Computers ∩ Vision ∩ Language

Jeffrey Mark Siskind, qobi@purdue.edu

PURDUE UNIVERSITY

CVPR Workshop on Vision and Language, Thursday 11 June 2015

Daniel Paul Barrett Scott Alan Bronikowski Zachary Burchill Haonan Yu

This research was sponsored, in part, by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0060. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

Computers ⊕ Vision ⊕ Language

Jeffrey Mark Siskind, qobi@purdue.edu

PURDUE UNIVERSITY

CVPR Workshop on Vision and Language, Thursday 11 June 2015

Daniel Paul Barrett Scott Alan Bronikowski Zachary Burchill Haonan Yu

This research was sponsored, in part, by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0060. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

Outline

Outline

- Computers and Vision, but no Language

Outline

- Computers and Vision, but no Language
- Computers and Language, but no Vision

Outline

- Computers and Vision, but no Language
- Computers and Language, but no Vision
- Vision and Language, but no Computers

Outline

- Computers and Vision, but ~~no~~ just a little bit of Language
- Computers and Language, but ~~no~~ just a little bit of Vision
- Vision and Language, but ~~no~~ just a little bit of Computers

Outline

Sentence-Directed Video Object Codetection

Sentence-Directed Video Object Codetection

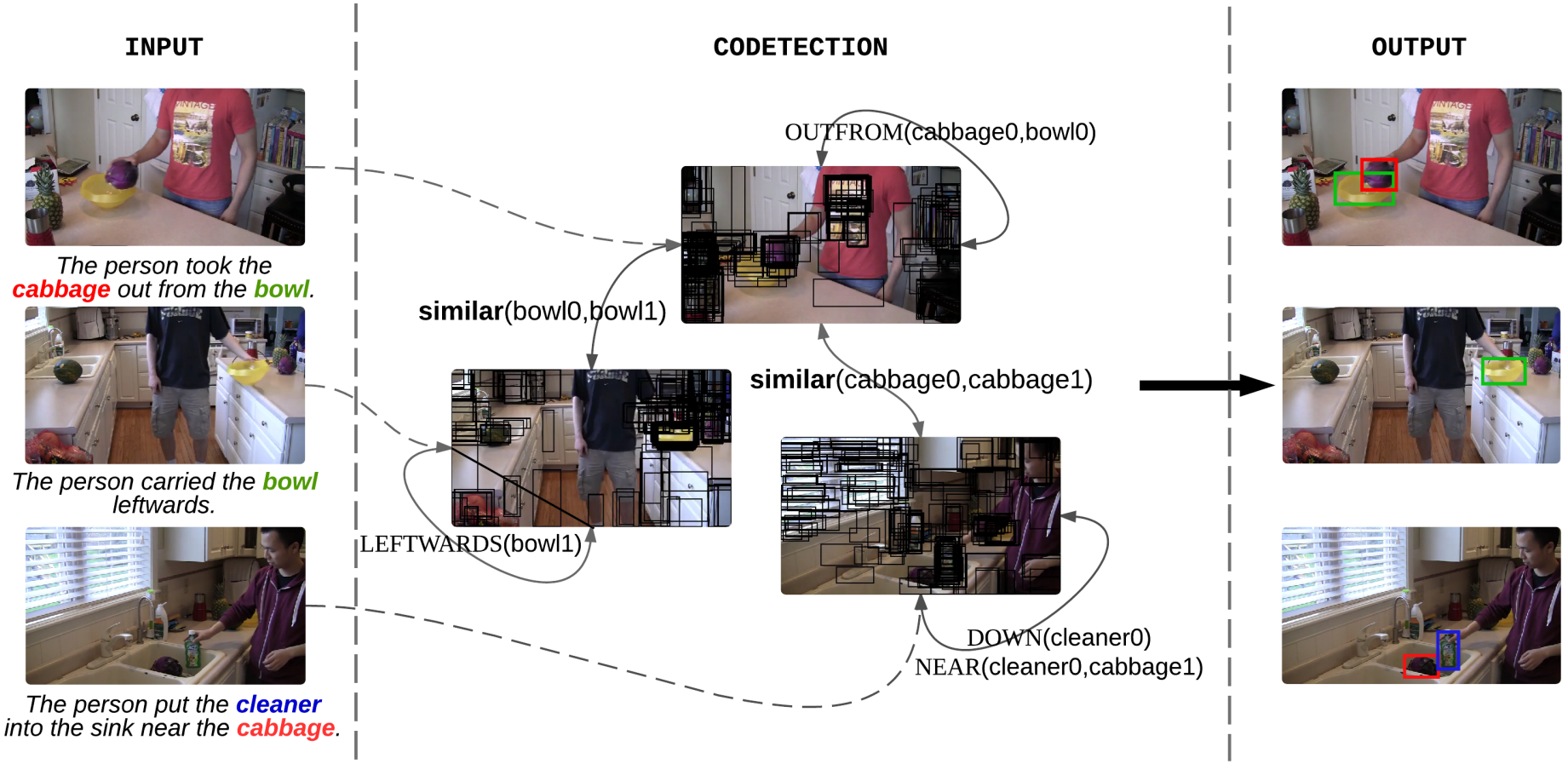
video captioning: video+detections→sentences

Sentence-Directed Video Object Codetection

video captioning: video+detections→sentences

inverse video captioning: video+sentences→detections

Overview


INPUT
CODETECTION
OUTPUT
The person took the cabbage out from the bowl.
The person carried the bowl leftwards.
The person put the cleaner into the sink near the cabbage.
OUTFROM(cabbage0,bowl0)
similar(bowl0,bowl1)
similar(cabbage0,cabbage1)
LEFTWARDS(bowl1)
DOWN(cleaner0)
NEAR(cleaner0,cabbage1)

The 7 ‘No’s

The 7 ‘No’s

- no background subtraction

The 7 ‘No’s

- no background subtraction
- no object detector

The 7 ‘No’s

- no background subtraction
- no object detector
- no object models

The 7 ‘No’s

- no background subtraction
- no object detector
- no object models
- no per-object-class parameters

The 7 ‘No’s

- no background subtraction
- no object detector
- no object models
- no per-object-class parameters
- no learning

The 7 ‘No’s

- no background subtraction
- no object detector
- no object models
- no per-object-class parameters
- no learning
- no training data

The 7 ‘No’s

- no background subtraction
- no object detector
- no object models
- no per-object-class parameters
- no learning
- no training data
- no human-annotated bounding boxes

The person put the cleaner into the sink near the cabbage.

The person put the cleaner into the sink near the cabbage.

$$\text{DOWN}(\textit{cleaner}) \wedge \text{NEAR}(\textit{cleaner}, \textit{cabbage})$$

The person put the cleaner into the sink near the cabbage.

$$\text{DOWN}(cleaner) \wedge \text{NEAR}(cleaner, cabbage)$$

Generated using Stanford parser (Socher et al. ACL 2013) and methods of Lin et al. (CVPR 2014).

The person put the cleaner into the sink near the cabbage.

$$\text{DOWN}(cleaner) \wedge \text{NEAR}(cleaner, cabbage)$$

Generated using Stanford parser (Socher et al. ACL 2013) and methods of Lin et al. (CVPR 2014).
Predicates are soft.

The person put the cleaner into the sink near the cabbage.

$$\text{DOWN}(cleaner) \wedge \text{NEAR}(cleaner, cabbage)$$

Generated using Stanford parser (Socher et al. ACL 2013) and methods of Lin et al. (CVPR 2014).
Predicates are soft.
Some are unary, some are binary.

Method

- generate proposals with EdgeBoxes (Zitnick et al. ECCV 2014) and MCG (Arbelaez et al. CVPR 2014)
- sample MOVING and STATIONARY proposals from sampled frames
- track sampled MOVING proposal with CamShift (Bradski 1998) in HSV and STATIONARY proposals with MeanShift (Comaniciu et al. 2000) in RGB forward and backward over whole clip
- rotate proposal multiples of 90°
- graphical model
  - tracks as vertices
    - unary predicates
  - track pairs as edges
    - χ^2 of PHOW (Bosch et al. ICCV 2007) and L_2 HOG (Dalal & Triggs CVPR 2005) to measure similarity
    - binary predicates

Four Variants

	SIM (variant 1)	**FLOW** (variant 2)	**SIM+FLOW** (variant 3)	**SENT** (variant 4)	**SIM+SENT** (our full method)
Flow score?	no	yes	yes	yes	yes
Similarity score?	yes	no	yes	no	yes
Sentence score?	no	partial	partial	yes	yes

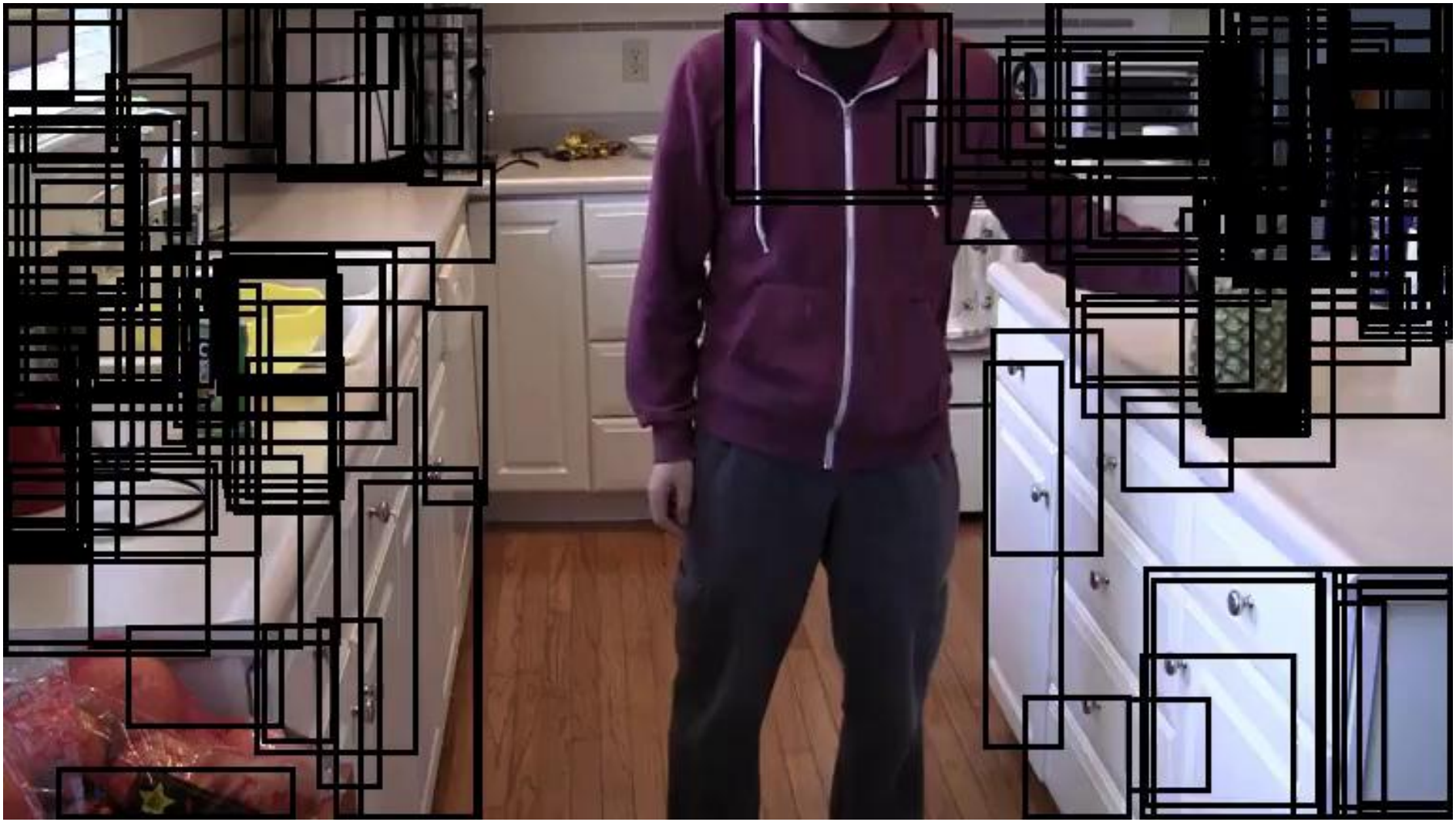
Proposals

The person put the cleaner into the sink near the cabbage.

Proposals

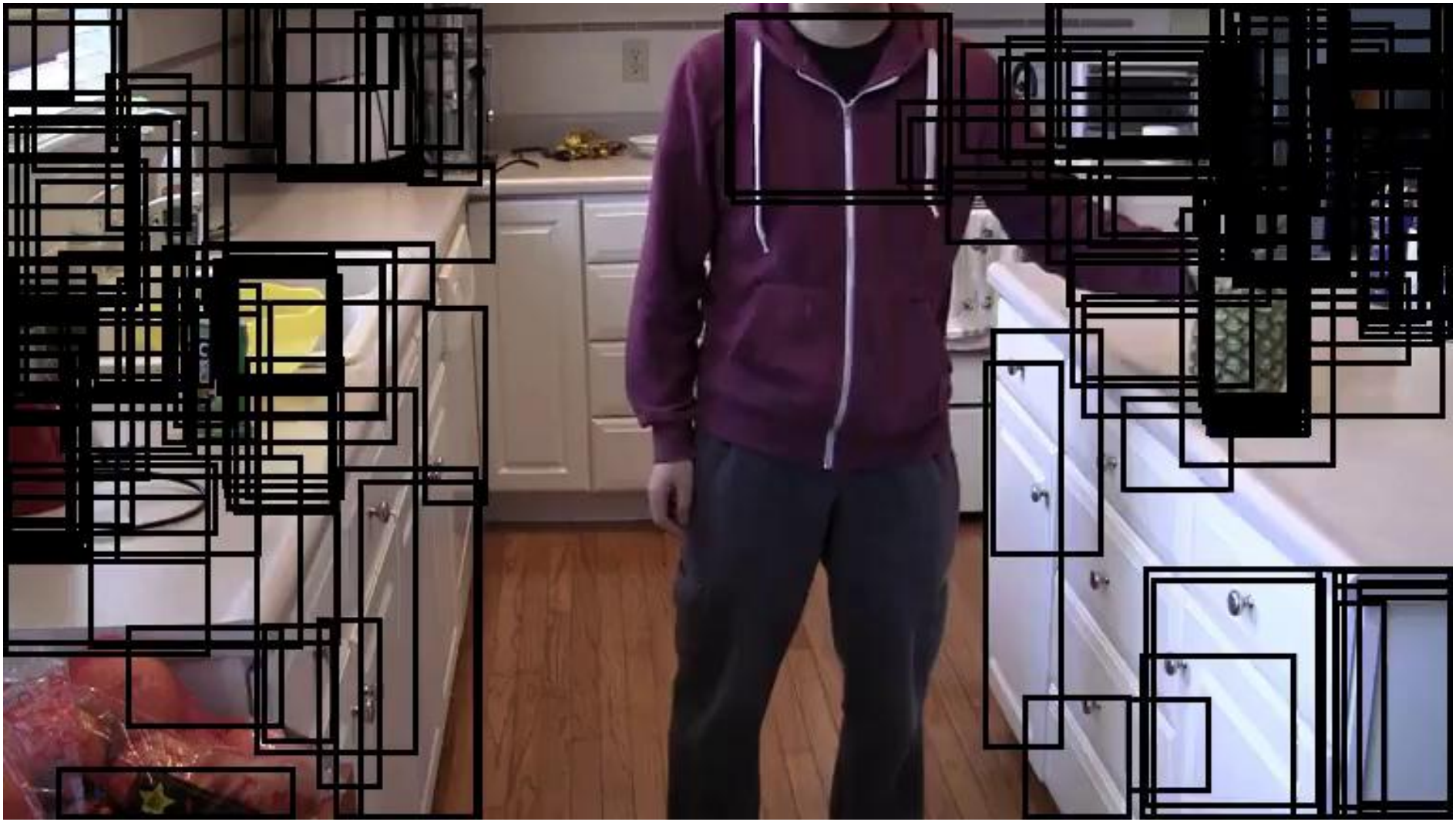
The person put the cleaner into the sink near the cabbage.

Proposals



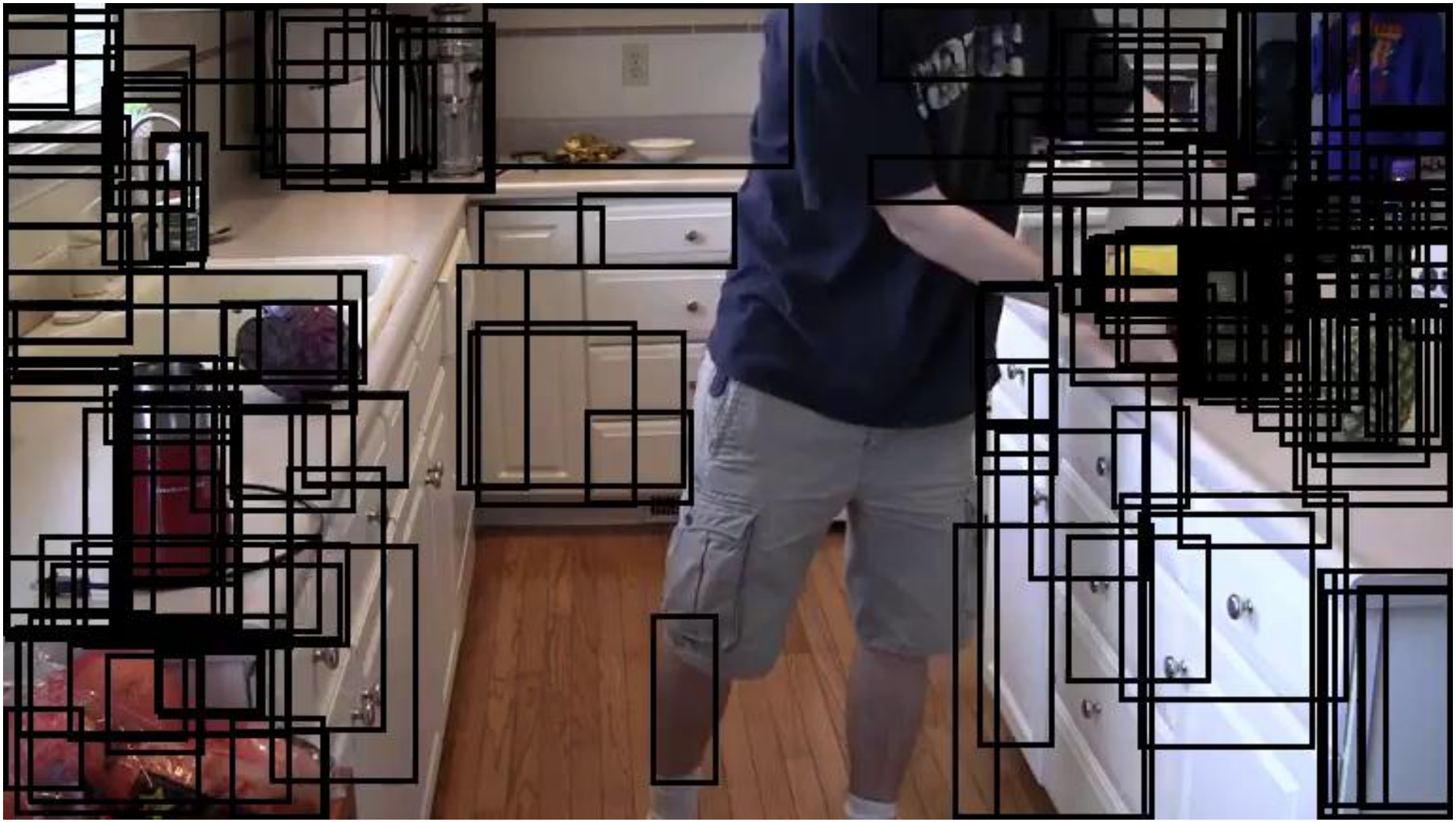
The person carried the pineapple towards the cleaner.

Proposals



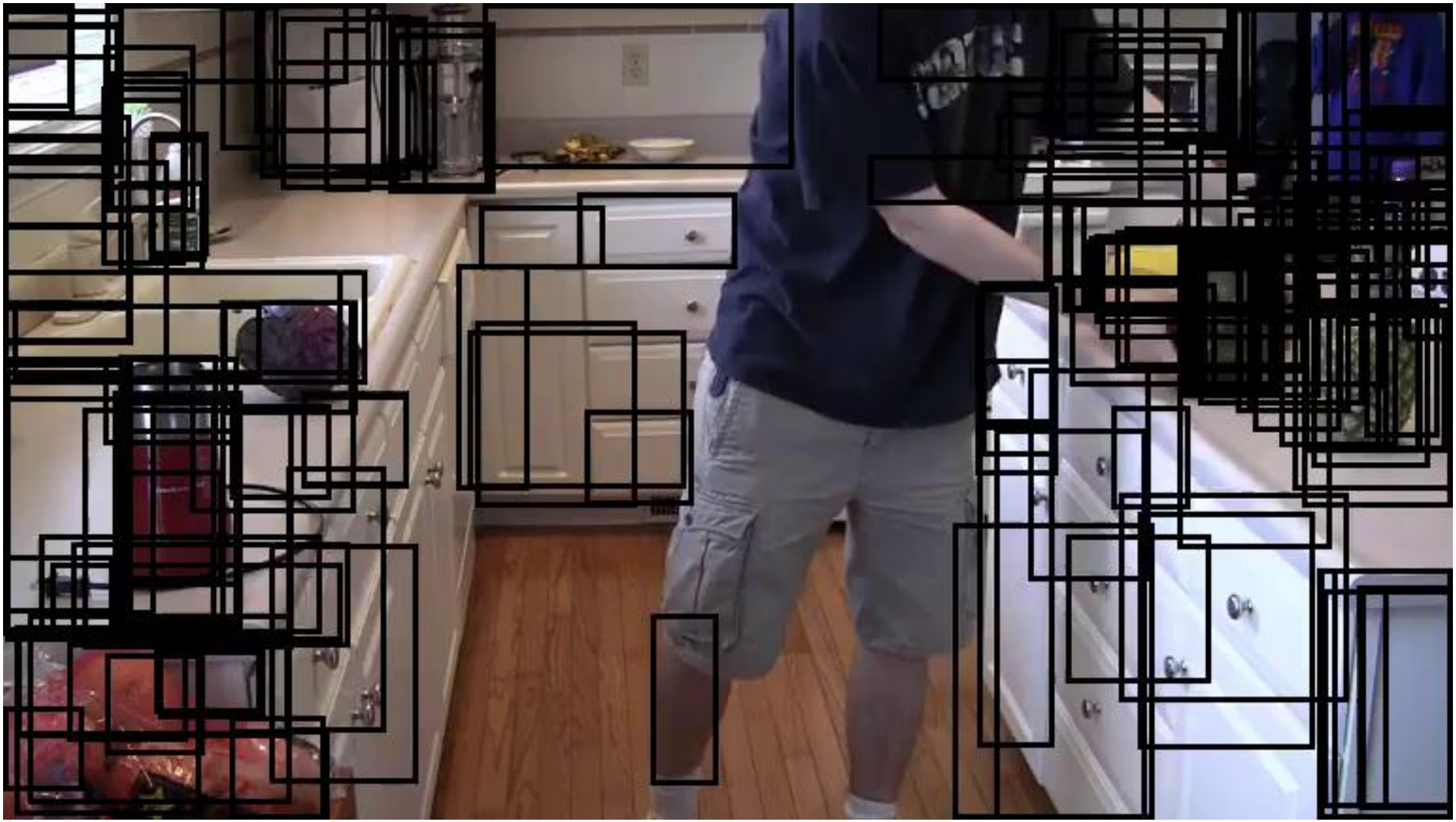
The person carried the pineapple towards the cleaner.

Proposals



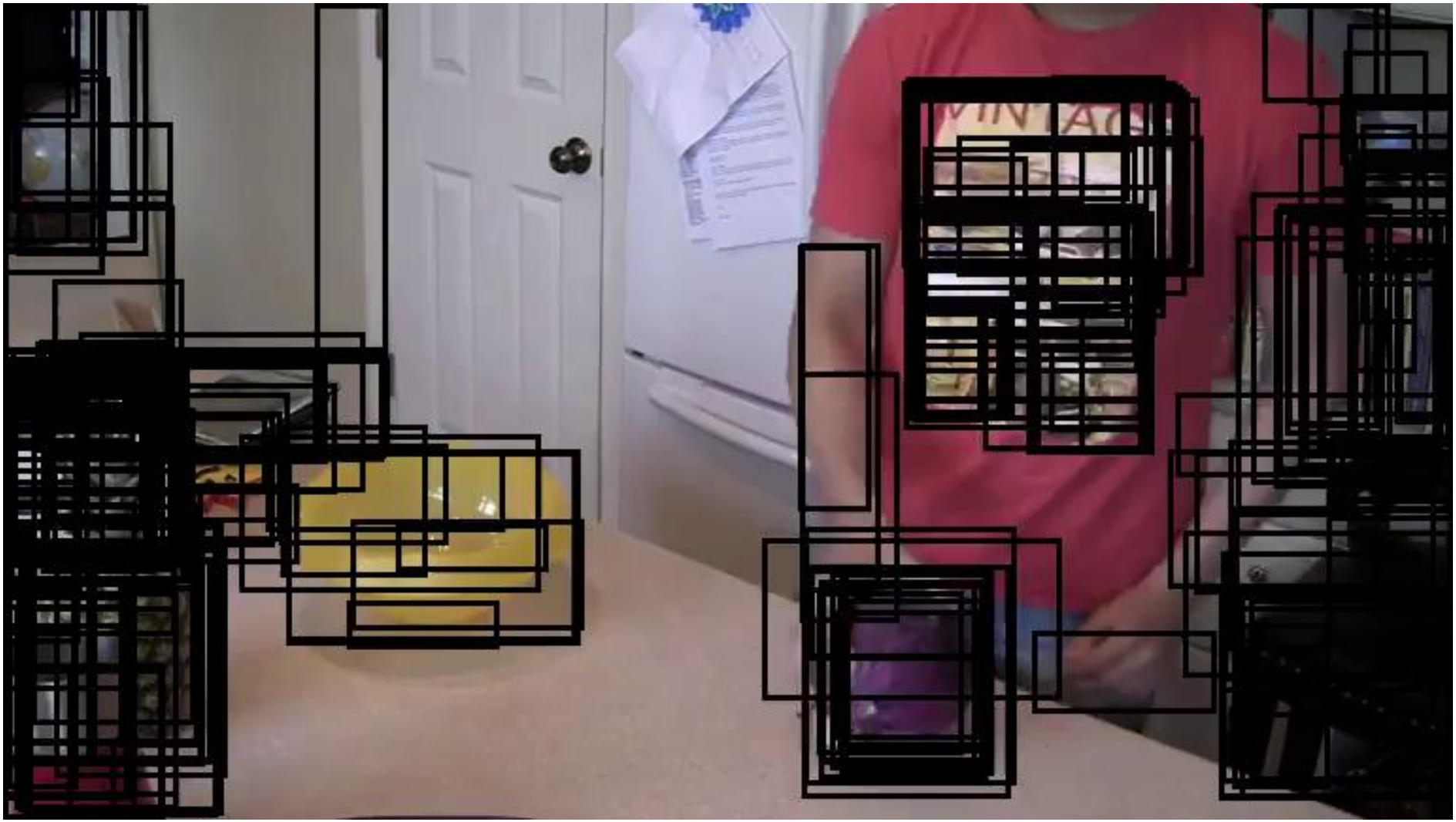
The person too the squash away from the pineapple and put it near the coffee.

Proposals



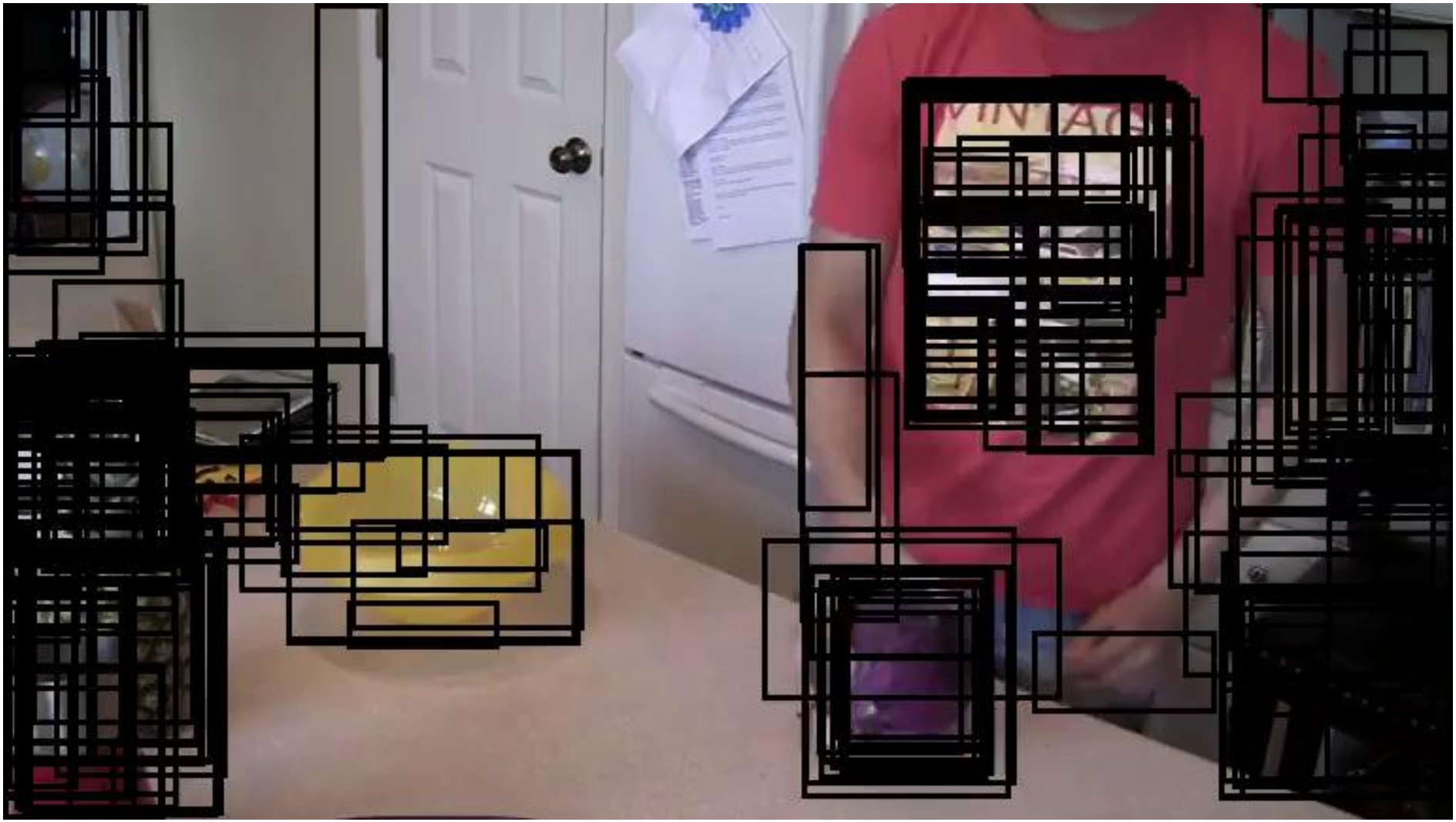
The person too the squash away from the pineapple and put it near the coffee.

Proposals



The person put the cabbage into the bowl.

Proposals



The person put the cabbage into the bowl.

SIM

The person put the cleaner into the sink near the cabbage.

SIM

The person put the cleaner into the sink near the cabbage.

SIM

The person carried the pineapple towards the cleaner.

SIM

The person carried the pineapple towards the cleaner.

SIM

The person too the squash away from the pineapple and put it near the coffee.

SIM

The person too the squash away from the pineapple and put it near the coffee.

SIM

The person put the cabbage into the bowl.

SIM

The person put the cabbage into the bowl.

FLOW

The person put the cleaner into the sink near the cabbage.

FLOW

The person put the cleaner into the sink near the cabbage.

FLOW

The person carried the pineapple towards the cleaner.

FLOW

The person carried the pineapple towards the cleaner.

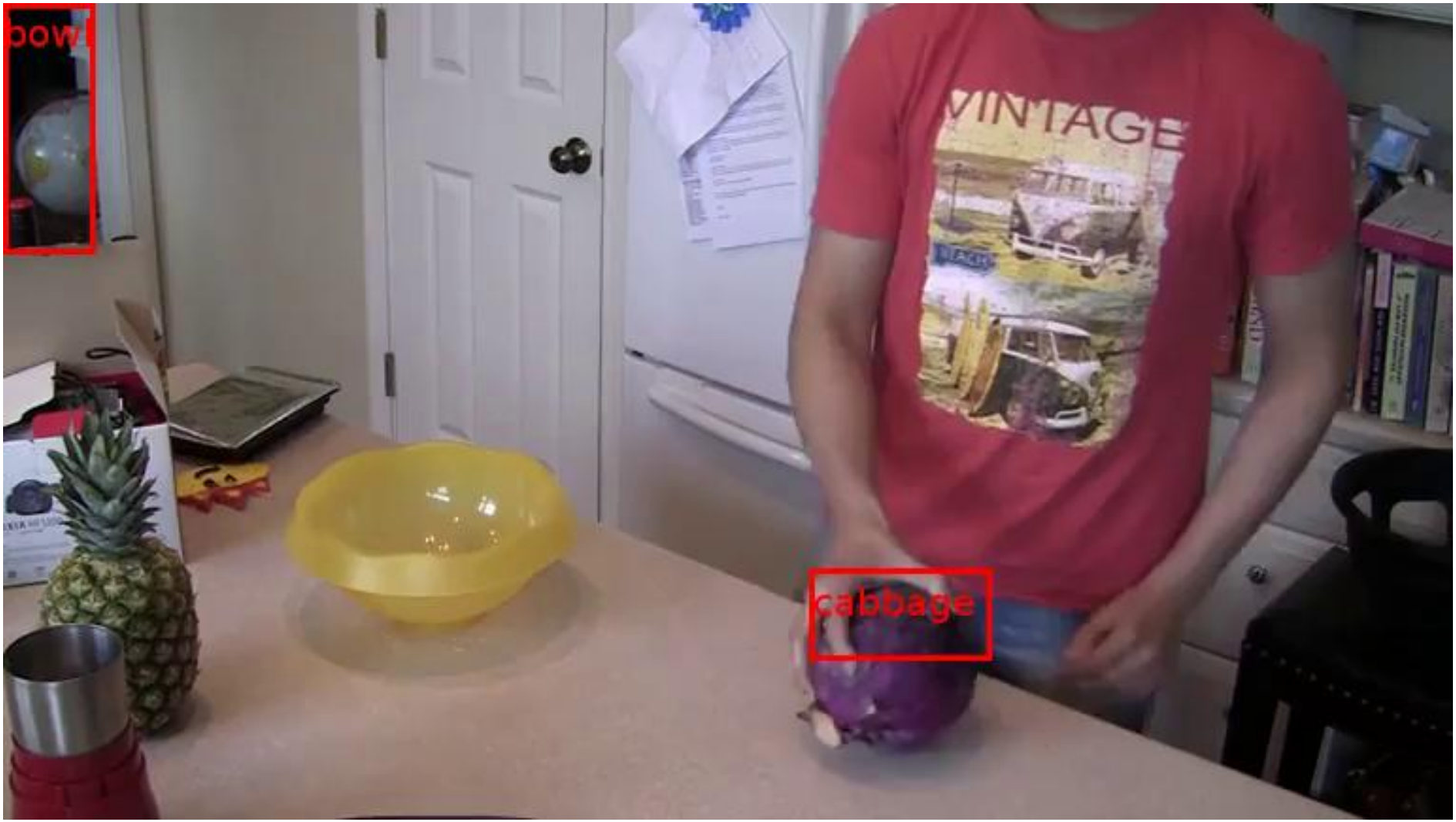
FLOW

The person too the squash away from the pineapple and put it near the coffee.

FLOW

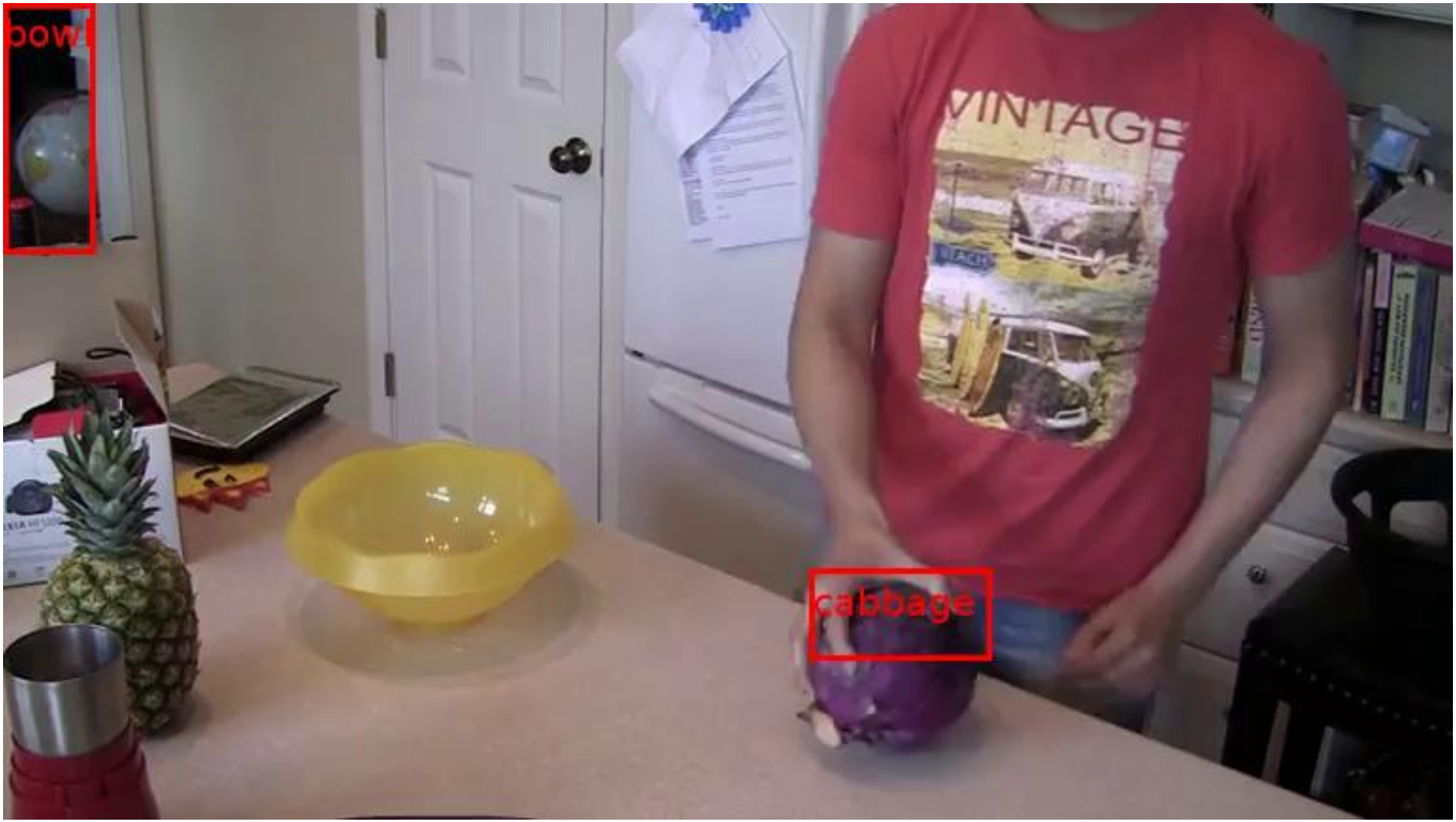
The person too the squash away from the pineapple and put it near the coffee.

FLOW



The person put the cabbage into the bowl.

FLOW



The person put the cabbage into the bowl.

SIM+FLOW

The person put the cleaner into the sink near the cabbage.

SIM+FLOW

The person put the cleaner into the sink near the cabbage.

SIM+FLOW

The person carried the pineapple towards the cleaner.

SIM+FLOW

The person carried the pineapple towards the cleaner.

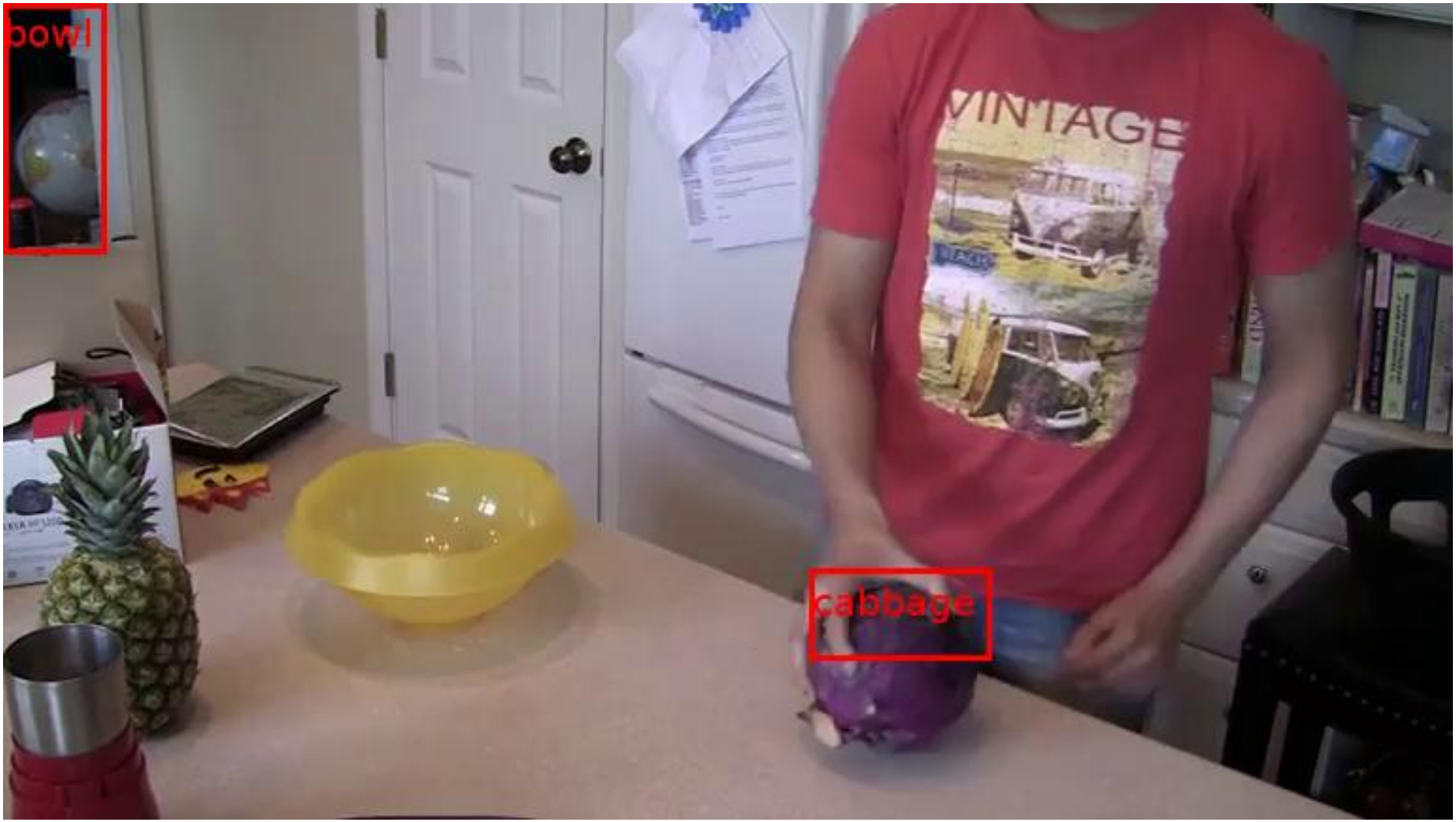
SIM+FLOW

The person too the squash away from the pineapple and put it near the coffee.

SIM+FLOW

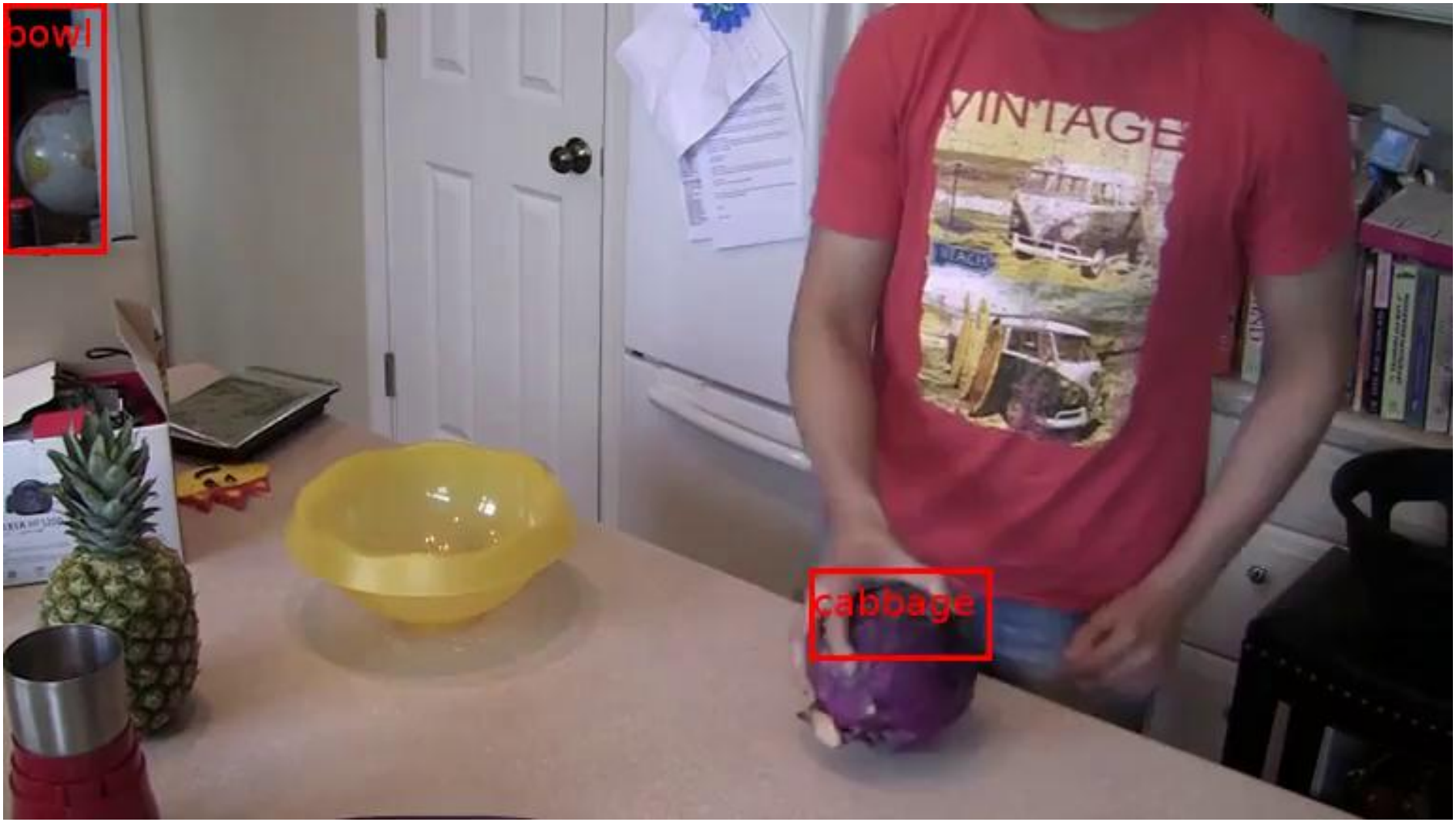
The person too the squash away from the pineapple and put it near the coffee.

SIM+FLOW



The person put the cabbage into the bowl.

SIM+FLOW



The person put the cabbage into the bowl.

SENT

The person put the cleaner into the sink near the cabbage.

SENT

The person put the cleaner into the sink near the cabbage.

SENT

The person carried the pineapple towards the cleaner.

SENT

The person carried the pineapple towards the cleaner.

SENT

The person too the squash away from the pineapple and put it near the coffee.

SENT

The person too the squash away from the pineapple and put it near the coffee.

SENT

The person put the cabbage into the bowl.

SENT

The person put the cabbage into the bowl.

SIM+SENT

The person put the cleaner into the sink near the cabbage.

SIM+SENT

The person put the cleaner into the sink near the cabbage.

SIM+SENT

The person carried the pineapple towards the cleaner.

SIM+SENT

The person carried the pineapple towards the cleaner.

SIM+SENT

The person too the squash away from the pineapple and put it near the coffee.

SIM+SENT

The person too the squash away from the pineapple and put it near the coffee.

SIM+SENT

The person put the cabbage into the bowl.

SIM+SENT

The person put the cabbage into the bowl.

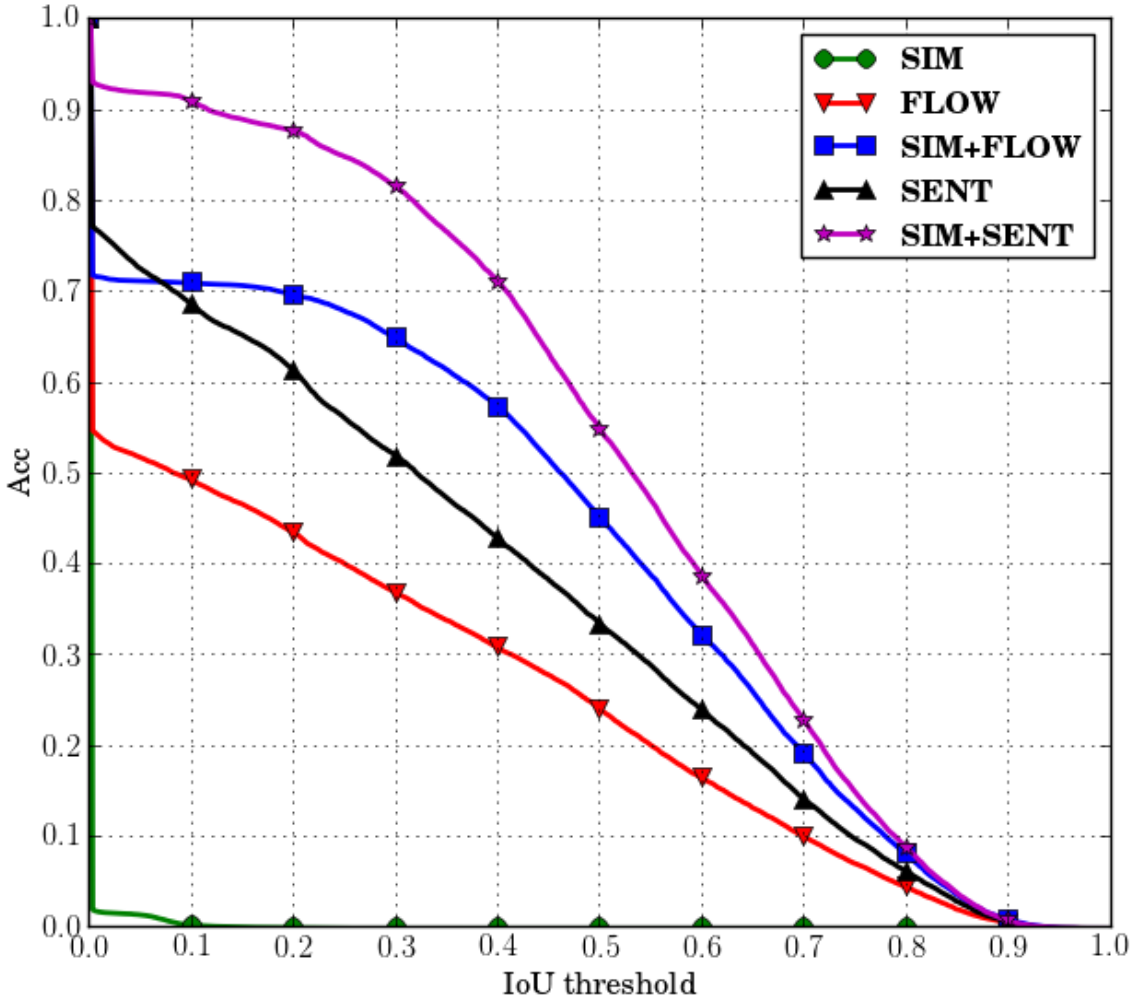
More Examples

More Examples

IoU Scores

Method	IoU_{fold}										Average
	1	2	3	4	5	6	7	8	9	10	
SIM	0.0011	0.0000	0.0003	0.0000	0.0005	0.0027	0.0000	0.0036	0.0003	0.0013	0.0010
FLOW	0.1881	0.2396	0.2252	0.1003	0.2151	0.3176	0.2431	0.1853	0.2308	0.1528	0.2098
SIM+FLOW	0.3074	0.3740	0.2828	0.1915	0.3773	0.4356	0.5018	0.2859	**0.4633**	0.2309	0.3450
SENT	0.2500	0.2885	**0.3613**	0.2486	0.2268	0.3745	0.3058	0.3238	0.2874	0.2654	0.2932
SIM+SENT	**0.3921**	**0.4738**	0.3600	**0.3499**	**0.4033**	**0.5240**	**0.5027**	**0.3733**	0.4553	**0.4281**	**0.4262**
Human-Human	0.6939	0.7024	0.7820	0.6546	0.6916	0.7861	0.7757	0.7590	0.7295	0.7187	0.7294

Codetection Accuracy



Outline

Grounded Semantics via a Unified Scoring Function

$$\mathcal{S} : (\mathbf{B}, \mathbf{s}, \Lambda) \mapsto (\tau, \mathbf{J})$$

- $\mathbf{B}$: video
- $\mathbf{s}$: sentence
- Λ: lexicon
- τ: score
- $\mathbf{J}$: tracks

Yu, H. and Siskind, J.M., 'Grounded Language Learning from Video Described with Sentences,', ACL 2013.
Siddharth, N., Barbu, A., and Siskind, J.M., 'Seeing What You're Told: Sentence-Guided Activity Recognition In Video,' CVPR 2014.
Yu, H. and Siddharth, N. and Barbu, A. and Siskind, J.M., 'A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video,' JAIR 2015.

Four Uses of Unified Scoring Function

Four Uses of Unified Scoring Function

- **Comprehension**: video $\times$ sentence $\times$ lexicon $\rightarrow$ track collection

$$\begin{aligned} \mathbf{J}_1 &= \mathcal{S}_{\mathbf{J}}(\mathbf{B}, \mathbf{s}_1, \Lambda) \\ \mathbf{J}_2 &= \mathcal{S}_{\mathbf{J}}(\mathbf{B}, \mathbf{s}_2, \Lambda) \end{aligned}$$

Four Uses of Unified Scoring Function

- **Comprehension**: video $\times$ sentence $\times$ lexicon $\rightarrow$ track collection

$$\begin{aligned} \mathbf{J}_1 &= \mathcal{S}_{\mathbf{J}}(\mathbf{B}, \mathbf{s}_1, \Lambda) \\ \mathbf{J}_2 &= \mathcal{S}_{\mathbf{J}}(\mathbf{B}, \mathbf{s}_2, \Lambda) \end{aligned}$$

- **Generation**: video $\times$ lexicon $\rightarrow$ sentence

$$\arg \max_{\mathbf{s}} \mathcal{S}_{\tau}(\mathbf{B}, \mathbf{s}, \Lambda)$$

Four Uses of Unified Scoring Function

- **Comprehension**: video $\times$ sentence $\times$ lexicon $\rightarrow$ track collection

$$\begin{aligned} \mathbf{J}_1 &= \mathcal{S}_{\mathbf{J}}(\mathbf{B}, \mathbf{s}_1, \Lambda) \\ \mathbf{J}_2 &= \mathcal{S}_{\mathbf{J}}(\mathbf{B}, \mathbf{s}_2, \Lambda) \end{aligned}$$

- **Generation**: video $\times$ lexicon $\rightarrow$ sentence

$$\arg \max_{\mathbf{s}} \mathcal{S}_{\tau}(\mathbf{B}, \mathbf{s}, \Lambda)$$

- **Retrieval**: sentence $\times$ lexicon $\rightarrow$ video

$$\arg \max_{i} \mathcal{S}_{\tau}(\mathbf{B}_i, \mathbf{s}, \Lambda)$$

Four Uses of Unified Scoring Function

- **Comprehension**: video $\times$ sentence $\times$ lexicon $\rightarrow$ track collection

$$\begin{aligned} \mathbf{J}_1 &= \mathcal{S}_{\mathbf{J}}(\mathbf{B}, \mathbf{s}_1, \Lambda) \\ \mathbf{J}_2 &= \mathcal{S}_{\mathbf{J}}(\mathbf{B}, \mathbf{s}_2, \Lambda) \end{aligned}$$

- **Generation**: video $\times$ lexicon $\rightarrow$ sentence

$$\arg \max_{\mathbf{s}} \mathcal{S}_{\tau}(\mathbf{B}, \mathbf{s}, \Lambda)$$

- **Retrieval**: sentence $\times$ lexicon $\rightarrow$ video

$$\arg \max_{i} \mathcal{S}_{\tau}(\mathbf{B}_i, \mathbf{s}, \Lambda)$$

- **Acquisition**: video $\times$ sentence $\rightarrow$ lexicon

$$\arg \max_{\Lambda} \sum_{m=1}^{M} \mathcal{S}_{\tau}(\mathbf{B}_m, \mathbf{s}_m, \Lambda)$$

Grounded Semantics via a Unified Scoring Function

$$\mathcal{R} : (\mathbf{p}, \mathbf{s}, \Lambda) \mapsto \tau$$

- $\mathbf{p}$: path
- $\mathbf{s}$: sentence
- Λ: lexicon
- τ: score

Three Uses of the Unified Scoring Function

Three Uses of the Unified Scoring Function

- **Acquisition**: sentence $\times$ path $\rightarrow$ lexicon

$$\arg \max_{\Lambda} \sum_{m=1}^{M} \mathcal{R}(\mathbf{p}_m, \mathbf{s}_m, \Lambda)$$

Three Uses of the Unified Scoring Function

- **Acquisition**: sentence $\times$ path $\rightarrow$ lexicon

$$\arg \max_{\Lambda} \sum_{m=1}^{M} \mathcal{R}(\mathbf{p}_m, \mathbf{s}_m, \Lambda)$$

- **Generation**: path $\times$ lexicon $\rightarrow$ sentence

$$\arg \max_{\mathbf{s}} \mathcal{R}(\mathbf{p}, \mathbf{s}, \Lambda)$$

Three Uses of the Unified Scoring Function

- **Acquisition**: sentence $\times$ path $\rightarrow$ lexicon

$$\arg \max_{\Lambda} \sum_{m=1}^{M} \mathcal{R}(\mathbf{p}_m, \mathbf{s}_m, \Lambda)$$

- **Generation**: path $\times$ lexicon $\rightarrow$ sentence

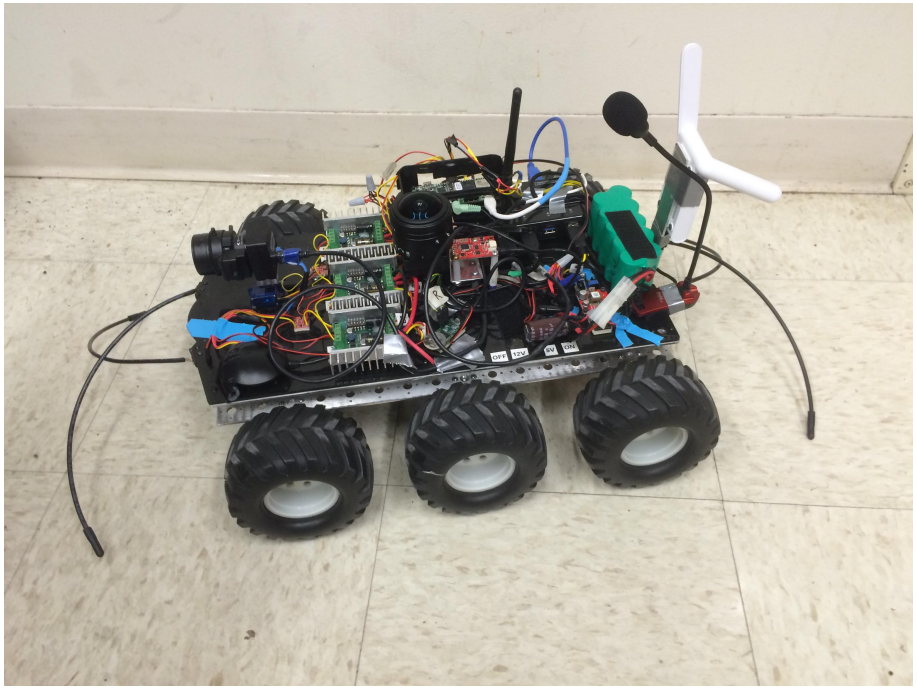
$$\arg \max_{\mathbf{s}} \mathcal{R}(\mathbf{p}, \mathbf{s}, \Lambda)$$

- **Comprehension**: sentence $\times$ lexicon $\rightarrow$ path

$$\arg \max_{\mathbf{p}} \mathcal{R}(\mathbf{p}, \mathbf{s}, \Lambda)$$

Our Custom Mobile Robot

- IMU (3-axis accelerometers, gyros, and magnetometers)
- GPS
- 6 independently controllable wheel motors
- 2 shaft encoders with Teensy controller
- Gumstix Overo FireSTORM + Summit running Linux
- Bluetooth, WiFi, and 4G LTE
- front and rear bump sensors
- ultrasonic rangefinder
- pan-tilt front-facing camera (Point Grey)
- omnidirectional camera (Point Grey)
- audio input and output
- touchscreen
- Logitech Wireless Gamepad
- custom firmware on IMU and Teensy
- synchronized timestamped logging of sensor and control data



Logical Form

The robot went toward the left side of the stool, then toward the cone which is behind the stool, then toward the table which is left of the cone, then went back toward the stool and to the left of the stool.

Logical Form

The robot went toward the left side of the stool, then toward the cone which is behind the stool, then toward the table which is left of the cone, then went back toward the stool and to the left of the stool.

$$[\alpha,\beta,\gamma,\delta]\{t,u,v,w,x,y,z\}\left(\begin{array}{l}\text{LEFT}(\alpha,t)\wedge\text{STOOL}(t)\wedge\\ \text{TOWARD}(\beta,u)\wedge\text{CONE}(u)\wedge\text{BEHIND}(u,v)\wedge\text{STOOL}(v)\wedge\\ \text{TOWARD}(\gamma,w)\wedge\text{TABLE}(w)\wedge\text{LEFT}(w,x)\wedge\text{CONE}(x)\wedge\\ \text{TOWARD}(\delta,y)\wedge\text{LEFT}(\delta,z)\wedge\text{STOOL}(y)\wedge\text{STOOL}(z)\wedge\end{array}\right)$$

Logical Form

The robot went toward the left side of the stool, then toward the cone which is behind the stool, then toward the table which is left of the cone, then went back toward the stool and to the left of the stool.

$$[\alpha, \beta, \gamma, \delta]\{t, u, v, w, x, y, z\} \left(\begin{array}{l} \text{LEFT}(\alpha, t) \wedge \text{STOOL}(t) \wedge \\ \text{TOWARD}(\beta, u) \wedge \text{CONE}(u) \wedge \text{BEHIND}(u, v) \wedge \text{STOOL}(v) \wedge \\ \text{TOWARD}(\gamma, w) \wedge \text{TABLE}(w) \wedge \text{LEFT}(w, x) \wedge \text{CONE}(x) \wedge \\ \text{TOWARD}(\delta, y) \wedge \text{LEFT}(\delta, z) \wedge \text{STOOL}(y) \wedge \text{STOOL}(z) \wedge \end{array} \right)$$

- all sentences naturally elicited from humans through AMT

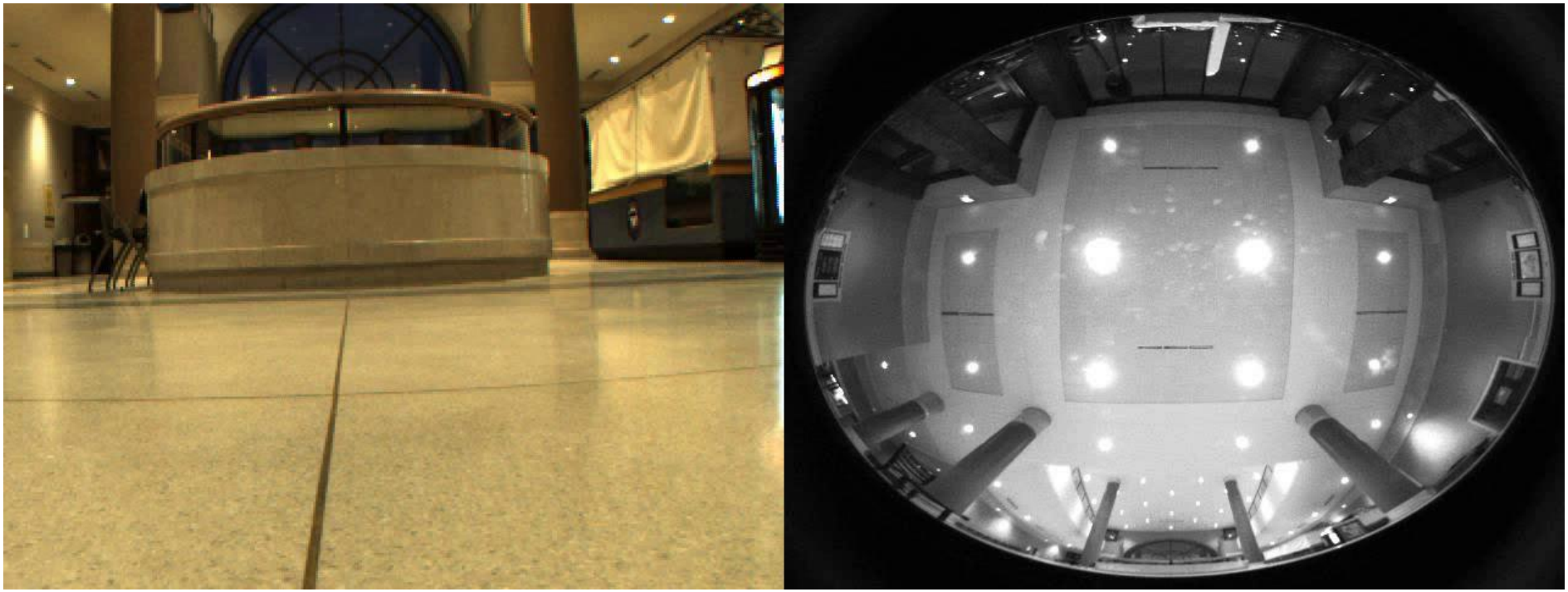
Logical Form

The robot went toward the left side of the stool, then toward the cone which is behind the stool, then toward the table which is left of the cone, then went back toward the stool and to the left of the stool.

$$[\alpha, \beta, \gamma, \delta]\{t, u, v, w, x, y, z\} \left(\begin{array}{l} \text{LEFT}(\alpha, t) \wedge \text{STOOL}(t) \wedge \\ \text{TOWARD}(\beta, u) \wedge \text{CONE}(u) \wedge \text{BEHIND}(u, v) \wedge \text{STOOL}(v) \wedge \\ \text{TOWARD}(\gamma, w) \wedge \text{TABLE}(w) \wedge \text{LEFT}(w, x) \wedge \text{CONE}(x) \wedge \\ \text{TOWARD}(\delta, y) \wedge \text{LEFT}(\delta, z) \wedge \text{STOOL}(y) \wedge \text{STOOL}(z) \wedge \end{array} \right)$$

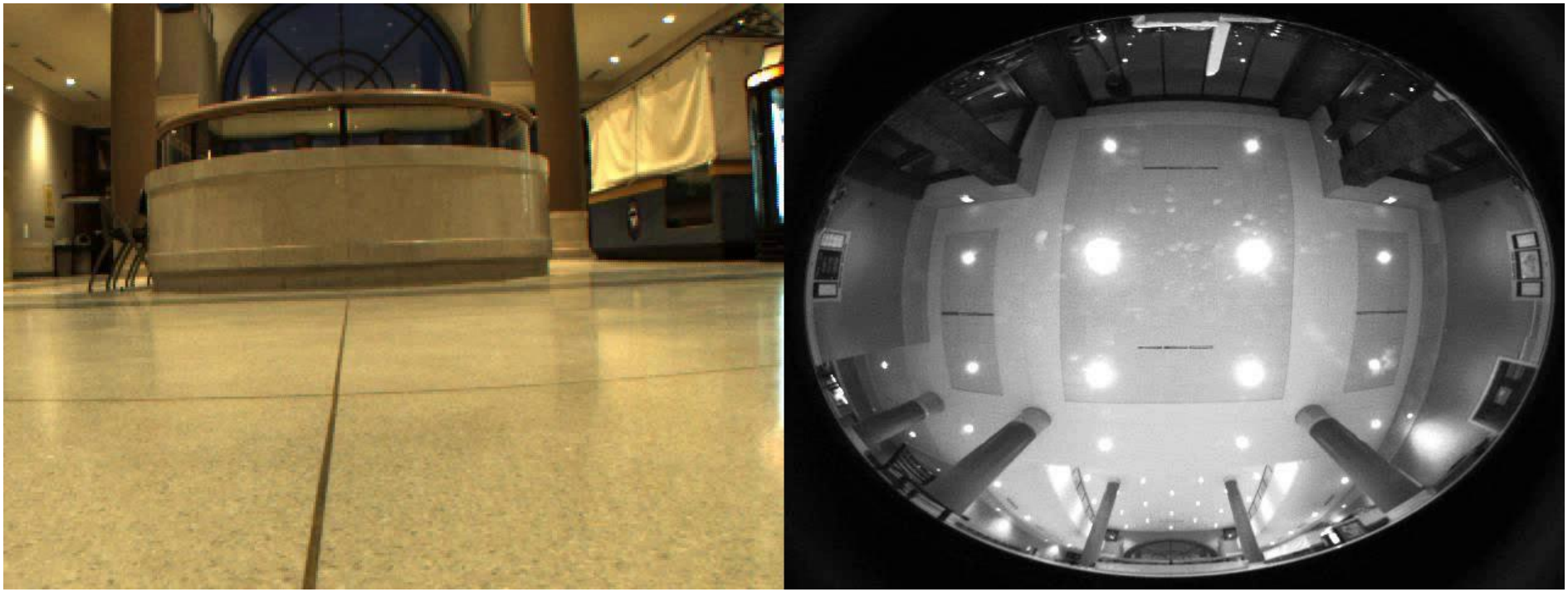
- all sentences naturally elicited from humans through AMT
- no grammar or parse trees at all

Acquisition



The robot began heading towards the right of the bag on the right, once behind the bag on the right, it turned around, went to the left of the bag on the right and headed towards the chair, ending in front of the chair.

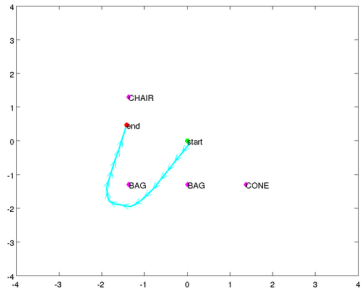
Acquisition



The robot began heading towards the right of the bag on the right, once behind the bag on the right, it turned around, went to the left of the bag on the right and headed towards the chair, ending in front of the chair.

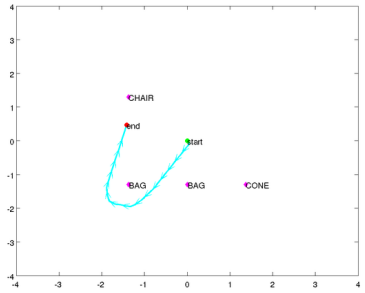
Acquisition

input: *The robot began heading towards the right of the bag on the right, once behind the bag on the right, it turned around, went to the left of the bag on the right and headed towards the chair, ending in front of the chair.*
. . . (749 more)

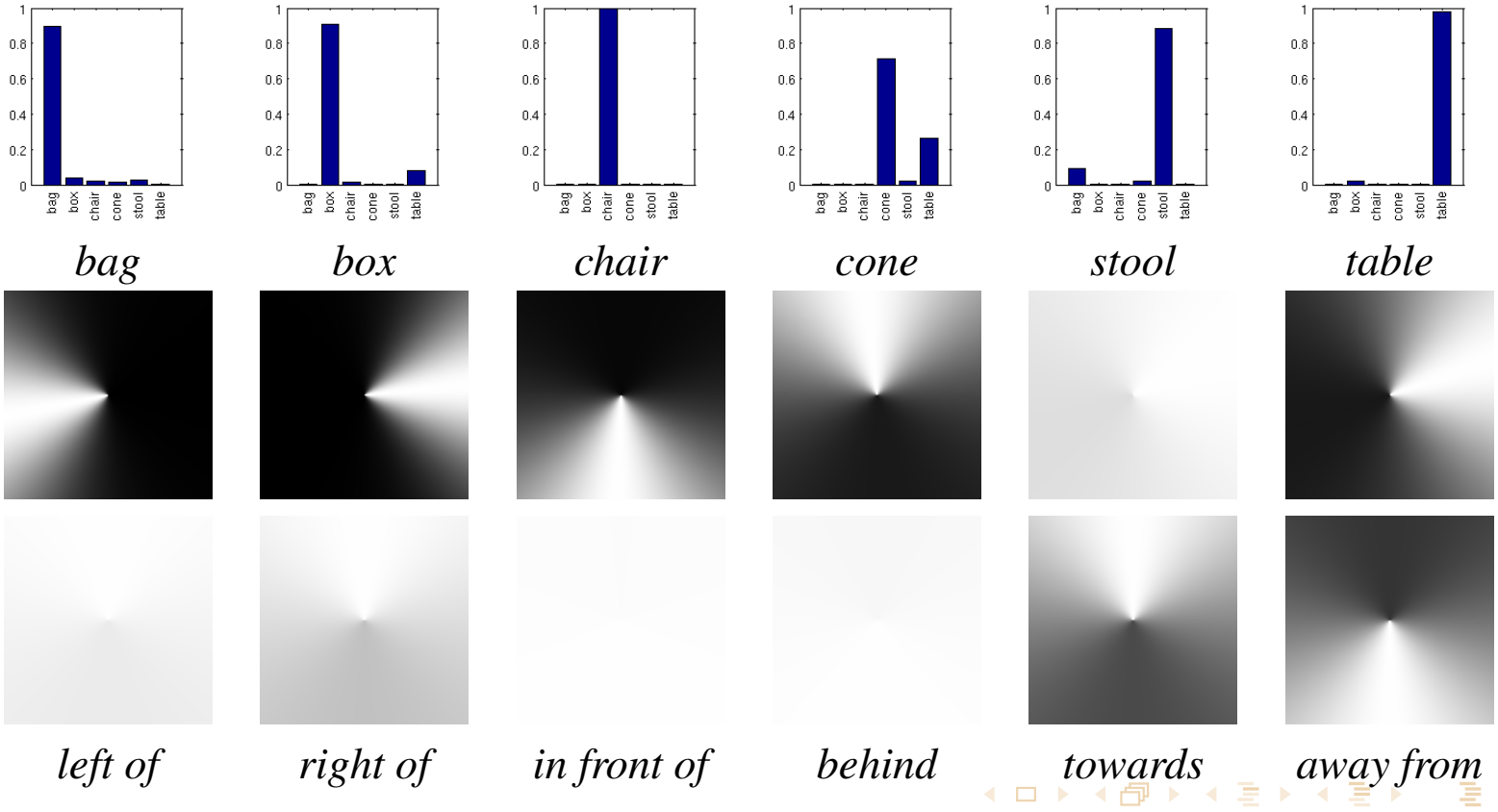


Acquisition

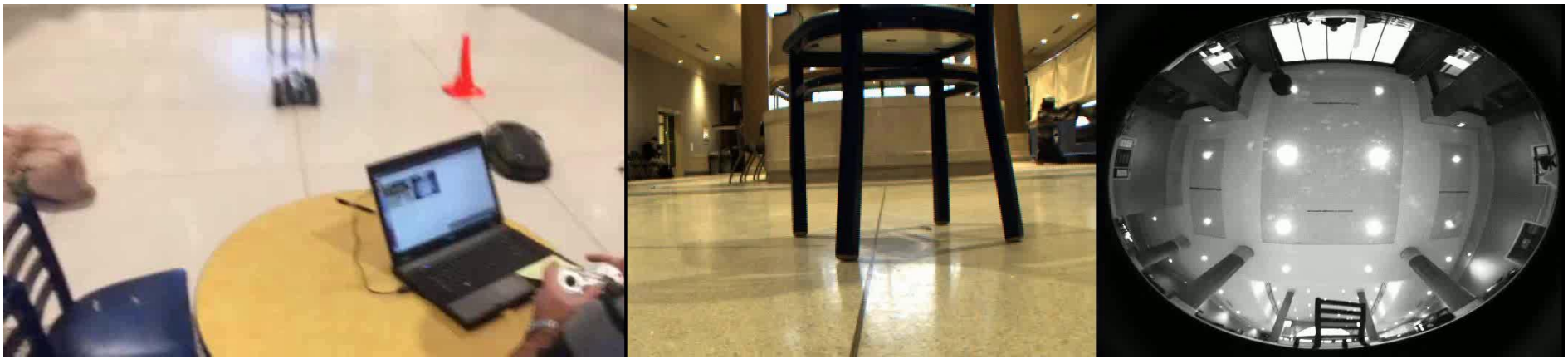
input: *The robot began heading towards the right of the bag on the right, once behind the bag on the right, it turned around, went to the left of the bag on the right and headed towards the chair, ending in front of the chair.*
. . . (749 more)



output:

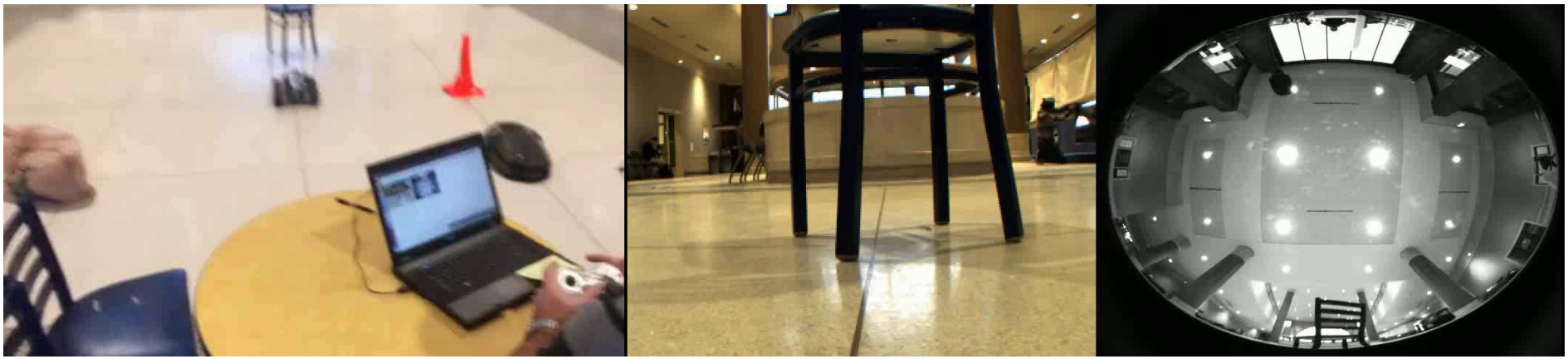


Generation



The robot went in front of the chair then went away from the chair and behind the cone then went right of the bag which is left of the cone then went left of the bag which is in front of the cone then went away from the cone and away from the chair.

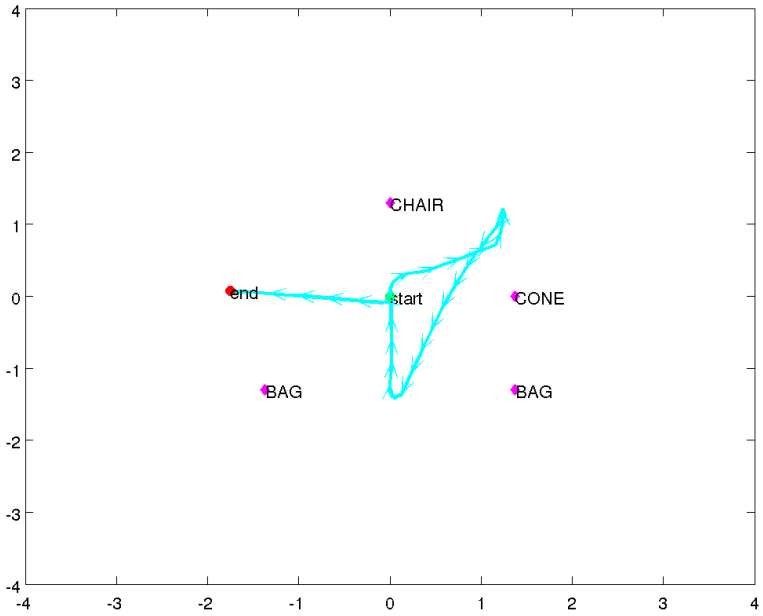
Generation



The robot went in front of the chair then went away from the chair and behind the cone then went right of the bag which is left of the cone then went left of the bag which is in front of the cone then went away from the cone and away from the chair.

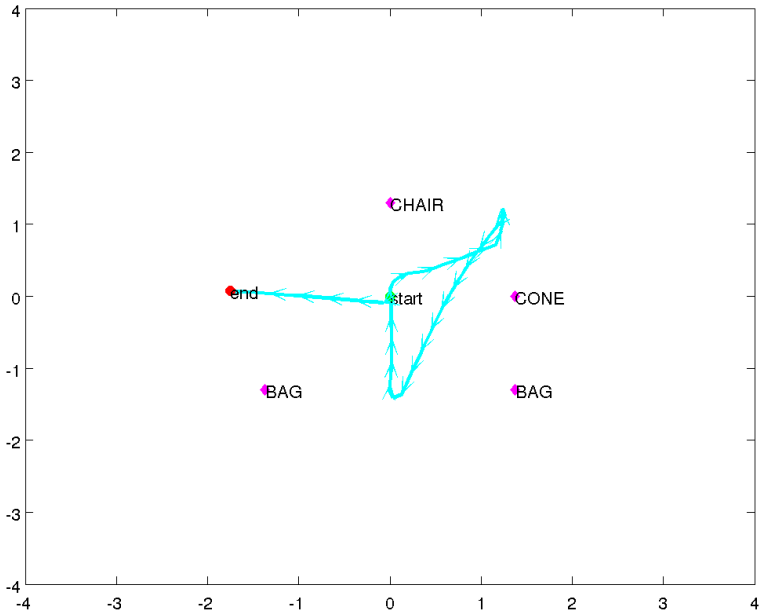
Generation

input:



Generation

input:



output: *The robot went in front of the chair then went away from the chair and behind the cone then went right of the bag which is left of the cone then went left of the bag which is in front of the cone then went away from the cone and away from the chair.*

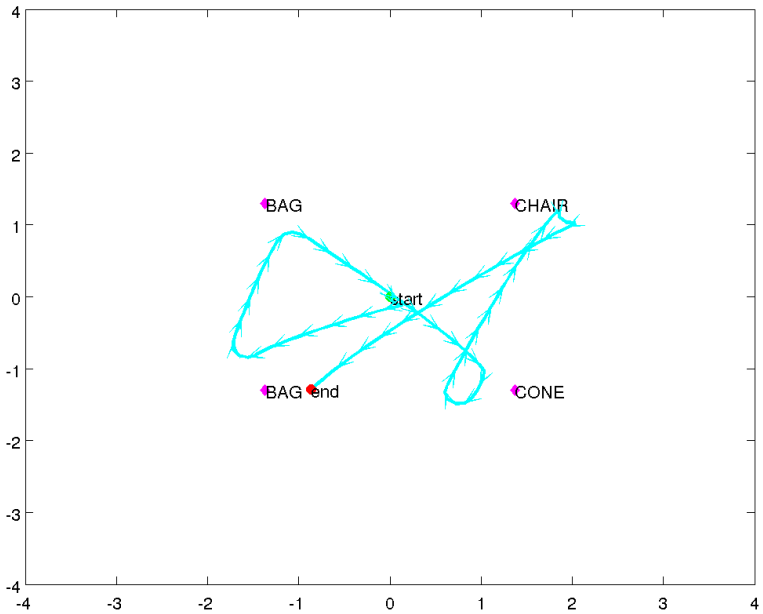
Comprehension

input: *The robot went behind the bag which is in front of the bag then went in front of the bag which is left of the chair then went towards the cone then went away from the chair then went right of the chair then went right of the bag which is left of the cone.*

Comprehension

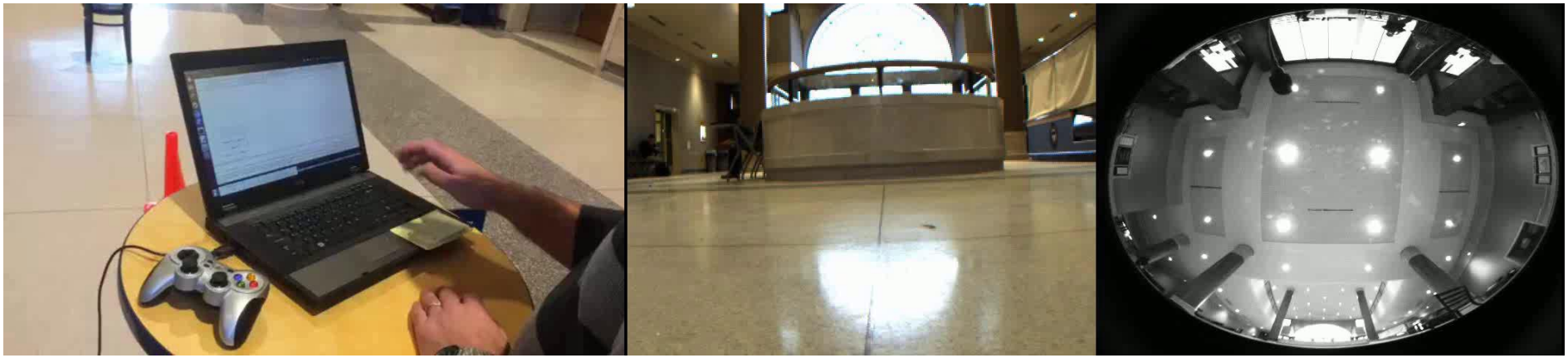
input: *The robot went behind the bag which is in front of the bag then went in front of the bag which is left of the chair then went towards the cone then went away from the chair then went right of the chair then went right of the bag which is left of the cone.*

output:



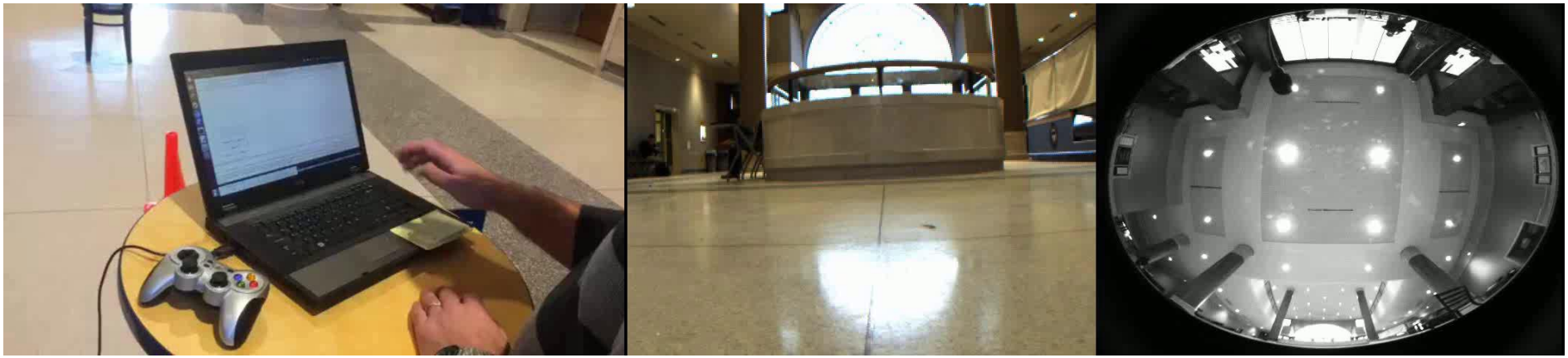
Comprehension

The robot went behind the bag which is in front of the bag then went in front of the bag which is left of the chair then went towards the cone then went away from the chair then went right of the chair then went right of the bag which is left of the cone.

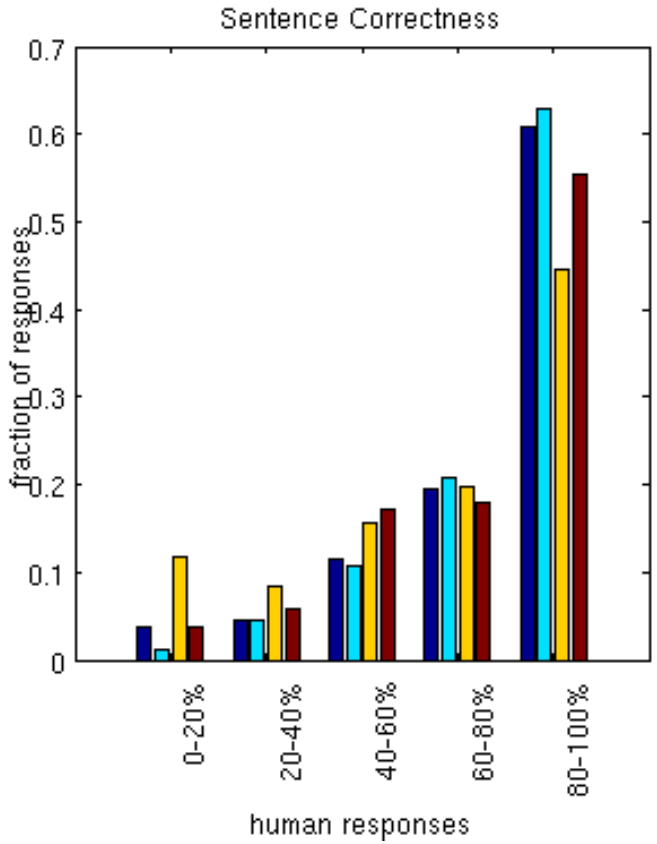


Comprehension

The robot went behind the bag which is in front of the bag then went in front of the bag which is left of the chair then went towards the cone then went away from the chair then went right of the chair then went right of the bag which is left of the cone.

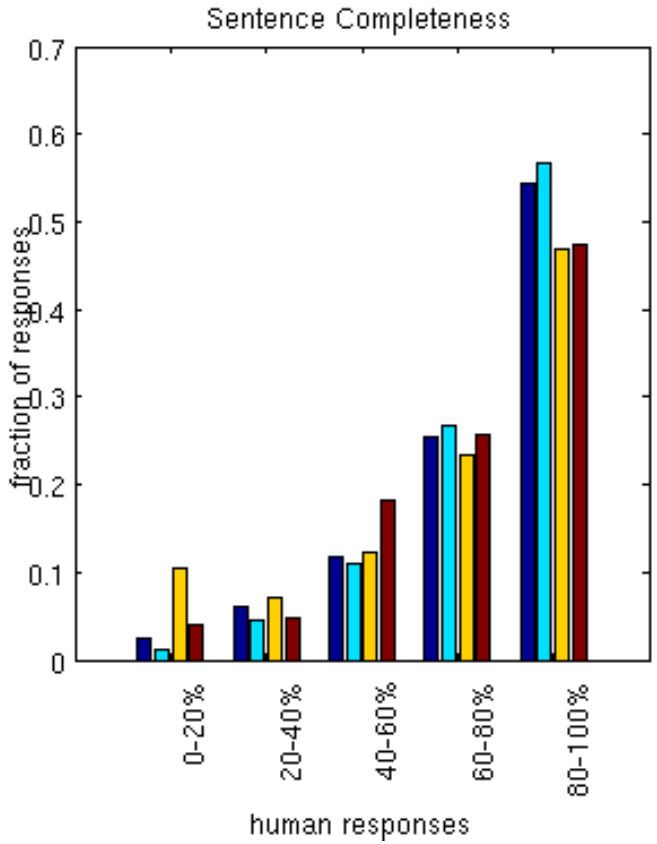


Sentence Correctness



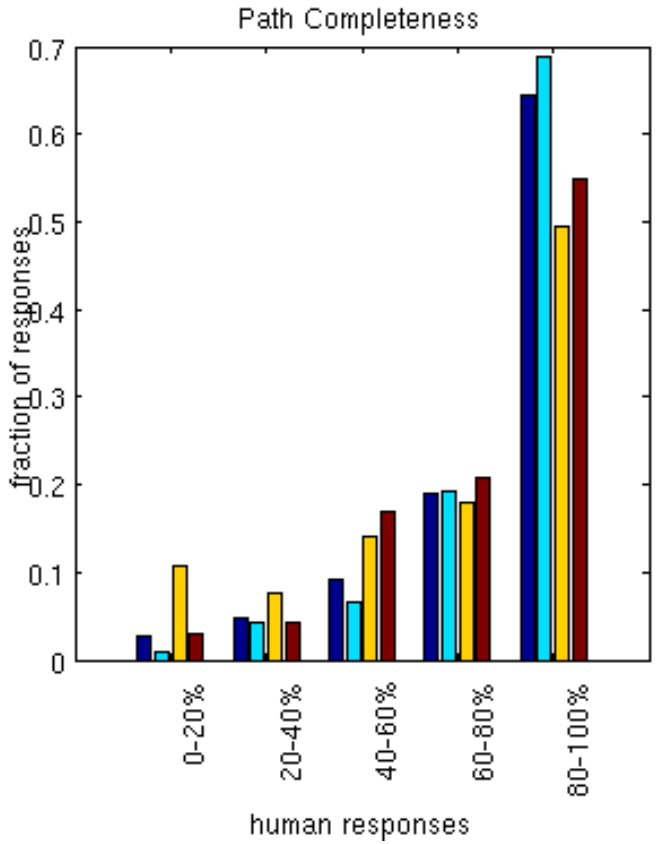
- acquisition (human sentences vs. human driven paths)
- comprehension (human sentences vs. human driven paths)
- comprehension (human sentences vs. machine driven paths)
- generation (machine sentences vs. human driven paths)

Sentence Completeness



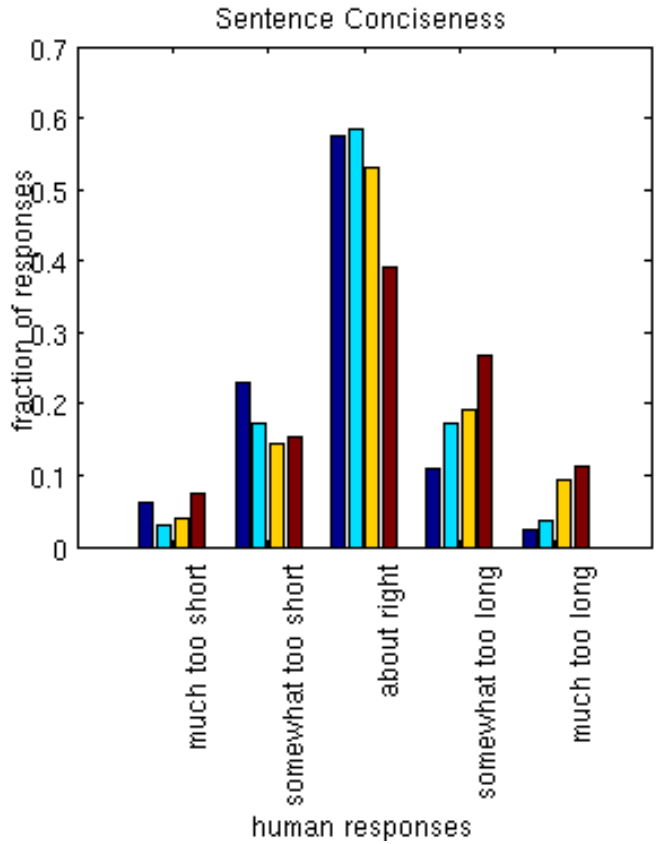
- acquisition (human sentences vs. human driven paths)
- comprehension (human sentences vs. human driven paths)
- comprehension (human sentences vs. machine driven paths)
- generation (machine sentences vs. human driven paths)

Path Completeness



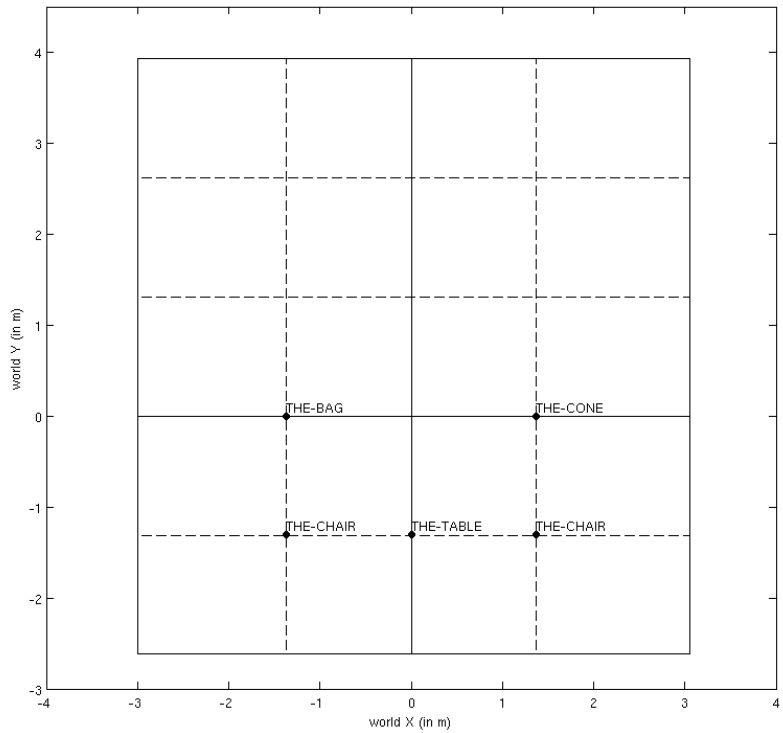
- acquisition (human sentences vs. human driven paths)
- comprehension (human sentences vs. human driven paths)
- comprehension (human sentences vs. machine driven paths)
- generation (machine sentences vs. human driven paths)

Sentence Conciseness



- acquisition (human sentences vs. human driven paths)
- comprehension (human sentences vs. human driven paths)
- comprehension (human sentences vs. machine driven paths)
- generation (machine sentences vs. human driven paths)

Using Codetection to Recover Floorplans



1. Codetection objects in video from robot's perspective
2. Use SFM and odometry to map codetections to world coordinates
3. Cluster world coordinates of codetections
4. Learn mapping from coordinate clusters to noun labels

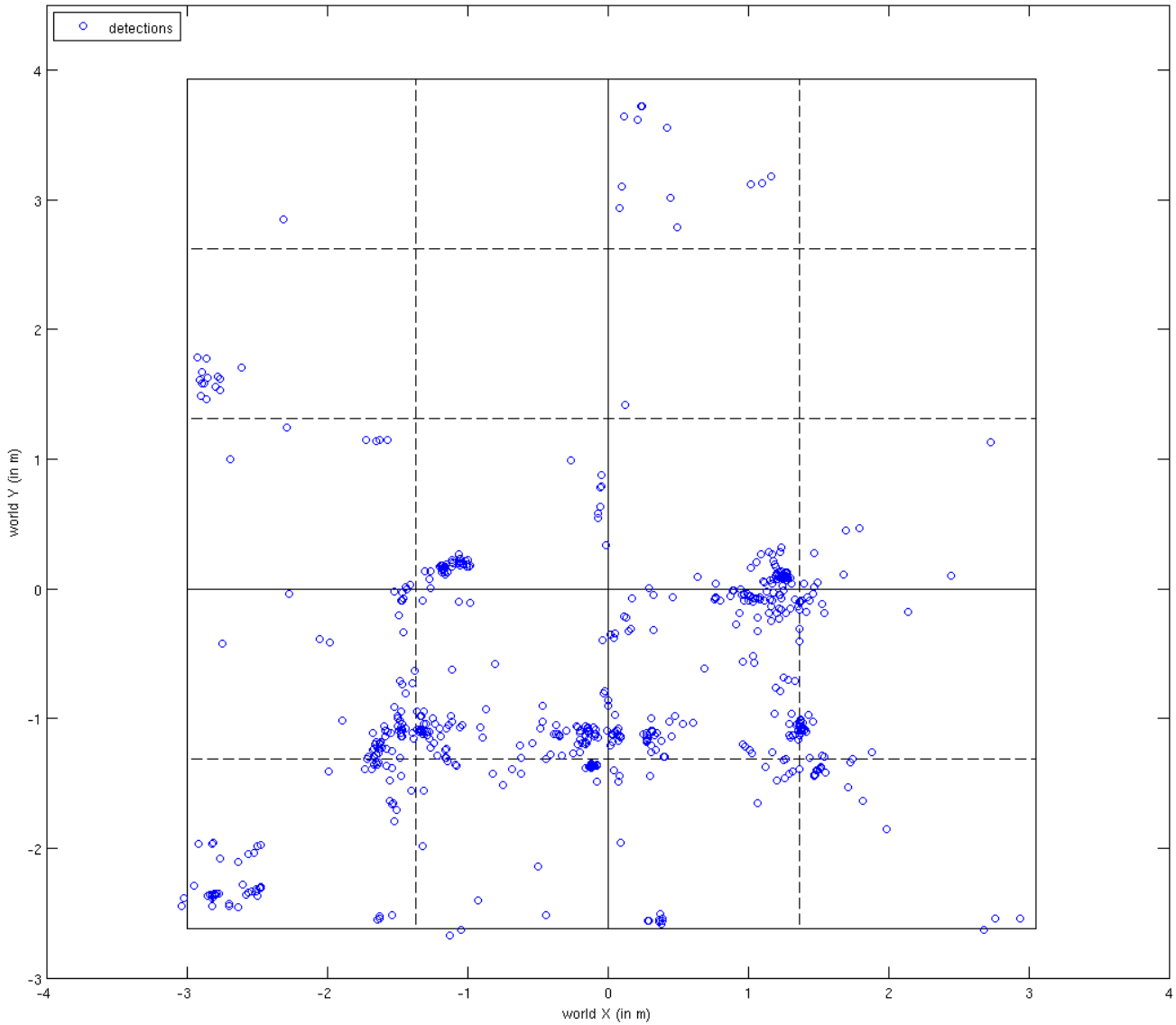
Using Codetection to Recover Floorplans

1. Generate proposals with MCG (Arbelaez et al. 2014)
2. Graphical model
    1. vertex for object in each frame, plus dummy
    2. clique for several adjacent frames
    3. proposal score as vertex score, penalized by implausibility of world size and position recovered with SFM
    4. edge score is weighted sum of
        1. similarity of SIFT descriptors
        2. similarity of world size and position as determined by SFM

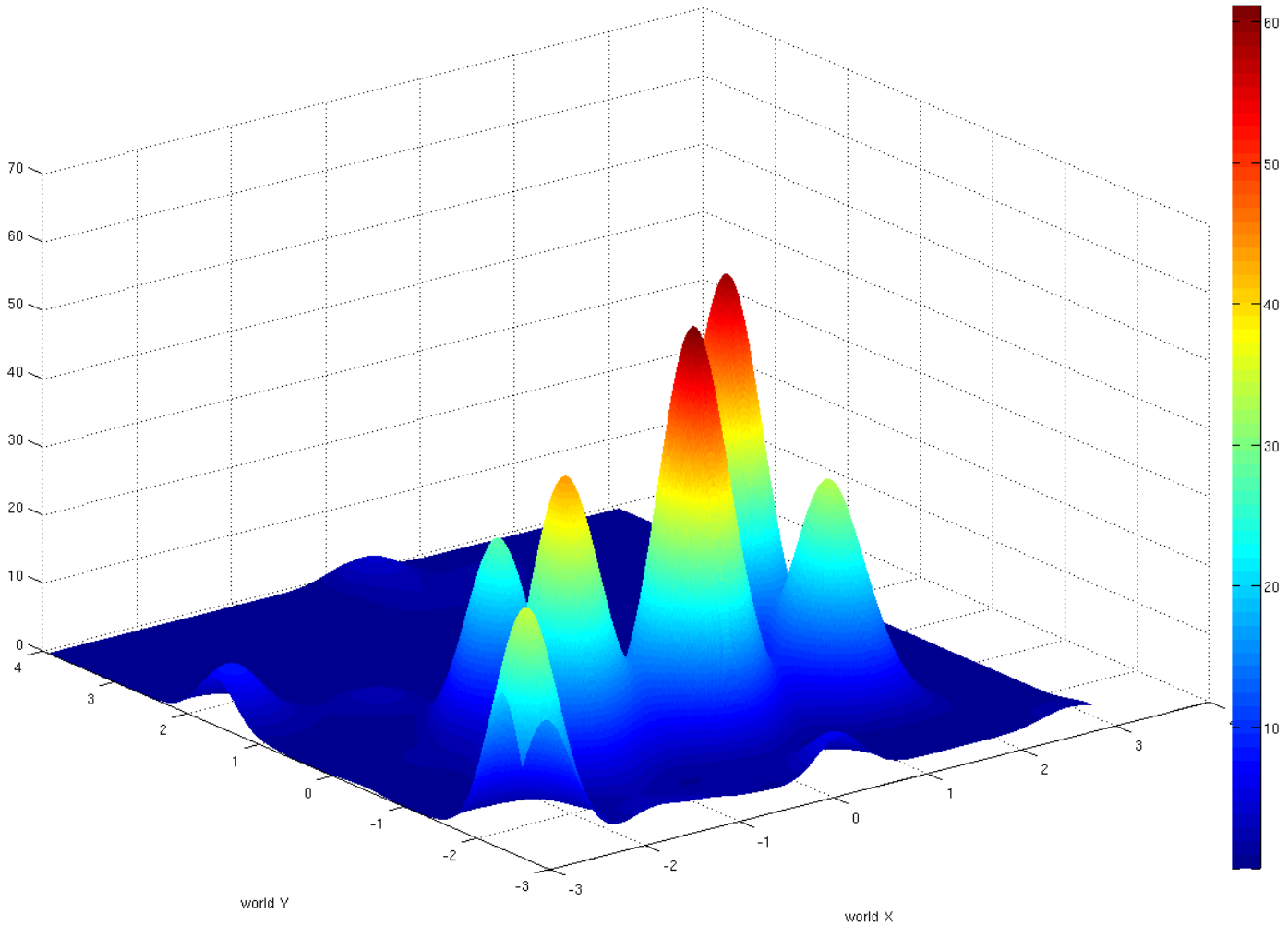
Using Codetection to Recover Floorplans

Using Codetection to Recover Floorplans

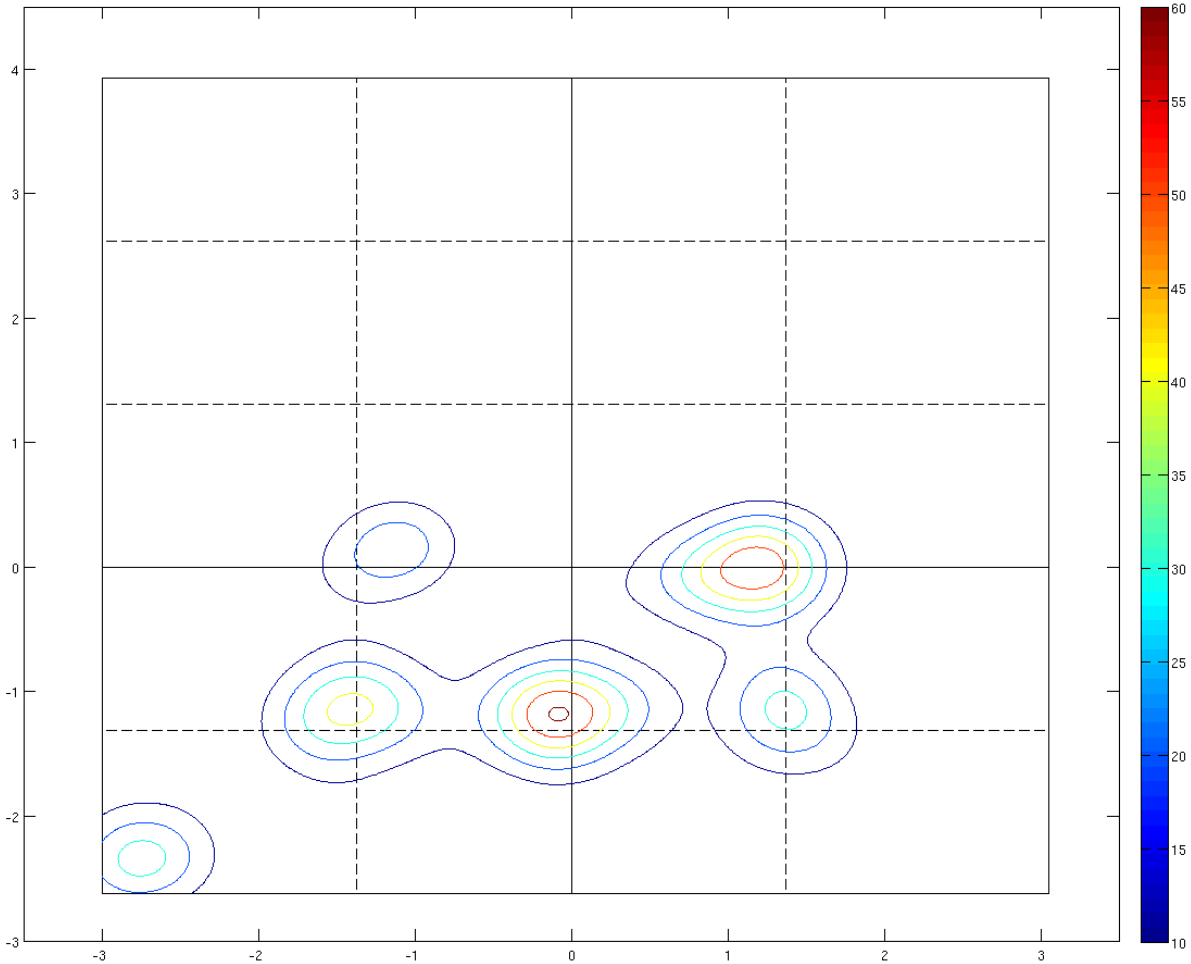
Using Codetection to Recover Floorplans



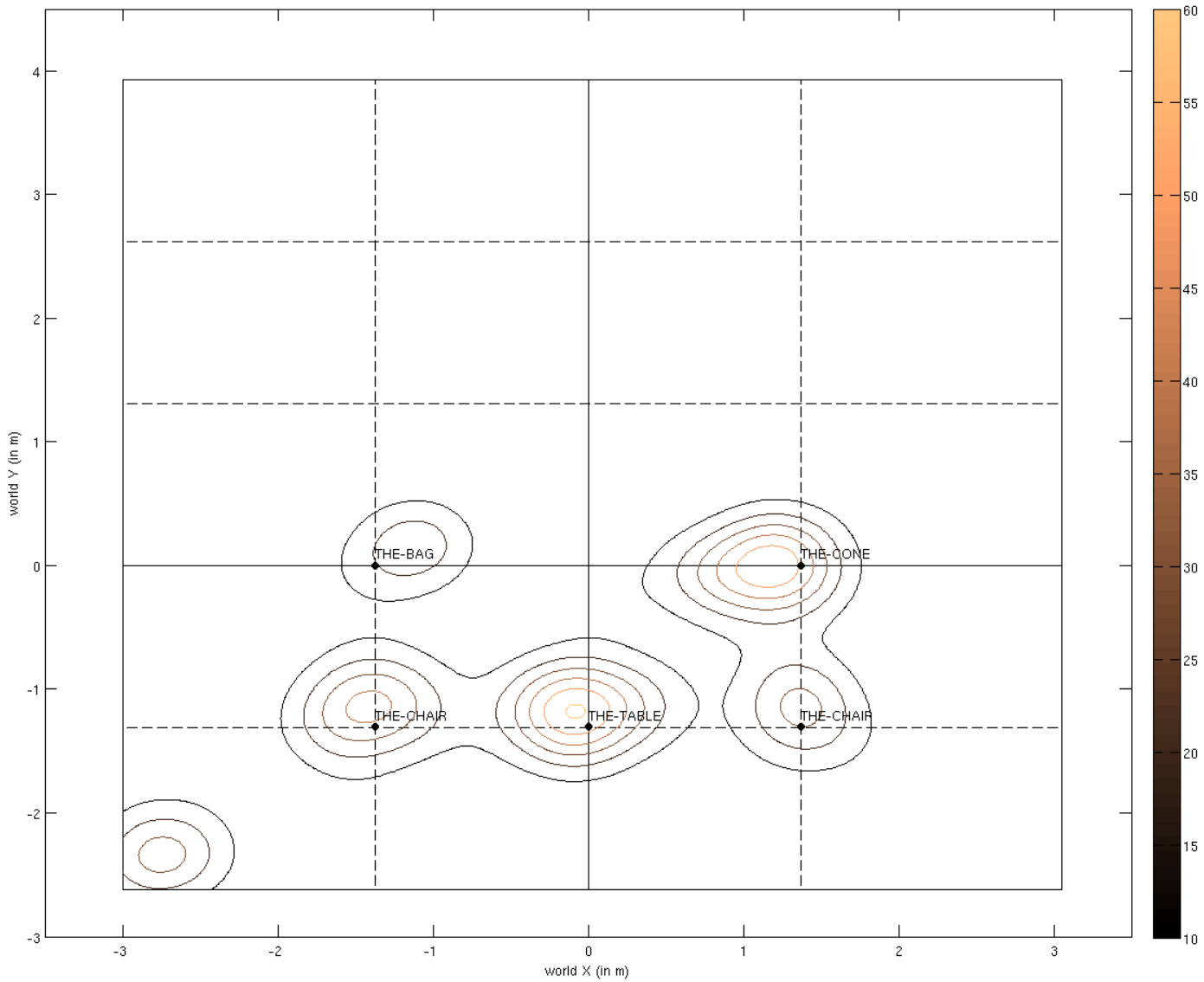
Using Codetection to Recover Floorplans



Using Codetection to Recover Floorplans



Using Codetection to Recover Floorplans



Outline

One Slide Tells it All

video

One Slide Tells it All

video

computer

One Slide Tells it All


video


computer

≈50%

accuracy

One Slide Tells it All


video


computer

≈50%

accuracy


video

One Slide Tells it All


video computer

≈50%

accuracy


video subject

One Slide Tells it All


video computer

$\approx 50\%$

accuracy



video subject fMRI

One Slide Tells it All


video


computer

≈50%

accuracy


video


subject


fMRI


computer

One Slide Tells it All


video computer → ≈50%

accuracy




video subject fMRI computer → ≈70%

accuracy

Stimuli

Barbu, A., Barrett, D.P., Chen, W., Siddharth, N., Xiong, C., Corso, J.J., Fellbaum, C.D., Hanson, C., Hanson, S.J., Hélie, S., Malaia, E., Pearlmutter, B.A., Siskind, J.M., Talavage, T.M., and Wilbur, R.B., 'Seeing is Worse than Believing: Reading People's Minds Better than Computer-Vision Methods Recognize Actions,' ECCV 2014.

Stimuli

Barbu, A., Barrett, D.P., Chen, W., Siddharth, N., Xiong, C., Corso, J.J., Fellbaum, C.D., Hanson, C., Hanson, S.J., Hélie, S., Malaia, E., Pearlmutter, B.A., Siskind, J.M., Talavage, T.M., and Wilbur, R.B., 'Seeing is Worse than Believing: Reading People's Minds Better than Computer-Vision Methods Recognize Actions,' ECCV 2014.

Stimuli

$$\Longrightarrow \left\{ \begin{array}{l} \textit{carry}, \\ \textit{dig}, \\ \textit{hold}, \\ \textit{pick up}, \\ \textit{put down}, \\ \textit{walk} \end{array} \right.$$

Barbu, A., Barrett, D.P., Chen, W., Siddharth, N., Xiong, C., Corso, J.J., Fellbaum, C.D., Hanson, C., Hanson, S.J., Hélie, S., Malaia, E., Pearlmutter, B.A., Siskind, J.M., Talavage, T.M., and Wilbur, R.B., 'Seeing is Worse than Believing: Reading People's Minds Better than Computer-Vision Methods Recognize Actions,' ECCV 2014.

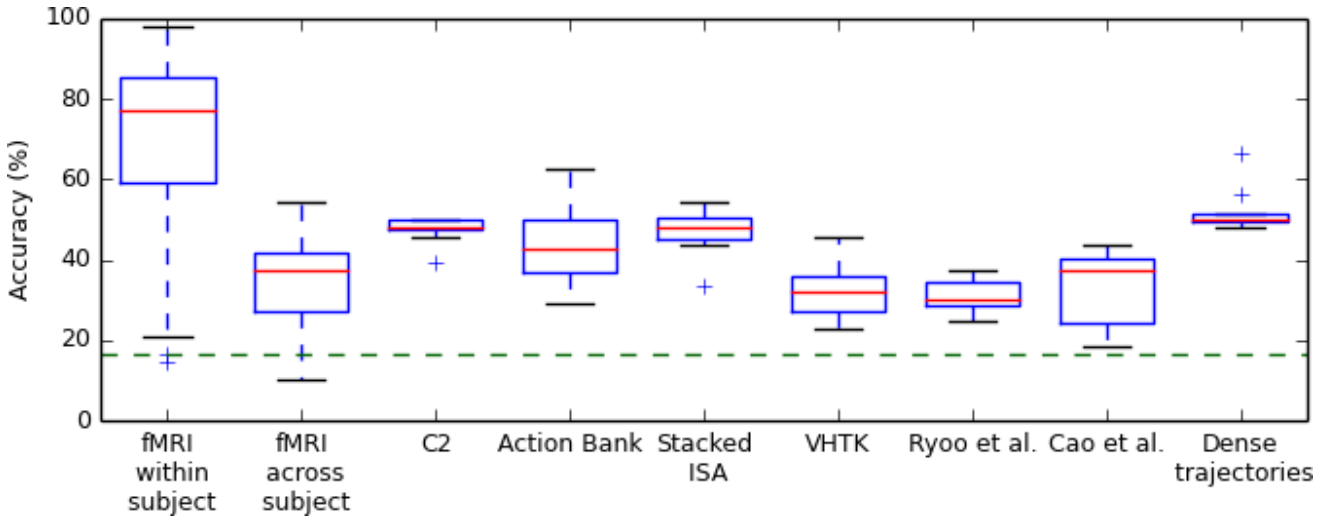
Stimuli

$$\Longrightarrow \begin{cases} \textit{carry}, \\ \textit{dig}, \\ \textit{hold}, \\ \textit{pick up}, \\ \textit{put down}, \\ \textit{walk} \end{cases}$$

Barbu, A., Barrett, D.P., Chen, W., Siddharth, N., Xiong, C., Corso, J.J., Fellbaum, C.D., Hanson, C., Hanson, S.J., Hélie, S., Malaia, E., Pearlmutter, B.A., Siskind, J.M., Talavage, T.M., and Wilbur, R.B., 'Seeing is Worse than Believing: Reading People's Minds Better than Computer-Vision Methods Recognize Actions,' ECCV 2014.

Classification Accuracies

fMRI within subject	69.7%
fMRI cross subject	34.8%
C2	47.4%
Action Bank	44.2%
Stacked ISA	46.8%
VHTK	32.5%
Ryoo et al.	31.2%
Cao et al.	33.3%
Dense Trajectories	52.3%



Stimuli

HOLLYWOOD-2 Marszałek et al. (2009)

Stimuli

HOLLYWOOD-2 Marszałek et al. (2009)

Stimuli

$$\Longrightarrow \left\{ \begin{array}{l} \textit{AnswerPhone}, \\ \textit{DriveCar}, \\ \textit{Eat}, \\ \textit{FightPerson}, \\ \textit{GetOutCar}, \\ \textit{HandShake}, \\ \textit{HugPerson}, \\ \textit{Kiss}, \\ \textit{Run}, \\ \textit{SitDown}, \\ \textit{SitUp}, \\ \textit{StandUp} \end{array} \right.$$

HOLLYWOOD-2 Marszałek et al. (2009)

Stimuli

$\Longrightarrow$

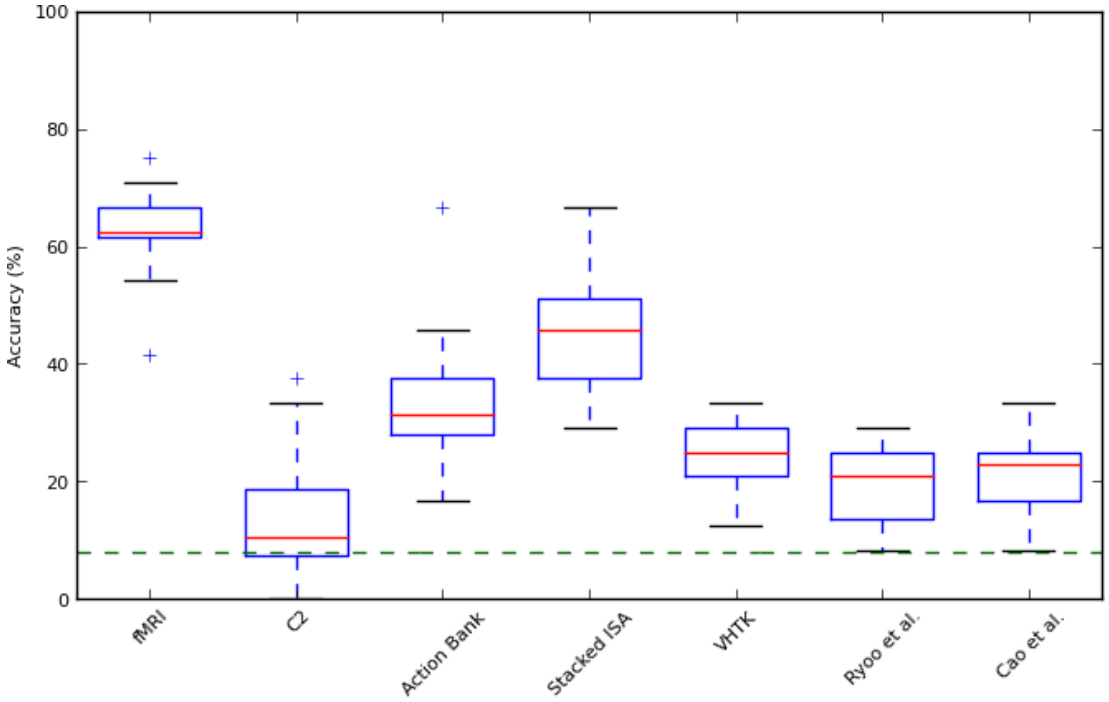
$$\left\{\begin{array}{l} \textit{AnswerPhone}, \\ \textit{DriveCar}, \\ \textit{Eat}, \\ \textit{FightPerson}, \\ \textit{GetOutCar}, \\ \textit{HandShake}, \\ \textit{HugPerson}, \\ \textit{Kiss}, \\ \textit{Run}, \\ \textit{SitDown}, \\ \textit{SitUp}, \\ \textit{StandUp} \end{array}\right.$$

HOLLYWOOD-2 Marszałek et al. (2009)

Classification Accuracies

Subject 1

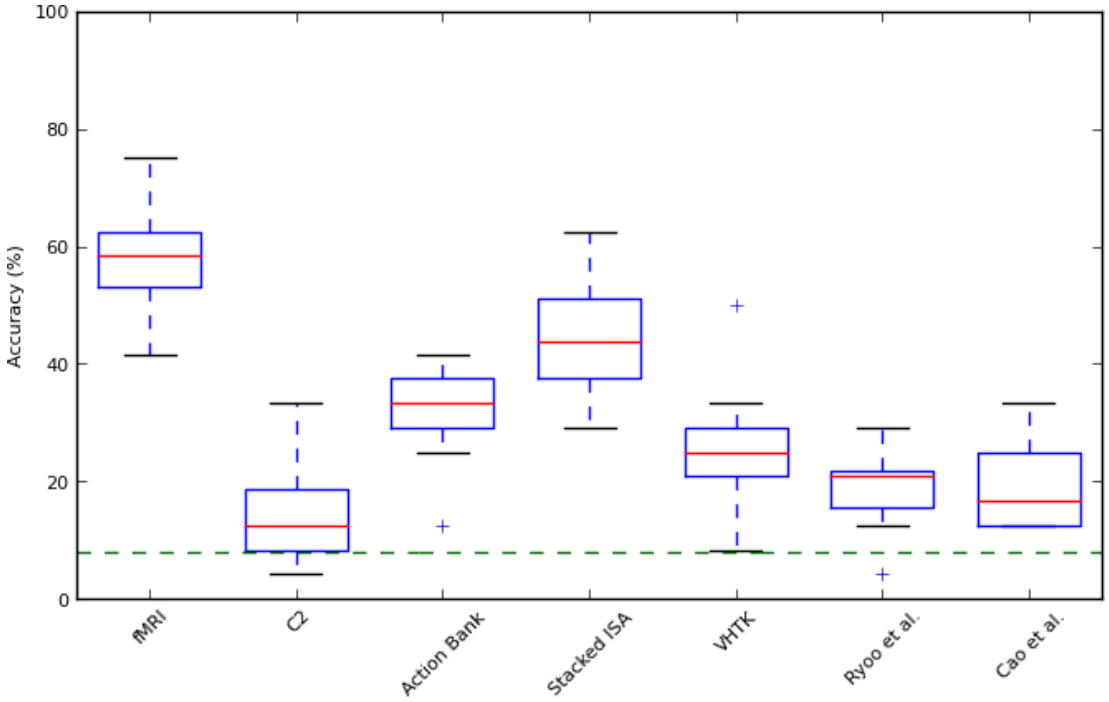
fMRI within subject	69.7%
fMRI cross subject	34.8%
C2	47.4%
Action Bank	44.2%
Stacked ISA	46.8%
VHTK	32.5%
Ryoo et al.	31.2%
Cao et al.	33.3%
Dense Trajectories	52.3%



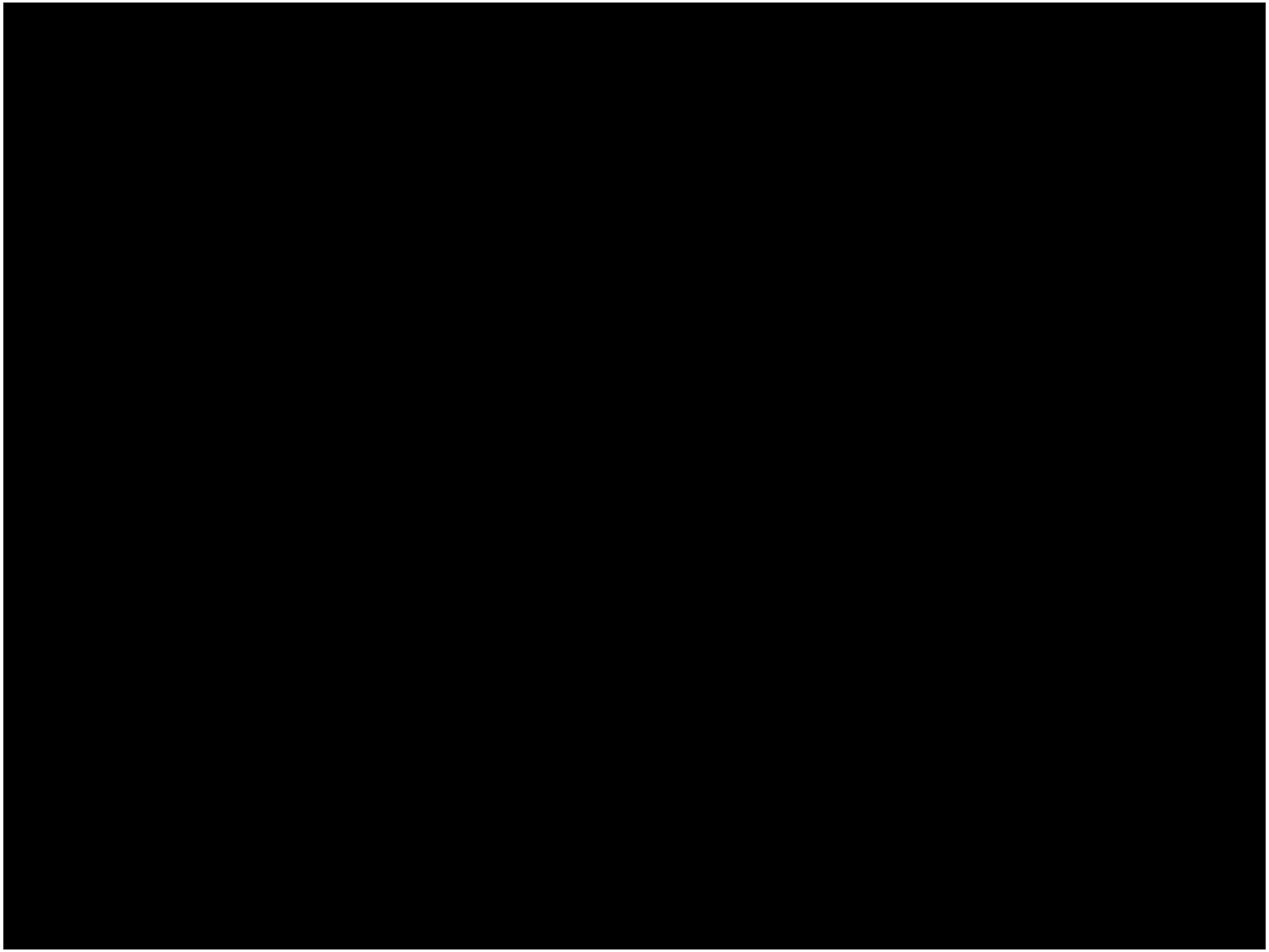
Classification Accuracies

Subject 2

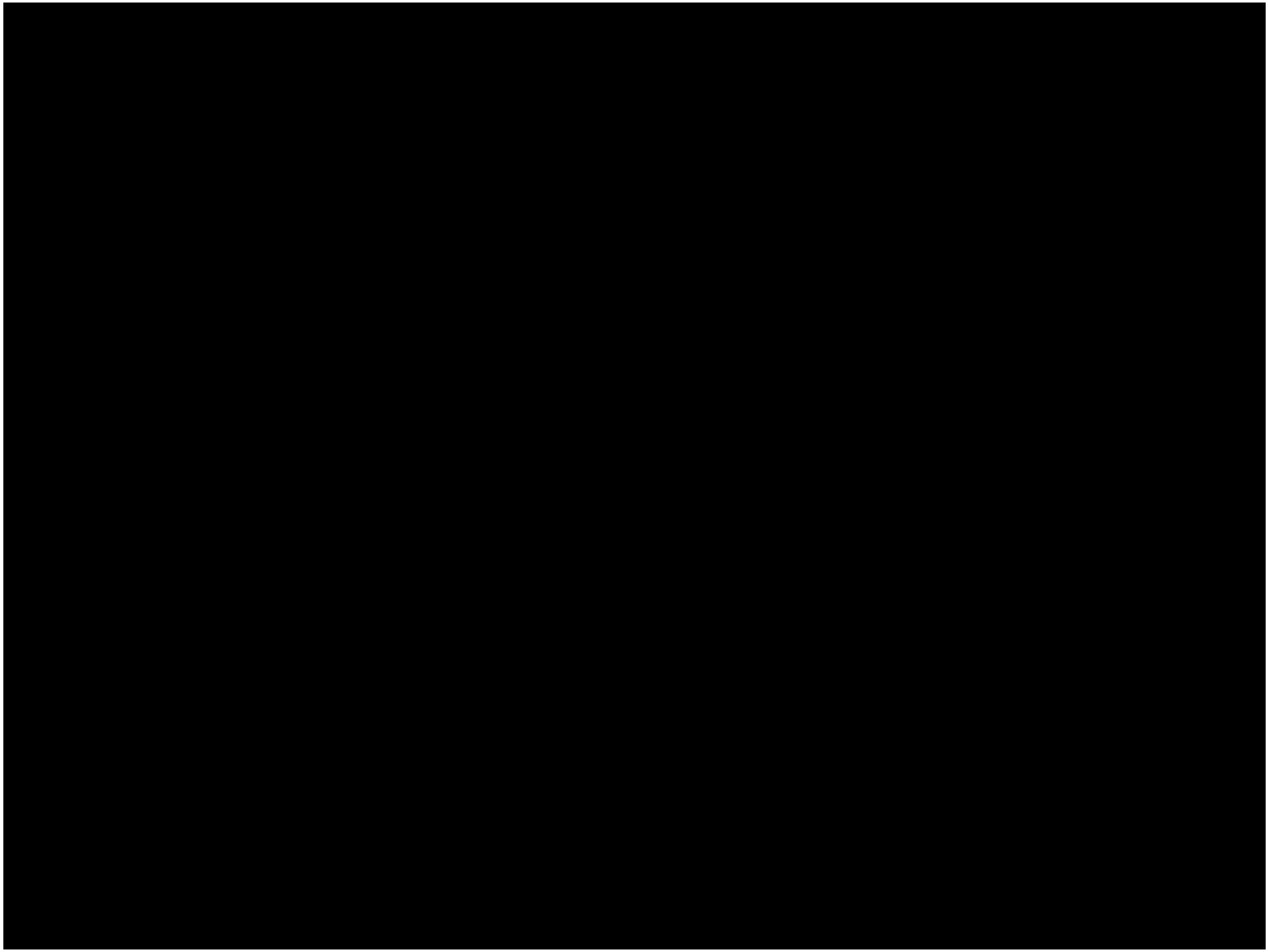
fMRI within subject	69.7%
fMRI cross subject	34.8%
C2	47.4%
Action Bank	44.2%
Stacked ISA	46.8%
VHTK	32.5%
Ryoo et al.	31.2%
Cao et al.	33.3%
Dense Trajectories	52.3%



Stimuli



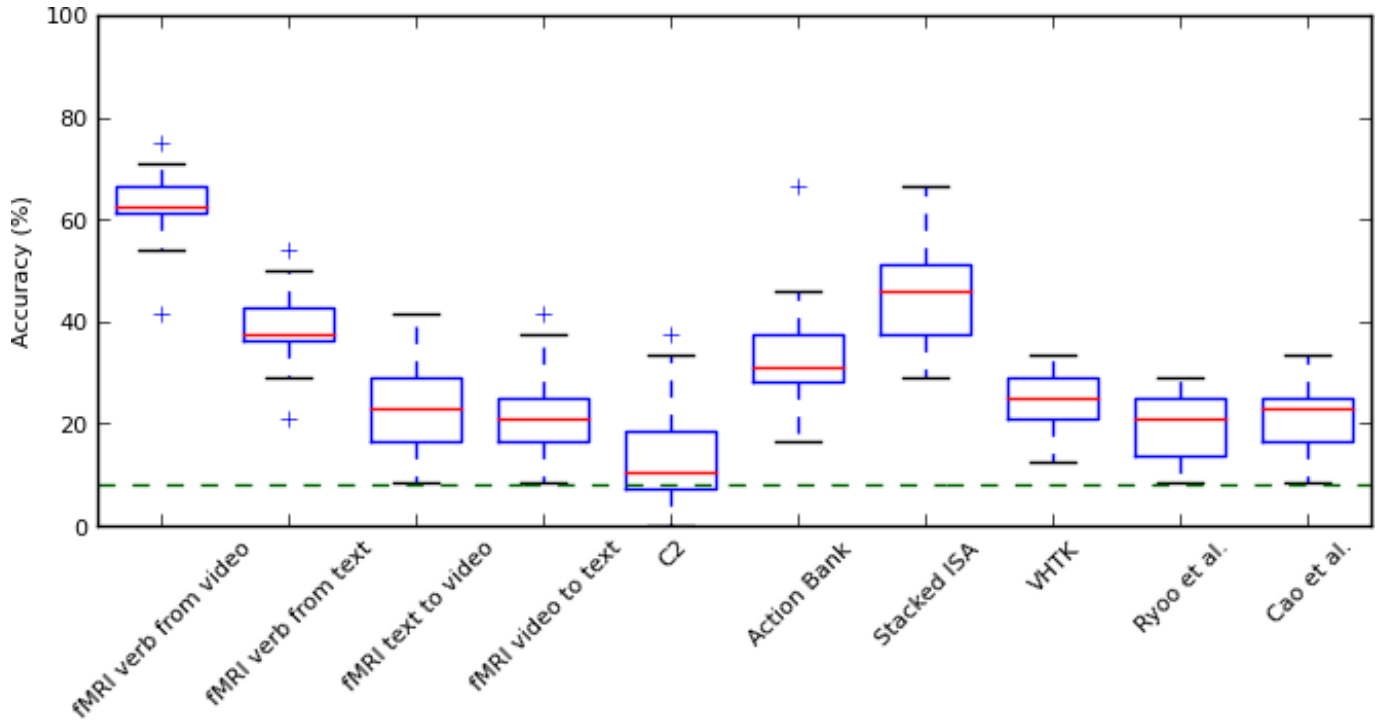
Stimuli



Classification Accuracies

Subject 1

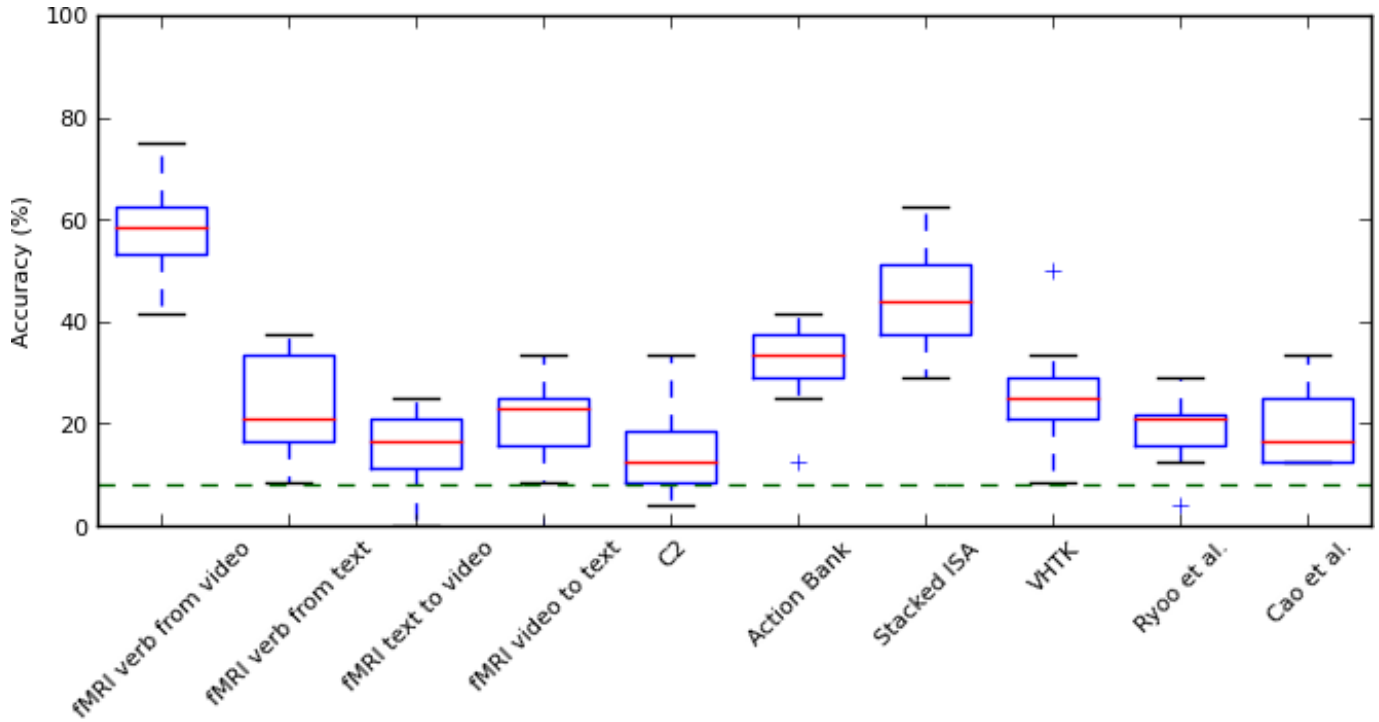
modality	100.0%
verb-from-video	62.7%
verb-from-text	39.0%
verb	53.1%
verb-modality	43.8%
text-to-video	23.4%
video-to-text	22.9%
C2	13.8%
Action Bank	33.8%
Stacked ISA	44.7%
VHTK	25.0%
Ryoo et al.	20.3%
Cao et al.	21.0%



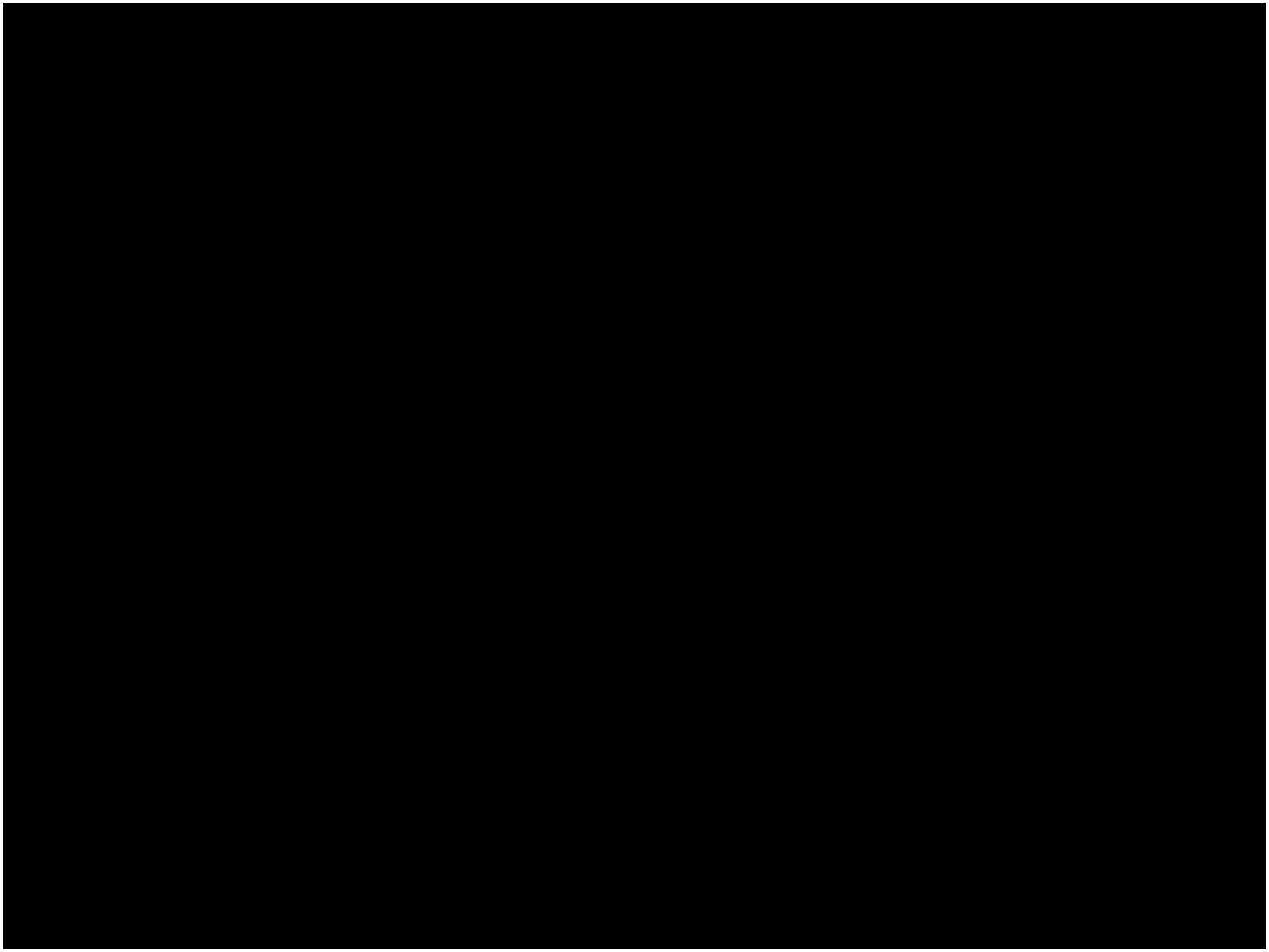
Classification Accuracies

Subject 2

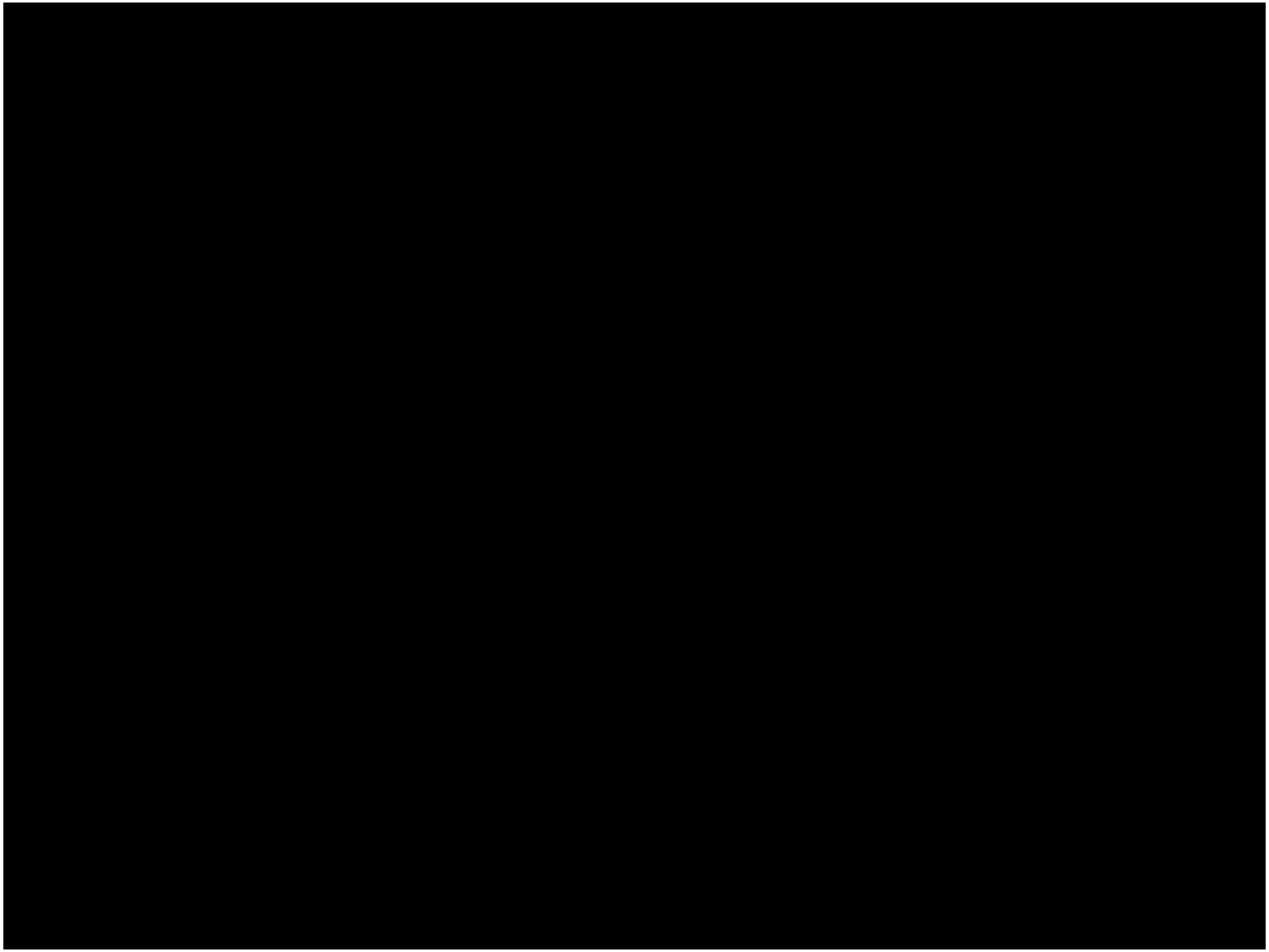
modality	100.0%
verb-from-video	57.8%
verb-from-text	23.6%
verb	37.8%
verb-modality	39.8%
text-to-video	15.6%
video-to-text	20.0%
C2	15.1%
Action Bank	32.5%
Stacked ISA	44.7%
VHTK	25.0%
Ryoo et al.	18.7%
Cao et al.	18.7%



Stimuli



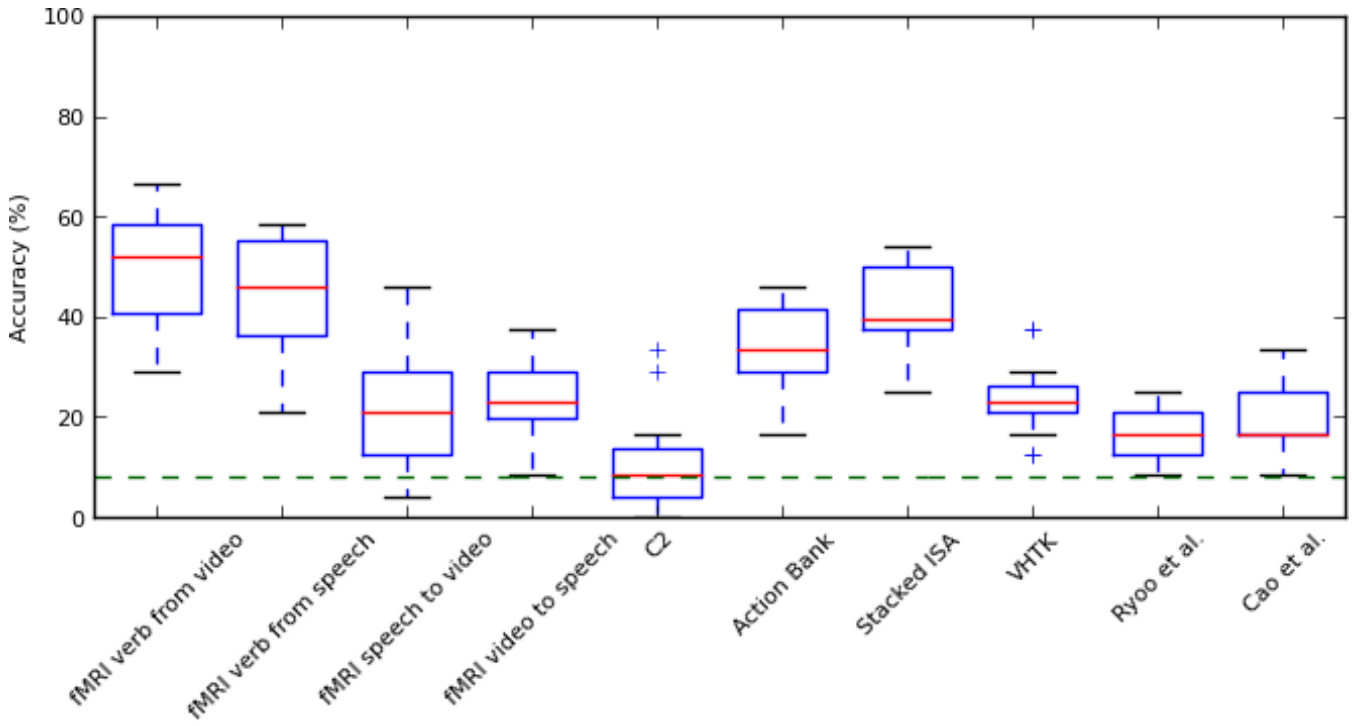
Stimuli



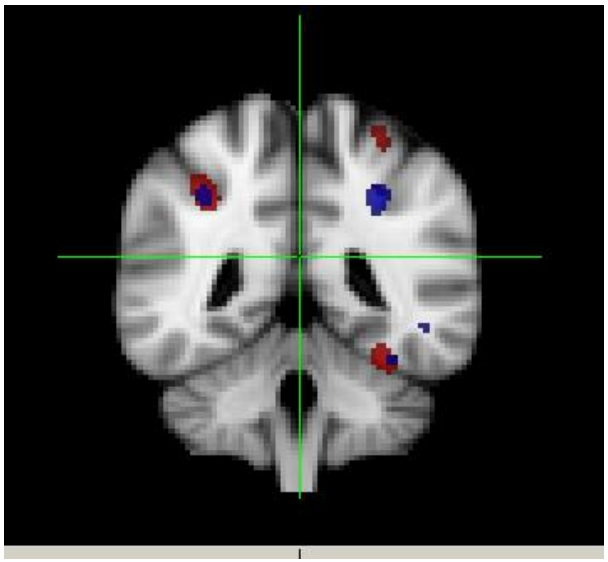
Classification Accuracies

Subject 1

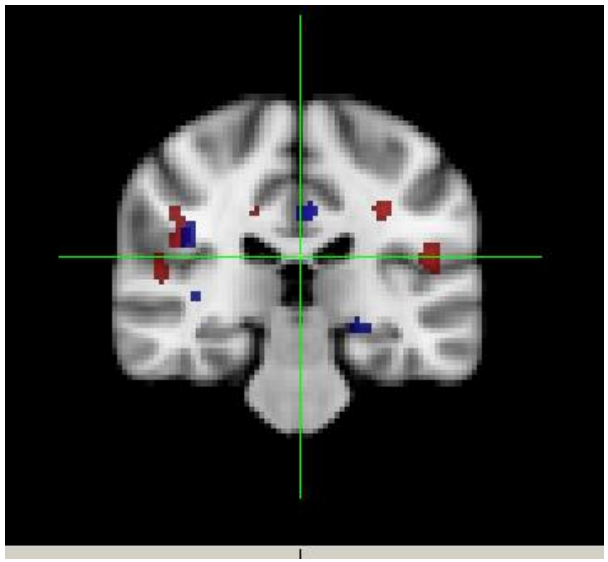
modality	100.0%
verb-from-video	49.2%
verb-from-speech	45.0%
verb	45.0%
verb-modality	43.0%
speech-to-video	21.0%
video-to-speech	23.6%
C2	10.9%
Action Bank	34.1%
Stacked ISA	42.1%
VHTK	23.6%
Ryoo et al.	17.4%
Cao et al.	19.2%



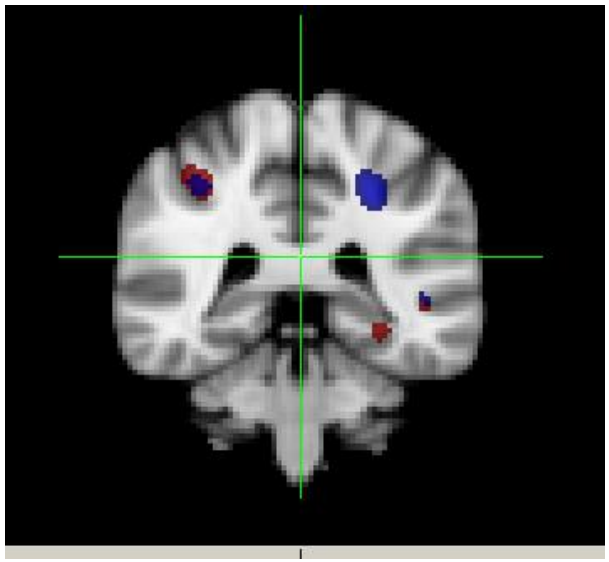
Brain Regions for Vision and Language



Subject 1 text



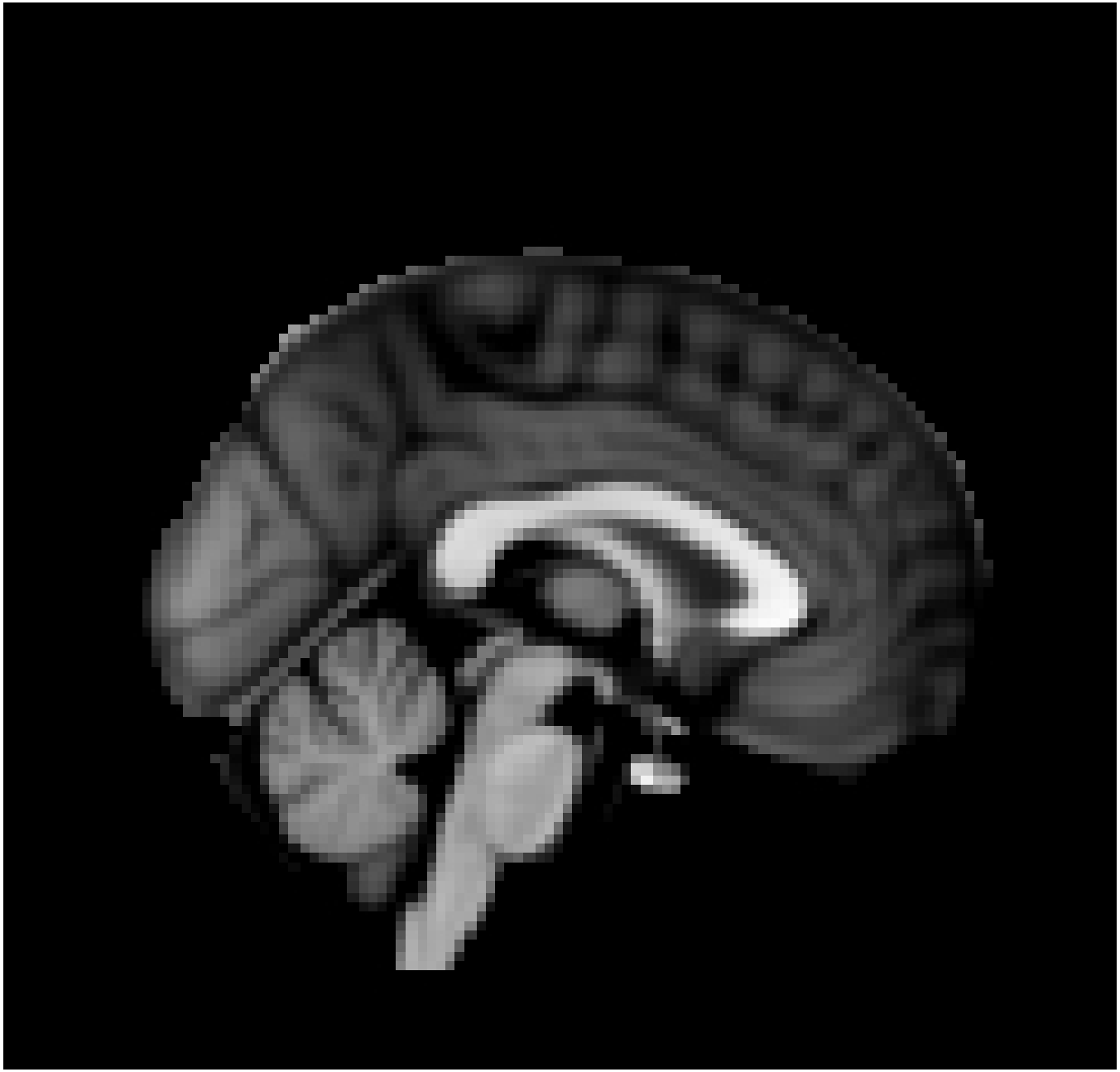
Subject 2 text

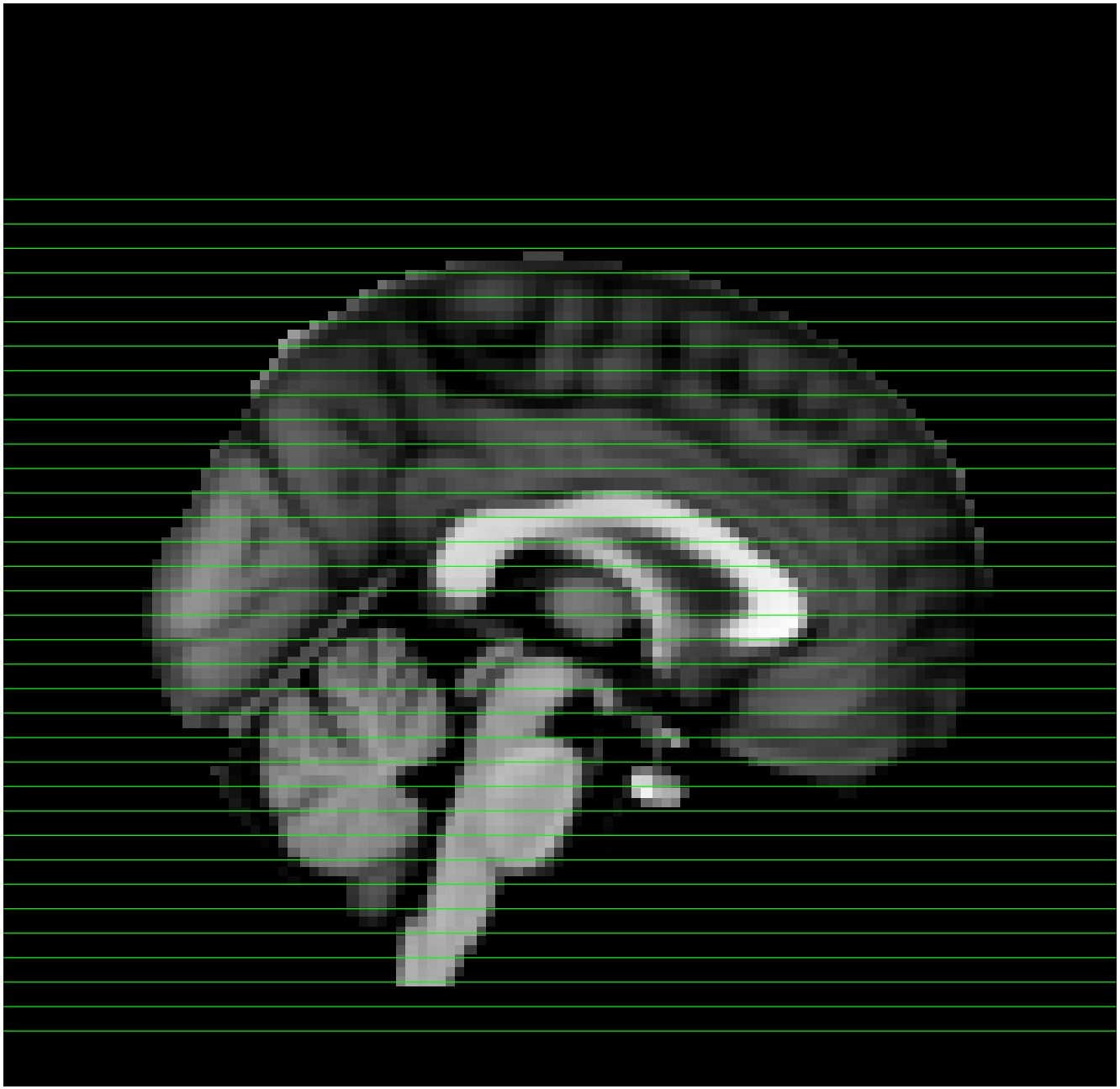


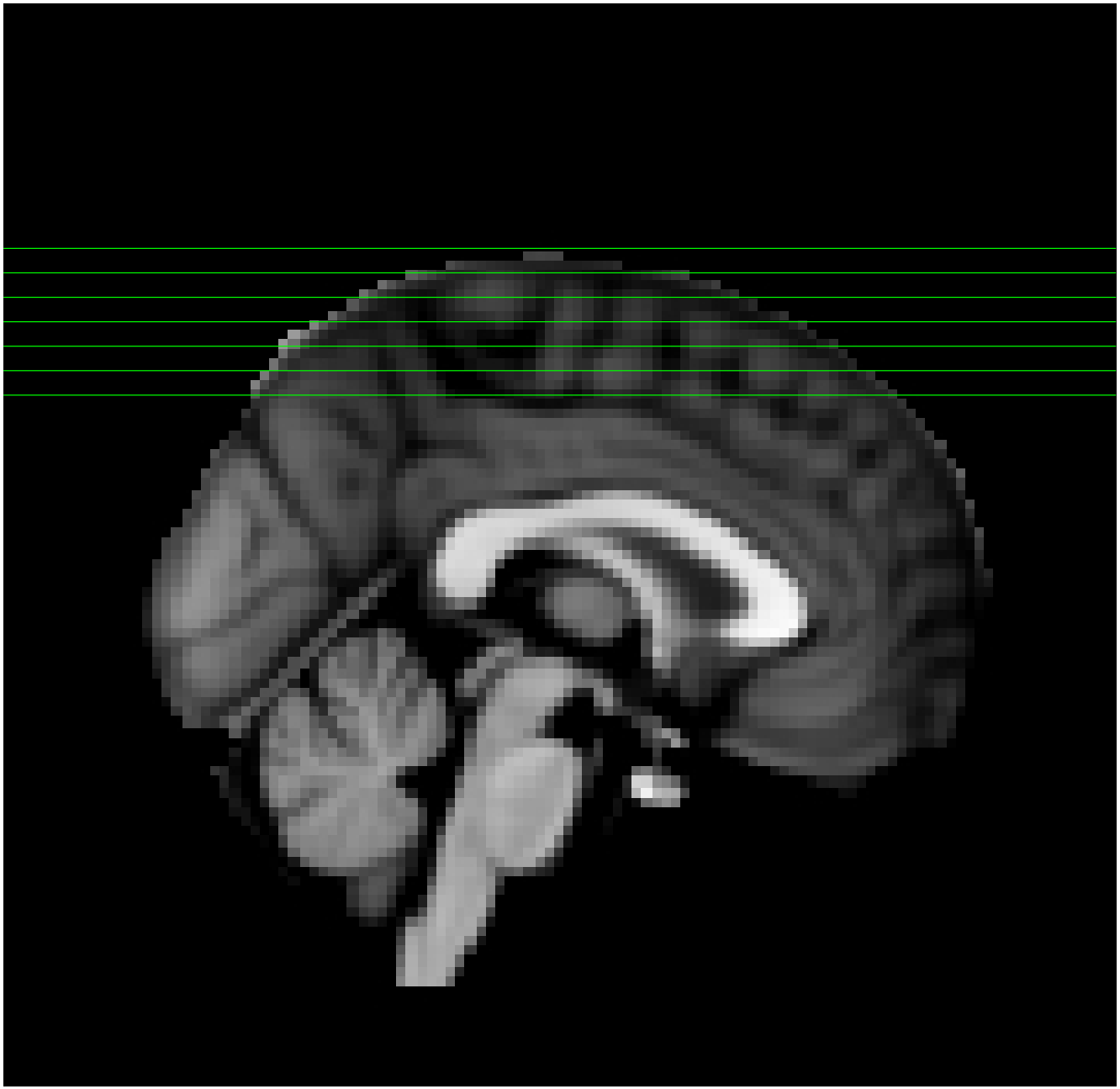
Subject 1 speech

video

text/speech

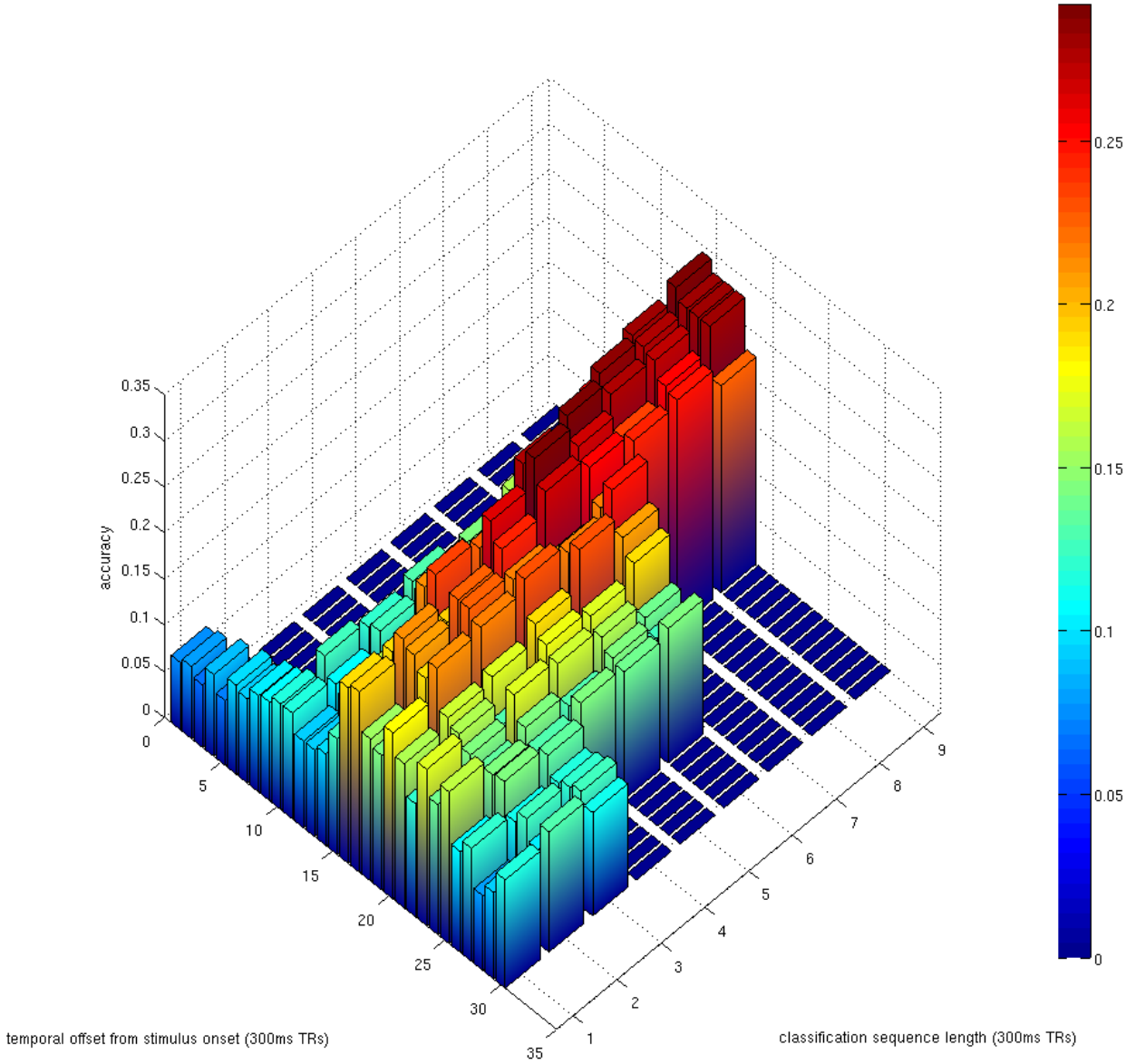










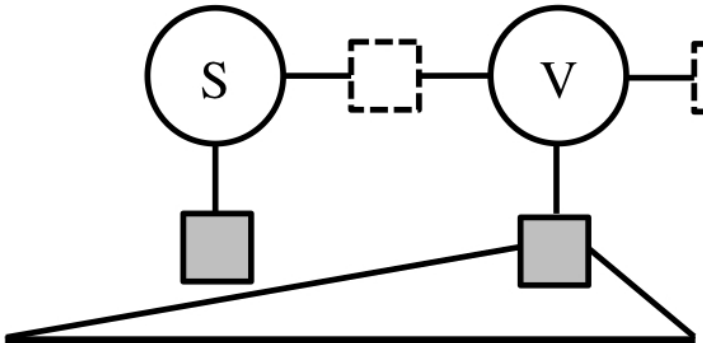


accuracy
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
35
temporal offset from stimulus onset (300ms TRs)
1
2
3
4
5
6
7
8
9
classification sequence length (300ms TRs)
0.25
0.2
0.15
0.1
0.05
0

Video Captioning

How do modern video-captioning systems work?

We use [. . .] priors learned from web-scale natural language corpora to penalize unlikely combinations of actors/actions/objects

Guadarrama et al. (2013)



Activity	Vision Confidence
slice	0.1909
chop	0.1098
play	0.0856
⋮	⋮
speak	0.0000

Entity	Vision Confidence
egg	0.3108
onion	0.2145
potato	0.2061
⋮	⋮
piano	0.0000

Thomason et al. (2014)

Does the brain condition perception of constituents with a joint model?

$$\left\{\begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array}\right\} \times \left\{\begin{array}{l} carried \\ folded \\ left \end{array}\right\} \times \left\{\begin{array}{l} the\ chair \\ the\ shirt \\ the\ tortilla \end{array}\right\} \times \left\{\begin{array}{l} [on\ the]left[ward] \\ [on\ the]right[ward] \end{array}\right\}$$

$$\left\{\begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array}\right\}$$

actor

$$\left\{ \begin{array}{l} carried \\ folded \\ left \end{array} \right\}$$

verb

$$\left\{ \begin{array}{l} \textit{the chair} \\ \textit{the shirt} \\ \textit{the tortilla} \end{array} \right\}$$

object

$$\left\{ \begin{array}{c} \\ left[ward] \\ right[ward] \end{array} \right\}$$

direction

$$\left\{ \begin{array}{l} [on\ the]left \\ [on\ the]right \end{array} \right\}$$

location

$$\left\{\begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array}\right\} \times \left\{\begin{array}{l} carried \\ folded \\ left \end{array}\right\}$$

actor-verb

$$\left\{\begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array}\right\} \times \left\{\begin{array}{l} the\ chair \\ the\ shirt \\ the\ tortilla \end{array}\right\}$$

actor-object

$$\left\{ \begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array} \right\} \times \left\{ \begin{array}{r} left[ward] \\ right[ward] \end{array} \right\}$$

actor-direction

$$\left\{\begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array}\right\} \times \left\{\begin{array}{l} [on\ the]left \\ [on\ the]right \end{array}\right\}$$

actor-location

$$\left\{ \begin{array}{l} carried \\ folded \\ left \end{array} \right\} \times \left\{ \begin{array}{l} the\ chair \\ the\ shirt \\ the\ tortilla \end{array} \right\}$$

verb-object

$$\left\{ \begin{array}{l} carried \\ \\ left \end{array} \right\} \times \left\{ \begin{array}{l} left[ward] \\ right[ward] \end{array} \right\}$$

verb-direction

$$\left\{ \begin{array}{l} the\ chair \\ the\ shirt \\ the\ tortilla \end{array} \right\} \times \left\{ \begin{array}{l} left[ward] \\ right[ward] \end{array} \right\}$$

object-direction

$$\left\{ \begin{array}{l} the\ chair \\ the\ shirt \\ the\ tortilla \end{array} \right\} \times \left\{ \begin{array}{l} [on\ the]left \\ [on\ the]right \end{array} \right\}$$

object-location

$$\left\{\begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array}\right\} \times \left\{\begin{array}{l} carried \\ folded \\ left \end{array}\right\} \times \left\{\begin{array}{l} the\ chair \\ the\ shirt \\ the\ tortilla \end{array}\right\}$$

actor-verb-object

$$\left\{\begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array}\right\} \times \left\{\begin{array}{l} carried \\ \\ left \end{array}\right\} \times \left\{\begin{array}{l} left[ward] \\ right[ward] \end{array}\right\}$$

actor-verb-direction

$$\left\{\begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array}\right\} \times \left\{\begin{array}{l} the\ chair \\ the\ shirt \\ the\ tortilla \end{array}\right\} \times \left\{\begin{array}{l} left[ward] \\ right[ward] \end{array}\right\}$$

actor-object-direction

$$\left\{\begin{array}{l} carried \\ \\ left \end{array}\right\} \times \left\{\begin{array}{l} the\ chair \\ the\ shirt \\ the\ tortilla \end{array}\right\} \times \left\{\begin{array}{l} \\ left[ward] \\ right[ward] \end{array}\right\}$$

verb-object-direction

$$\left\{\begin{array}{l} Andrei \\ Dan \\ Siddharth \\ Jeff \end{array}\right\} \times \left\{\begin{array}{l} carried \\ folded \\ left \end{array}\right\} \times \left\{\begin{array}{l} the\ chair \\ the\ shirt \\ the\ tortilla \end{array}\right\} \times \left\{\begin{array}{l} [on\ the]left[ward] \\ [on\ the]right[ward] \end{array}\right\}$$

sentence

Jeff carried the chair leftward.

Jeff carried the chair leftward.

Andrei carried the tortilla leftward.

Andrei carried the tortilla leftward.

Siddharth folded the chair on the right.

Siddharth folded the chair on the right.

Siddharth left the chair leftward.

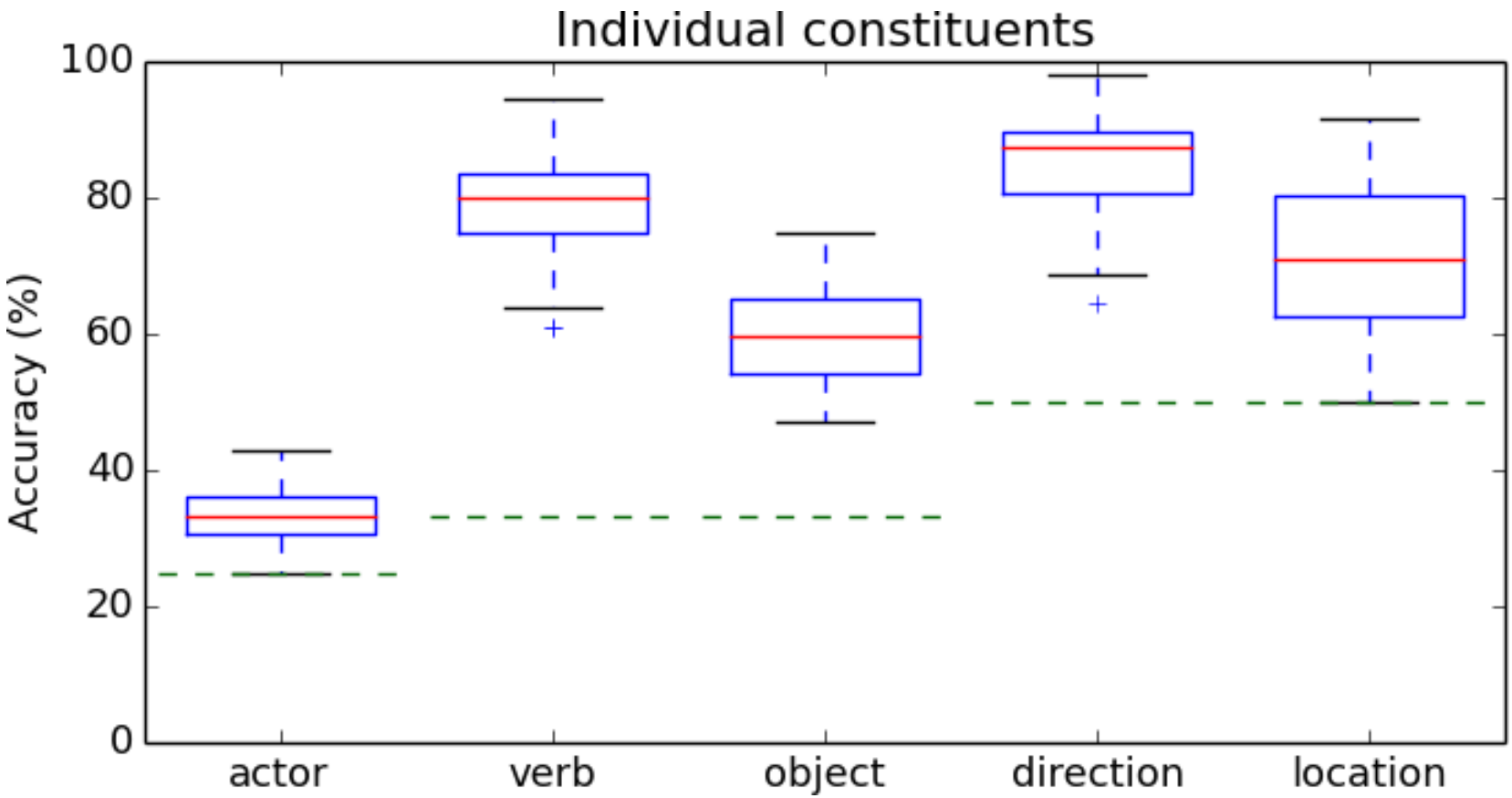
Siddharth left the chair leftward.

Dan left the shirt rightward.

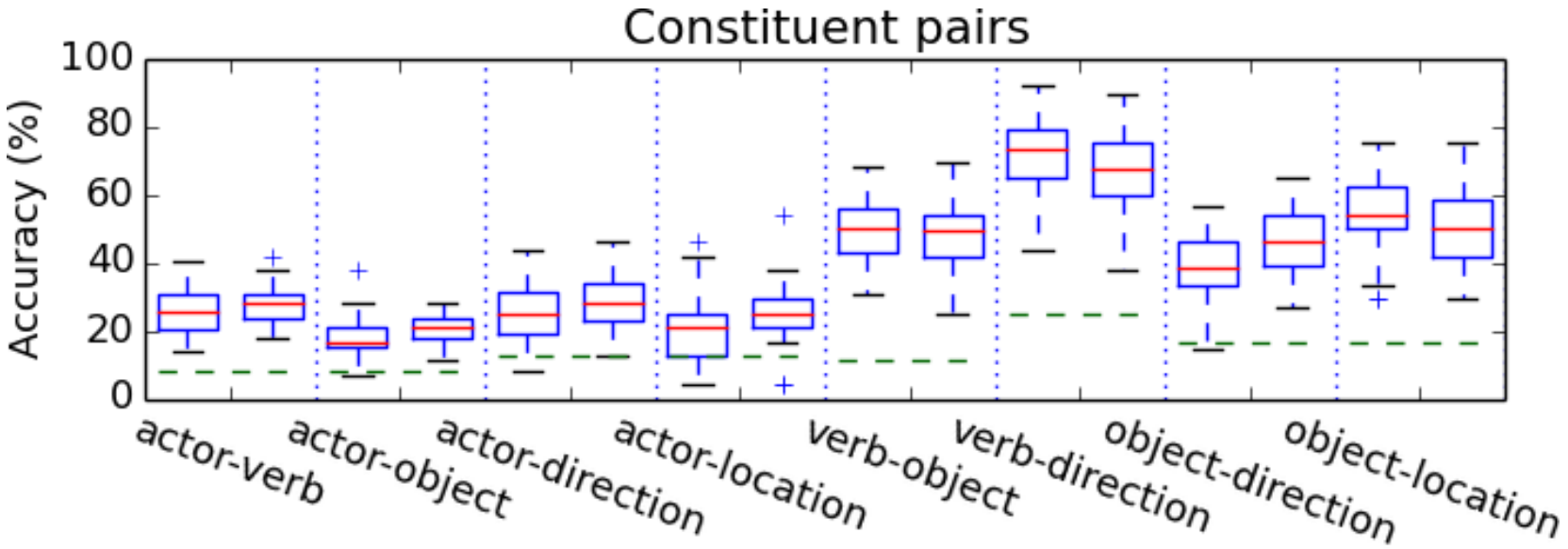
Dan left the shirt rightward.

independent Train on individual constituents, test on constituent pairs and triples.

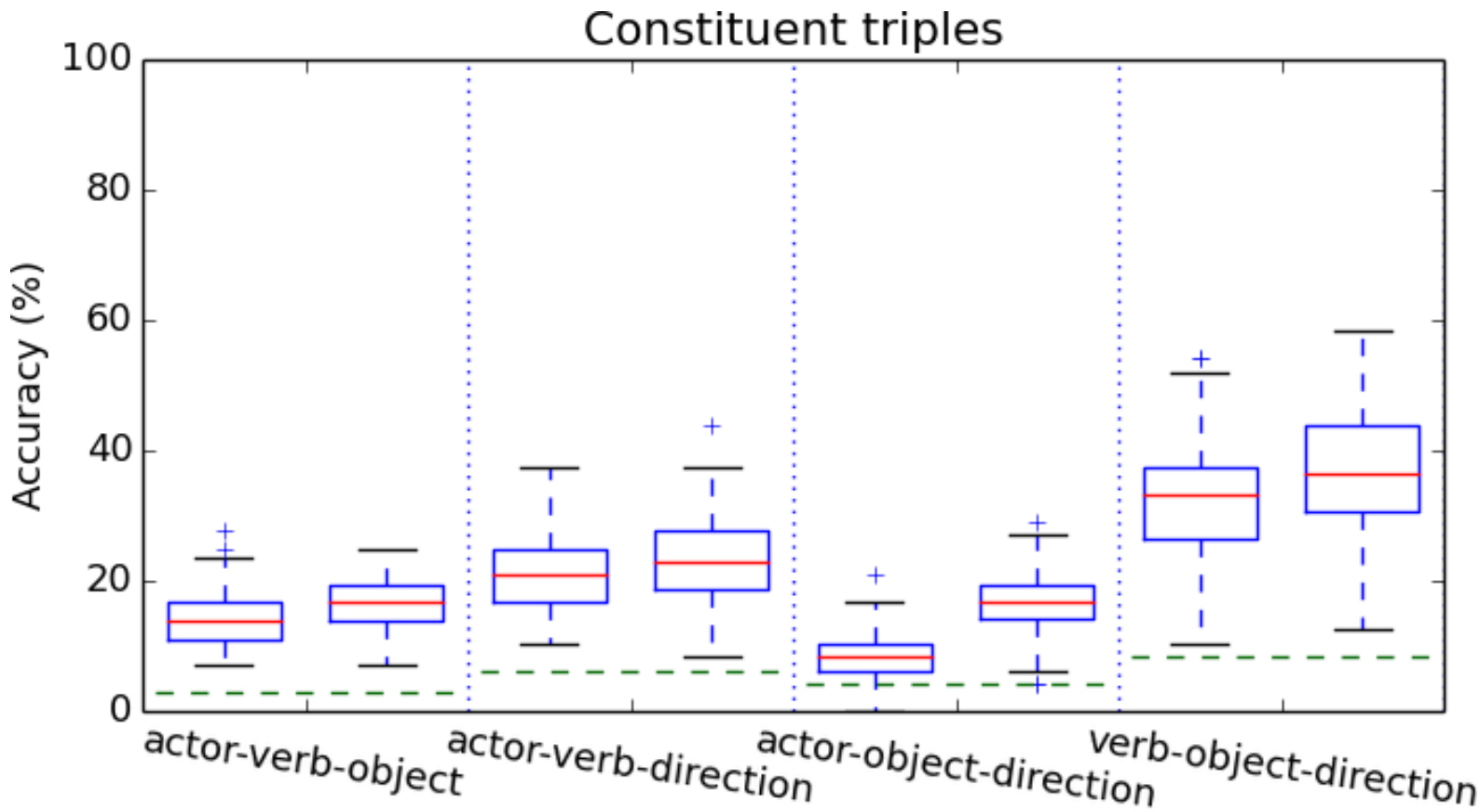
joint Train and test on constituent pairs and triples.



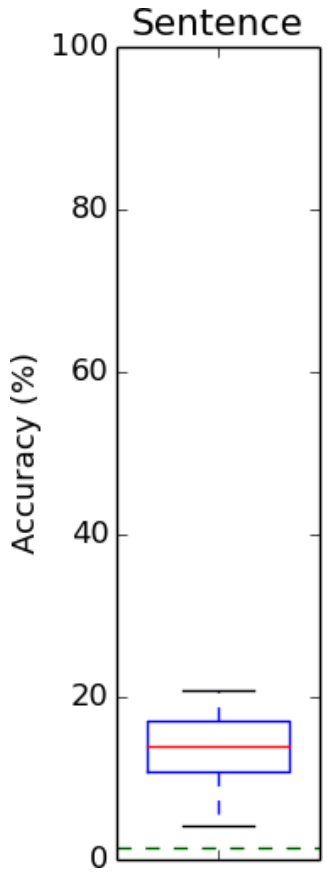
Aggregated across subject, class, and run



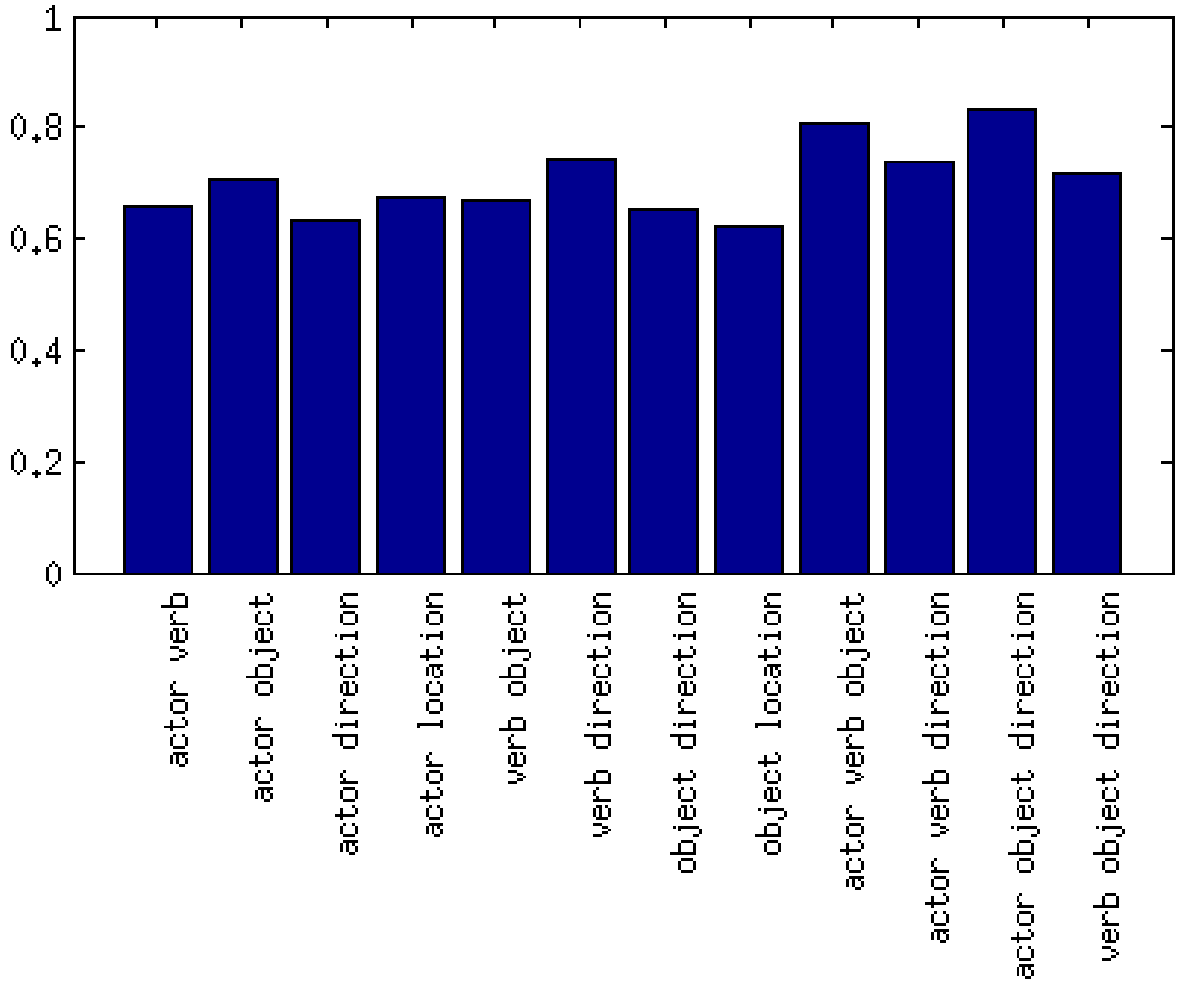
Aggregated across subject, class, and run



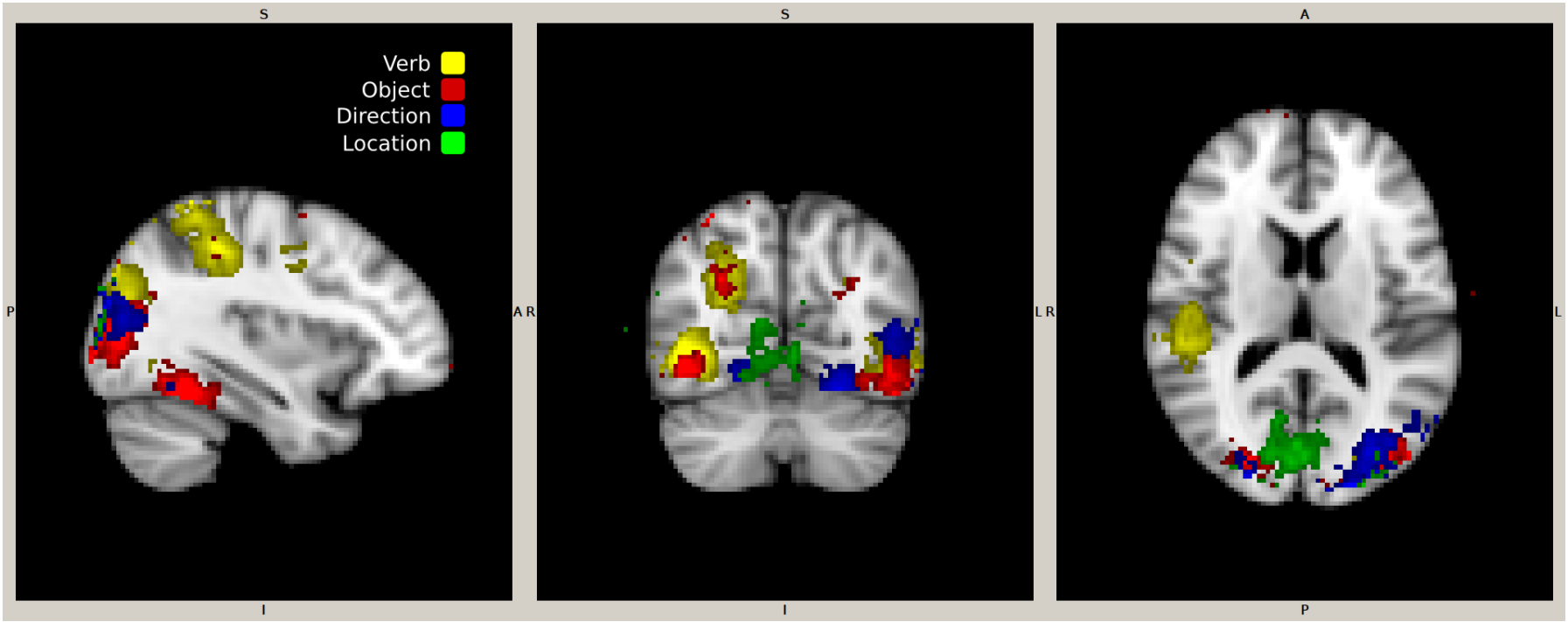
Aggregated across subject, class, and run



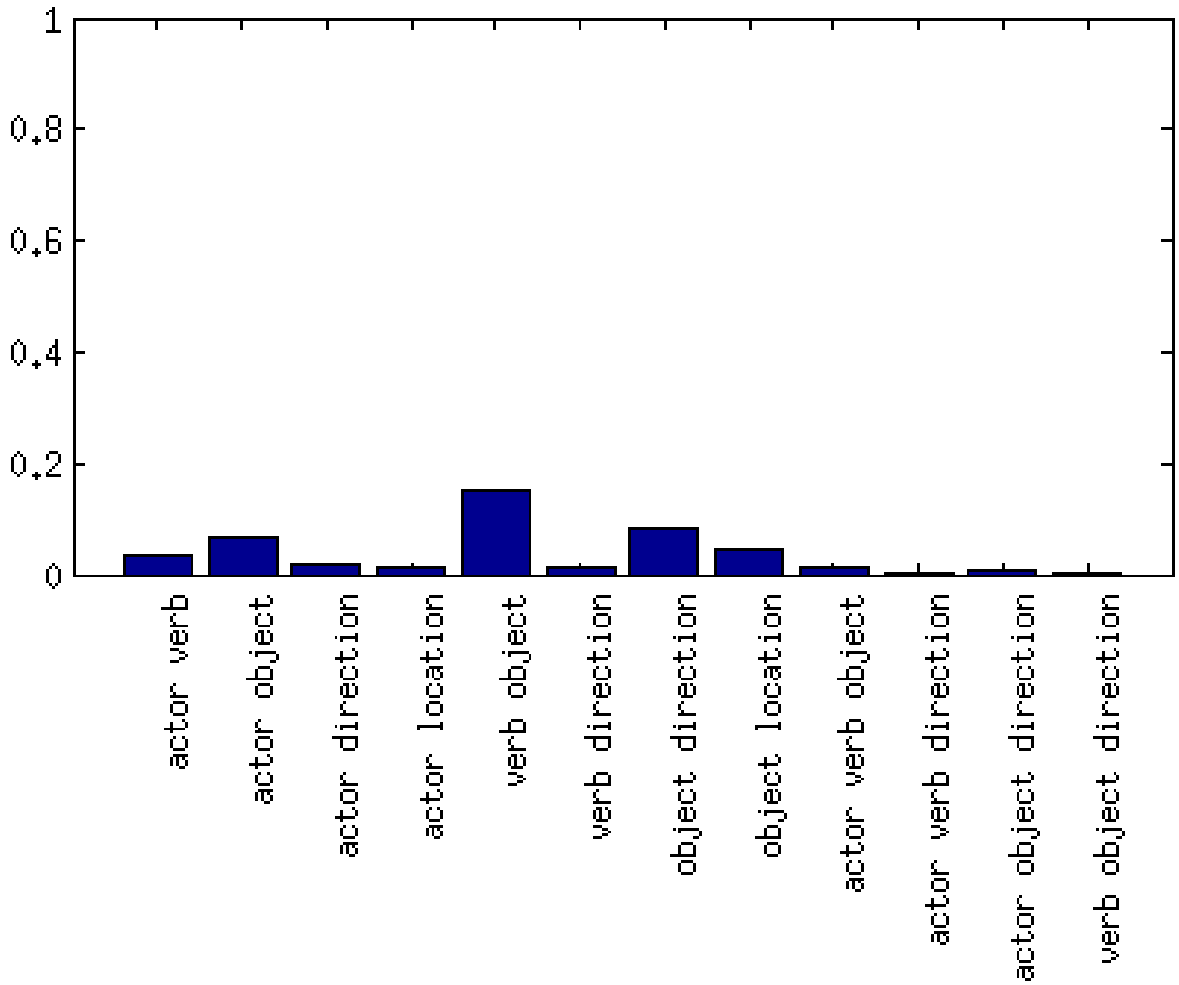
Aggregated across subject, class, and run



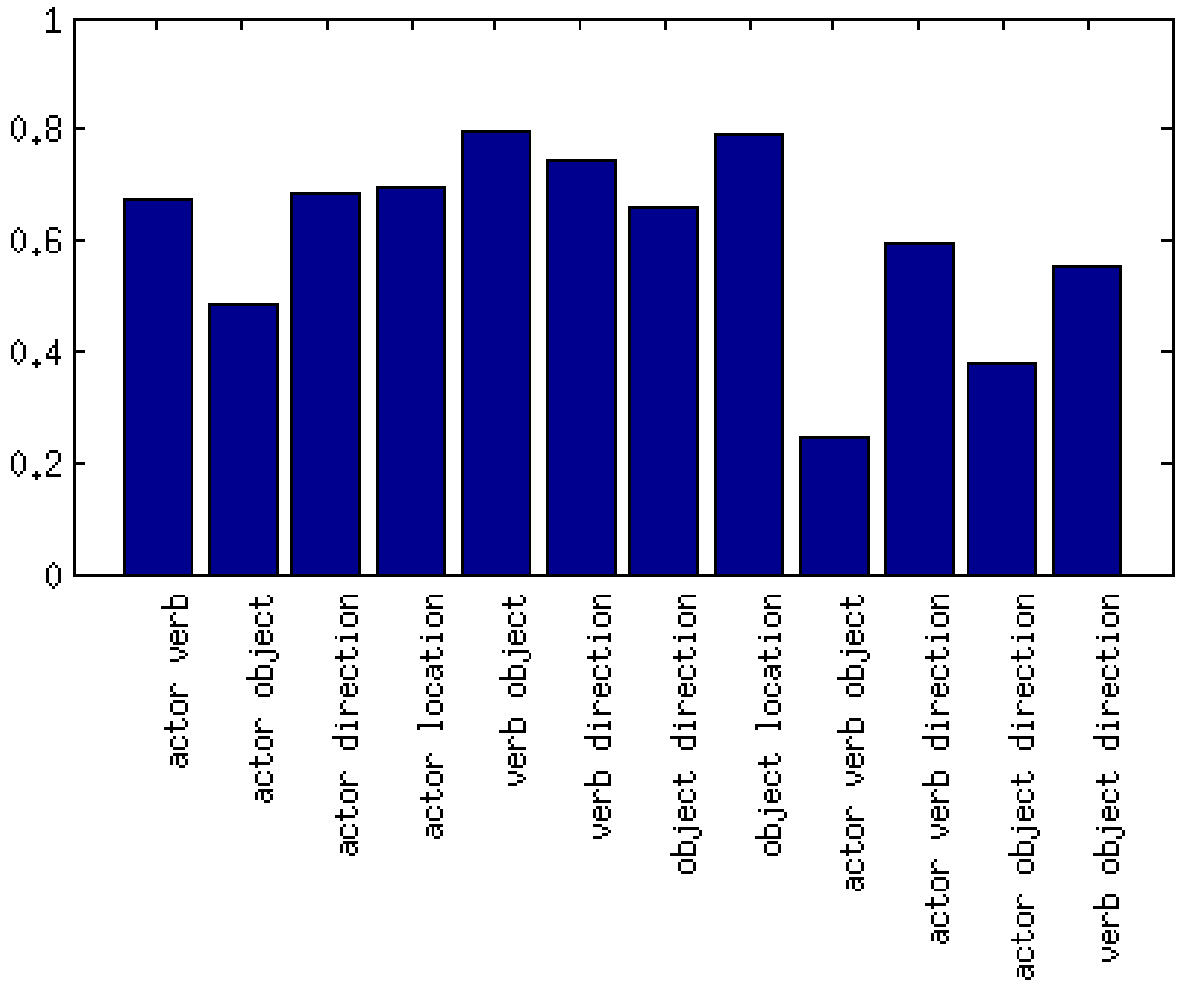
Correlation between independent and joint classifier judgments



Subject 1



Fraction of overlap between all of the independent classifier regions



Fraction of the joint classifier region covered by the union of the independent classifier regions

Does the brain condition perception of constituents with a joint model?

Thank you