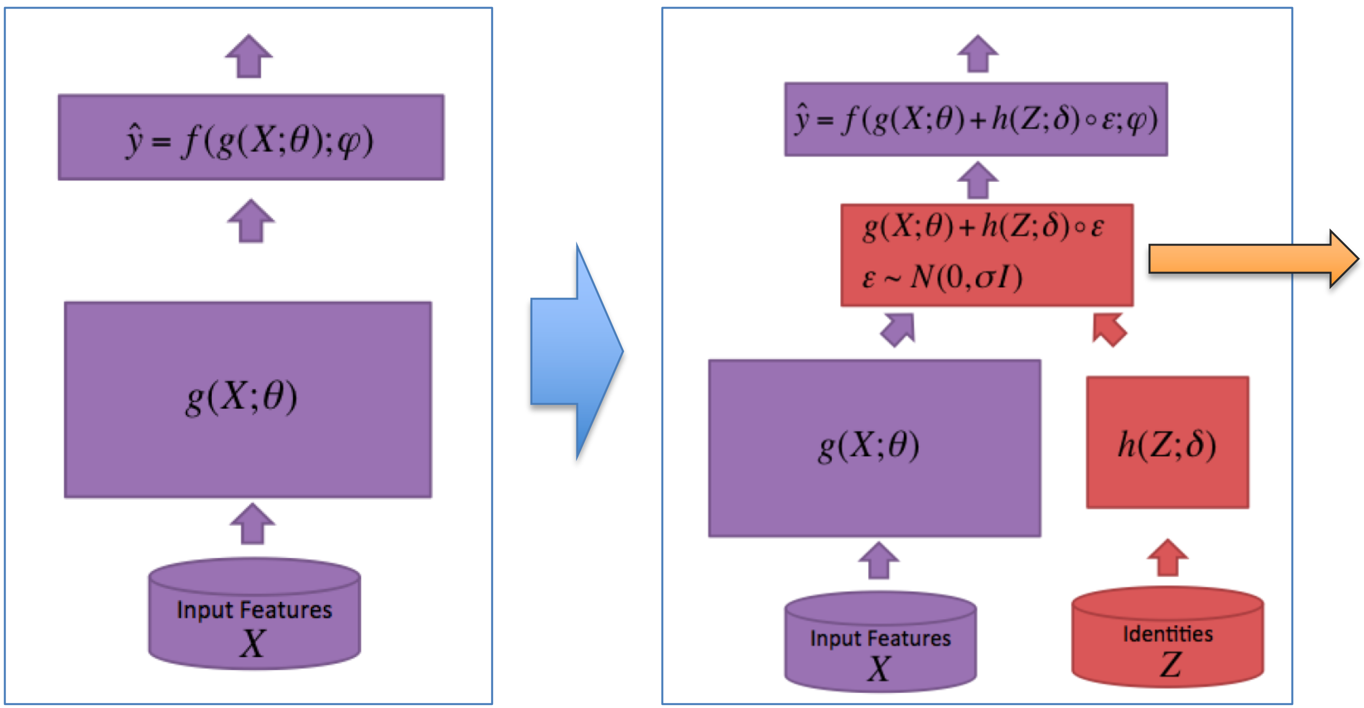
Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency and Eric P. Xing, Select-additive Learning: Improving Generalization In Multimodal Sentiment Analysis, ICME 2017, https://arxiv.org/abs/1609.05244

Project Example: Select-Additive Learning

Solution: Learning representations that reduce the effect of user identity

"Conventional" representation learning

Select-Additive Learning



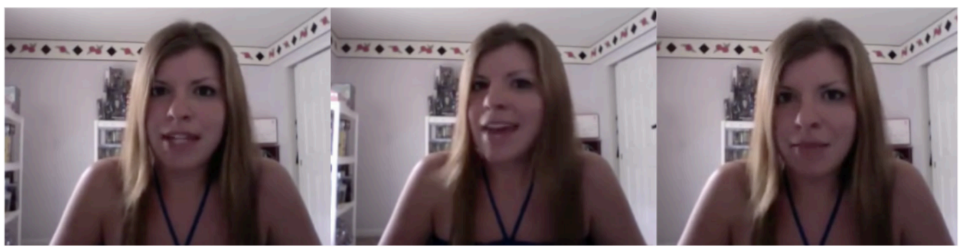
Hypothesis: the representation is a mixture from the person-independent factor g(X) and the person-dependent factor h(Z).

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency and Eric P. Xing, Select-additive Learning: Improving Generalization In Multimodal Sentiment Analysis, ICME 2017, https://arxiv.org/abs/1609.05244

Project Example: Word-Level Gated Fusion

Research task: Multimodal sentiment analysis
Datasets: MOSI, YouTube, MOUD

Main idea: Estimating importance of each modality at the word-level in a video.



Visual Gate: Reject Pass Reject

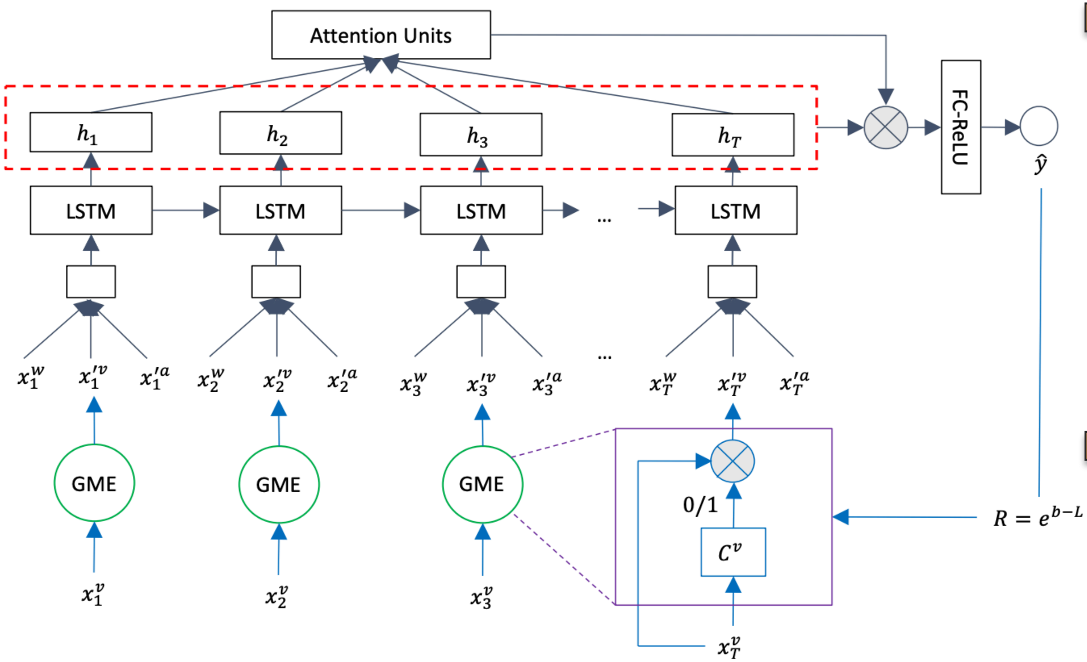
Visual modality: Hands cover mouth

How can we build an interpretable model that estimates modality and temporal importance, and learns to attend to important information?

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, Louis-Philippe Morency, Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning, ICMI 2017, https://arxiv.org/abs/1802.00924

Project Example: Word-Level Gated Fusion

Solution:
- Word-level alignment
- Temporal attention over words
- Gated attention over modalities



Hypothesis: attention weights represent contribution of each modality at each time step

Modality gates that determine importance and contribution of each modality – trained with reinforcement learning

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, Louis-Philippe Morency, Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning, ICMI 2017, https://arxiv.org/abs/1802.00924

Media Description

Media description

- Given a media (image, video, audio-visual clips) provide a free form text description

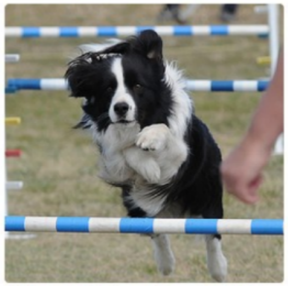
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."



"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

Large-Scale Image Captioning Dataset

- Microsoft Common Objects in COntext (MS COCO)
- 120000 images
- Each image is accompanied with five free form sentences describing it (at least 8 words)
- Sentences collected using crowdsourcing (Mechanical Turk)
- Also contains object detections, boundaries and keypoints

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Evaluating Image Caption Generations

- Has an evaluation server
  - Training and validation - 80K images (400K captions)
  - Testing – 40K images (380K captions), a subset contains more captions for better evaluation, these are kept privately (to avoid over-fitting and cheating)
- Evaluation is difficult as there is no one “correct” answer for describing an image in a sentence
- Given a candidate sentence it is evaluated against a set of “ground truth” sentences

Evaluating Image Captioning Results

- A challenge was done with actual human evaluations of the captions (CVPR 2015)

M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	Average correctness of the captions on a scale 1-5 (incorrect - correct).
M4	Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5	Percentage of captions that are similar to human description.

Evaluating Image Captioning Results

- A challenge was done with actual human evaluations of the captions (CVPR 2015)

	M1	M2	M3	M4	M5
Human[5]	0.638	0.675	4.836	3.428	0.352
Google[4]	0.273	0.317	4.107	2.742	0.233
MSR[8]	0.268	0.322	4.137	2.662	0.234
Montreal/Toronto[10]	0.262	0.272	3.932	2.832	0.197
MSR Captivator[9]	0.250	0.301	4.149	2.565	0.233
Berkeley LRCN[2]	0.246	0.268	3.924	2.786	0.204
m-RNN[15]	0.223	0.252	3.897	2.595	0.202
Nearest Neighbor[11]	0.216	0.255	3.801	2.716	0.196

Evaluating Image Captioning Results

	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2
Google[4]	0.943	0.254	0.53	0.713	0.542
MSR Captivator[9]	0.931	0.248	0.526	0.715	0.543
m-RNN[15]	0.917	0.242	0.521	0.716	0.545
MSR[8]	0.912	0.247	0.519	0.695	0.526
Nearest Neighbor[11]	0.886	0.237	0.507	0.697	0.521
m-RNN (Baidu/ UCLA)[16]	0.886	0.238	0.524	0.72	0.553
Berkeley LRCN[2]	0.869	0.242	0.517	0.702	0.528
Human[5]	0.854	0.252	0.484	0.663	0.469

Video captioning

AD: Abby gets in the basket.

Mike leans over and sees how high they are.

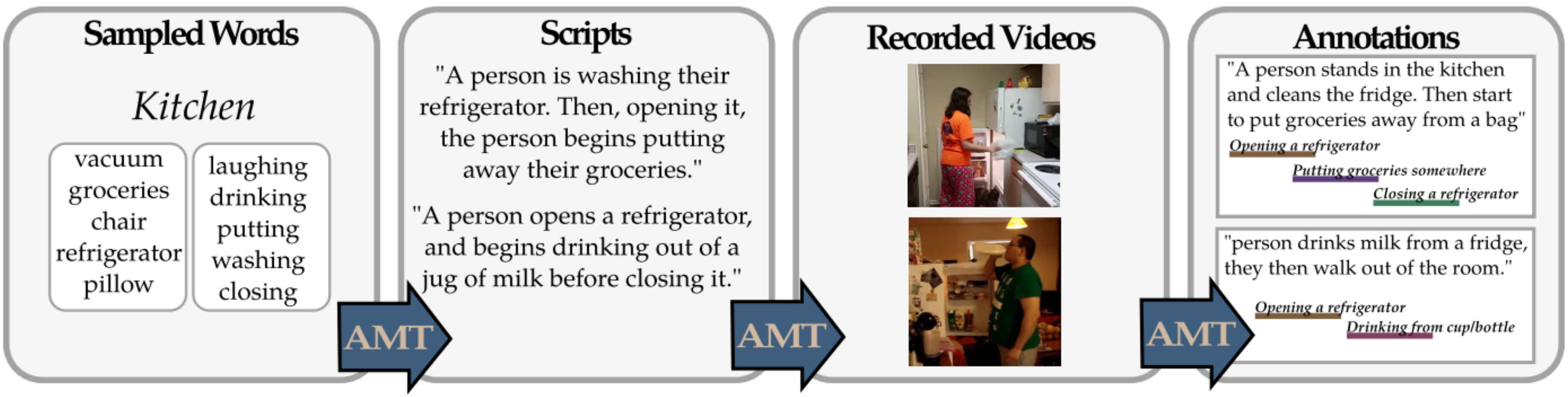
Abby clasps her hands around his face and kisses him passionately.

Based on audio descriptions for the blind (Descriptive Video Service – DVS)

- Alignment is a challenge since description can happen after the video segment
- Only one single caption per clip – Challenge with evaluation

Video Description and Alignment

Let's ask MTurk users to "act" the description!

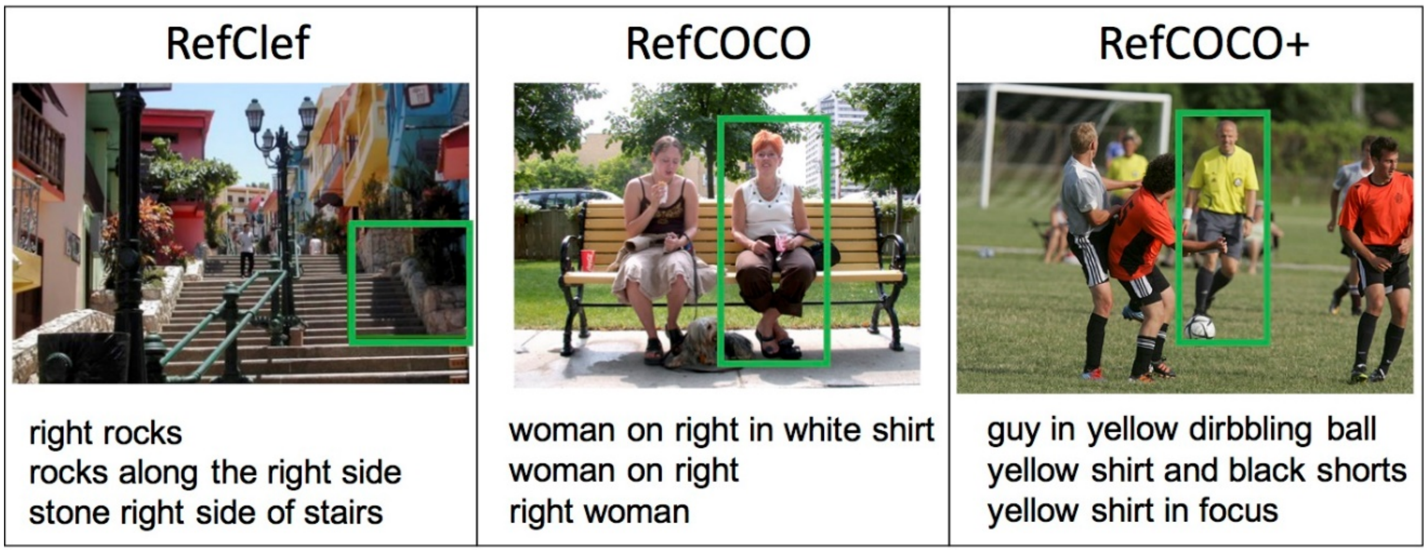


Charade Dataset: http://allenai.org/plato/charades/

First author was student in first edition of MMML course!

How to Address the Challenge of Evaluation?

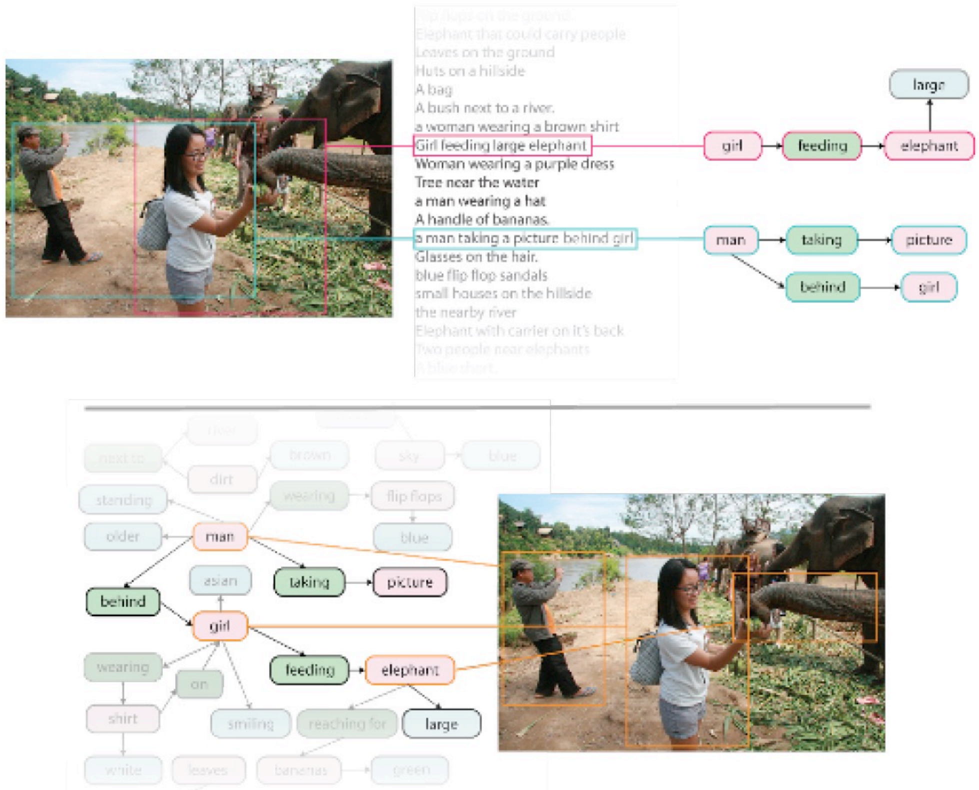
Referring Expressions: Generate / Comprehend a noun phrase which identifies a particular object in an image



This is related to “grounding” which links linguistic elements to the shared environment (in this case, it’s an image)

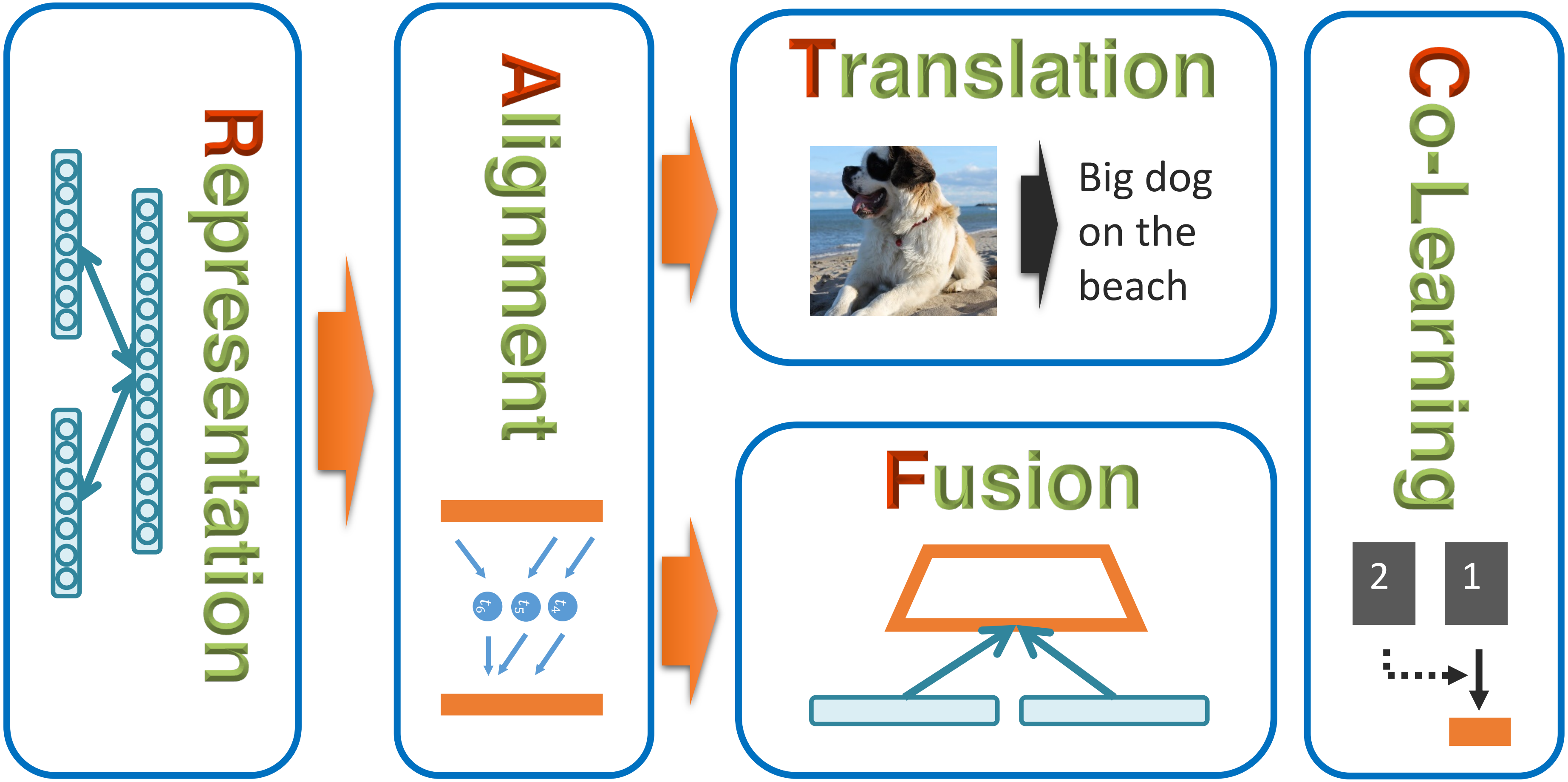
Large-Scale Description and Grounding Dataset

Visual Genome Dataset



https://visualgenome.org/

What are the Core Challenges Most Involved in Media Description?



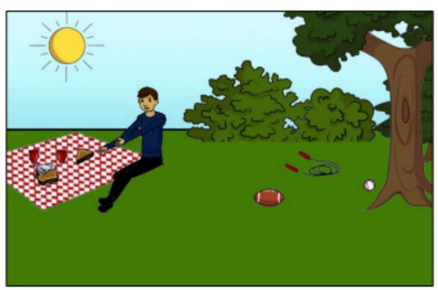
Multimodal QA

Visual

- Task - Given an image and a question, answer the question (http://www.visualqa.org/)


What color are her eyes?
What is the mustache made of?

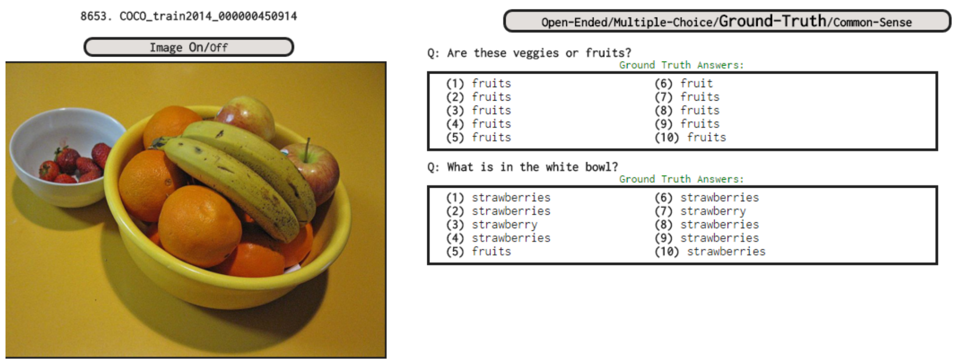

How many slices of pizza are there?
Is this a vegetarian pizza?

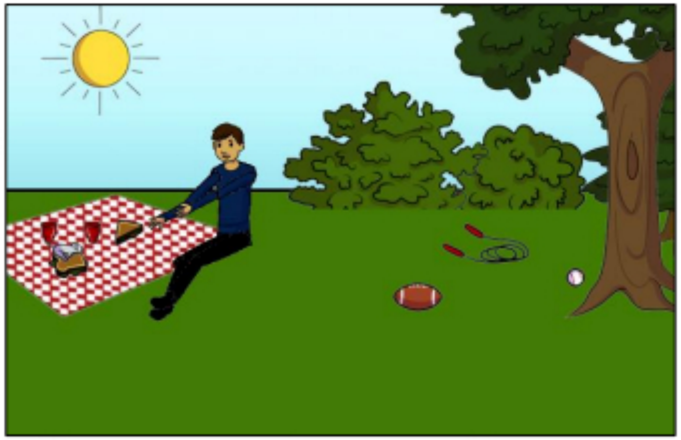

Is this person expecting company?
What is just under the tree?


Does it appear to be rainy?
Does this person have 20/20 vision?

Multimodal QA dataset 1 – VQA (C1)

- Real images
  - 200k MS COCO images
  - 600k questions
  - 6M answers
  - 1.8M plausible answers
- Abstract images
  - 50k scenes
  - 150k questions
  - 1.5M answers
  - 450k plausible answers





Is this person expecting company?
What is just under the tree?

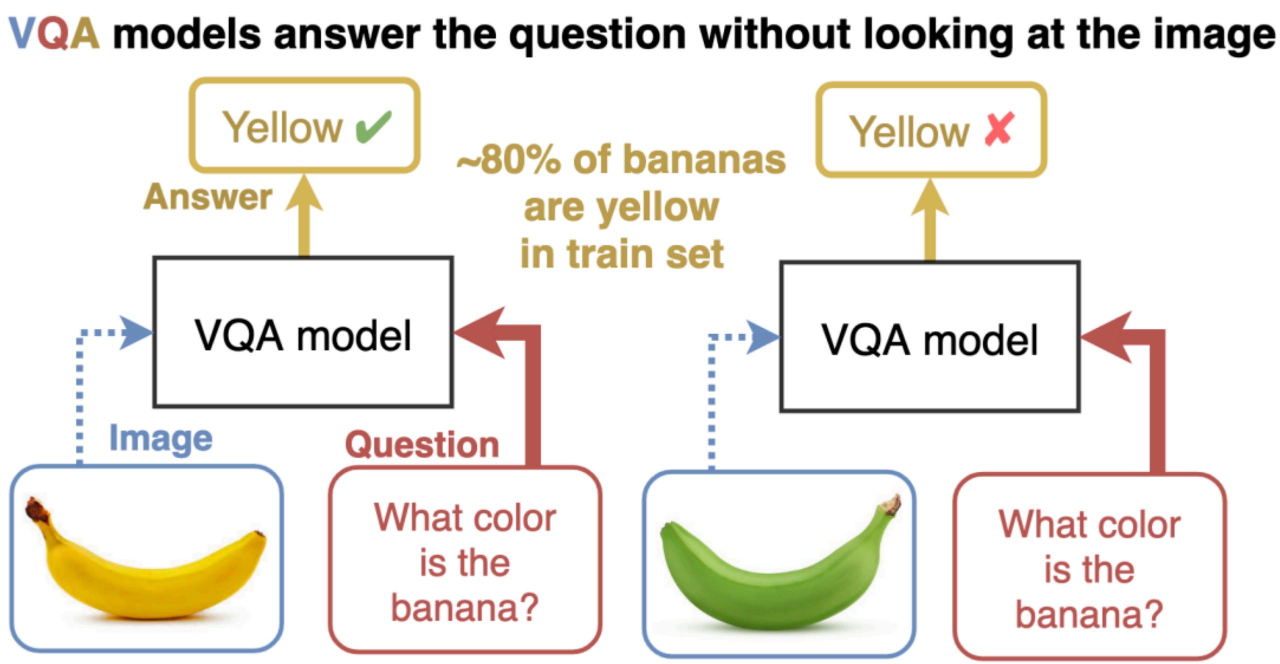
VQA Challenge 2016 and 2017 (C1)

- Two challenges organized these past two years (link)
- Currently good at yes/no question, not so much free form and counting

	By Answer Type			Overall
	Yes/No	Number	Other	
UC Berkeley & Sony[14]	83.79	38.9	58.64	66.9
Naver Labs[10]	83.78	37.67	54.74	64.89
DLAIT[5]	83.65	39.18	52.62	63.97
snubi-naverlabs[25]	83.64	38.43	51.61	63.4
POSTECH[11]	81.85	38.02	53.12	63.35
Brandeis[3]	82.53	36.54	51.71	62.8
VTComputerVison[19]	80.31	37.87	52.16	62.23
MIL-UT[7]	82.39	36.7	49.76	61.82

VQA 2.0

- Just guessing without an image lead to ~51% accuracy
  - So the V in VQA "only" adds 14% increase in accuracy



VQA 2.0

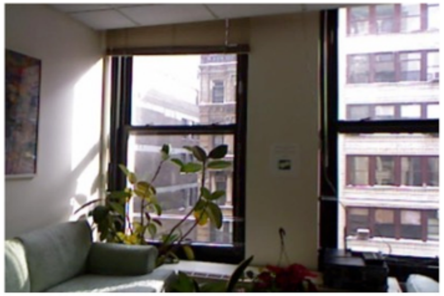
- Just guessing without an image lead to ~51% accuracy
  - So the V in VQA “only” adds 14% increase in accuracy
- [VQA v2.0](#) is attempting to address this

Multimodal QA – other VQA datasets

COCOQA
Q: What is the color of the desk?
A: white
Q: What are on the white desk?
A: computers

COCOQA
Q: What is the color of the dresses?
A: purple
Q: What are three women dressed up and on?
A: phones

DAQUAR
Q: What is the object close to the wall?
A: whiteboard
Q: What is the object in front of the sofa?
A: table



DAQUAR
Q: What is the largest object?
A: sofa
Q: How many windows are there?
A: 2

VQA
Q: How many bikes are there?
A: 2
Q: What number is the bus?
A: 48

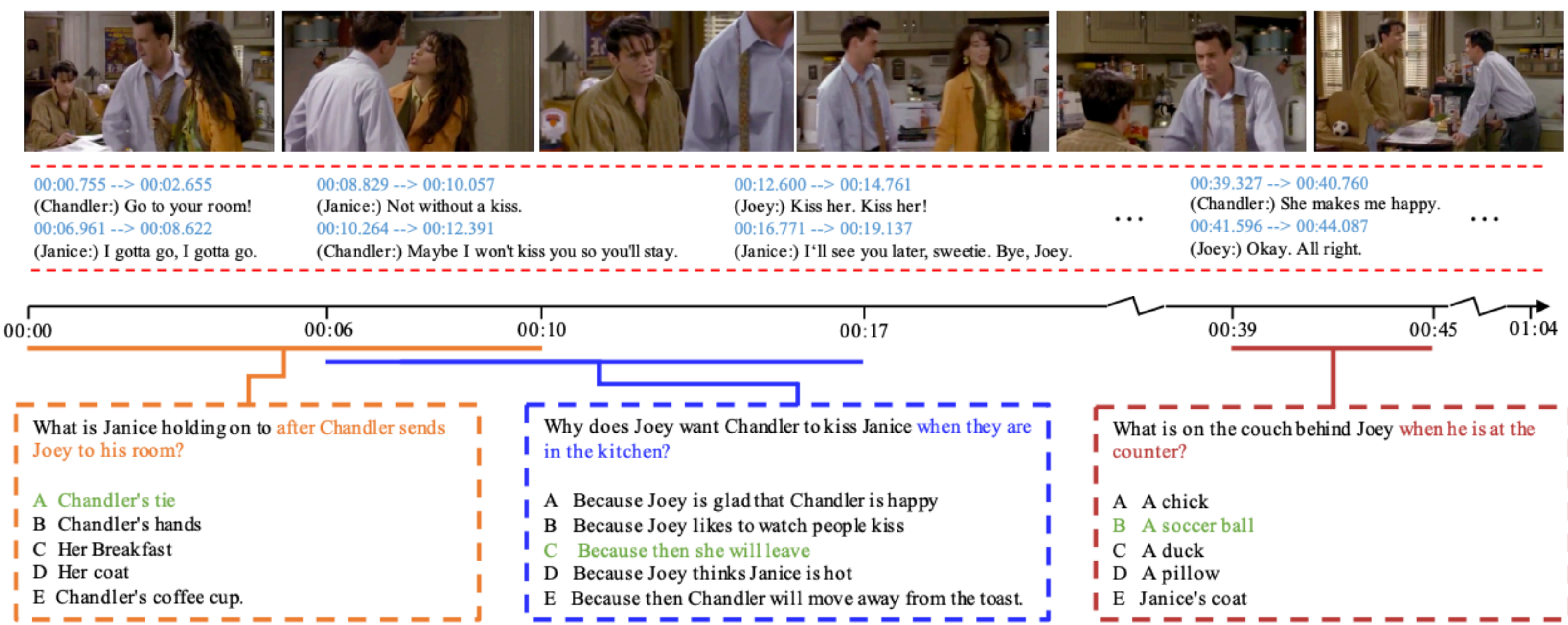
VQA
Q: How many pickles are on the plate?
A: 1
Q: What is the shape of the plate?
A: round

VQA
Q: What does the sign say?
A: stop
Q: What shape is this sign?
A: octagon

VQA
Q: What type of trees are here?
A: palm
Q: Is the skateboard airborne?
A: yes

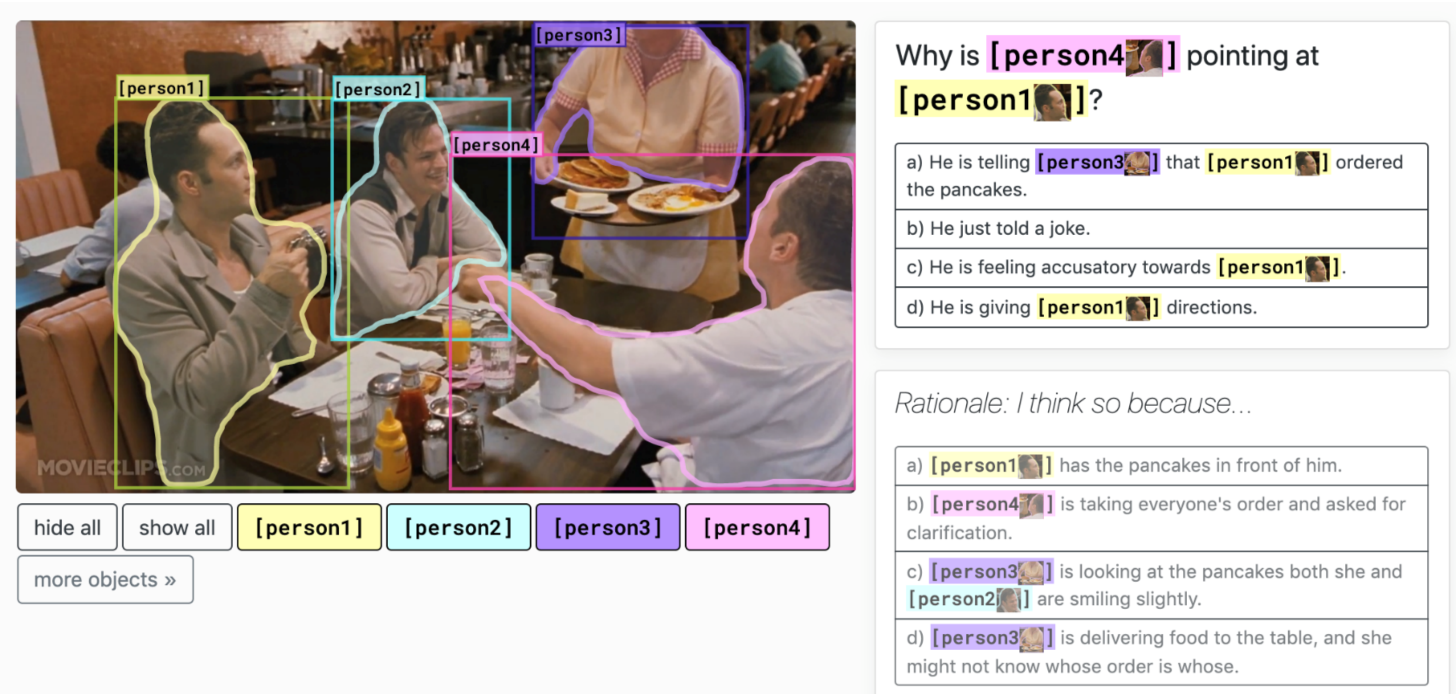
Multimodal QA – other VQA datasets (C7)

- TVQA
  - Video QA dataset based on 6 popular TV shows
  - 152.5K QA pairs from 21.8K clips
  - Compositional questions



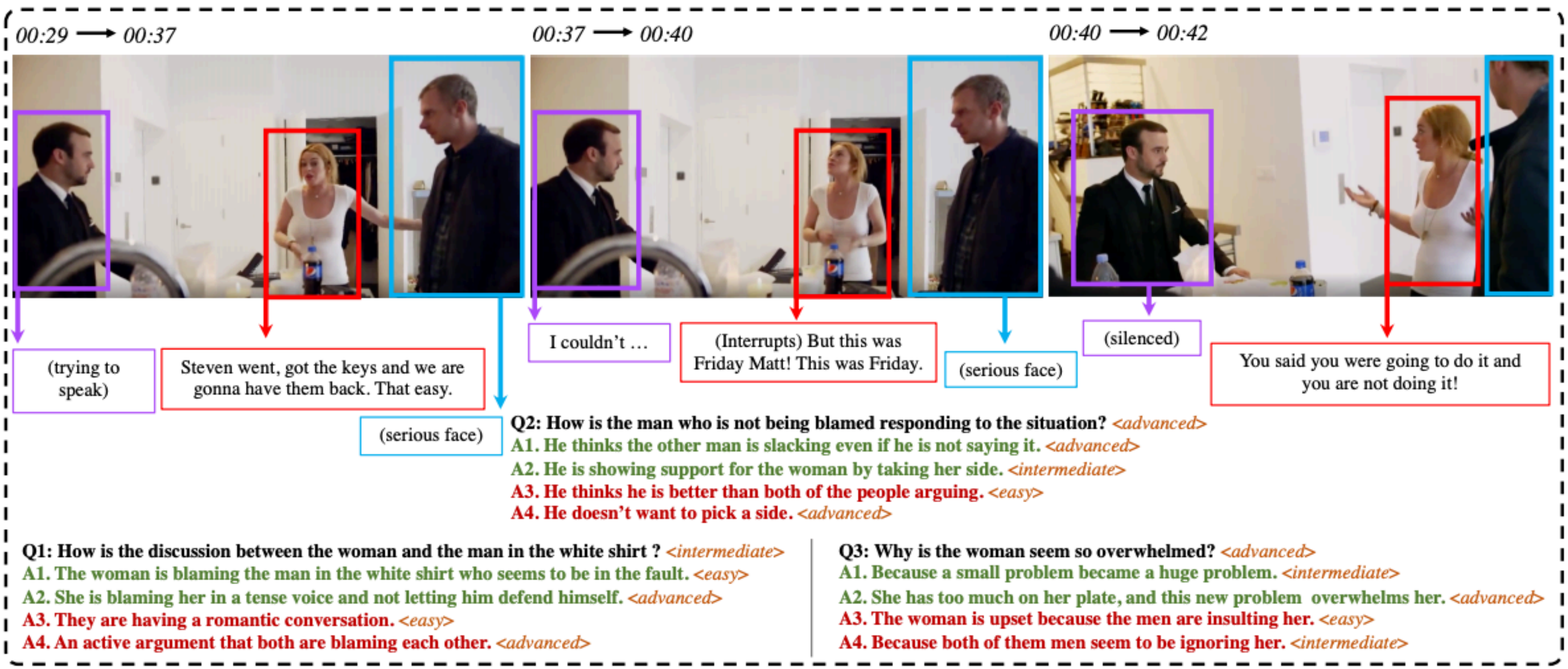
Multimodal QA – Visual Reasoning (C8)

- VCR: Visual Commonsense Reasoning
  - Model must answer challenging visual questions expressed in language
  - And provide a **rationale explaining why its answer is true**.

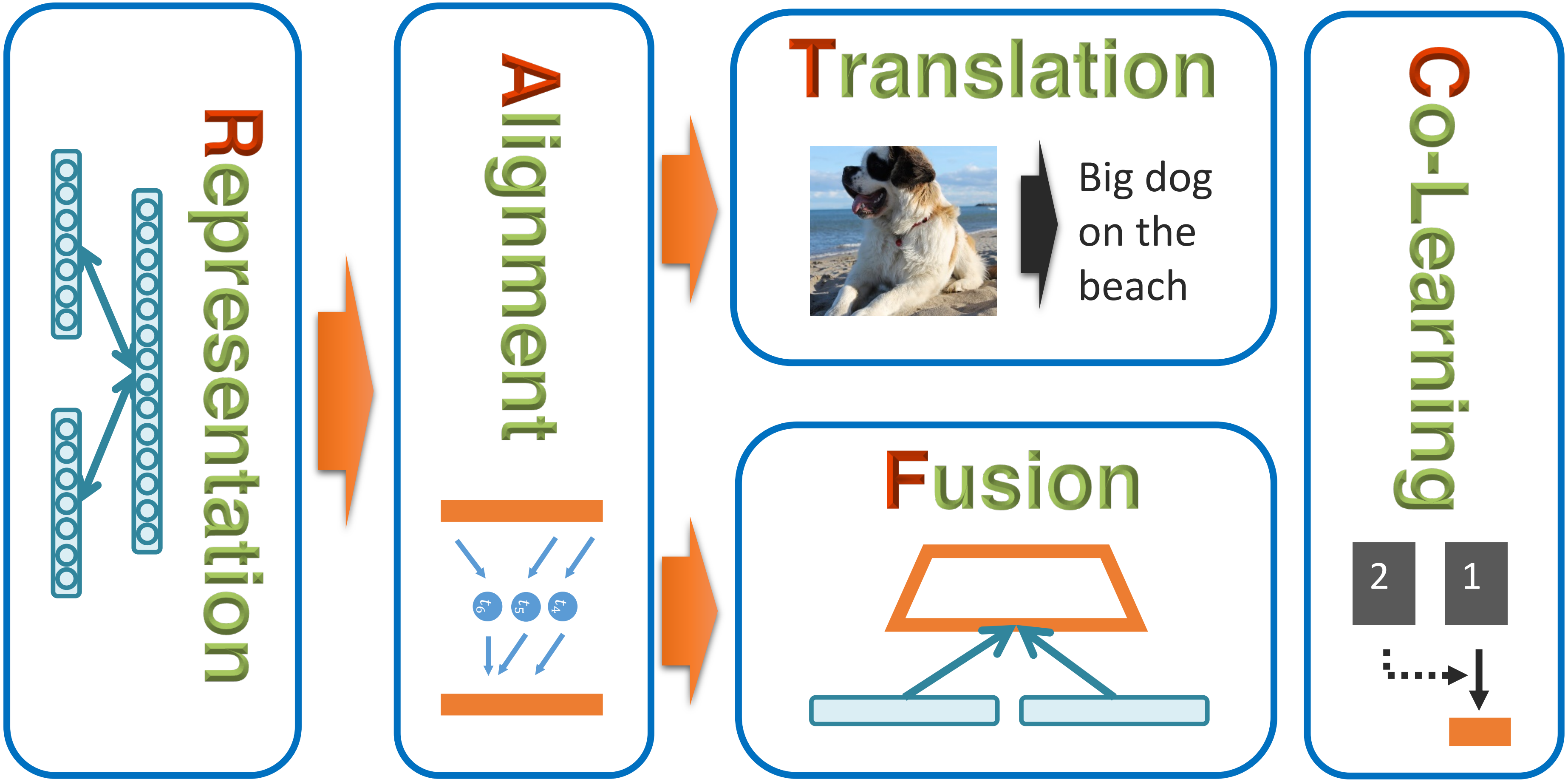


Social-IQ (A10)

- Social-IQ: 1.2k videos, 7.5k questions, 50k answers
- Questions and answers centered around social behaviors

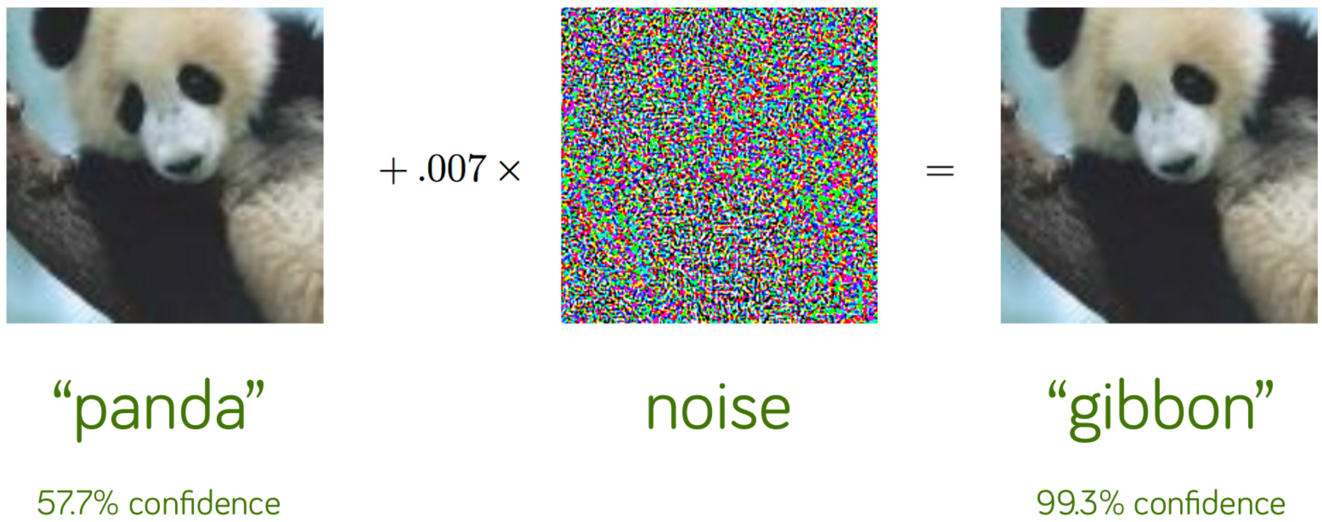


What are the Core Challenges Most Involved in Multimodal QA?



Project Example: Adversarial Attacks on VQA models

Research task: Adversarial Attacks on VQA models
Datasets: VQA
Main idea: Test the robustness of VQA models to adversarial attacks on the image.



Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, Louis-Philippe Morency, Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models. NeurIPS ViGIL workshop 2018.
https://nips2018vigil.github.io/static/papers/accepted/33.pdf

Project Example: Adversarial Attacks on VQA models

Research task: Adversarial Attacks on VQA models
Datasets: VQA
Main idea: Test the robustness of VQA models to adversarial attacks on the image.

+

Q: what kind of flowers are in the vase?

VQA model

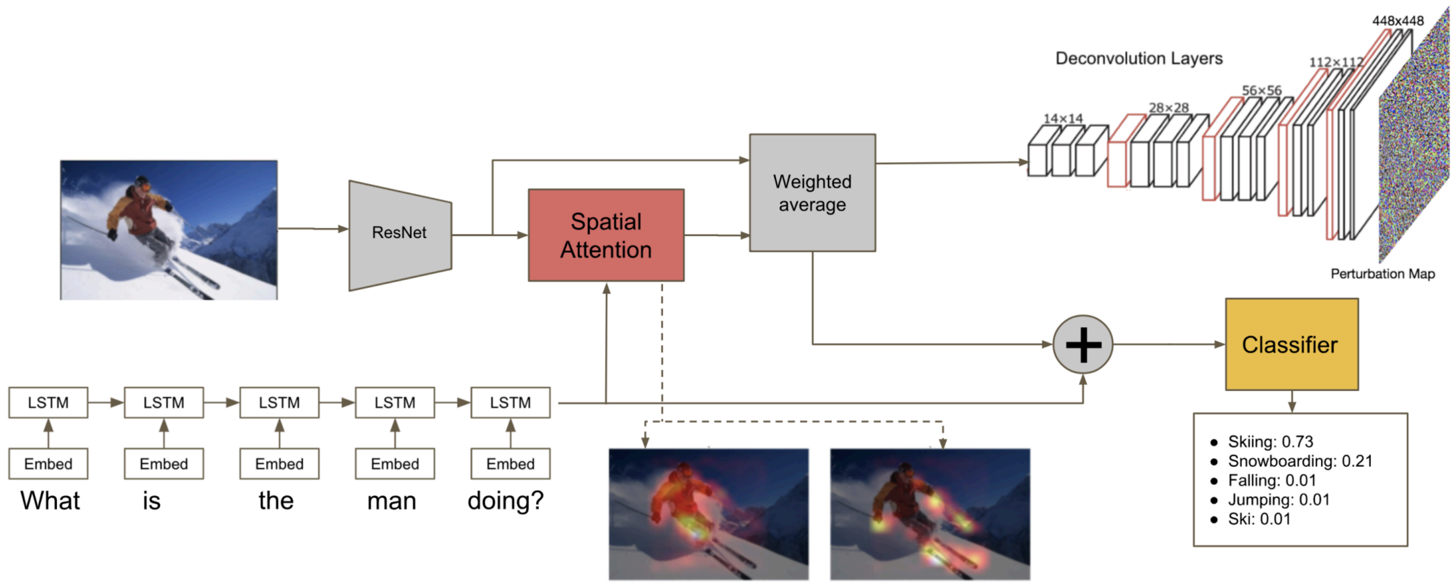
A: Roses to Sunflower

How can we design a targeted attack on images in VQA models, which will help in assessing robustness of existing models?

Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, Louis-Philippe Morency, Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models. NeurIPS ViGIL workshop 2018.
https://nips2018vigil.github.io/static/papers/accepted/33.pdf

Project Example: Adversarial Attacks on VQA models

Solution: Use fusion over original image and question to generate an **adversarial perturbation map** over the image



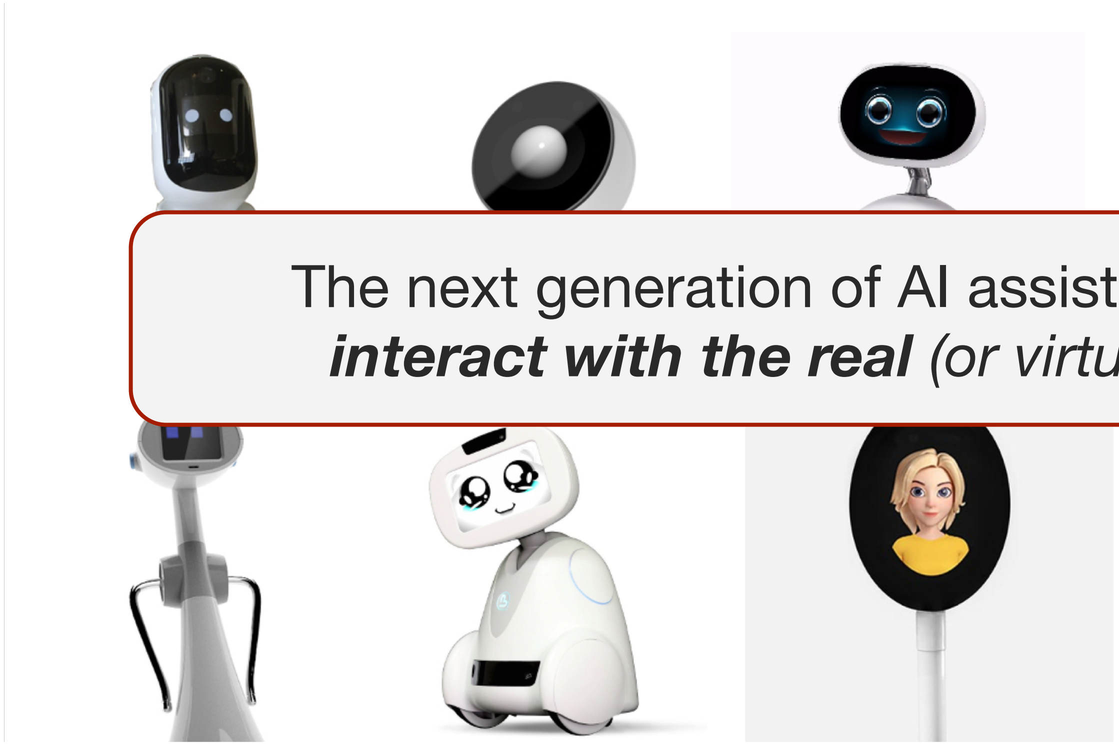
Hypothesis: question helps to localize important visual regions for targeted adversarial attacks

Adversarial perturbation map

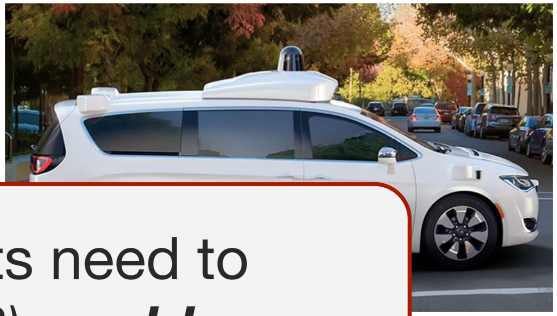
Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, Louis-Philippe Morency, Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models. NeurIPS ViGIL workshop 2018.
https://nips2018vigil.github.io/static/papers/accepted/33.pdf

Multimodal Navigation

Embedded Assistive Agents



The next generation of AI assistants need to ***interact with the real*** *(or virtual?)* ***world.***



Language, Vision and Actions

User: **Go** to the **entrance** of the **lounge area**.

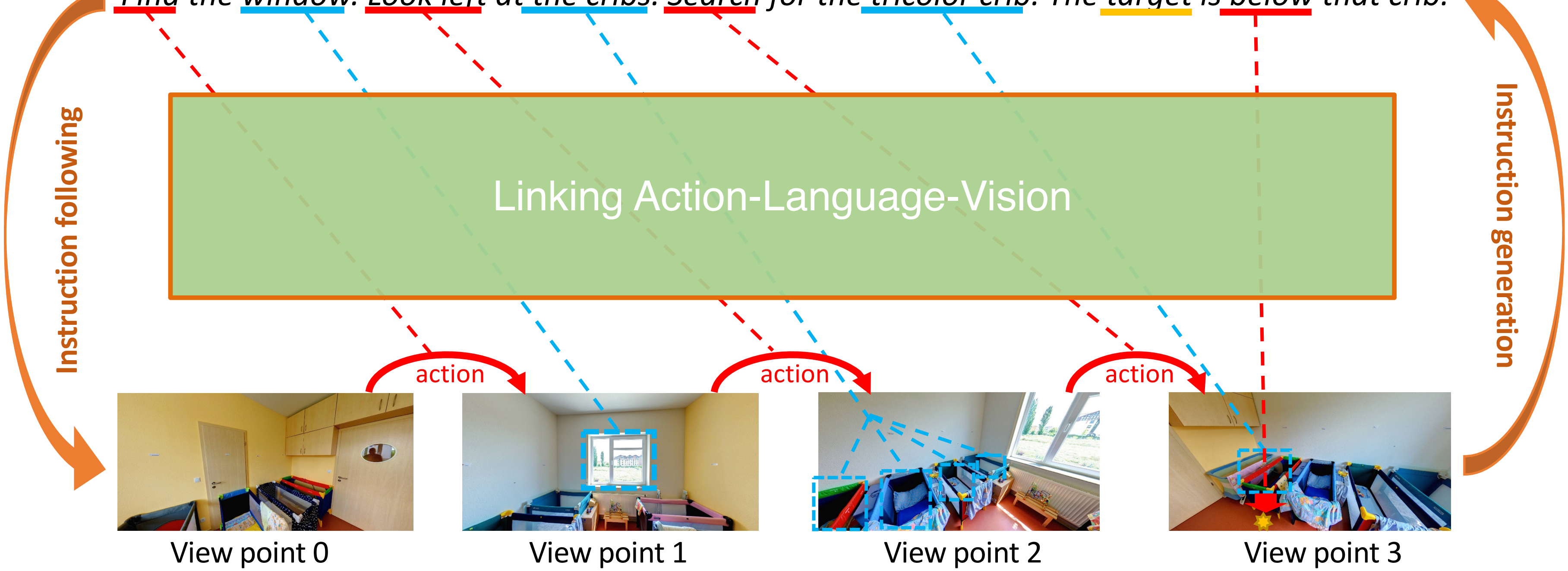
Robot: Sure. I think I'm **there**. What else?

User: **On your right** there will be **a bar**. **On top** of the **counter**, you will see **a box**. **Bring** me **that**.

Many Technical Challenges

Instruction:

Find the window. Look left at the cribs. Search for the tricolor crib. The target is below that crib.



Navigating in a Virtual House

Visually-grounded natural language navigation in real buildings

- Room-2-Room: 21,567 open vocabulary, crowd-sourced navigation instructions

Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

Multiple Step Instructions

Refer360 Dataset

Step1

place the door leading outside to center.

Step2

notice the silver and black coffee pot closest to you on the bar. see the black trash bin on the floor in front of the coffee pot

Step3

waldo is on the face of the trash bin about 1 foot off the floor and also slightly on the brown wood

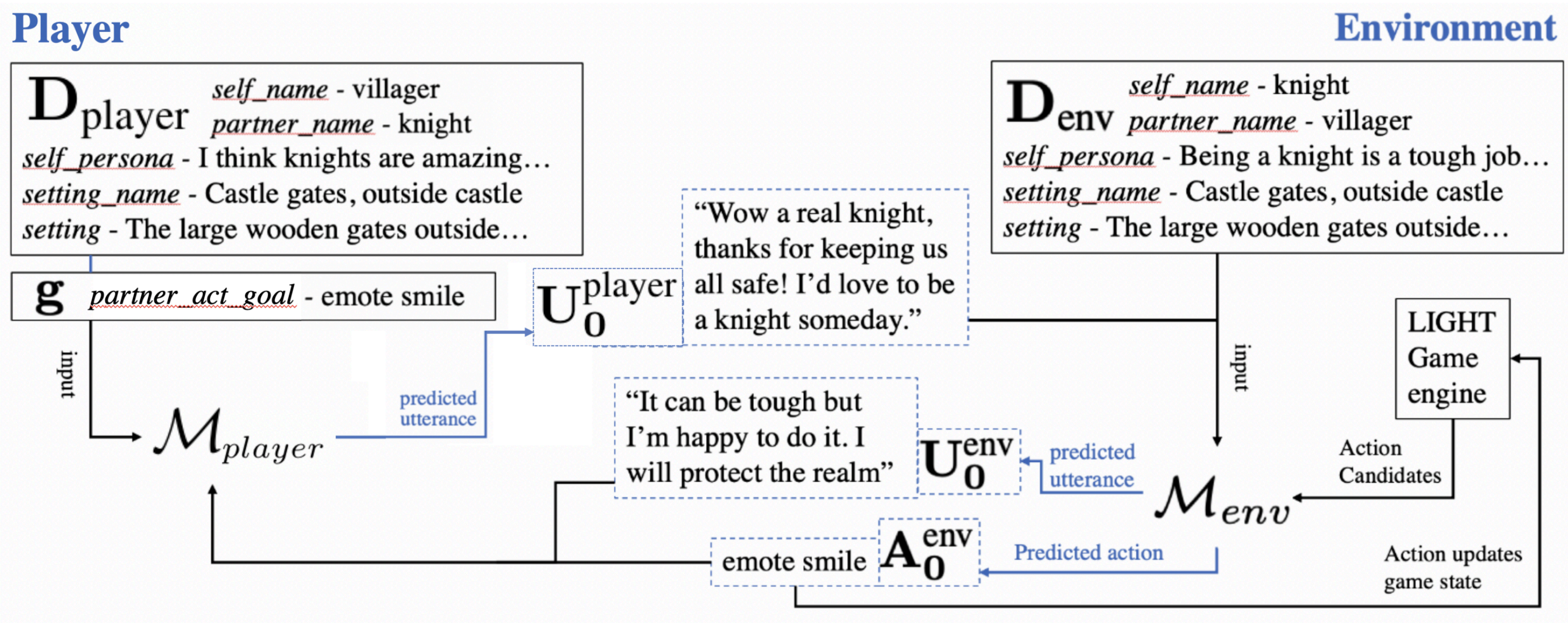
Language meets Games

Interactive game playing RL agents with language input

Heinrich Kuttler and Nantas Nardelli and Alexander H. Miller and Roberta Raileanu and Marco Selvatici and Edward Grefenstette and Tim Rocktaschel, The Nethack Learning Environment. https://arxiv.org/abs/2006.13760

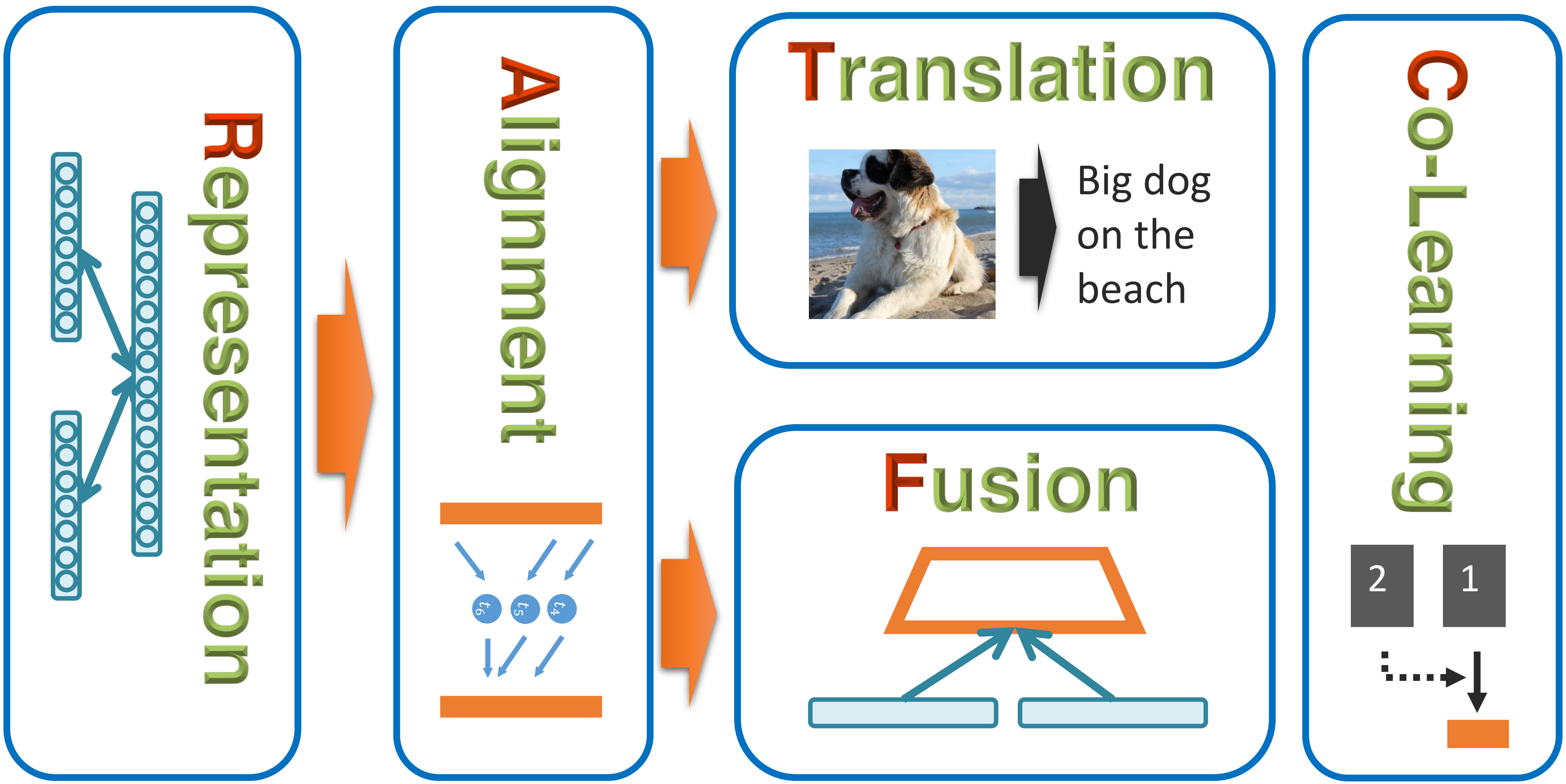
Language meets Games

Agents who must **speak** and **act** in a game



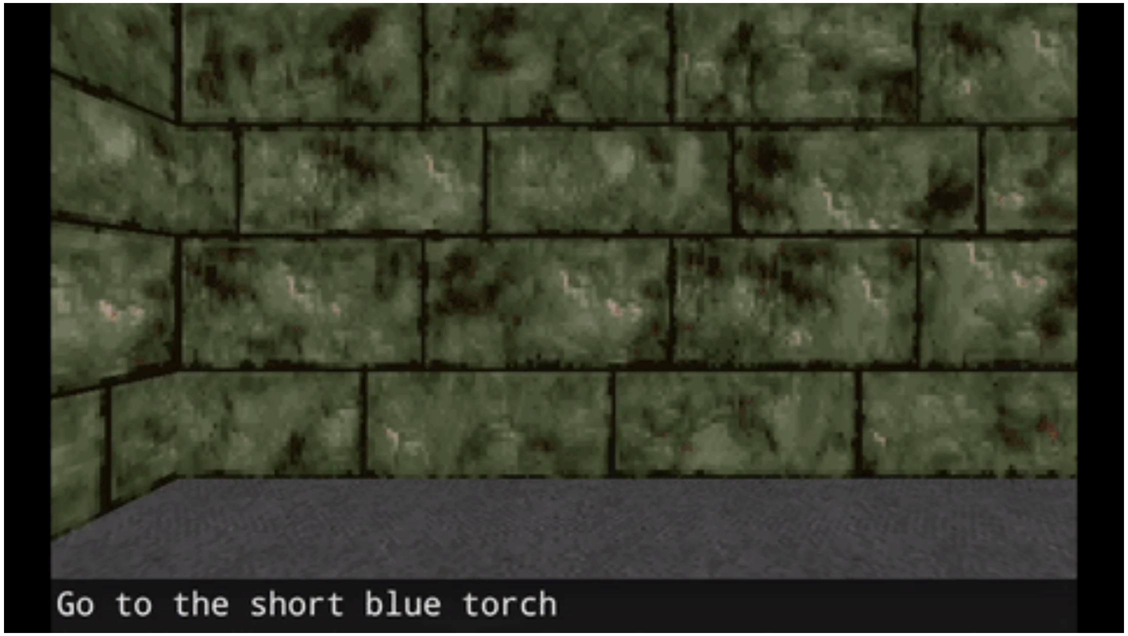
Shrimai Prabhumoye, Margaret Li, Jack Urbanek, Emily Dinan, Douwe Kiela, Jason Weston, Arthur Szlam. I love your chain mail! Making knights smile in a fantasy game world: Open-domain goal-oriented dialogue agents. https://arxiv.org/abs/2002.02878

What are the Core Challenges Most Involved in Multimodal Navigation?



Project Example: Instruction Following

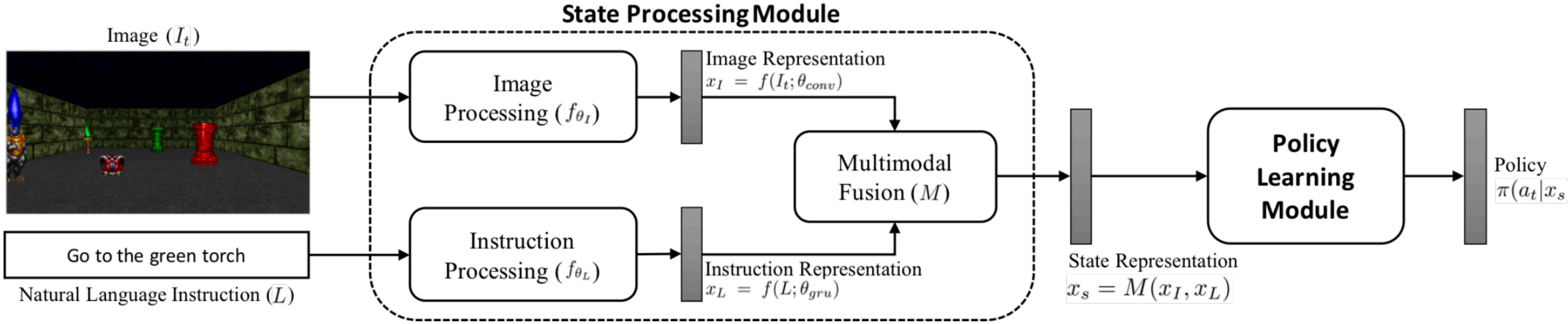
Research task: Task-Oriented Language Grounding in an Environment
Datasets: ViZDoom, based on the Doom video game
Main idea: Build a model that comprehends natural language instructions, grounds the entities and relations to the environment, and execute the instruction.



Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, Ruslan Salakhutdinov, Gated-Attention Architectures for Task-Oriented Language Grounding. AAAI 2018 https://arxiv.org/abs/1706.07230

Project Example: Instruction Following

Solution: Gated attention architecture to attend to instruction and states

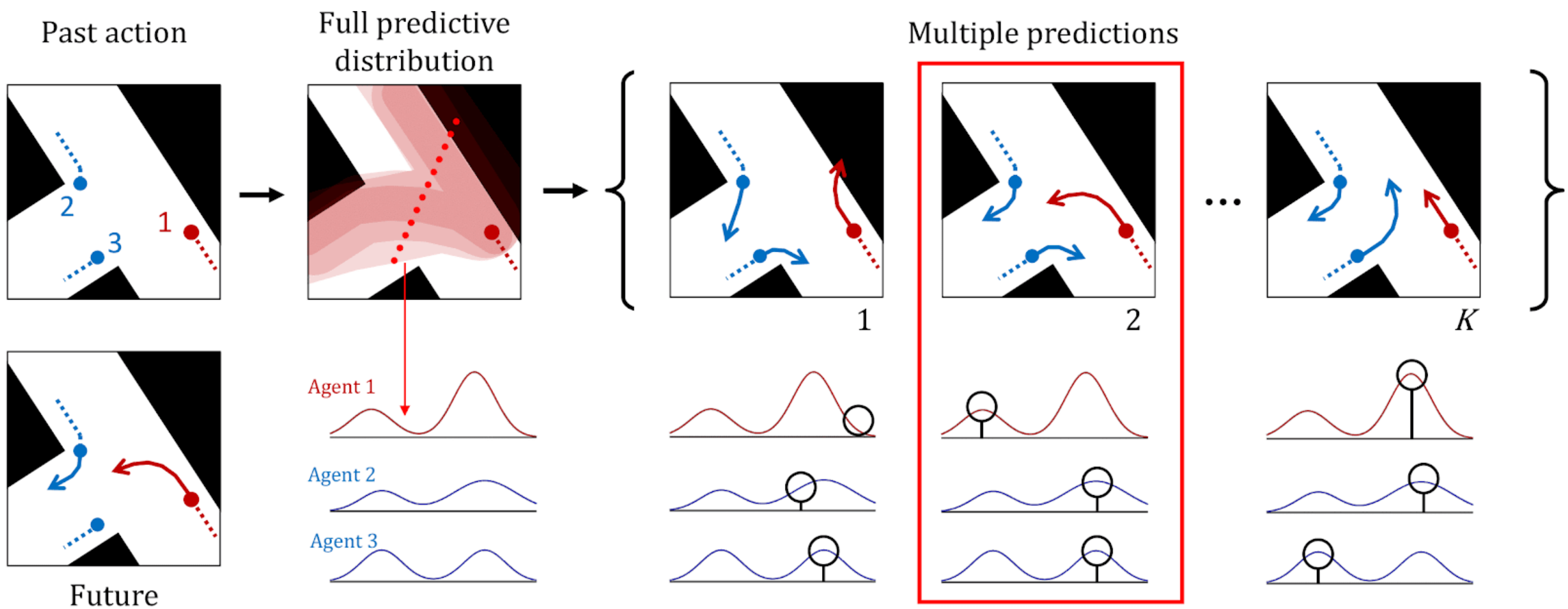


Hypothesis: Gated attention learns to ground and compose attributes in natural language with the image features. e.g. learning grounded representations for 'green' and 'torch'.

Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, Ruslan Salakhutdinov, Gated-Attention Architectures for Task-Oriented Language Grounding. AAAI 2018 https://arxiv.org/abs/1706.07230

Project Example: Multiagent Trajectory Forecasting

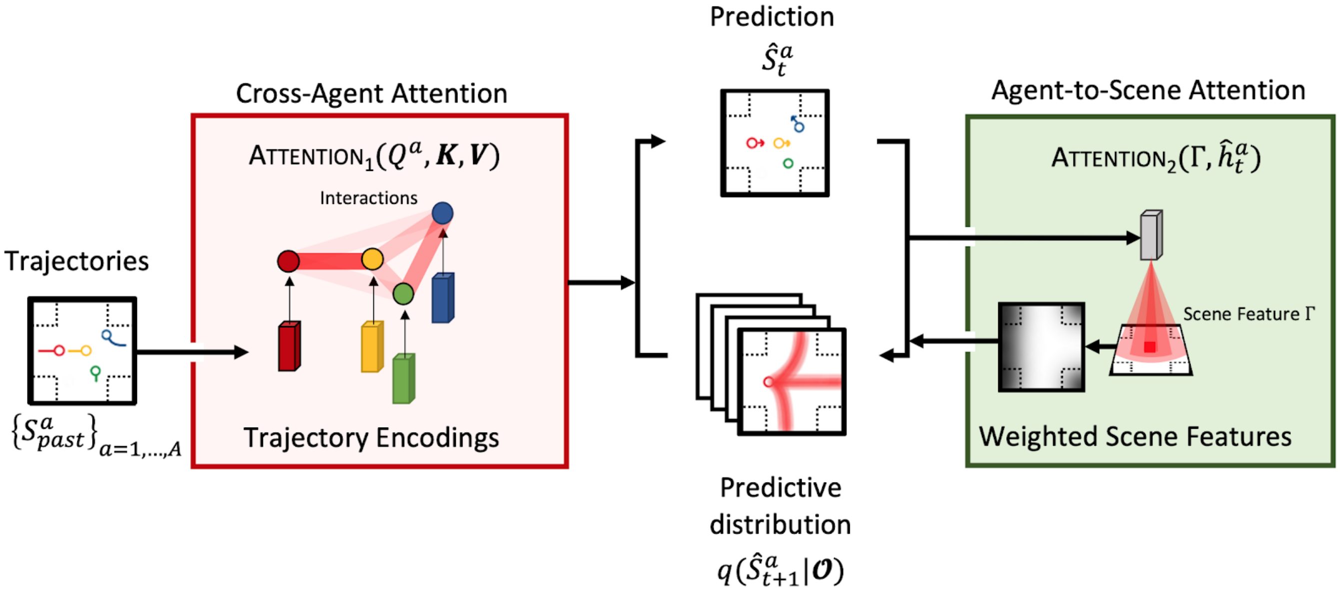
Research task: Multiagent trajectory forecasting for autonomous driving
Datasets: Argoverse and Nuscenes autonomous driving datasets
Main idea: Build a model that understands the environment and multiagent trajectories and predicts a set of multimodal future trajectories for each agent.



Seong Hyeon Park, Gyubok Lee, Manoj Bhat, Jimin Seo, Minseok Kang, Jonathan Francis, Ashwin R. Jadhav, Paul Pu Liang, Louis-Philippe Morency, Diverse and Admissible Trajectory Forecasting through Multimodal Context Understanding. ECCV 2020
https://arxiv.org/abs/1706.07230

Project Example: Multiagent Trajectory Forecasting

Solution: Modeling the environment and multiple agents to learn a distribution of future trajectories for each agent.



Hypothesis: both agent-agent interactions and agent-scene interactions are important!

Seong Hyeon Park, Gyubok Lee, Manoj Bhat, Jimin Seo, Minseok Kang, Jonathan Francis, Ashwin R. Jadhav, Paul Pu Liang, Louis-Philippe Morency, Diverse and Admissible Trajectory Forecasting through Multimodal Context Understanding. ECCV 2020
https://arxiv.org/abs/1706.07230

Project Examples, Advice and Support

Our Latest List of Multimodal Datasets

A. Affect Recognition

AFEW	A1
AVEC	A2
IEMOCAP	A3
POM	A4
MOSI	A5
CMU-MOSEI	A6
TUMBLR	A7
AMHUSE	A8
VGD	A9
Social-IQ	A10
MELD	A11
MUStARD	A12
DEAP	A14
MAHNOB	A15
Continuous LIRIS-ACCEDE	A16
DECAF	A17
ASCERTAIN	A18
AMIGOS	A19

B. Media Description

MSCOCO	B1
MPII	B2
MONTREAL	B3
LSMDC	B4
CHARADES	B5
REFEXP	B6
GUESSWHAT	B7
FLICKR30K	B8
CSI	B9
MVSQ	B10
NeuralWalker	B11
Visual Relation	B12
Visual Genome	B13
Pinterest	B14
Movie Graph	B15
Nocaps	B16
CrossTalk	B17
Refer360	B18

Our Latest List of Multimodal Datasets

C. Multimodal QA

VQA	C1
DAQUAR	C2
COCO-QA	C3
MADLIBS	C4
TEXTBOOK	C5
VISUAL7W	C6
TVQA	C7
VCR	C8
Cornell NLVR	C9
CLEVR	C10
EQA	C11
TextVQA	C12
GQA	C13
CompGuessWhat	C14

D. Multimodal Navigation

Room-2-Room	D1
RERERE	D2
VNLA	D3
nuScenese	D4
Waymo	D5
CARLA	D6
Argoverse	D7
ALFRED	D8

Our Latest List of Multimodal Datasets

E. Multimodal Dialog

VISDIAL	E1
Talk the Walk	E2
Vision-and-Dialog Navigation	E3
CLEVR-Dialog	E4
Fashion Retrieval	E5

F. Event Detection

WHATS-COOKING	F1
TACOS	F2
TACOS-MULTI	F3
YOU-COOK	F4
MED	F5
TITLE-VIDEO-SUMM	F6
MEDIA-EVAL	F7
CRISSMMD	F8

G. Cross-media Retrieval

IKEA	G1
MIRFLICKR	G2
NUS-WIDE	G3
YAHOO-FLICKR	G4
YOUTUBE-8M	G5
YOUTUBE-BOUNDING	G6
YOUTUBE-OPEN	G7
VIST	G8
Recipe1M+	G9
VATEX	G10

… and please let us know (via Piazza) when you find more!

More Project Examples

See the course website:

https://cmu-multicomp-lab.github.io/mmml-course/fall2020/projects/

cmu-multicomp-lab.github.io/mmml-course/fall2020/projects/

11-777 MMML   home schedule syllabus projects

Examples of Previous Project Reports

Project reports from student teams who participated in previous editions of the MMML course

We list here only project reports that were publicly released by students. It should be noted that some of these links are for the follow-up submissions to conferences, after some revisions of the original project reports.

Fall 2019:

1. Seong Hyeon Park, Gyubok Lee, Manoj Bhat, Jimin Seo, Minseok Kang, Jonathan Francis, Ashwin R. Jadhav, Paul Pu Liang, Louis-Philippe Morency. Diverse and Admissible Trajectory Prediction through Multimodal Context Understanding. ECCV 2020

Fall 2018:

1. Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, Louis-Philippe Morency. Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models. NeurIPS ViGIL workshop 2018
2. Yash Patel, Lluis Gomez, Marçal Rusiñol, Dimosthenis Karatzas, C.V. Jawahar. Self-Supervised Visual Representations for Cross-Modal Retrieval

Fall 2017:

1. Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018

Some Advice About Multimodal Research

- Think more about the research problems, and less about the datasets themselves
  - Aim for generalizable models across several datasets
  - Aim for models inspired by existing research e.g. psychology
- Some areas to consider beyond performance:
  - Robustness to missing/noisy modalities, adversarial attacks
  - Studying social biases and creating fairer models
  - Interpretable models
  - Faster models for training/storage/inference
- Theoretical projects are welcome too – make sure there are also experiments to validate theory

Some Advice About Multimodal Datasets

- If you are used to deal with text or speech
  - Space will become an issue working with image/video data
  - Some datasets are in 100s of GB (compressed)
- Memory for processing it will become an issue as well
  - Won't be able to store it all in memory
- Time to extract features and train algorithms will also become an issue
- Plan accordingly!
  - Sometimes tricky to experiment on a laptop (might need to do it on a subset of data)

Available Tools

- Use available tools in your research groups
  - Or pair up with someone that has access to them
- Find some GPUs!
- We will be getting AWS credit for some extra computational power
- Google Cloud Platform credit as well

Google Cloud Platform

amazon web services™

Upcoming Course Assignments

Project preferences (deadline Tuesday 9/8 at 8pm ET)

- Let us know about your project preferences, including datasets, research topics and potential teammates
  - See instructions on Piazza
- We will reserve a moment for discussions on Thursday 9/10 to help you with finding project teammates

Reading Assignment (Summaries due Friday 9/11 at 8pm ET)

- We created the study groups in Piazza.
  - End of the discussion period: Monday 9/14 at 8pm ET

Lecture Highlights (for both lectures next week)

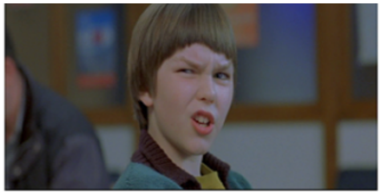
- Starting next week, you need to post your lecture highlights following each course lecture. See Piazza for detailed instructions.

END
of Today’s Lecture

Appendix: List of Multimodal datasets

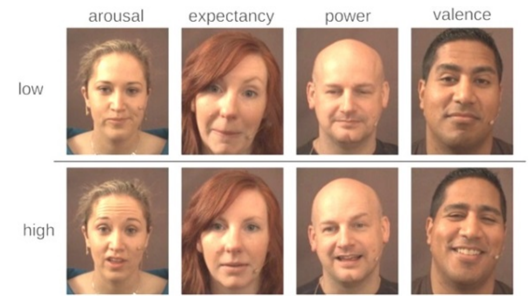
Affect recognition dataset 1 (A1)

- AFEW – Acted Facial Expressions in the Wild (part of EmotiW Challenge)
- Audio-Visual emotion labels – acted emotion clips from movies
  - 1400 video sequences of about 330 subjects
- Labelled for six basic emotions + neutral
- Movies are known, can extract the subtitles/script of the scenes
- Part of EmotiW challenge



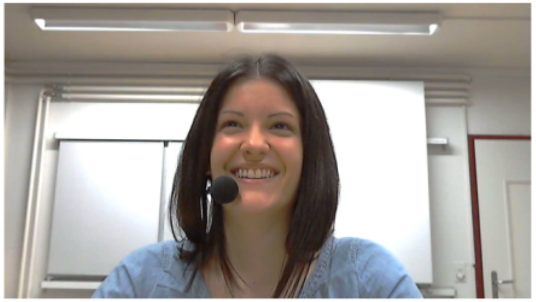
Affect recognition dataset 2 (A2)

- Three AVEC challenge datasets 2011/2012, 2013/2014, 2015, 2016, 2017, 2018
- Audio-Visual emotion recognition
- Labeled for dimensional emotion (per frame)
- 2011/2012 has transcripts
- 2013/2014/2016 also includes depression labels per subject
- 2013/2014 reading specific text in a subset of videos
- 2015/2016 includes physiological data
- 2017/2018 includes depression/bipolar



AVEC 2011/2012

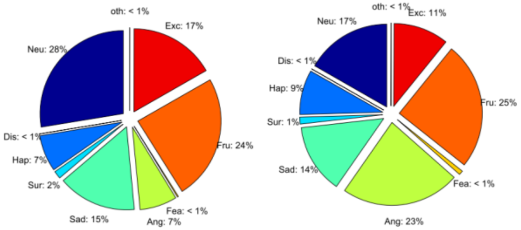
AVEC 2013/2014



AVEC 2015/2016

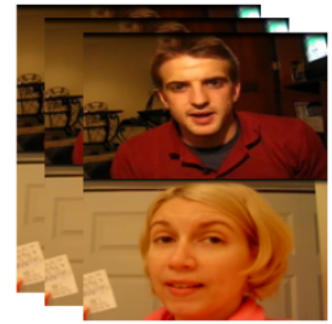
Affect recognition dataset 3 (A3)

- The Interactive Emotional Dyadic Motion Capture (IEMOCAP)
- 12 hours of data, but only 10 participants
- Video, speech, motion capture of face, text transcriptions
- Dyadic sessions where actors perform improvisations or scripted scenarios
- Categorical labels (6 basic emotions plus excitement, frustration) as well as dimensional labels (valence, activation and dominance)
- Focus is on speech

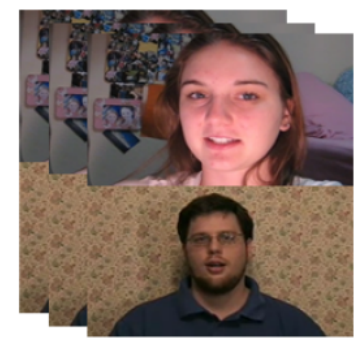


Affect recognition dataset 4 (A4)

- Persuasive Opinion Multimedia (POM)
- 1,000 online movie review videos
- A number of speaker traits/attributes labeled – confidence, credibility, passion, persuasion, big 5…
- Video, audio and text
- Good quality audio and video recordings



Positive opinions
(5-star ratings)



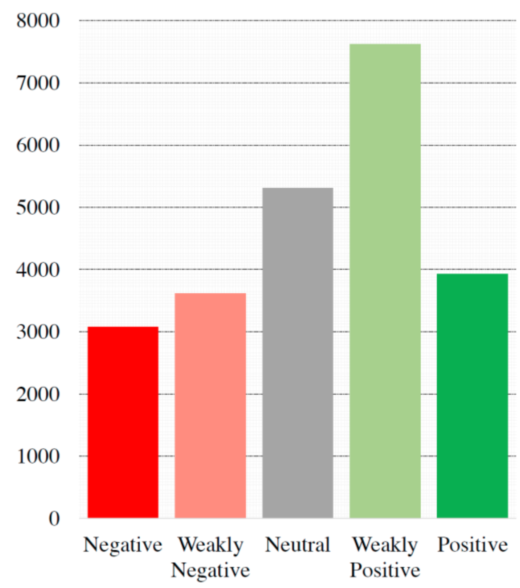
Negative opinions
(1- or 2-star ratings)

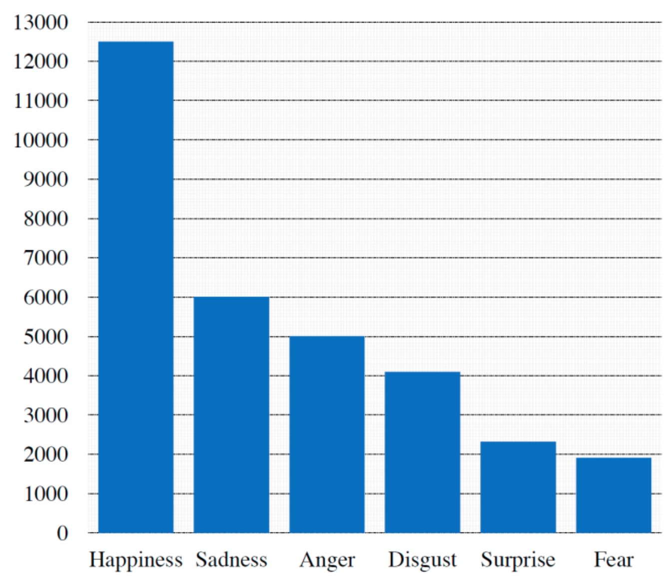
Affect recognition dataset 5 (A5)

- Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos (MOSI)
- 89 speakers with 2199 opinion segments
- Audio-visual data with transcriptions
- Labels for sentiment/opinion
  - Subjective vs objective
  - Positive vs negative

Affect Recognition: CMU-MOSEI (A6)

- Multimodal sentiment and emotion recognition
- CMU-MOSEI : 23,453 annotated video segments from 1,000 distinct speakers and 250 topics





Tumblr Dataset: Sentiment and Emotion Analysis (A7)

- Tumblr Dataset – Tumblr posts with images and emotion word tags.
- 256,897 posts with images.
- Labels obtained from 15 categories of emotion word tags.
- Dataset not directly available but code for collecting the dataset is provided.

Figure 1: Optimistic: "This reminds me that it doesn't matter how bad or sad do you feel, always the sun will come out." Source: travelingpilot [42]

Figure 2: Happy: "Just relax with this amazing view (at McWay Falls)" Source: fordosjulius [37]

AMHUSE Dataset: Multimodal Humor Sensing (A8)

- AMHUSE – Multimodal humor sensing.
- Include various modalities:
  - Video from RGB-d camera, **but no audio/language**
  - Sensory data: blood volume pulse, electrodermal activity, etc.
- Time series of 36 recipients during 4 different stimuli.
- Continuous annotations of arousal, dominance through out each time series. Case-level annotation of level of pleasure is also available.

Video Game Dataset: Multimodal Game Rating (A9)

- VGD – Video Game Dataset, game rating based on text and trailer screenshots.
- 1,950 game trailers.
- Labelled for score ranges of the game, based on online critics.

Super Mario Odyssey

Sample Game Trailer Frames

+

Game Summary

"Mario embarks on a new journey through unknown worlds, running and jumping through huge 3D worlds in the first sandbox-style Mario game since Super Mario 64 and Super Mario Sunshine."

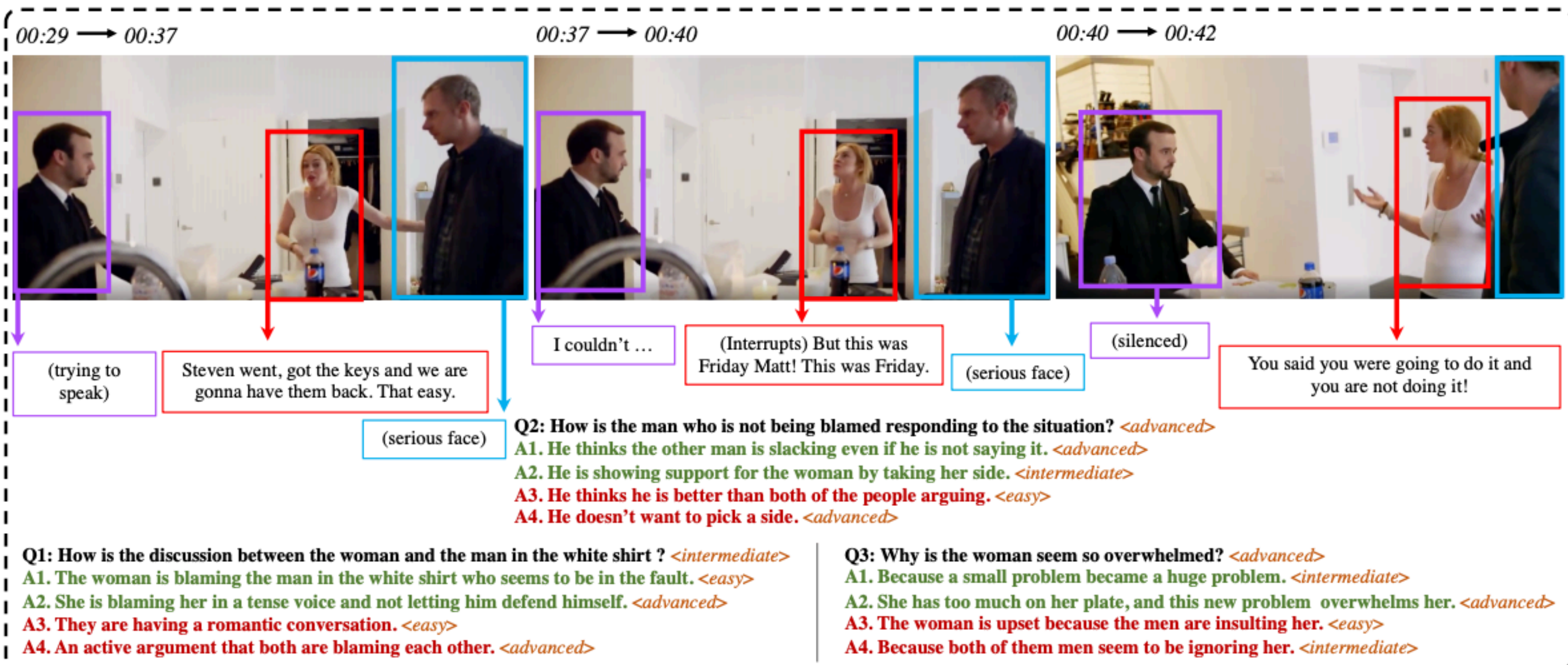
⇩

Predicted Score Class

90-100

Social-IQ (A10)

- Social-IQ: 1.2k videos, 7.5k questions, 50k answers
- Questions and answers centered around social behaviors

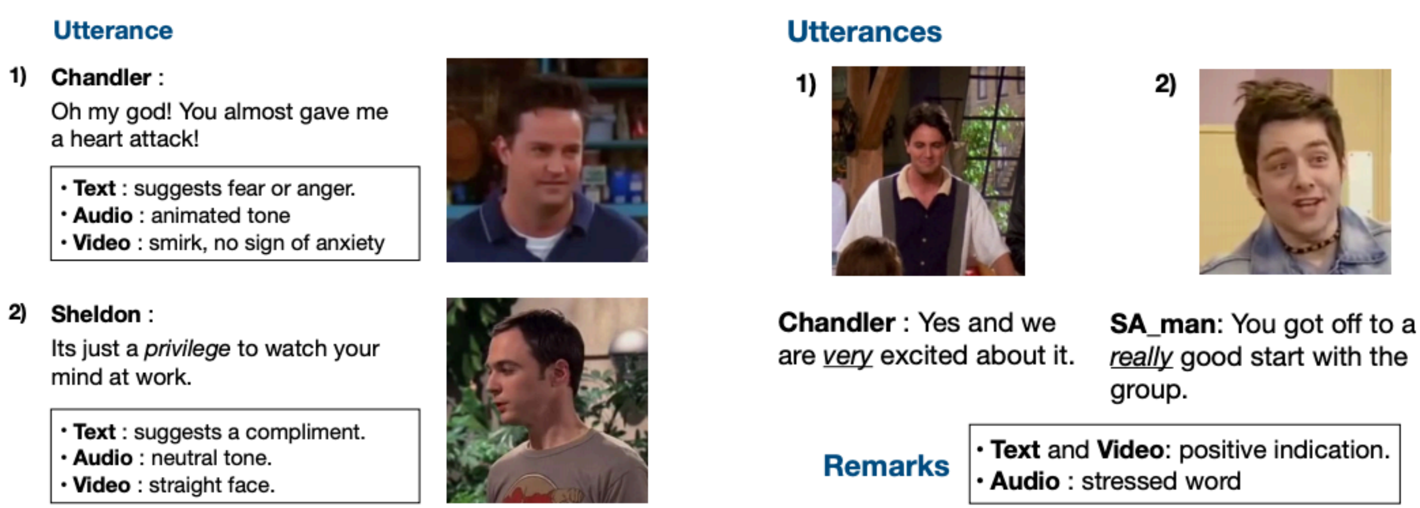


MELD (A11)

- MELD: Multi-party dataset for emotion recognition in conversations

MUStARD (A12)

- MUStARD: Multimodal sarcasm dataset



More affect recognition datasets (A13-A18)

- DEAP (A13)
  - Emotion analysis using EEG, physiological, and video signals
- MAHNOB (A14)
  - Laughter database
- Continuous LIRIS-ACCEDE (A15)
  - Induced valence and arousal self-assessments for 30 movies
- DECAF (A16)
  - MEG + near-infra-red facial videos + ECG + … signals
- ASCERTAIN (A17)
  - Personality and affect recognition from physiological sensors
- AMIGOS (A18)
  - Affect, personality, and mood from neuro-physiological signals
- EMOTIC (A19)
  - Context Based Emotion Recognition

Media description dataset 1 – MS COCO (B1)

- Microsoft Common Objects in COntext (MS COCO)
- 120000 images
- Each image is accompanied with five free form sentences describing it (at least 8 words)
- Sentences collected using crowdsourcing (Mechanical Turk)
- Also contains object detections, boundaries and keypoints

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Media description dataset 2 - Video captioning (B2&B3)

- MPII Movie Description dataset (B2)
  - A Dataset for Movie Description
- Montréal Video Annotation dataset (B3)
  - Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research

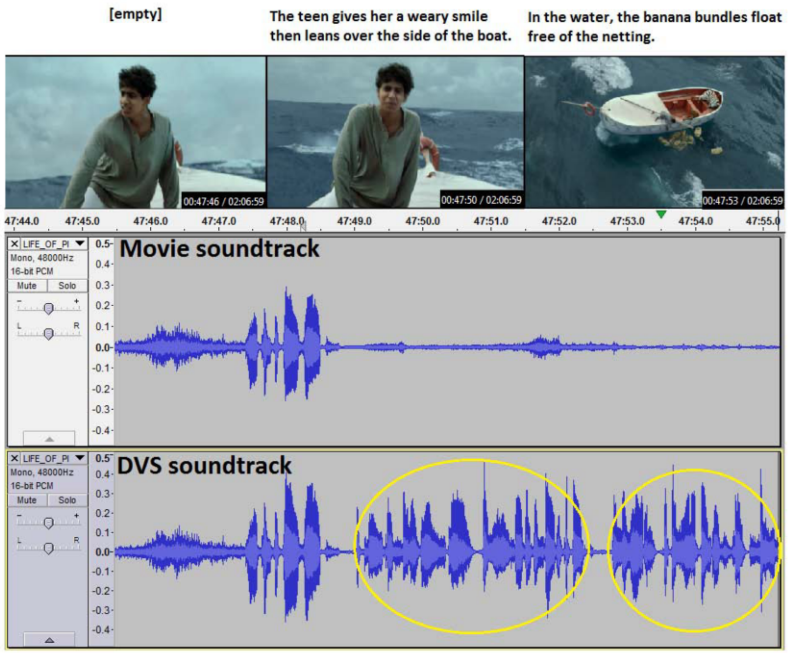
AD: Abby gets in the basket.

Mike leans over and sees how high they are.

Abby clasps her hands around his face and kisses him passionately.

Media description dataset 2 - Video captioning (B2&B3)

- Both based on audio descriptions for the blind (Descriptive Video Service - DVS tracks)
- MPII – 70k clips (~4s) with corresponding sentences from 94 movies
- Montréal – 50k clips (~6s) with corresponding sentences from 92 movies
- Not always well aligned
- Quite noisy labels
- Single caption per clip



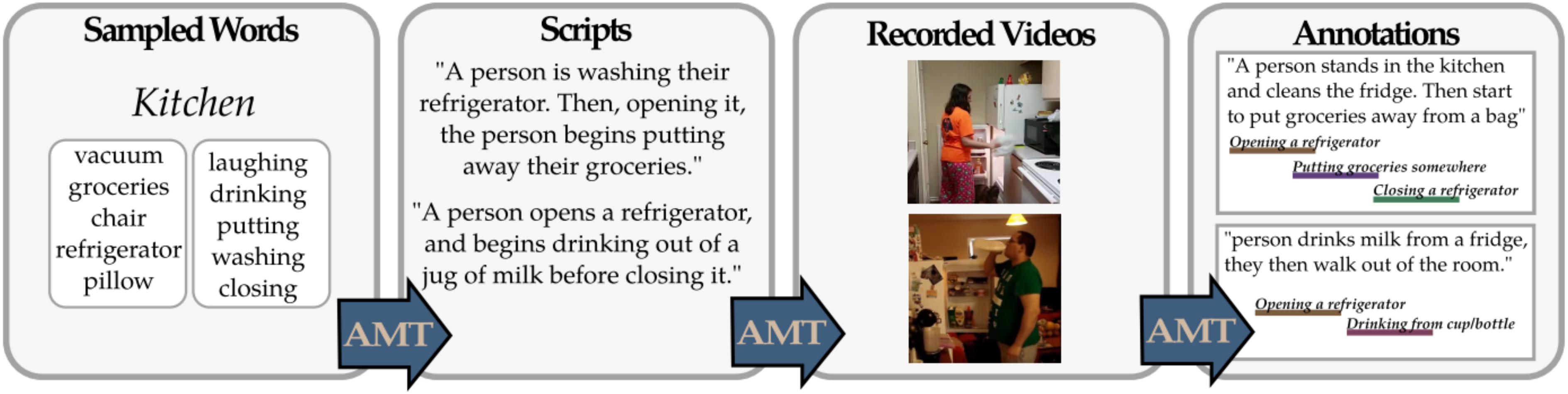
Media description dataset 2 - Video captioning (B4)

- Large Scale Movie Description and Understanding Challenge (LSMDC) hosted at ECCV 2016 and ICCV 2015
- Combines both of the datasets and provides three challenges
  - Movie description
  - Movie annotation and Retrieval
  - Movie Fill-in-the-blank
- Nice challenge, but beware
  - Need a lot of computational power
  - Processing will take space and time

QUERY: answering phone

Charades Dataset – video description dataset (B5)

- http://allenai.org/plato/charades/
- 9848 videos of daily indoors activities
- 267 different users
- Recording videos at home
- Home quality videos



Media Description – Referring Expression datasets (B6)

- Referring Expressions:
  - Generation (Bounding Box to Text) and Comprehension (Text to Bounding Box)
  - Generate / Comprehend a noun phrase which identifies a particular object in an image
  - Many datasets!
    - RefClef
    - RefCOCO (+, g)
    - GRef

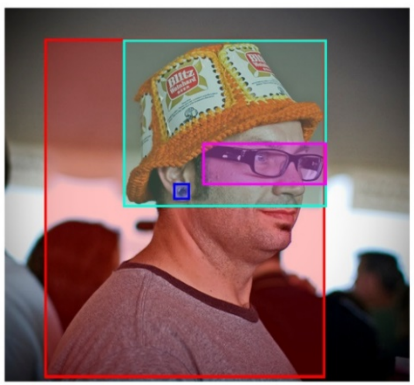
Media Description - Referring Expression datasets (B7)

- GuessWhat?!
  - Cooperative two-player guessing game for language grounding
  - Locate an unknown object in a rich image scene by asking a sequence of questions
  - 821,889 questions+answers
  - 66,537 images and 134,073 objects

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

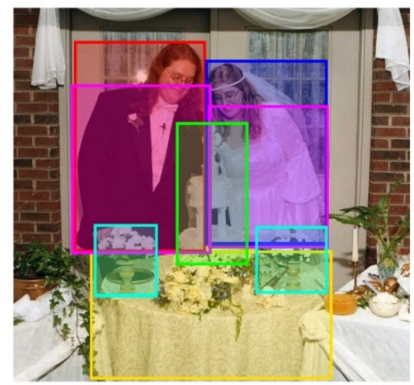
Media Description - other datasets (B8)

- Flickr30k Entities
  - Region-to-Phrase Correspondences for Richer Image-to-Sentence Models
  - 158k captions
  - 244k coreference chains
  - 276k manually annotated bounding boxes



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .



A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

CSI Corpus (B9)

- CSI-Corpus: 39 videos from the U.S. TV show “Crime Scene Investigation Las Vegas”
- Data: Sequence of inputs comprising information from different modalities such as text, video, or audio.The task is to predict for each input whether the perpetrator is mentioned or not.

Peter Berglund:
You're still going to have to convince a jury that I killed two strangers for no reason.

Grissom doesn't look worried.
He takes his gloves off and puts them on the table.

Grissom:
You ever been to the theater Peter?
There 's a play called six degrees of separation.

It 's about how all the people in the world are connected to each other by no more than six people.
All it takes to connect you to the victims is one degree.

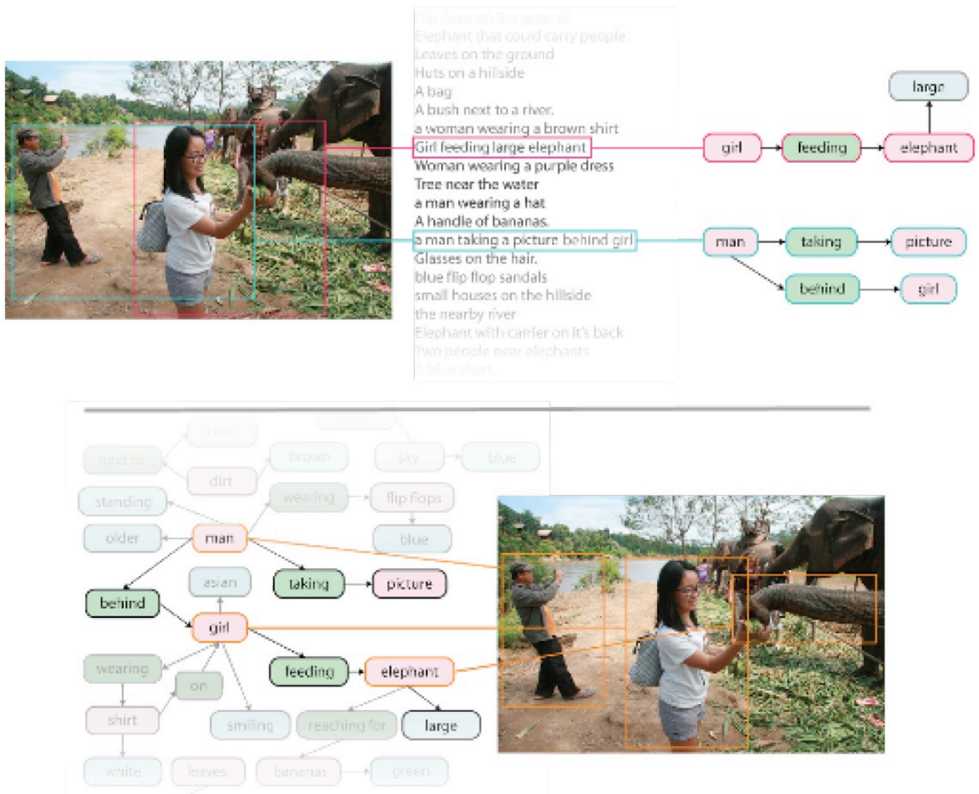
Camera holds on Peter Berglund's worried look.

Other Media Description Datasets (B10-B14)

- MVSO (B10): Multilingual Visual Sentiment Ontology. There are multiple derivatives of this as well
- NeuralWalker (B11): 'Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences’
- Visual Relation dataset (B12): learning relations between objects based on language priors.
- Visual genome (B13) Great resource for many multimodal problems.
- Pinterest (B14): Contains 300 million sentences describing over 40 million 'pins'
- nocaps (B16): novel object captioning at scale
- CrossTask (B17): procedure annotations in videos
- Refer360° (B18): Referring Expression Recognition in 360° Images

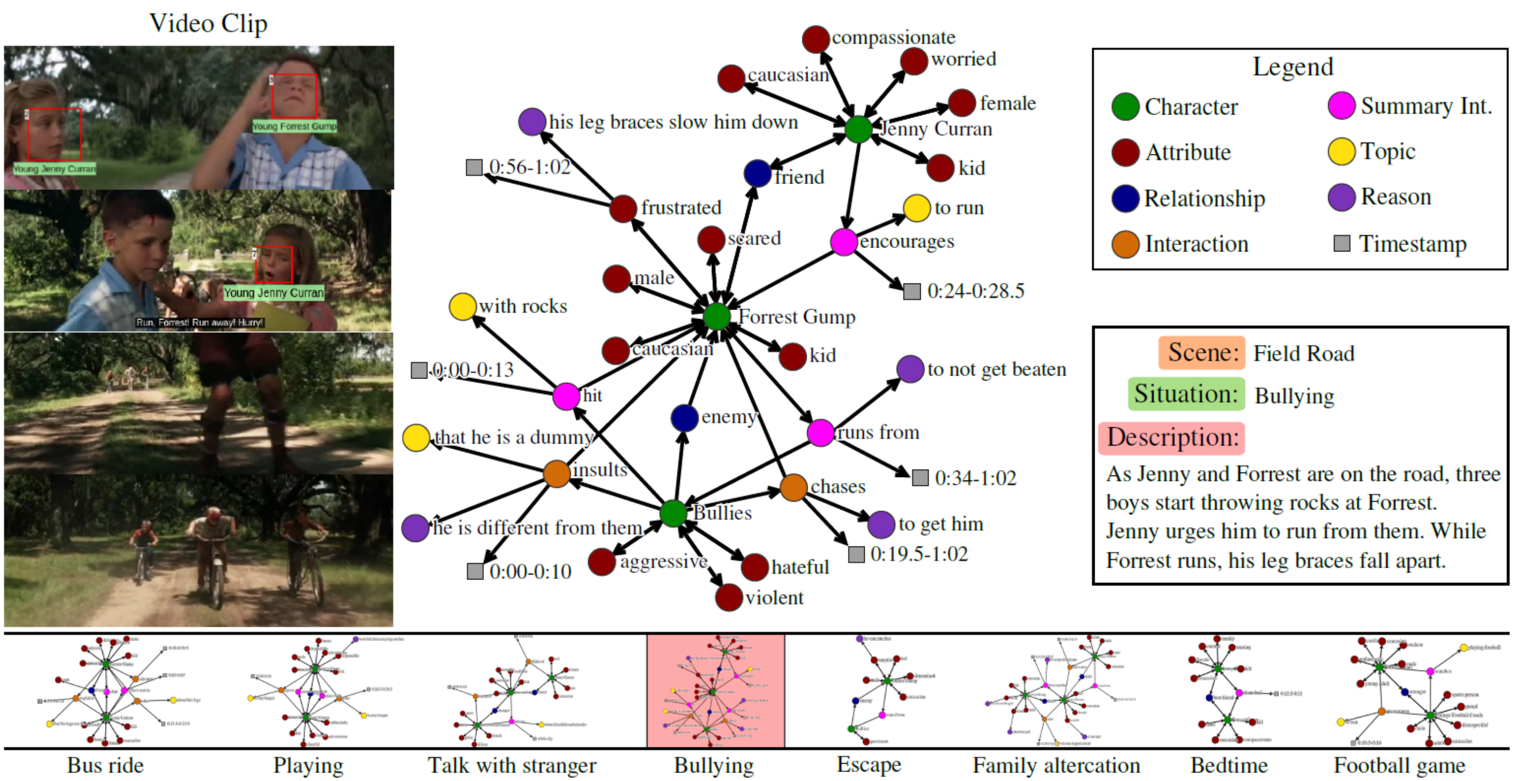
Visual Genome (B13)

- https://visualgenome.org/



MovieGraph dataset (B15)

- http://moviegraphs.cs.toronto.edu/



Media description technical challenges

- What technical problems could be addressed?
  - Translation
  - Representation
  - Alignment
  - Co-training/transfer learning
  - Fusion

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

AD: Abby gets in the basket.

Mike leans over and sees how high they are.

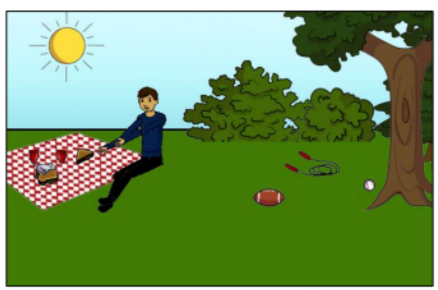
Abby clasps her hands around his face and kisses him passionately.

Multimodal QA dataset 1 – VQA (C1)

- Task - Given an image and a question, answer the question (http://www.visualqa.org/)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

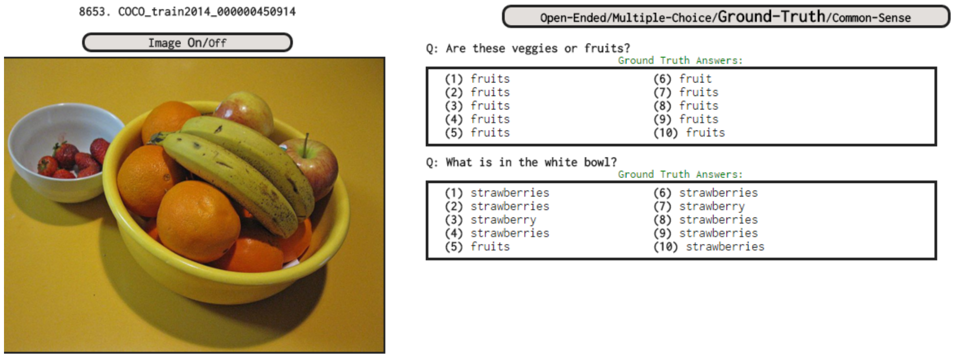


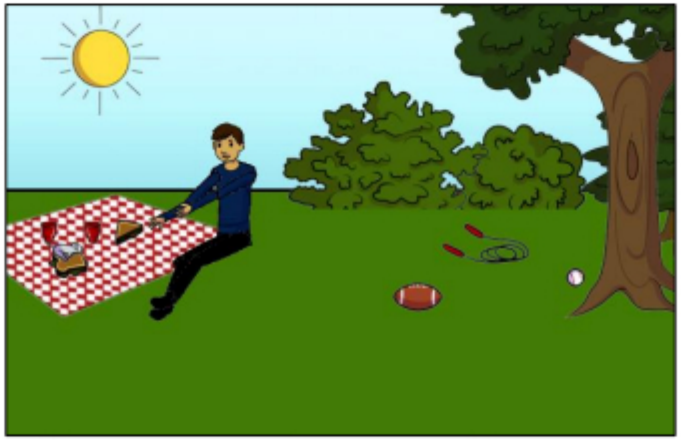
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Multimodal QA dataset 1 – VQA (C1)

- Real images
  - 200k MS COCO images
  - 600k questions
  - 6M answers
  - 1.8M plausible answers
- Abstract images
  - 50k scenes
  - 150k questions
  - 1.5M answers
  - 450k plausible answers





Is this person expecting company?
What is just under the tree?

VQA Challenge 2016 and 2017 (C1)

- Two challenges organized these past two years (link)
- Currently good at yes/no question, not so much free form and counting

	By Answer Type			Overall
	Yes/No	Number	Other	
UC Berkeley & Sony[14]	83.79	38.9	58.64	66.9
Naver Labs[10]	83.78	37.67	54.74	64.89
DLAIT[5]	83.65	39.18	52.62	63.97
snubi-naverlabs[25]	83.64	38.43	51.61	63.4
POSTECH[11]	81.85	38.02	53.12	63.35
Brandeis[3]	82.53	36.54	51.71	62.8
VTComputerVison[19]	80.31	37.87	52.16	62.23
MIL-UT[7]	82.39	36.7	49.76	61.82

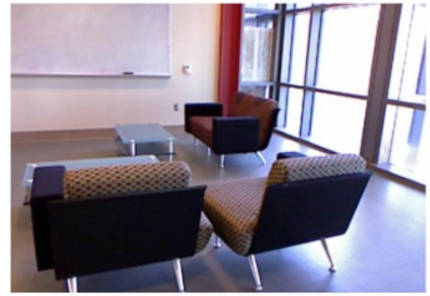
VQA 2.0

- Just guessing without an image lead to ~51% accuracy
  - So the V in VQA “only” adds 14% increase in accuracy
- VQA v2.0 is attempting to address this

Multimodal QA – other VQA datasets

COCOQA
Q: What is the color of the desk?
A: white
Q: What are on the white desk?
A: computers

COCOQA
Q: What is the color of the dresses?
A: purple
Q: What are three women dressed up and on?
A: phones



DAQUAR
Q: What is the object close to the wall?
A: whiteboard
Q: What is the object in front of the sofa?
A: table

DAQUAR
Q: What is the largest object?
A: sofa
Q: How many windows are there?
A: 2

VQA
Q: How many bikes are there?
A: 2
Q: What number is the bus?
A: 48

VQA
Q: How many pickles are on the plate?
A: 1
Q: What is the shape of the plate?
A: round

VQA
Q: What does the sign say?
A: stop
Q: What shape is this sign?
A: octagon

VQA
Q: What type of trees are here?
A: palm
Q: Is the skateboard airborne?
A: yes

Multimodal QA – other VQA datasets (C2&C3)

- DAQUAR (C2)
  - Synthetic QA pairs based on templates
  - 12468 human question-answer pairs
- COCO-QA (C3)
  - Object, Number, Color, Location
  - Training: 78736
  - Test: 38948

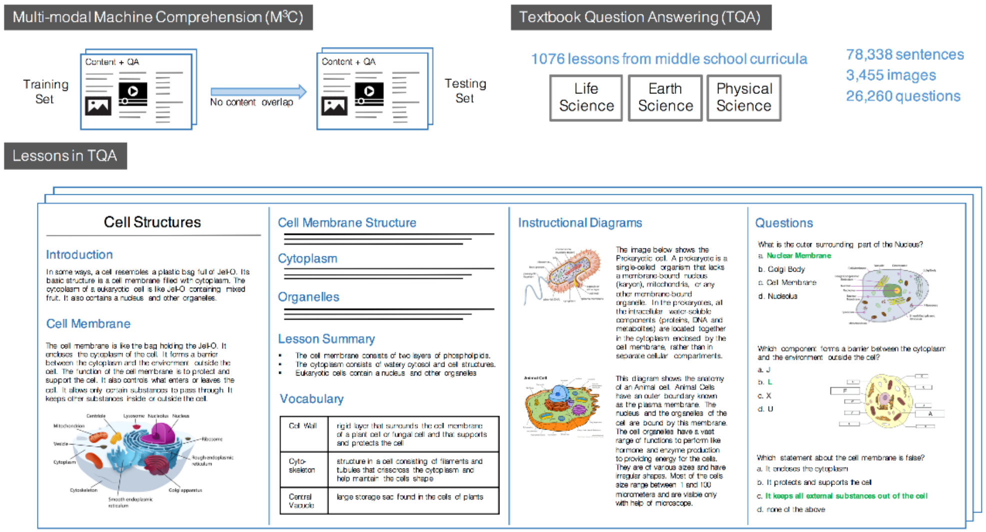
Multimodal QA – other VQA datasets (C4)

- Visual Madlibs
  - Fill in the blank Image Generation and Question Answering
  - 360,001 focused natural language descriptions for 10,738 images
  - collected using automatically produced fill-in-the-blank templates designed to gather targeted descriptions about: people and objects, their appearances, activities, and interactions, as well as inferences about the general scene or its broader context

1. This place is a park.
2. When I look at this picture, I feel competitive.
3. The most interesting aspect of this picture is the guys playing shirtless.
4. One or two seconds before this picture was taken, the person caught the frisbee.
5. One or two seconds after this picture was taken, the guy will throw the frisbee.
6. Person A is wearing blue shorts.
7. Person A is in front of person B.
8. Person A is blocking person B.
9. Person B is a young man wearing an orange hat.
10. Person B is on a grassy field.
11. Person B is holding a frisbee.
12. The frisbee is white and round.
13. The frisbee is in the hand of the man with the orange cap.
14. People could throw the frisbee.
15. The people are playing with the frisbee.

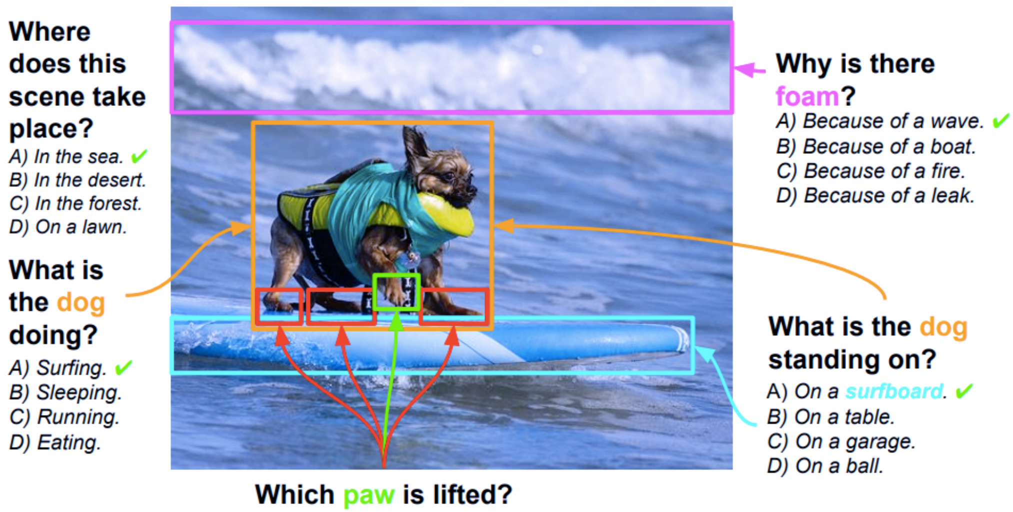
Multimodal QA – other VQA datasets (C5)

- [Textbook Question Answering]
  - Multi-Modal Machine Comprehension
  - Context needed to answer questions provided and composed of both text and images
  - 78338 sentences, 3455 images
  - 26260 questions



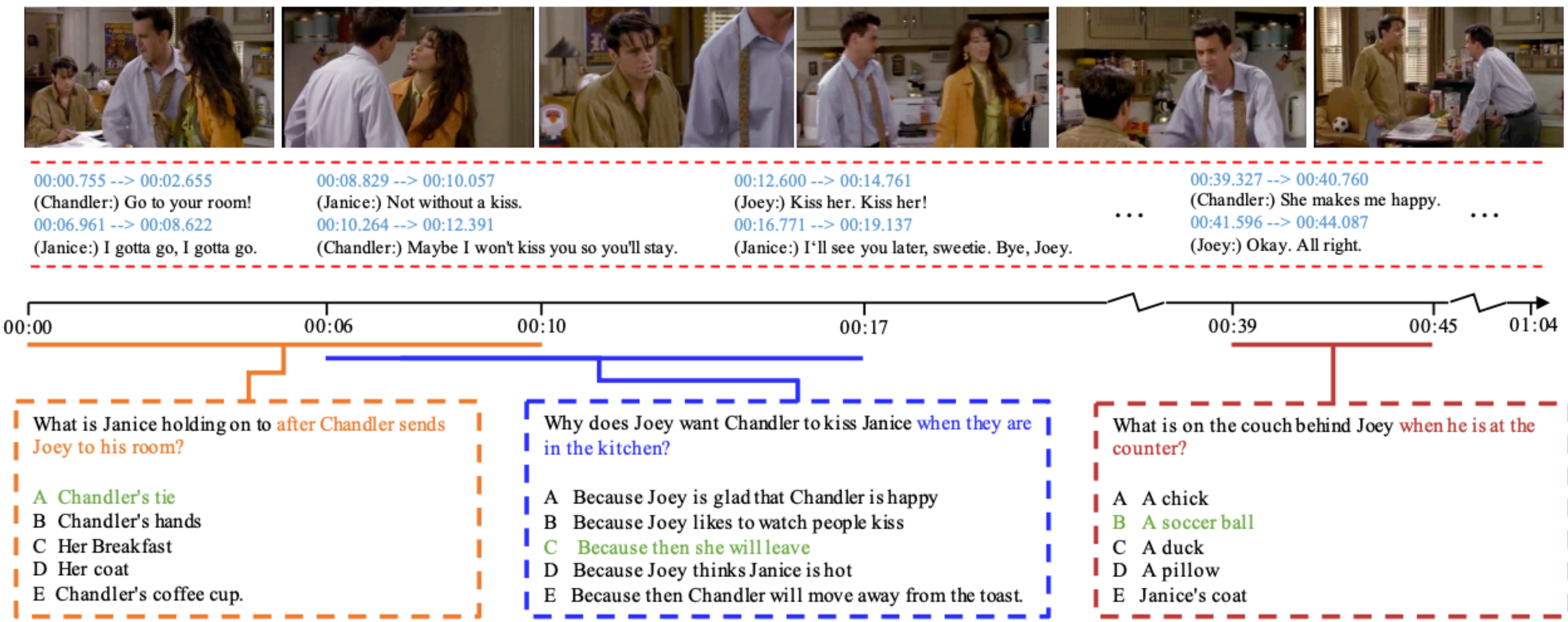
Multimodal QA – other VQA datasets (C6)

- Visual7W
  - Grounded Question Answering in Images
  - 327,939 QA pairs on 47,300 COCO images
  - 1,311,756 multiple-choices, 561,459 object groundings, 36,579 categories
  - what, where, when, who, why, how and which



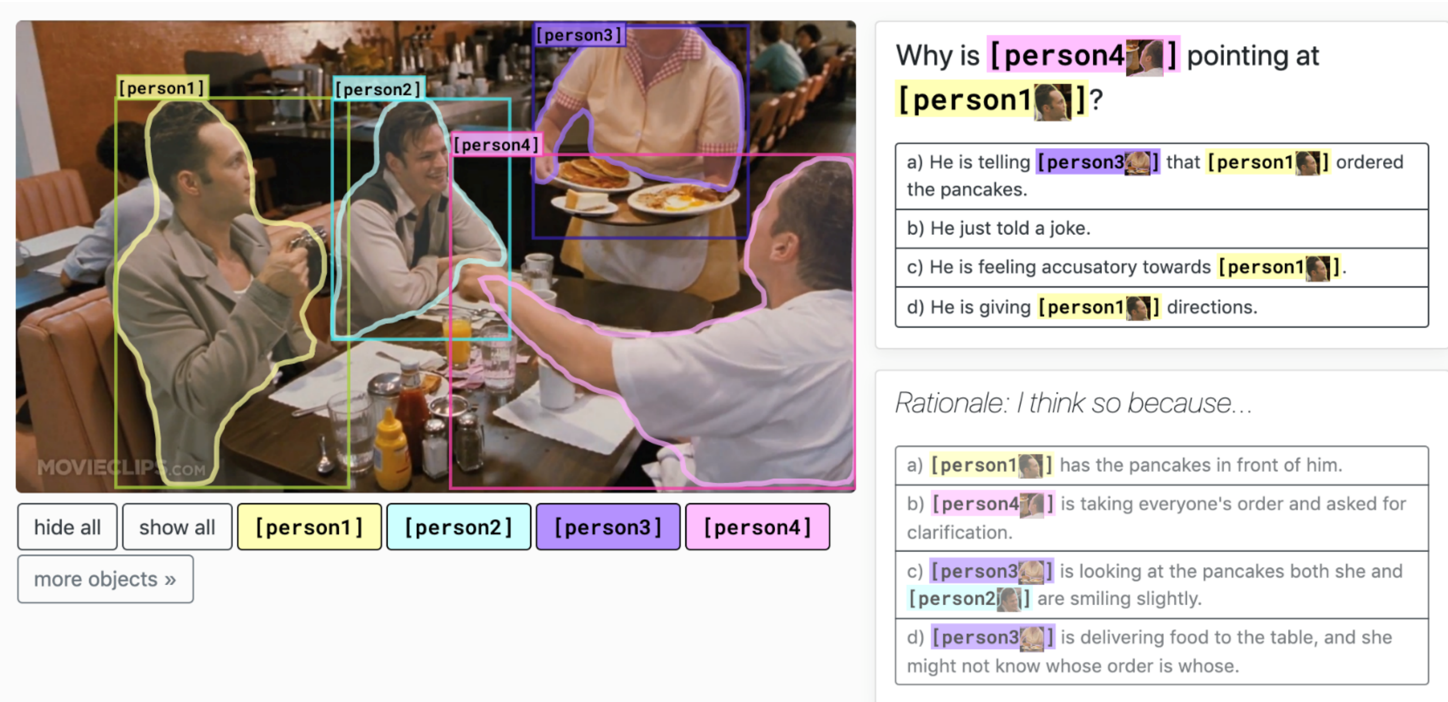
Multimodal QA – other VQA datasets (C7)

- TVQA
  - Video QA dataset based on 6 popular TV shows
  - 152.5K QA pairs from 21.8K clips
  - Compositional questions



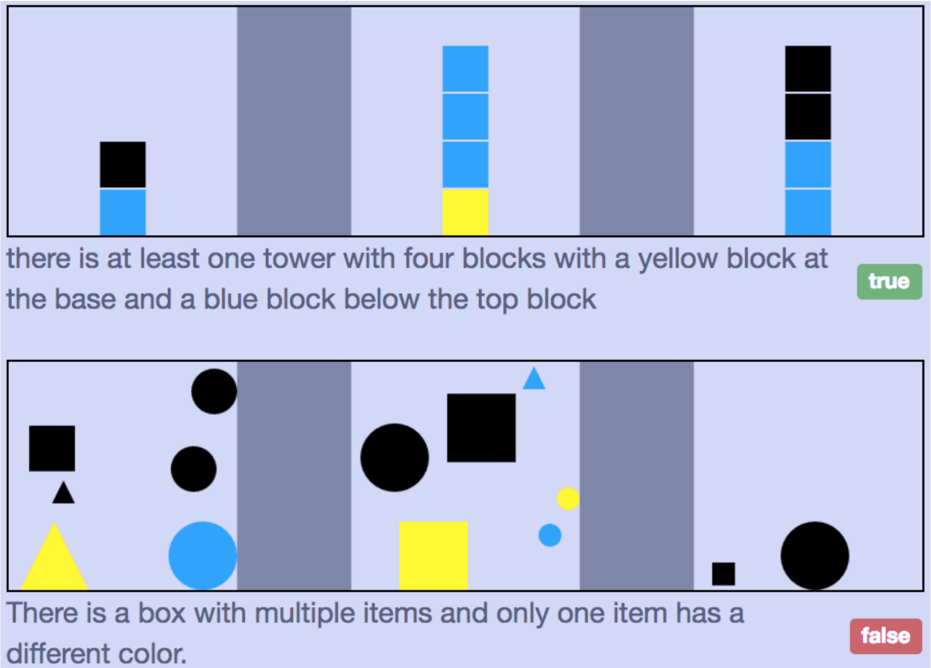
Multimodal QA – Visual Reasoning (C8)

- VCR: Visual Commonsense Reasoning
  - Model must answer challenging visual questions expressed in language
  - And provide a **rationale explaining why its answer is true**.



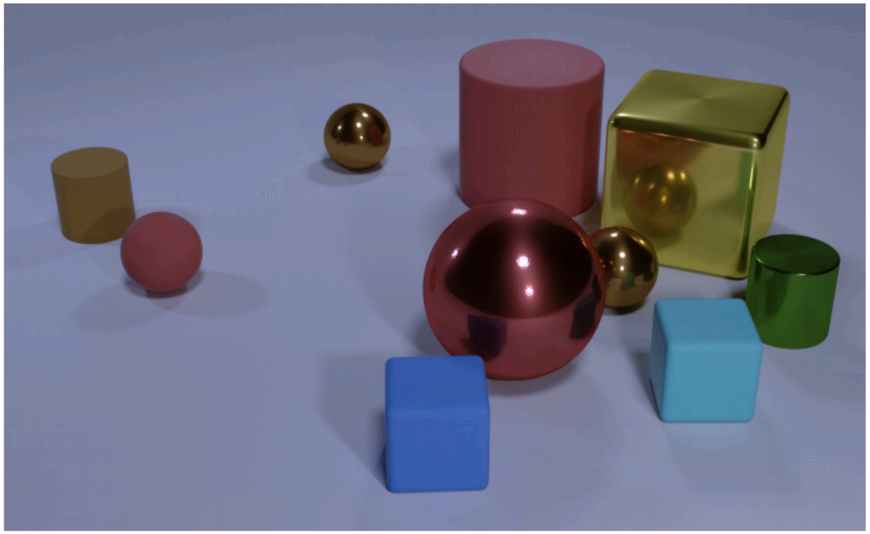
Multimodal QA – Visual Reasoning (C9)

- [Cornell NLVR]
  - 92,244 pairs of natural language statements grounded in synthetic images
  - Determine whether a sentence is true or false about an image



Multimodal QA – Visual Reasoning (C10)

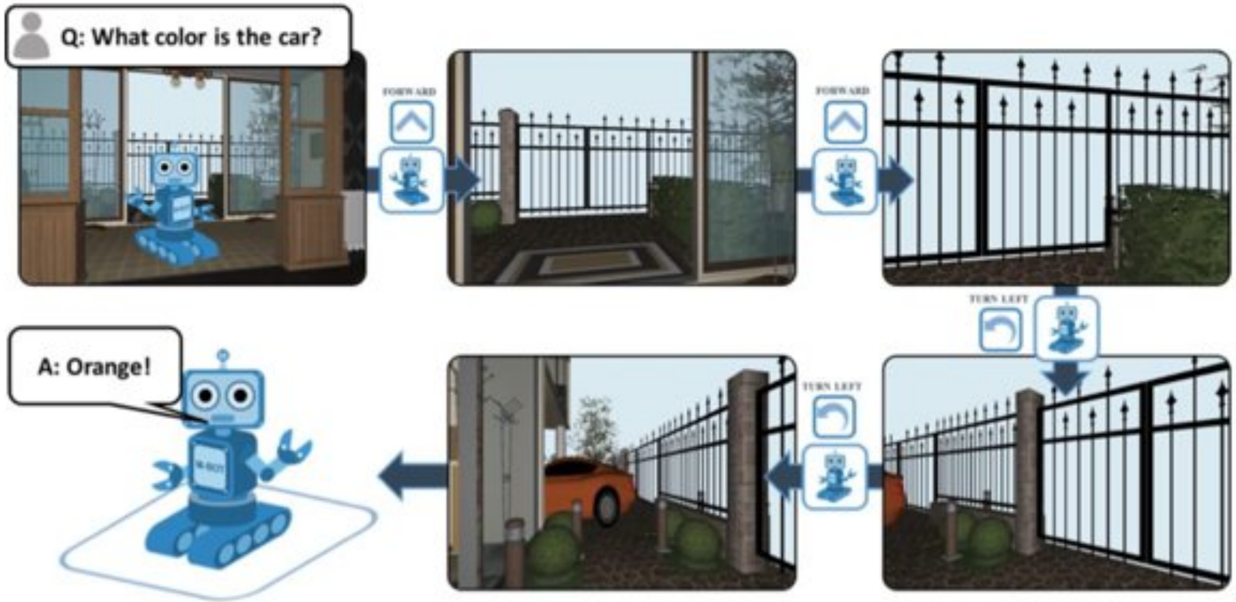
- CLEVR
  - A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning
  - Tests a range of different specific visual reasoning abilities
  - Training set: 70,000 images and 699,989 questions
  - Validation set: 15,000 images and 149,991 questions
  - Test set: 15,000 images and 14,988 questions



Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many objects are either small cylinders or metal things?

Embodied Question Answering (C11)

- An agent is spawned at a random location in a 3D environment and asked a question
- EQA v1.0: 9,000 questions from 774 environments



TextVQA (C12), GQA (C13), CompGuessWhat (C14)

- TextVQA requires models to read and reason about text in images to answer questions about them. Specifically, models need to incorporate a new modality of text present in the images and reason over it to answer TextVQA questions.

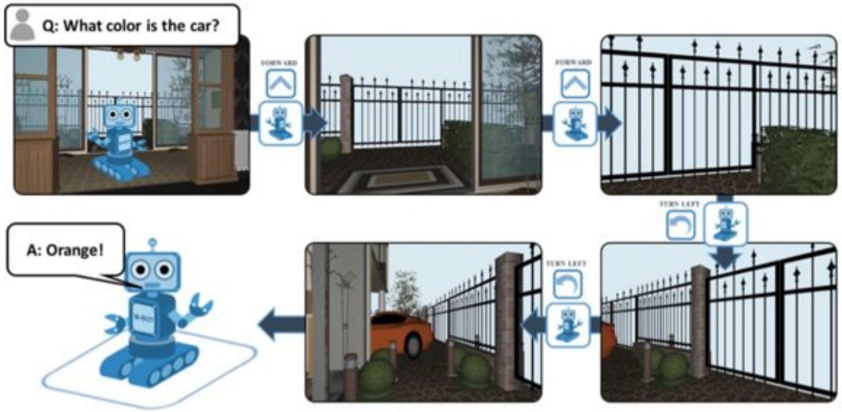
- GQA Real-World Visual Reasoning and Compositional Question Answering. A new dataset for real-world visual reasoning and compositional question answering, seeking to address key shortcomings of previous VQA datasets.
- CompGuessWhat Framework for evaluating the quality of learned neural representations, in particular concerning attribute grounding.

Multimodal QA technical challenges

- What technical problems could be addressed?
  - Translation
  - Representation
  - Alignment
  - Fusion
  - Co-training/transfer learning

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Room-2-Room Navigation with NL instructions (D1)

- Visually grounded natural language navigation in real buildings
- Room-2-Room: 21,567 open vocabulary, crowd-sourced navigation instructions

Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

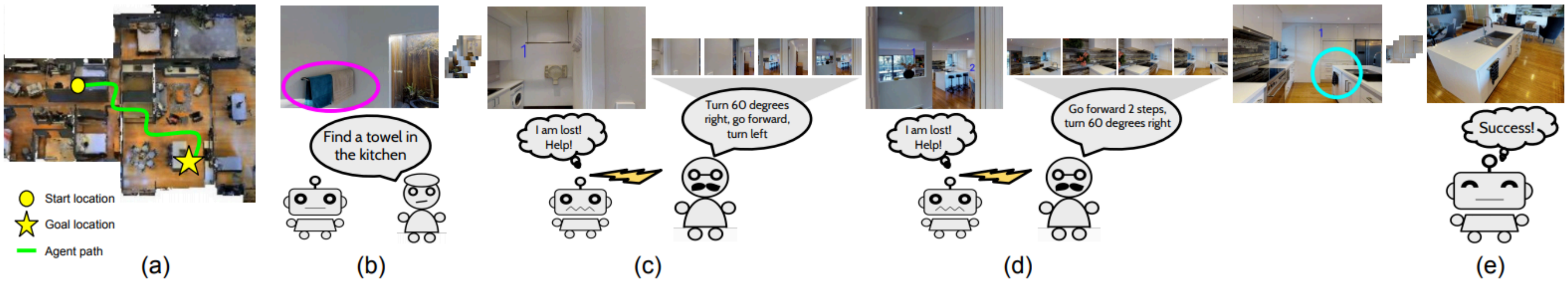
Multimodal Navigation: RERERE (D2)

- Remote embodied referring expressions in real indoor environments

Instruction: Go to the stairs on level one and bring me the bottom picture that is next to the top of the stairs.

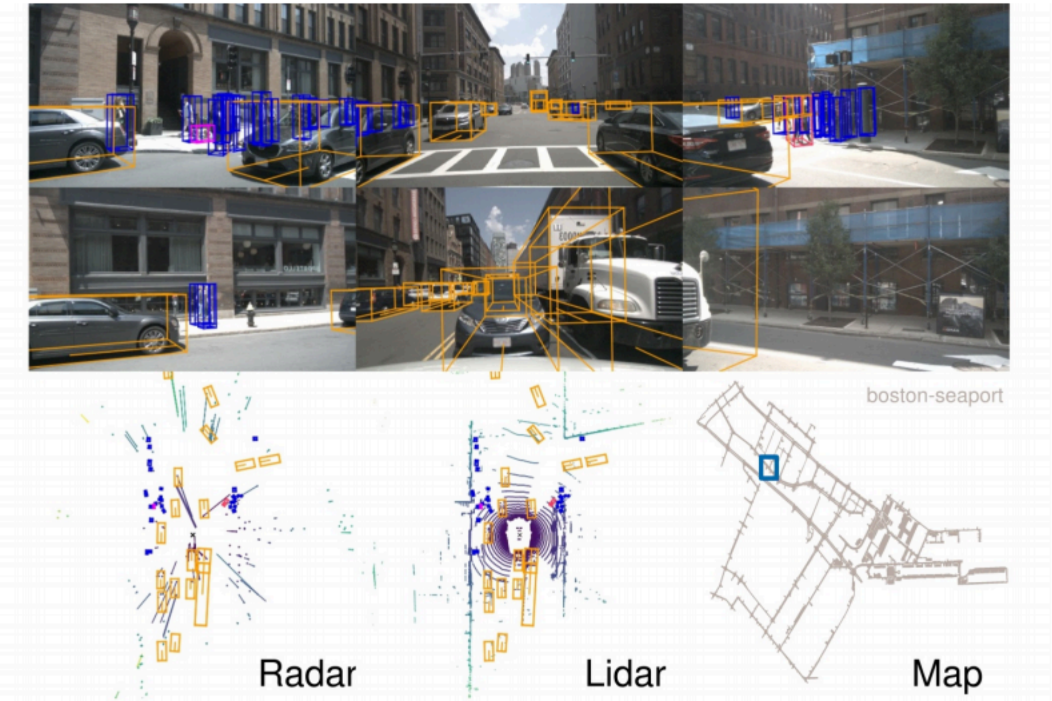
Multimodal Navigation: VNLA (D3)

- [Vision-based navigation with language-based assistance]



Autonomous driving: nuScenes (D4)

- [Multimodal dataset for autonomous driving]()



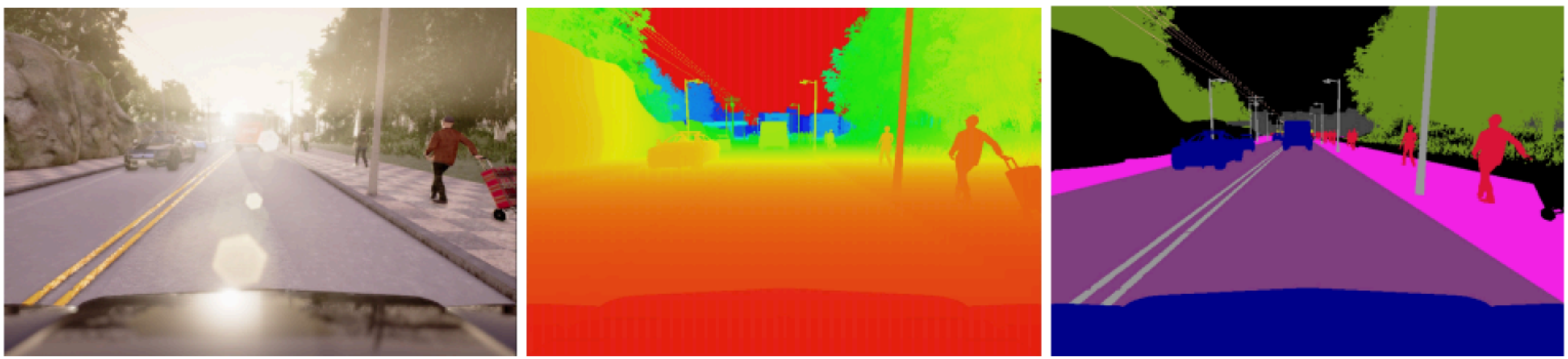
"Ped with pet, bicycle, car makes a u-turn, lane change, peds crossing crosswalk"

Autonomous driving: Waymo Open Dataset (D5)

- [Autonomous vehicle dataset]
- 1000 driving segments
- 5 cameras and 5 lidar inputs
- Dense labels for vehicles, pedestrians, cyclists, road signs.

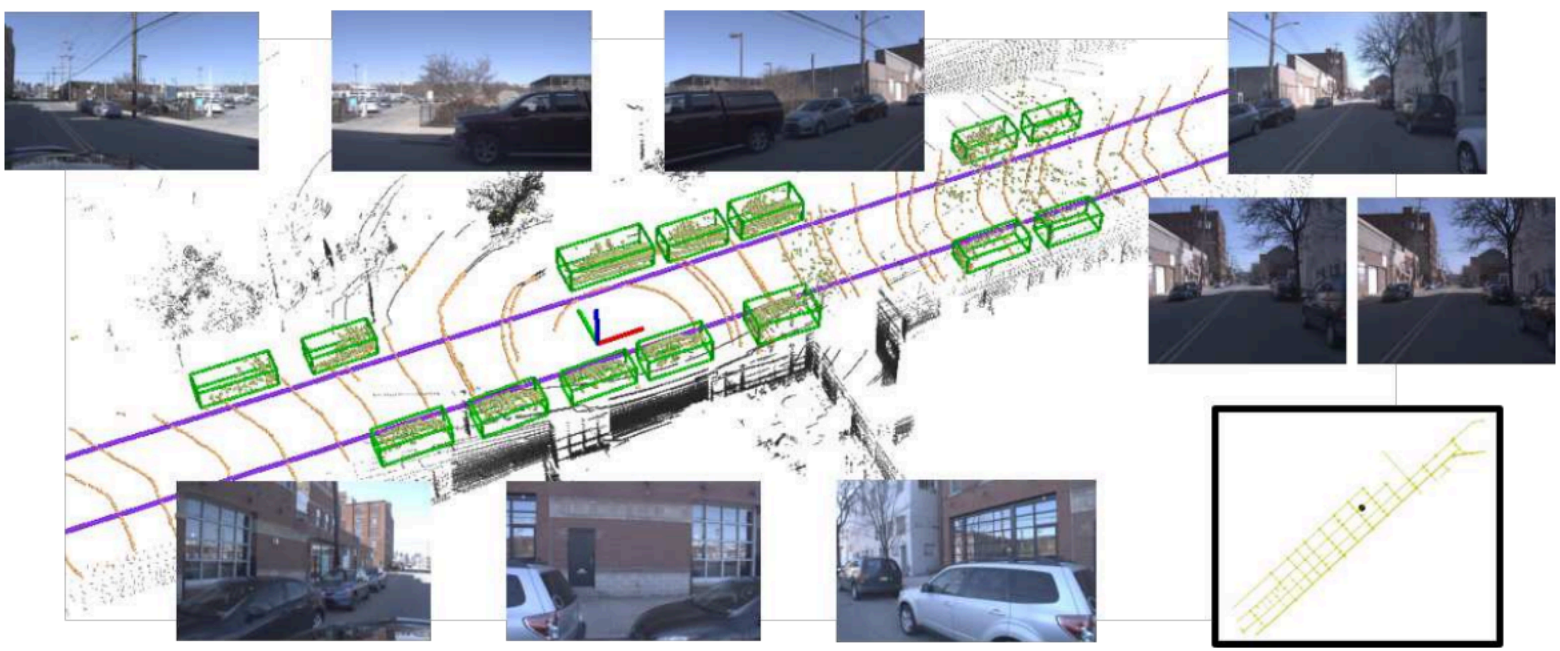
Autonomous driving: CARLA (D6)

- Simulator for autonomous driving research
- 3 sensing modalities: normal vision camera, ground-truth depth, and ground-truth semantic segmentation



Autonomous driving: Argoverse (D7)

- Autonomous vehicle dataset
- 3D tracking annotations for 113 scenes and 327,793 interesting vehicle trajectories for motion forecasting
- Input modalities: LiDAR measurements, 360◦ RGB video, front-facing stereo, and 6-dof localization



ALFRED (D8)

- ALFRED Instruction following with long trajectories and basic affordances

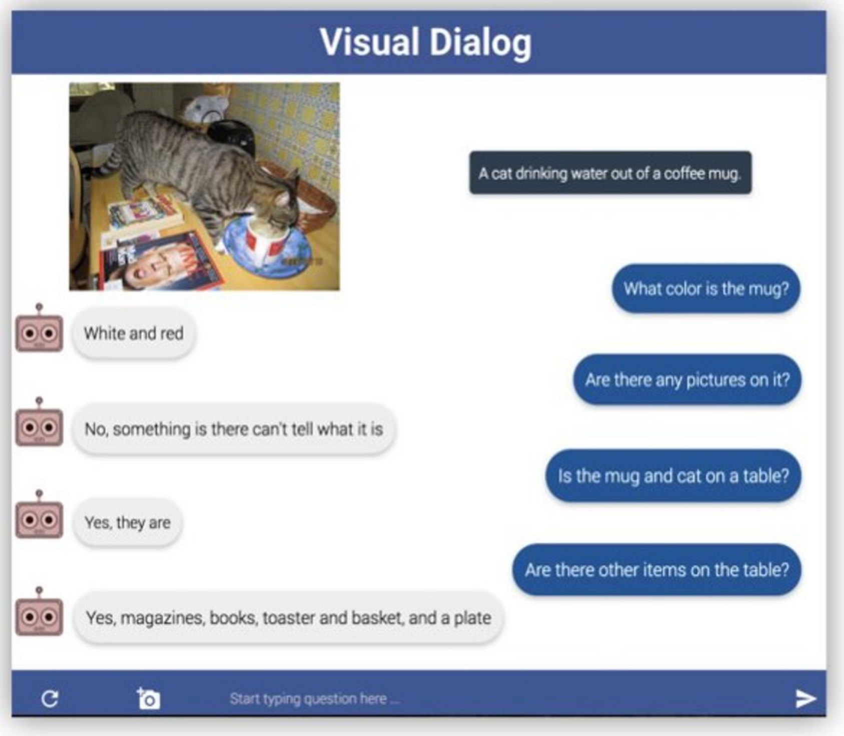
Multimodal Navigation technical challenges

- What technical problems could be addressed?
  - Translation
  - Representation
  - Alignment
  - Co-training/transfer learning
  - Fusion

Instruction: Go to the stairs on level one and bring me the bottom picture that is next to the top of the stairs.

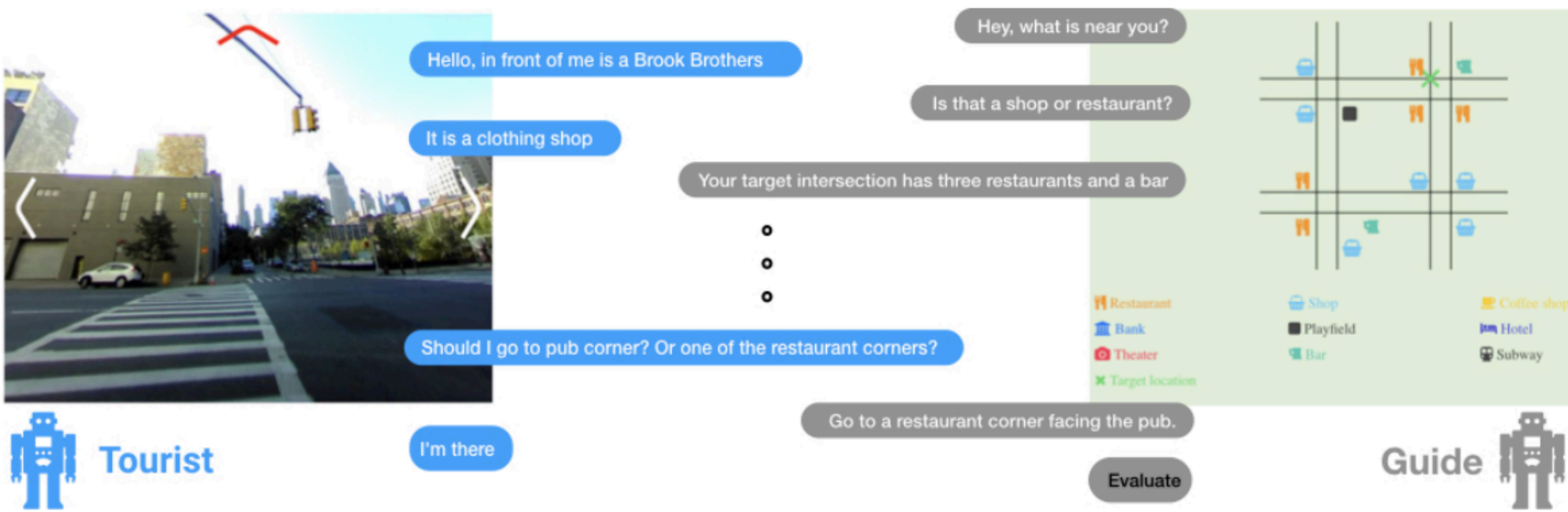
Multimodal Dialog: Visual Dialog (E1)

- VisDial v0.9: total of ~1.2M dialog question-answer pairs (1 dialog with 10 question-answer pairs on ~120k images from MS-COCO)
- VisDial v1.0 has also been released recently
- A Visual Dialog Challenge is organized at ECCV 2018



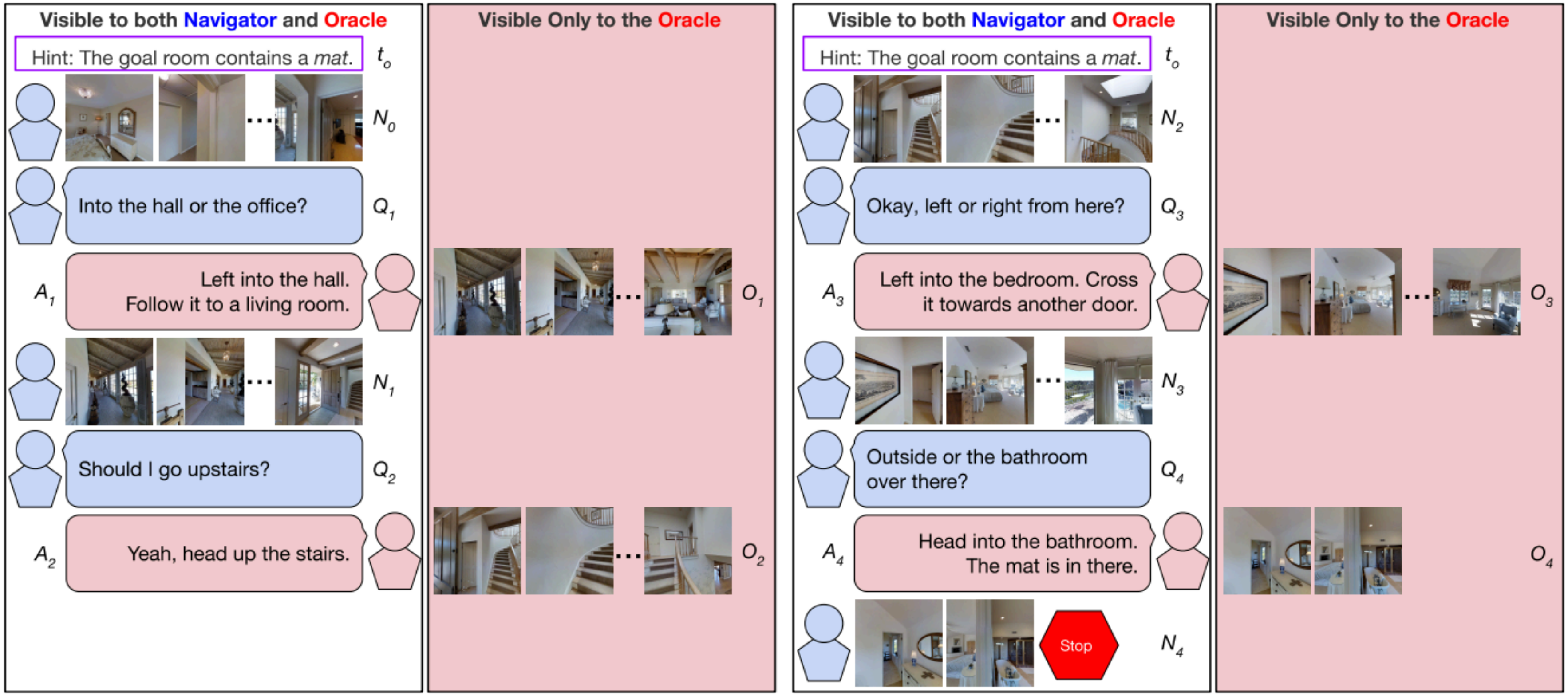
Multimodal Dialog: Talk the Walk (E2)

- A guide and a tourist communicate via natural language to navigate the tourist to a given target location. [(paper)]



Cooperative Vision-and-Dialog Navigation (E3)

- 2k embodied, human-human dialogs situated in simulated, photorealistic home environments. [(code+data)]
- Agent has to navigate towards the goal



Multimodal Dialog: CLEVR-Dialog (E4)

- Used to benchmark visual coreference resolution. [(code+data)]

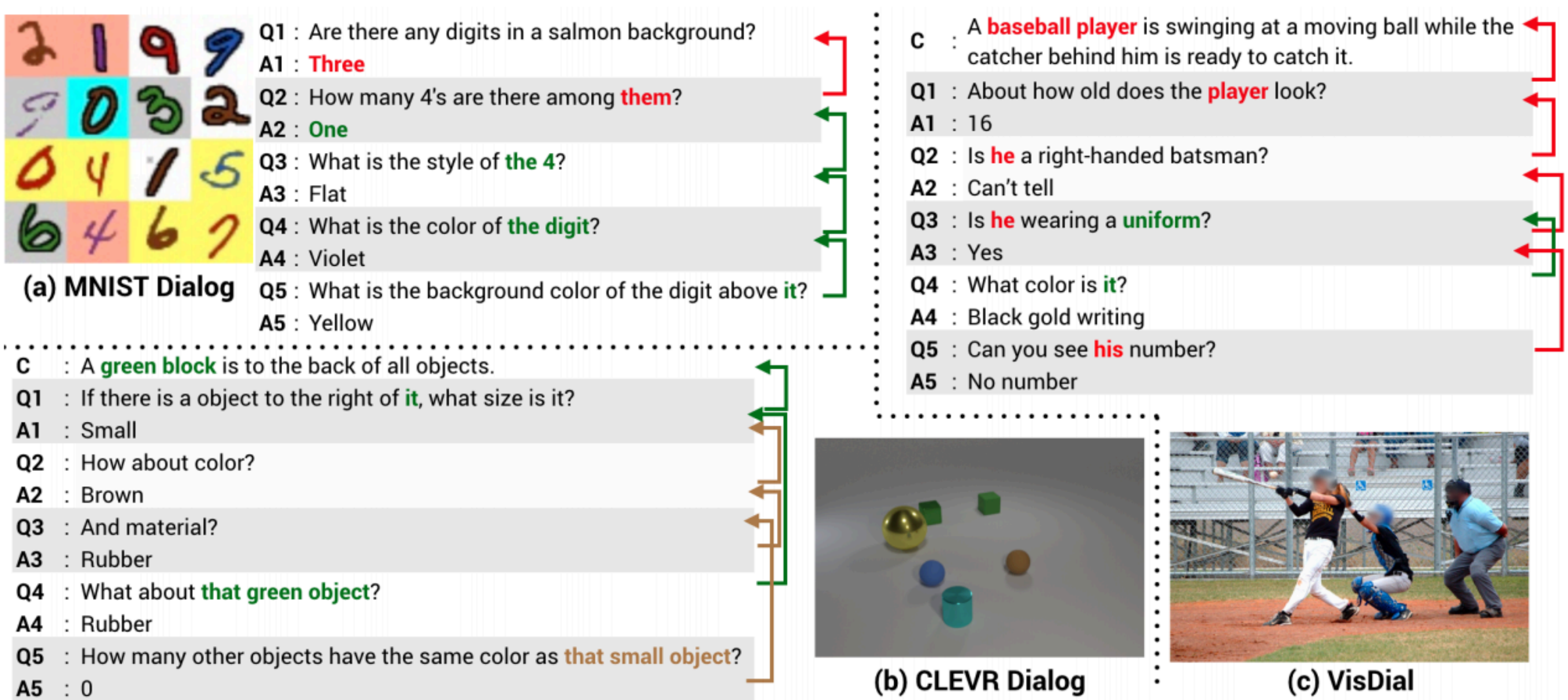


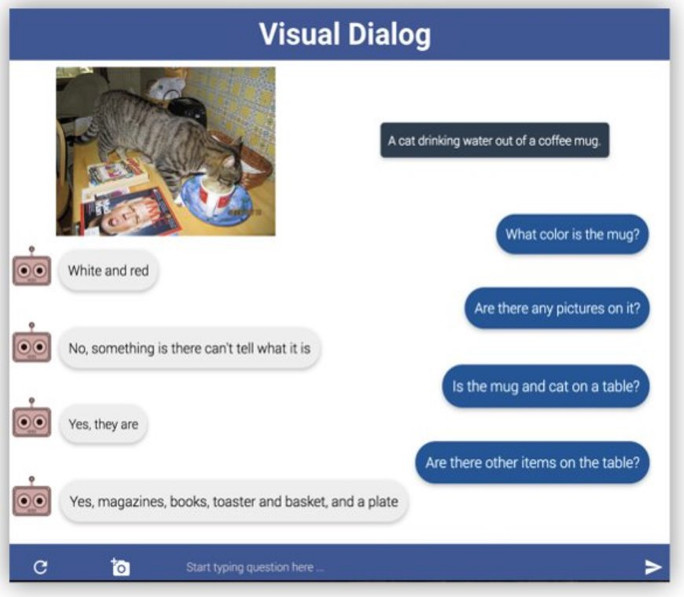
Figure 2: Example dialogs from MNIST Dialog, CLEVR-Dialog, and VisDial, with coreference chains manually marked for VisDial and automatically extracted for MNIST Dialog and CLEVR-Dialog.

Multimodal Dialog: Fashion Retrieval (E5)

- Fashion retrieval dataset
- Dialog-based interactive image retrieval

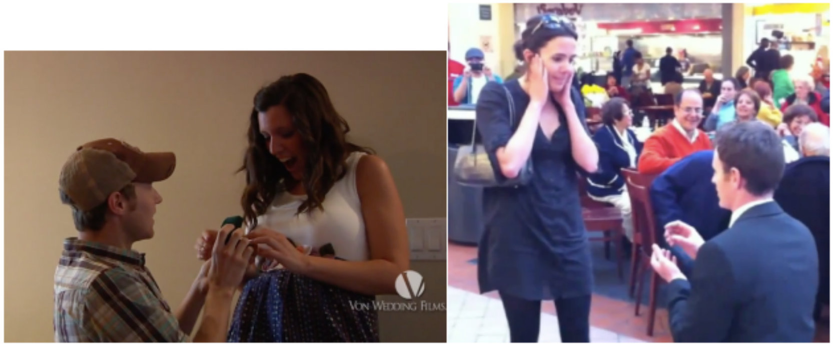
Multimodal Dialog technical challenges

- What technical problems could be addressed?
  - Representation
  - Alignment
  - Translation
  - Co-training/transfer learning
  - Fusion



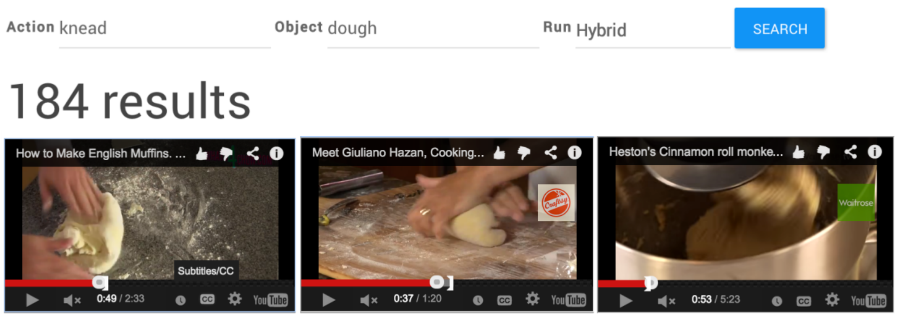
Event detection

- Given video/audio/ text detect predefined events or scenes
- Segment events in a stream
- Summarize videos



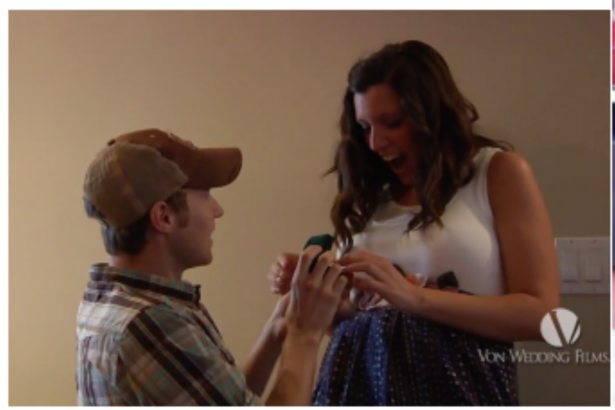
Event detection dataset 1 (F1, F2, F3 & F4)

- What's Cooking (F1)- cooking action dataset
  - **melt butter**, **brush oil**, etc.
  - **taste**, **bake** etc.
- Audio-visual, ASR captions
  - 365k clips
  - Quite noisy
- Surprisingly many cooking datasets:
  - TACoS (F2), TACoS Multi-Level (F3), YouCook (F4)



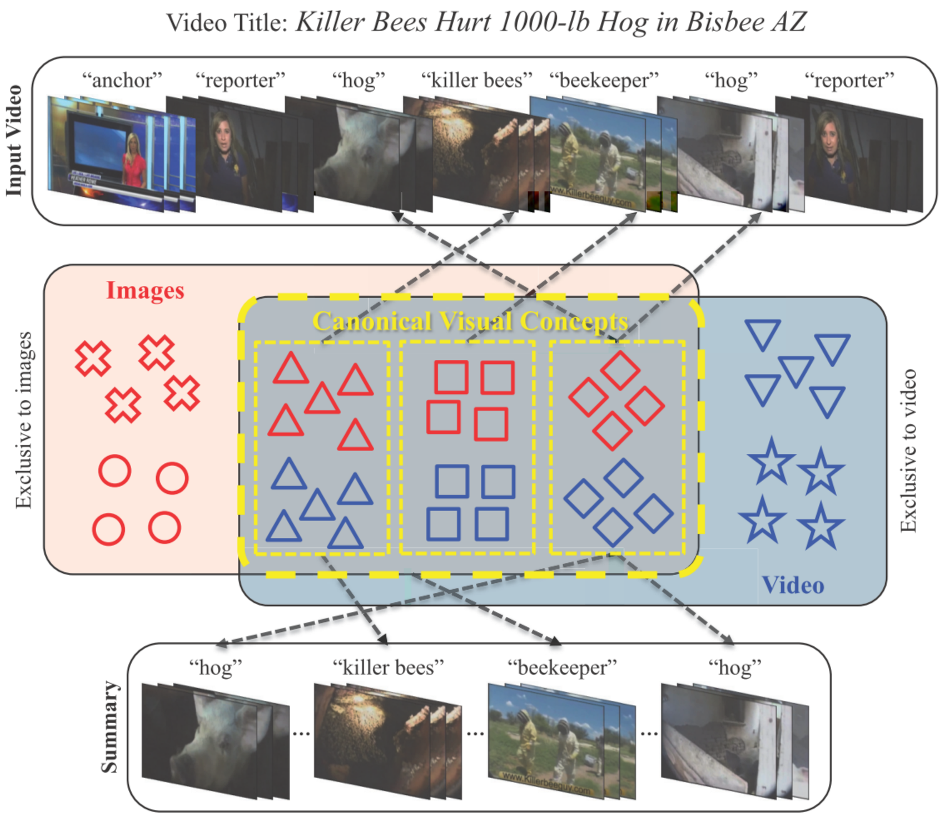
Event detection dataset 2 (F5)

- Multimedia event detection
  - TrecVid Multimedia Event Detection (MED) 2010-2015
  - One of the six TrecVid tasks
  - Audio-visual data
  - Event detection



Event detection dataset 3 (F6)

- [Title-based Video Summarization dataset]()
- 50 videos labeled for scene importance, can be used for summarization based on the title



Event detection dataset 4 (F7)

- MediaEval challenge datasets
  - Affective Impact of Movies (including Violent Scenes Detection)
  - Synchronization of Multi-User Event Media
  - Multimodal Person Discovery in Broadcast TV

CrisisMMD: Natural Disaster Assessment (F8)

- CrisisMMD – Multimodal Dataset for Natural Disasters
- 16,097 Twitter posts with one or more images
- Annotations comprises of 3 types:
  - Informative vs. Uninformative for humanitarian aid purposes
  - Humanitarian aid categories
  - Damage Assessment

Informative

(a) Hurricane Maria turns Dominica into 'giant debris field' https://t.co/rAISiAhMUy by #AJEnglish via @c0nvey https://t.co/I4zeuW4gkc

Not informative

(d) @SueAikens hi su o back againe big hug FROM PUERTO RICO love you https://t.co/HCEyIHB0QZ

Rescue & volunteering

(g) Puerto Rico donation drive going on until 4 p.m. today and again on Oct. 28! https://t.co/zXZBrHeLCQ https://t.co/2T9k2mTCIs

Event detection technical challenges

- What technical problems could be addressed?
  - Fusion
  - Representation
  - Co-learning
  - Mapping
  - Alignment (after misaligning)

Cross-media retrieval

- Given one form of media retrieve related forms of media, given text retrieve images, given image retrieve relevant documents
- Examples:
  - Image search
  - Similar image search
- Additional challenges
  - Space and speed considerations

Multimodal Retrieval: IKEA Interior Design Dataset (G1)

- Interior Design Dataset – Retrieve desired product using room photos and text queries.
- 298 room photos, 2193 product images/descriptions.

Room images:

Object images:

Description:

You sit comfortably thanks to the armrests.

There's a natural and living feeling of wood, as knots and other marks remain on the surface.

This lamp gives a pleasant light for dining and spreads a good directed light across your dining or bar table.

Cross-media retrieval datasets (G2, G3, G4)

- MIRFLICKR-1M (G2)
  - 1M images with associated tags and captions
  - Labels of general and specific categories
- NUS-WIDE dataset (G3)
  - 269,648 images and the associated tags from Flickr, with a total number of 5,018 unique tags;
- Yahoo Flickr Creative Commons 100M (G4)
  - Videos and images
- Can also use image and video captioning datasets
  - Just pose it as a retrieval task

Other Multimodal Datasets (G5, G6, G7, G8, G9, G10)

- 1) YouTube 8M (G5)
  - https://research.google.com/youtube8m/
- 2) YouTube Bounding Boxes (G6)
  - https://research.google.com/youtube-bb/
- 3) YouTube Open Images (G7)
  - https://research.googleblog.com/2016/09/introducing-open-images-dataset.html
- 4) VIST (G8)
  - http://visionandlanguage.net/VIST/
- 5) Recipe1M+ (G9)
  - http://pic2recipe.csail.mit.edu/
- 6) VATEX (G10)
  - https://eric-xw.github.io/vatex-website/

Cross-media retrieval challenges

- What technical problems could be addressed?
  - Representation
  - Translation
  - Alignment
  - Co-learning
  - Fusion