Grounded Language Learning from Video Described with Sentences

Haonan Yu Jeffrey Mark Siskind

PURDUE UNIVERSITY™

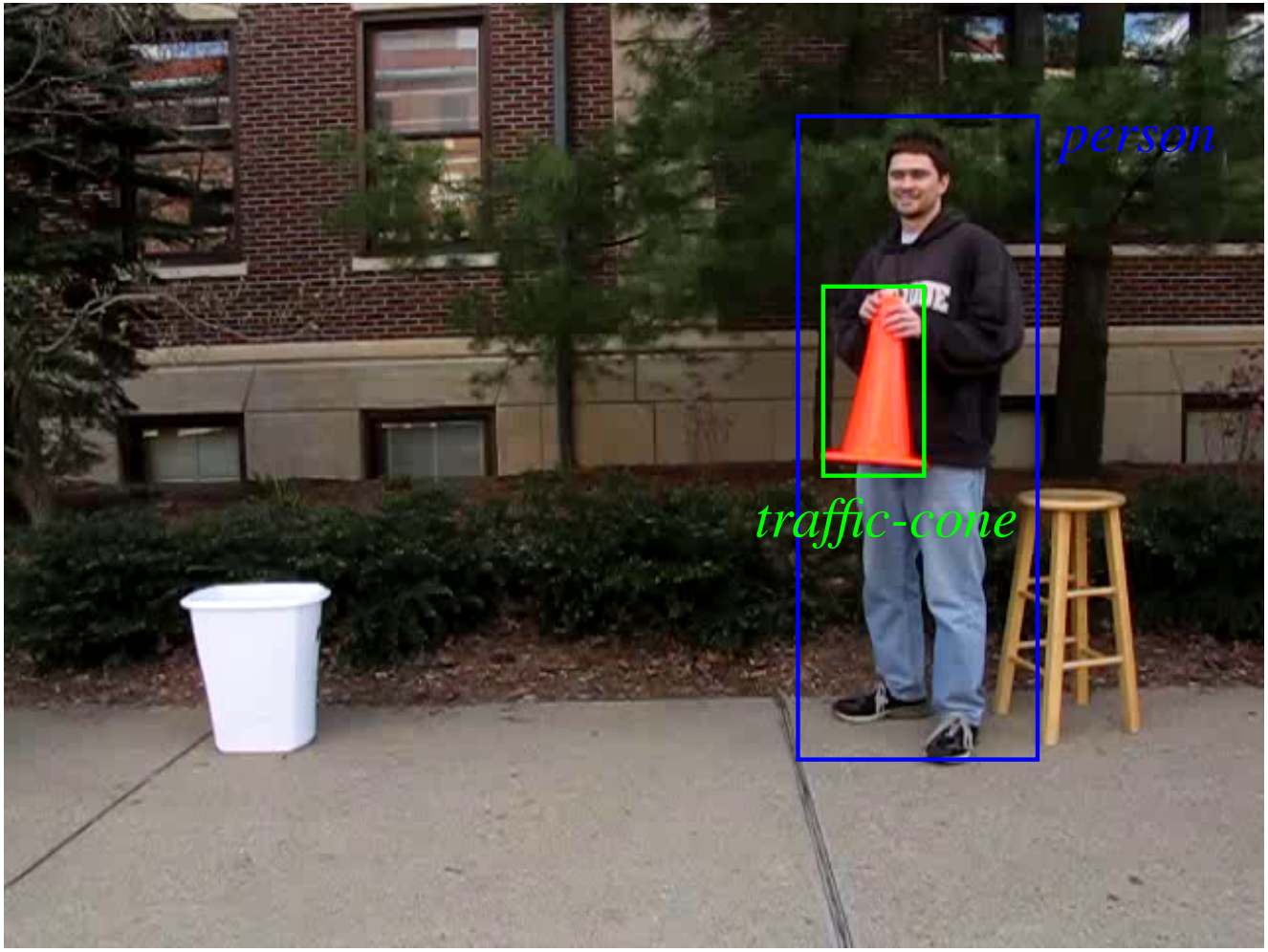
This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0060. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, 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.

Adults Can Recognize What Happened

Adults Can Recognize What Happened

The person carried the traffic-cone.

Adults Can Recognize What Happened



The person carried the traffic-cone.

Adults Can Describe What Happened

The person picked up the backpack.

The chair approached the trash-can.

The person carried the traffic-cone.

. . .

Adults Can Describe What Happened

The person picked up the backpack.

The chair approached the trash-can.

The person carried the traffic-cone.

. . .

Adults Can Describe What Happened

~~*The person picked up the backpack.*~~

~~*The chair approached the trash-can.*~~

The person carried the traffic-cone.

. . .

What Children Learn From

The person picked up the traffic-cone.

What Children Learn From

The person picked up the traffic-cone.

What Children Learn From

The person carried the chair away from the backpack.

What Children Learn From

The person carried the chair away from the backpack.

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

↓ training sentences → learned words

person	*traffic-cone*	*chair*	*stool*	*backpack*	*trash-can*
picked up	*carried*	*put down*	*approached*	*to the left of*	*quickly*

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

↓ training sentences → learned words

person	*traffic-cone*	*chair*	*stool*	*backpack*	*trash-can*
picked up	*carried*	*put down*	*approached*	*to the left of*	*quickly*

↓ learned words → new sentences

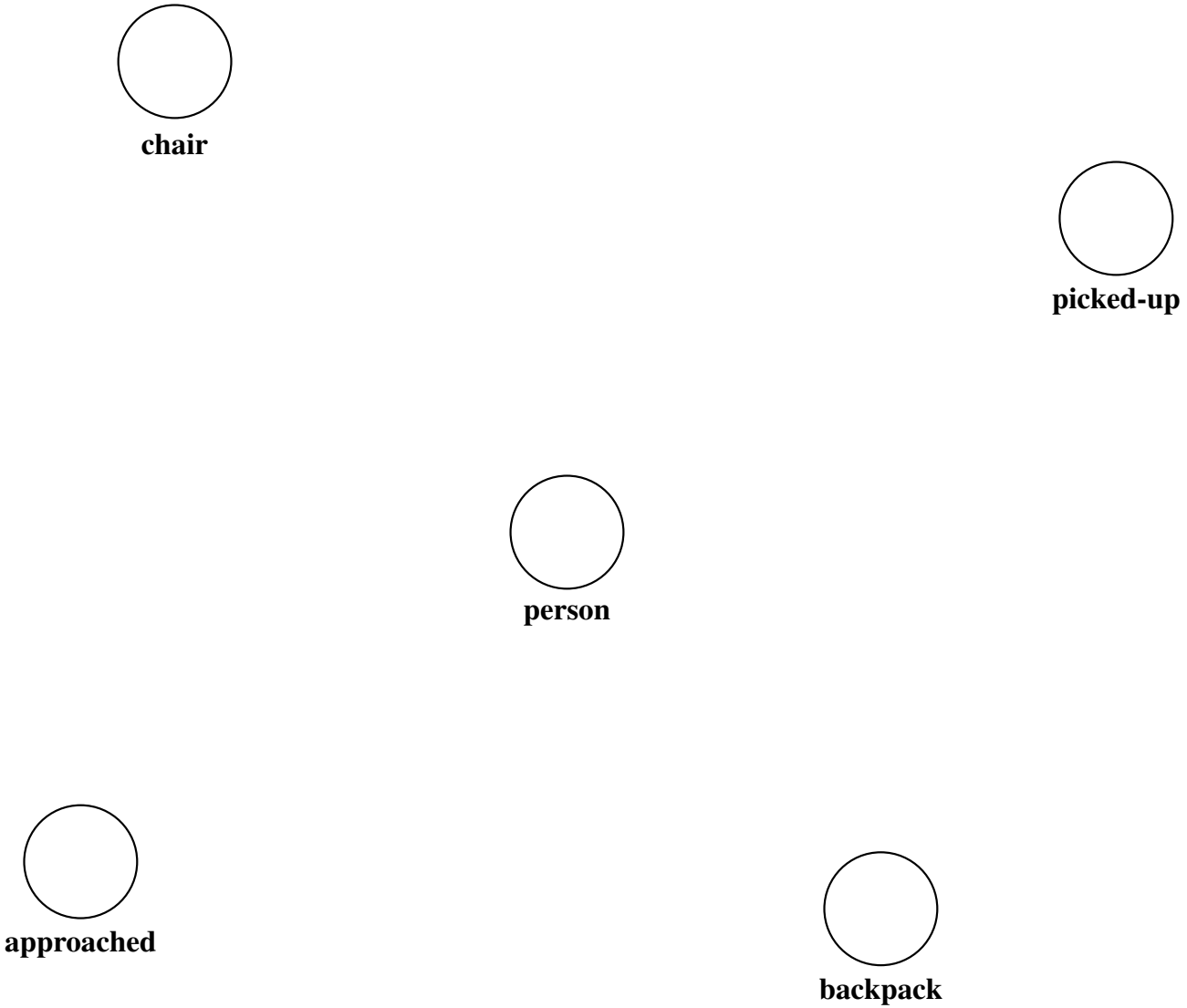
The backpack approached the person.
The person carried the trash-can.

The person put down the chair quickly.
. . .

What Children Learn From

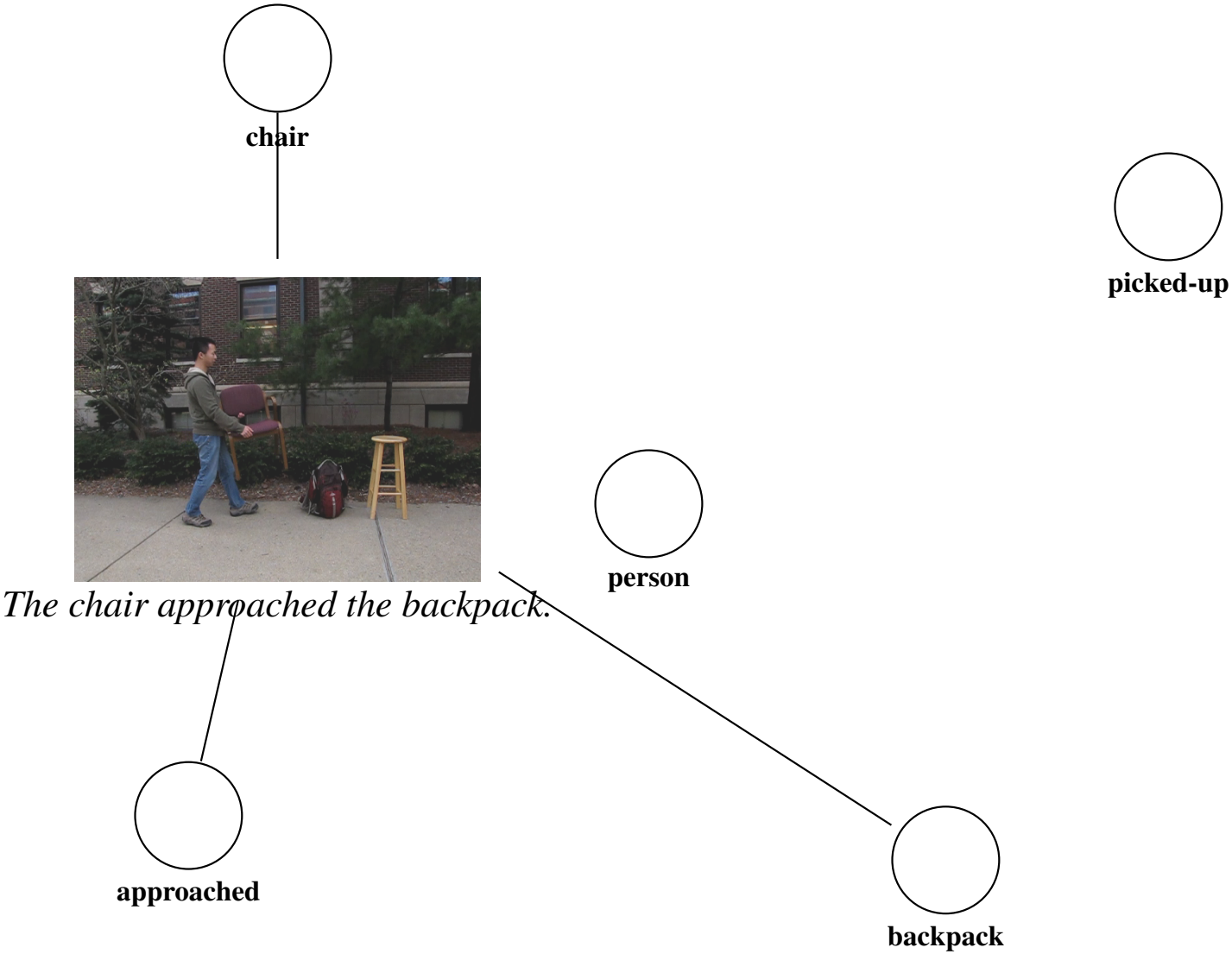
We're going to model this language acquisition problem.

How Can We Do This?

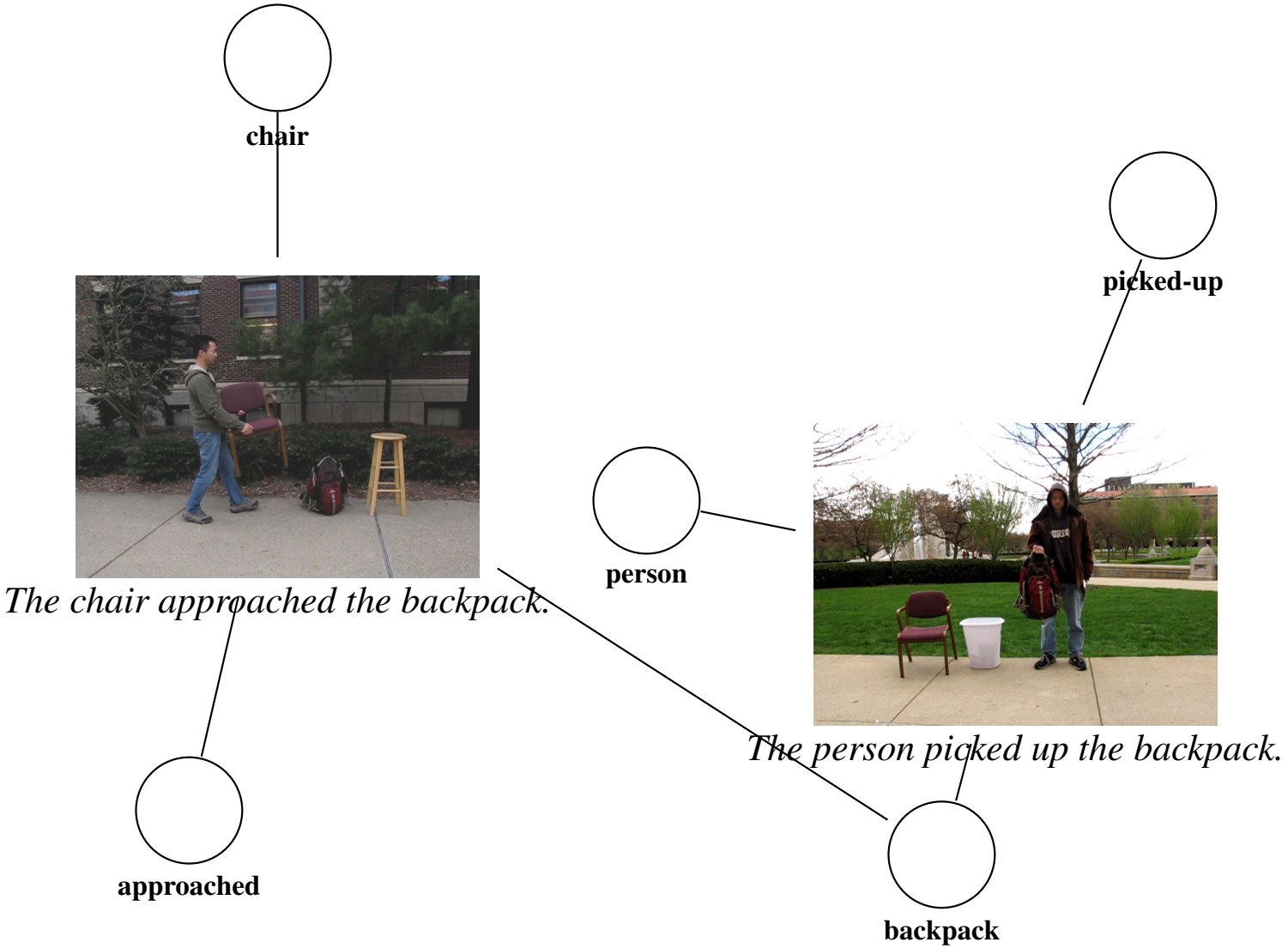


How Can We Do This?

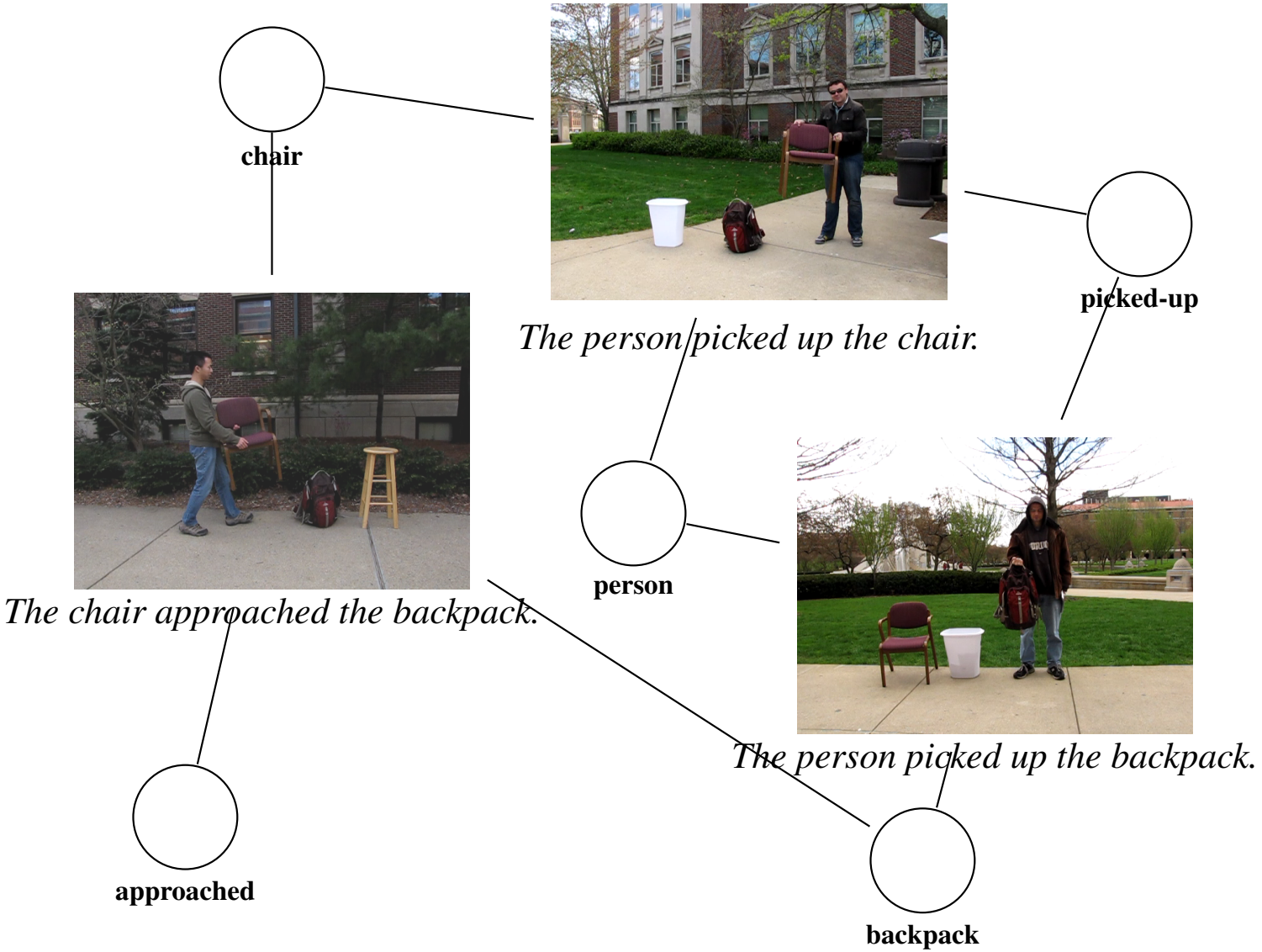
The chair approached the backpack.



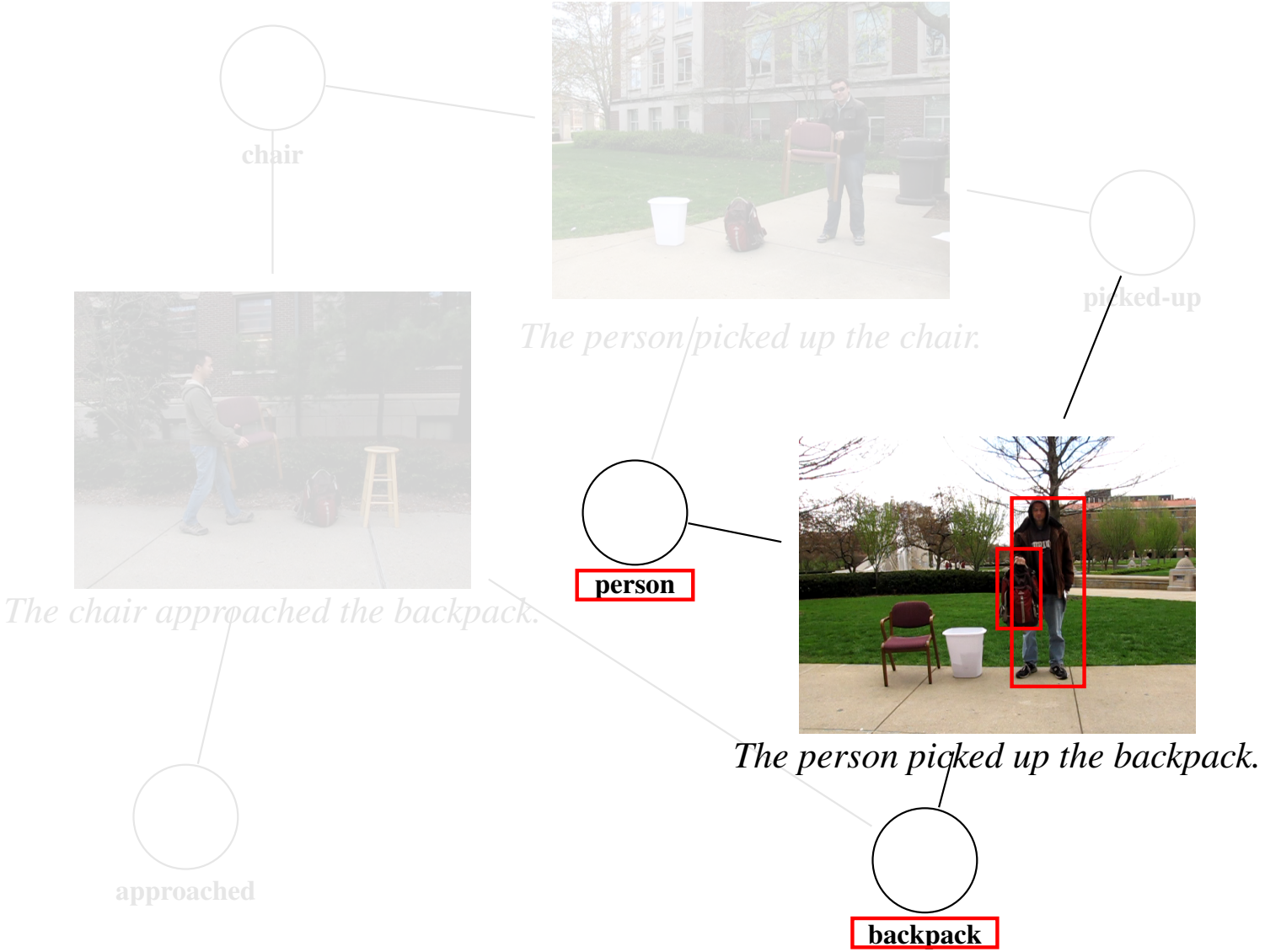
How Can We Do This?



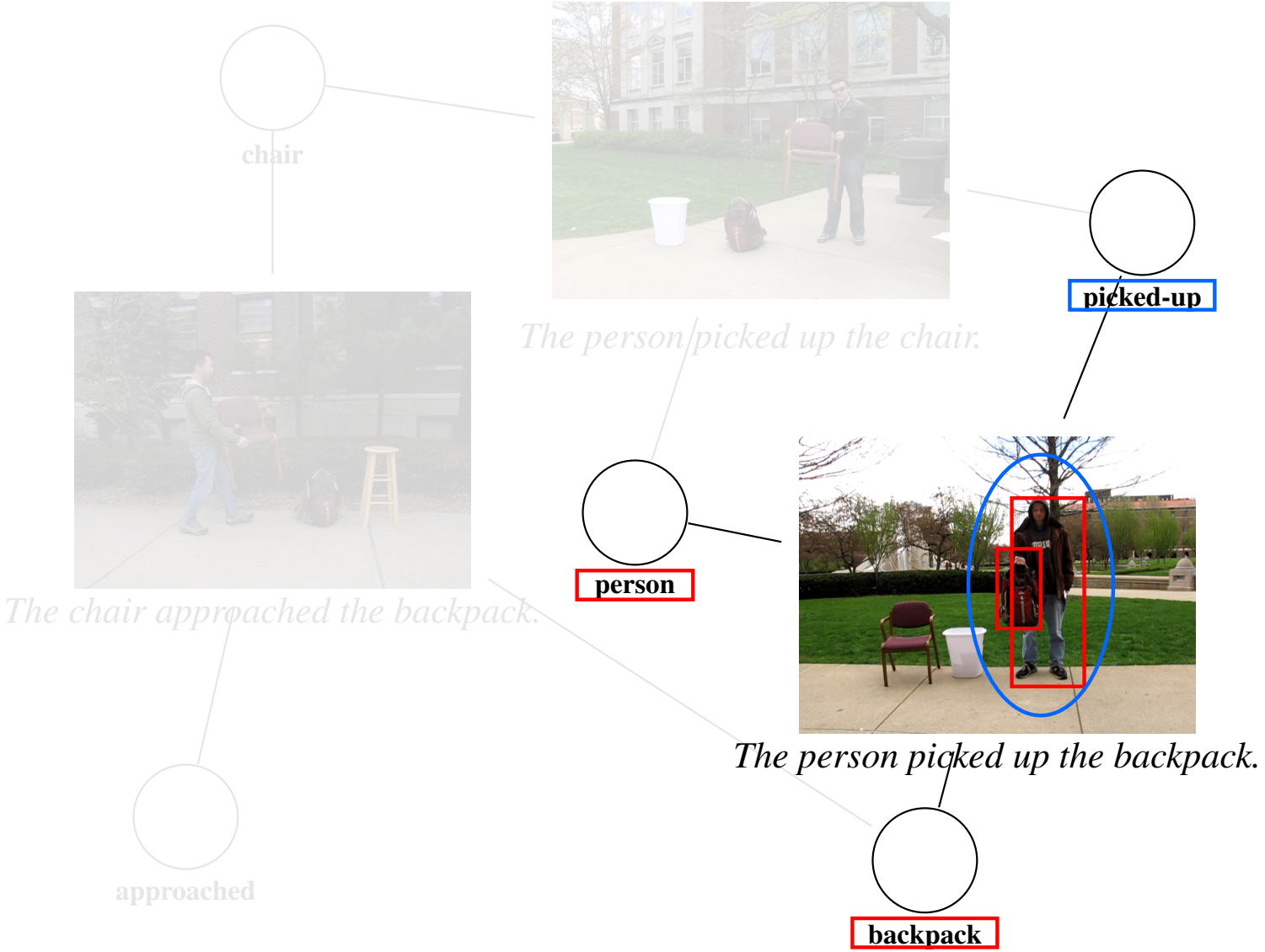
How Can We Do This?



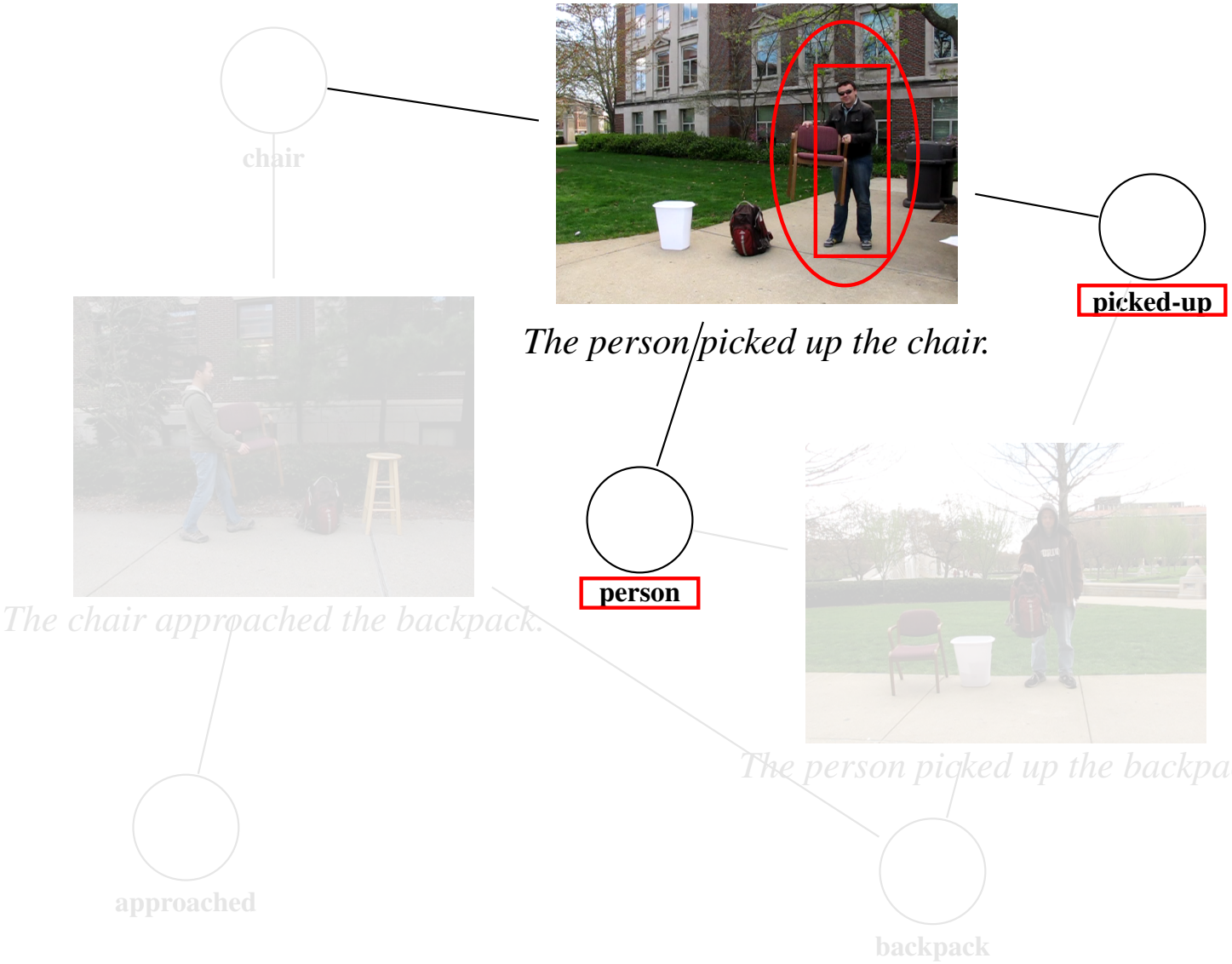
Inference Across *Different* Words in the *Same* Sentence



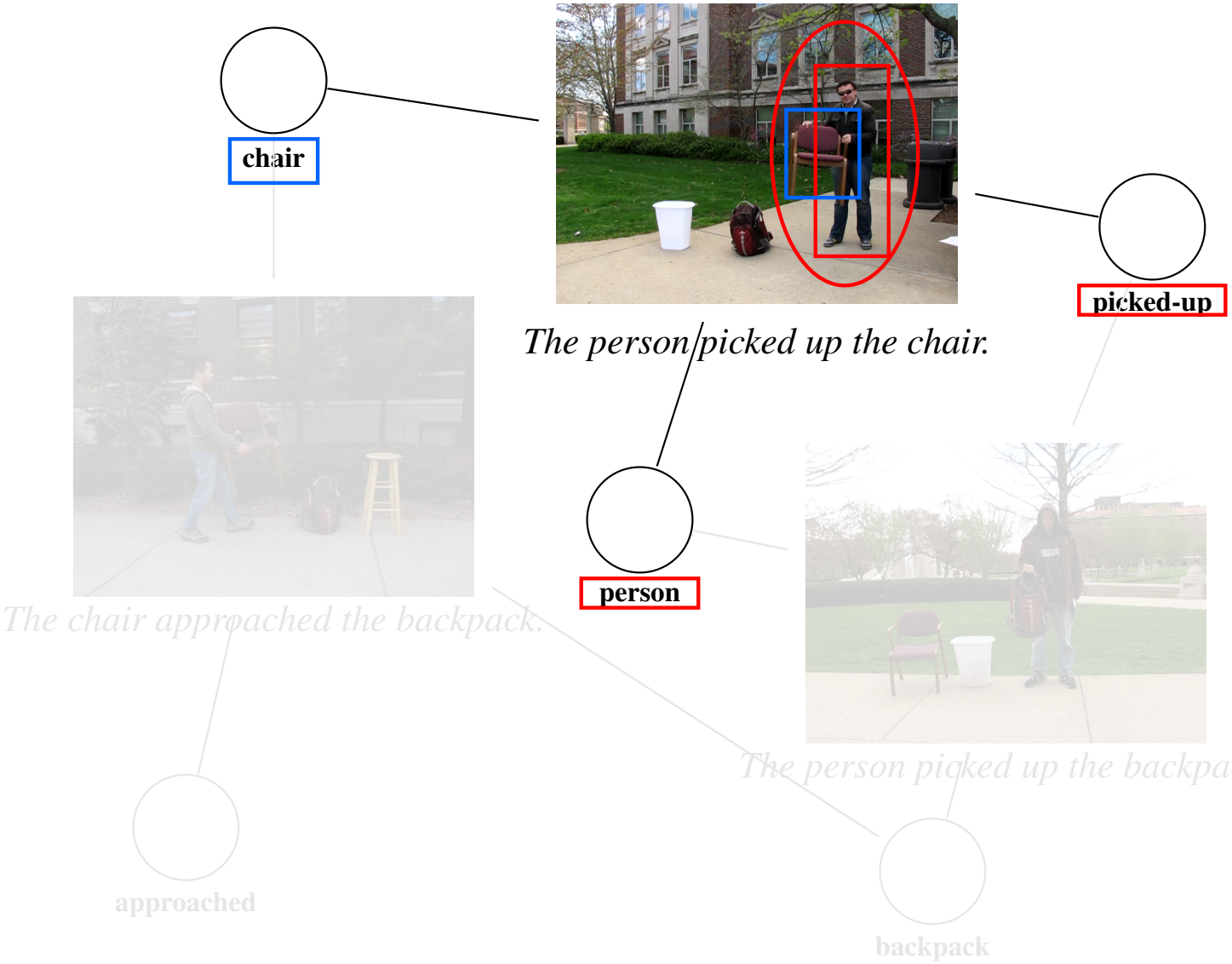
Inference Across *Different* Words in the *Same* Sentence



Inference Across the *Same* Word in *Different* Sentences

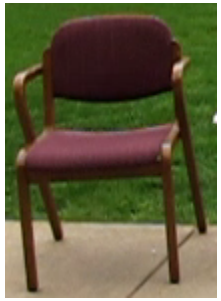


Inference Across the *Same* Word in *Different* Sentences



Existing Work on Modeling Events and Objects

Existing Work on Modeling Events and Objects



. . .

Existing Work on Modeling Events and Objects

chair model

e.g., [Felzenszwalb et al., 2010]

Existing Work on Modeling Events and Objects

chair model
e.g., [Felzenszwalb et al., 2010]

Existing Work on Modeling Events and Objects

chair model
e.g., [Felzenszwalb et al., 2010]

stand-up model
e.g., [Laptev et al., 2008]

Existing Work on Modeling Events and Objects

chair model
e.g., [Felzenszwalb et al., 2010]

stand-up model
e.g., [Laptev et al., 2008]

- usually only nouns and verbs, no other parts-of-speech
  - e.g., prepositions or adverbs

Existing Work on Modeling Events and Objects

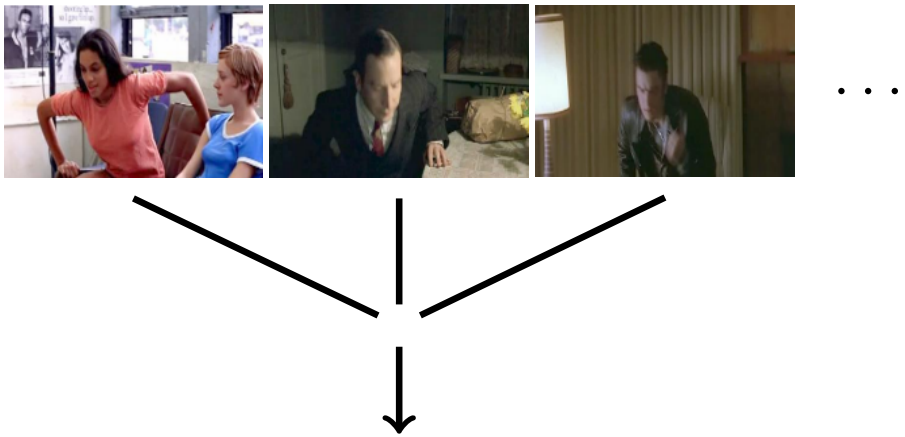
chair model
e.g., [Felzenszwalb et al., 2010]

stand-up model
e.g., [Laptev et al., 2008]

- usually only nouns and verbs, no other parts-of-speech
  - e.g., prepositions or adverbs
- We learn other parts-of-speech in addition to nouns and verbs.

Existing Work on Modeling Events and Objects

chair model
e.g., [Felzenszwalb et al., 2010]



stand-up model
e.g., [Laptev et al., 2008]

- usually only nouns and verbs, no other parts-of-speech
  - e.g., prepositions or adverbs
- We learn other parts-of-speech in addition to nouns and verbs.
- We learn all of them *simultaneously*.

Background Knowledge

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP}
\end{aligned}
$$

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP}
\end{aligned}
$$

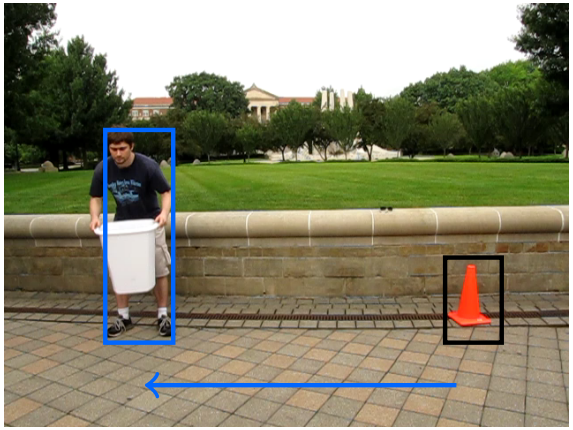
$$
\begin{aligned}
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

Examples of Wording Meanings

$\textsc{class}(\mathrm{O}) = \textsc{chair}$

$$
\begin{aligned}
\mathrm{D} &\rightarrow \textit{the} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\mathrm{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

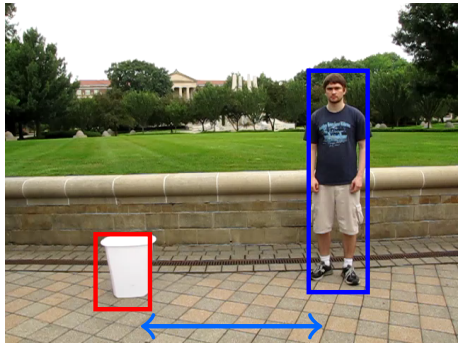
Examples of Wording Meanings



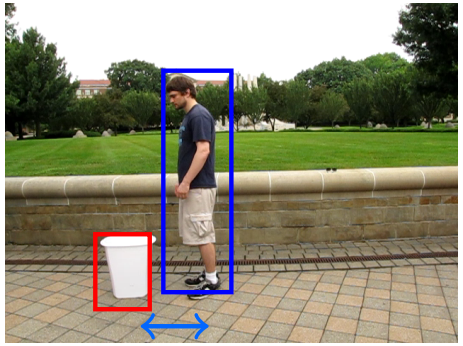
$$x(O1) < x(O2)$$

D	$\rightarrow$	*the*
N	$\rightarrow$	*person* | *backpack* | *trash-can* | *chair* | *traffic-cone* | *stool*
P	$\rightarrow$	*to the left of* | *to the right of*
V	$\rightarrow$	*picked up* | *put down* | *carried* | *approached*
ADV	$\rightarrow$	*quickly* | *slowly*
PM	$\rightarrow$	*towards* | *away from*

Examples of Wording Meanings



DISTANCE(O1,O2) > T



DISTANCE(O1,O2) < T

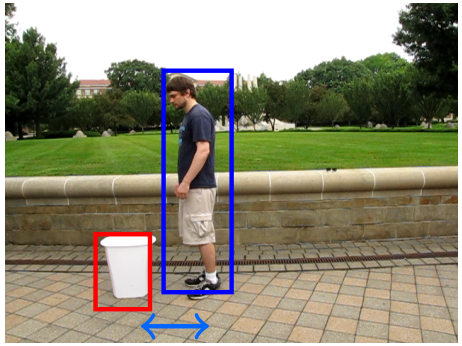
D → *the*
N → *person* | *backpack* | *trash-can* | *chair* | *traffic-cone* | *stool*
P → *to the left of* | *to the right of*
V → *picked up* | *put down* | *carried* | *approached*
ADV → *quickly* | *slowly*
PM → *towards* | *away from*

Examples of Wording Meanings

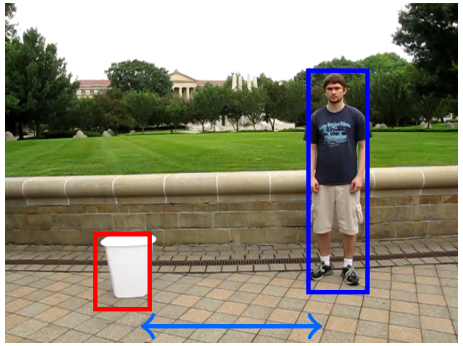
$$\mathrm{v}(\mathrm{O1}) > \mathrm{T}$$

$$\begin{aligned}
\mathrm{D} &\rightarrow \textit{the} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\mathrm{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}$$

Examples of Wording Meanings



DISTANCE(O1,O2) $<$ T



DISTANCE(O1,O2) $>$ T

D $\rightarrow$ *the*
N $\rightarrow$ *person* | *backpack* | *trash-can* | *chair* | *traffic-cone* | *stool*
P $\rightarrow$ *to the left of* | *to the right of*
V $\rightarrow$ *picked up* | *put down* | *carried* | *approached*
ADV $\rightarrow$ *quickly* | *slowly*
PM $\rightarrow$ *towards* | *away from*

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP}
\end{aligned}
$$

$$
\begin{aligned}
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

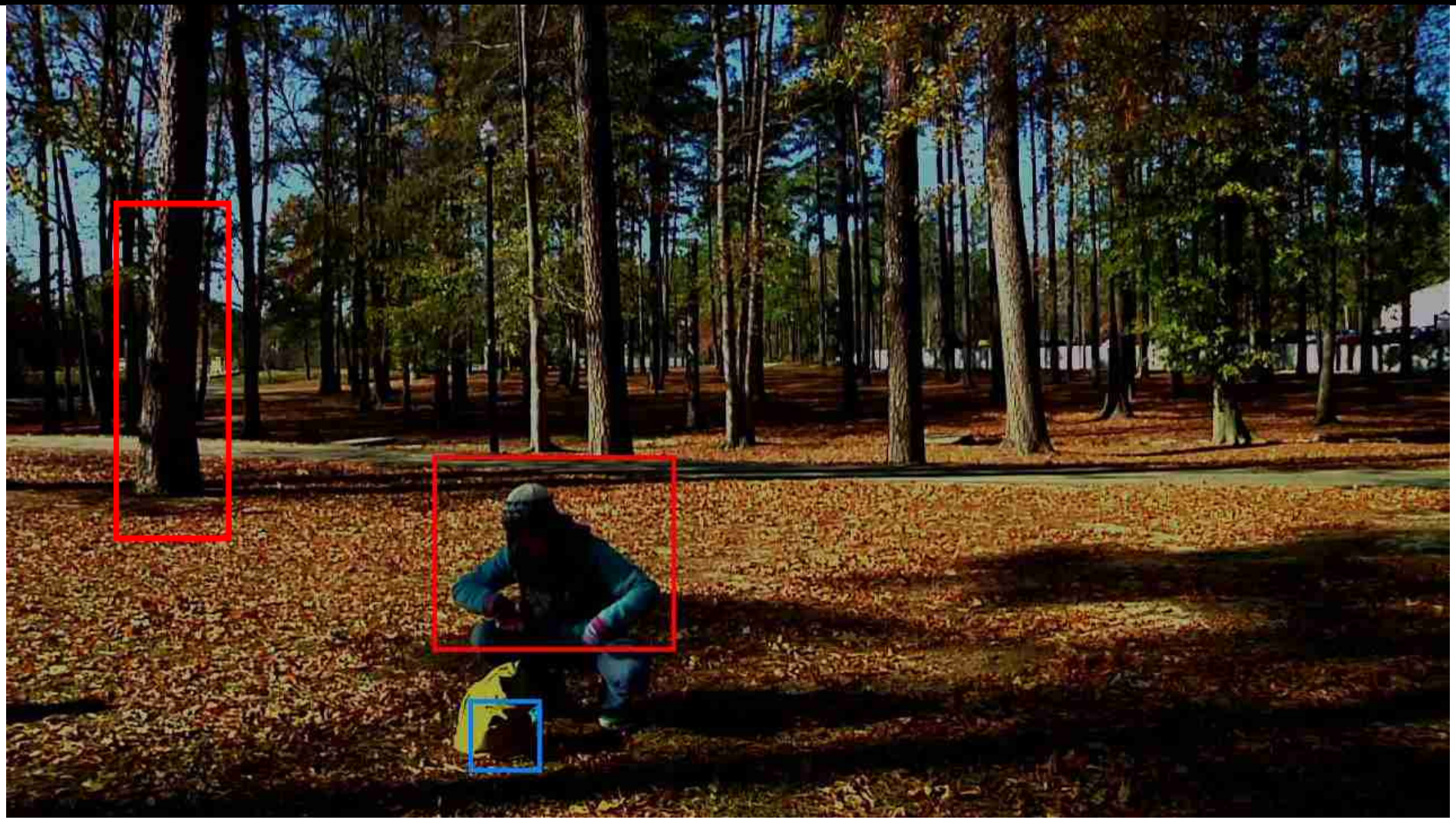
Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP}
\end{aligned}
$$

$$
\begin{aligned}
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

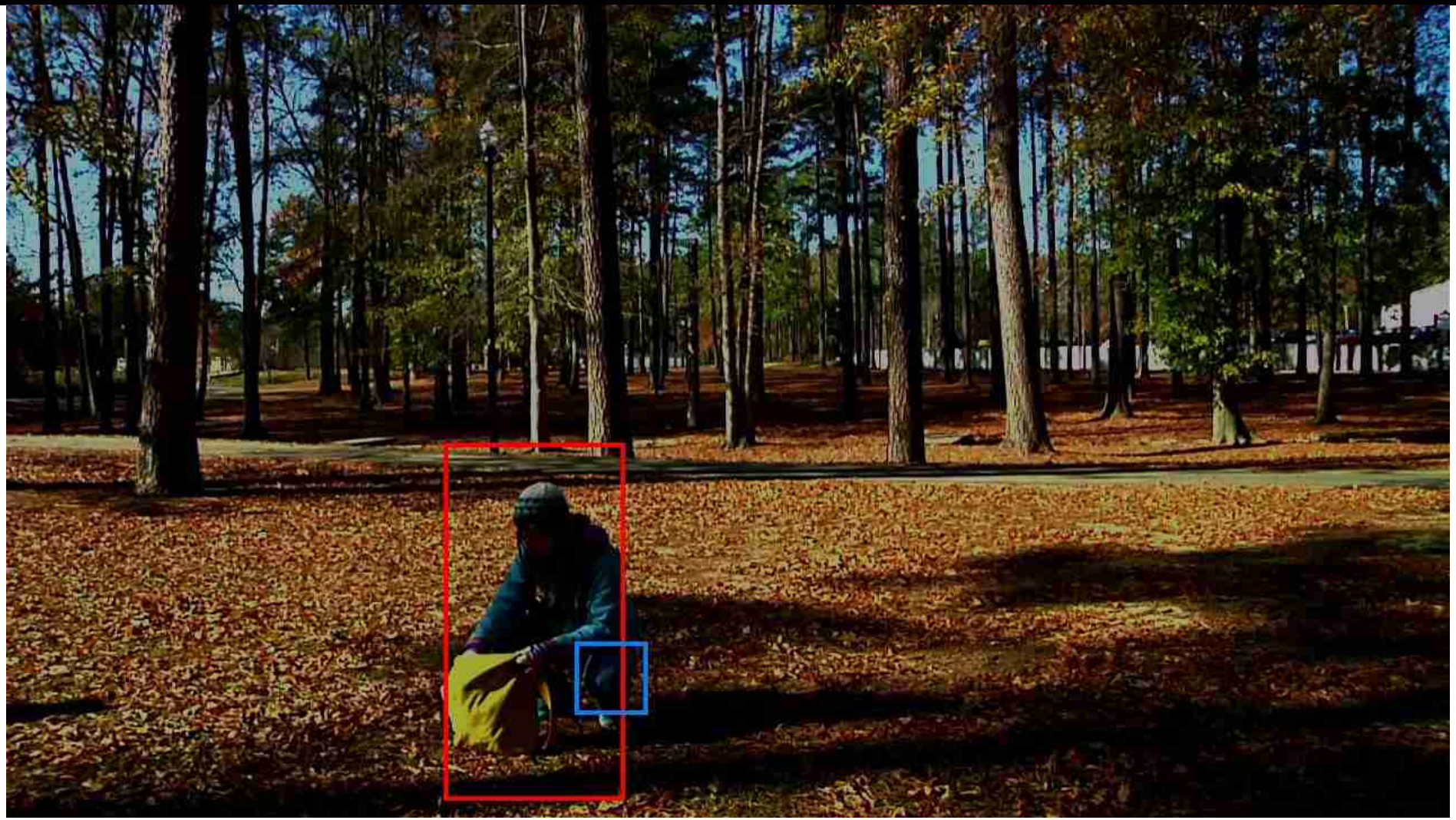
Object Detection

Object Detection



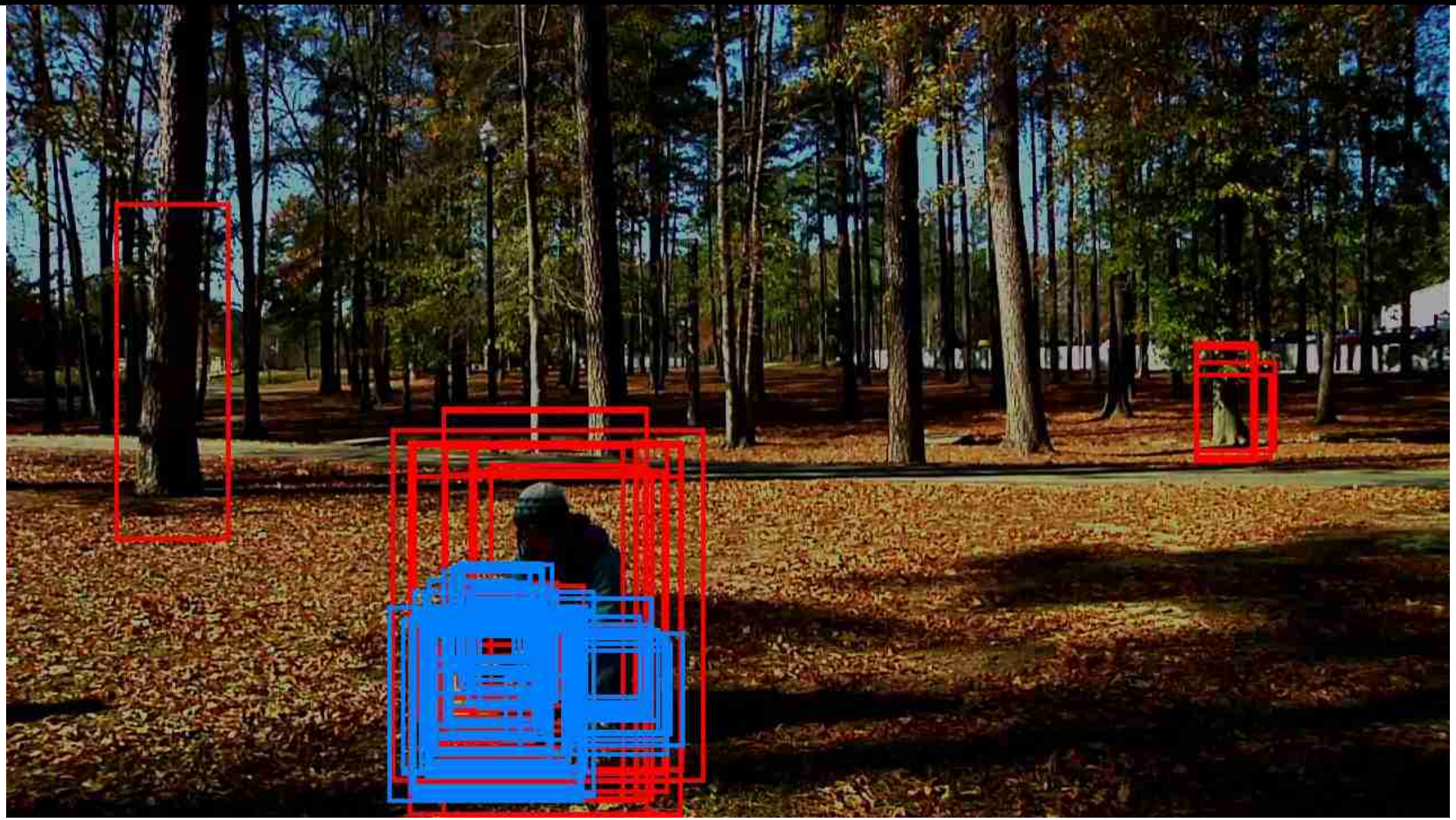
False positives

Object Detection



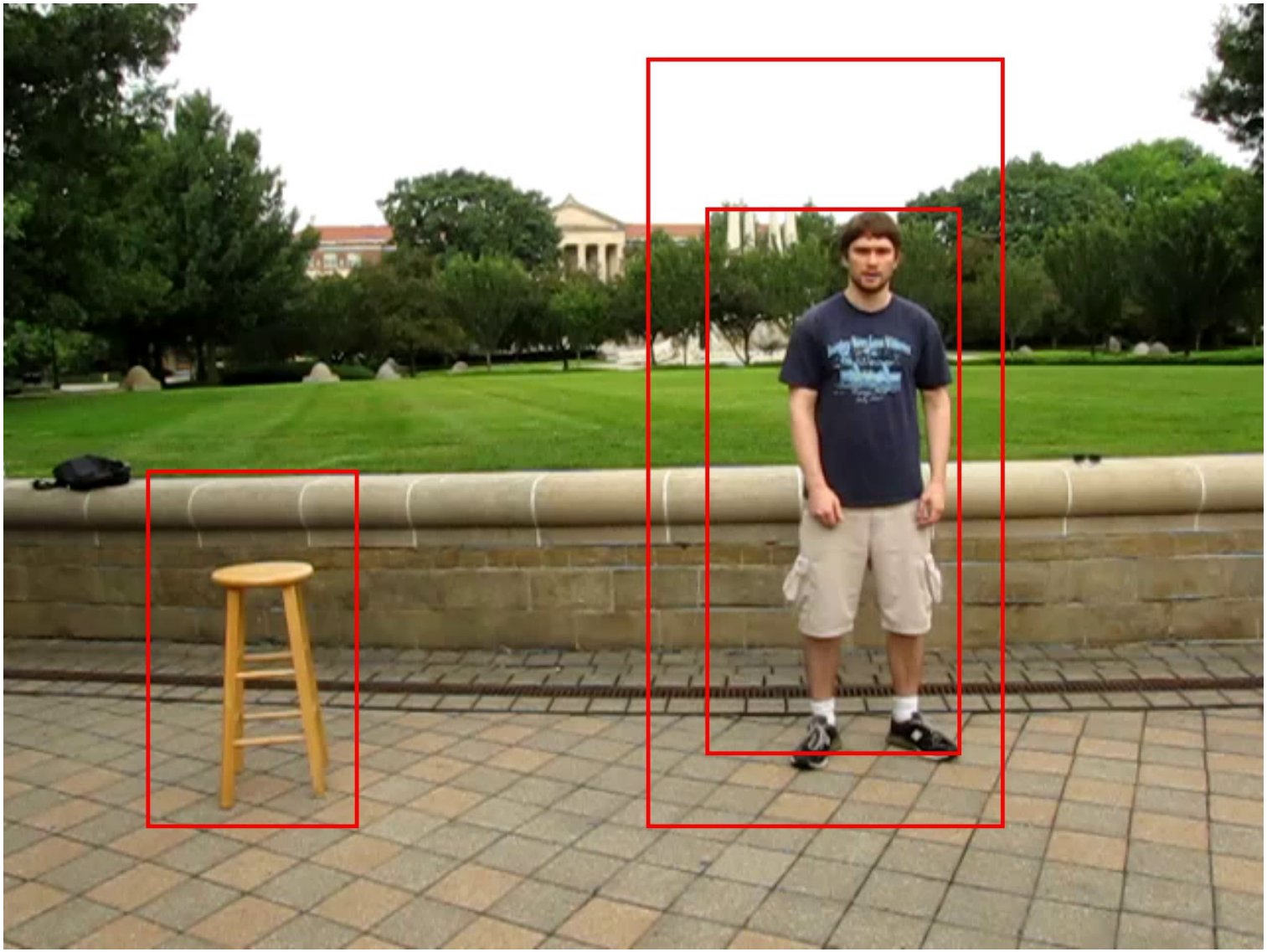
False negatives

Object Detection *and Overgeneration*

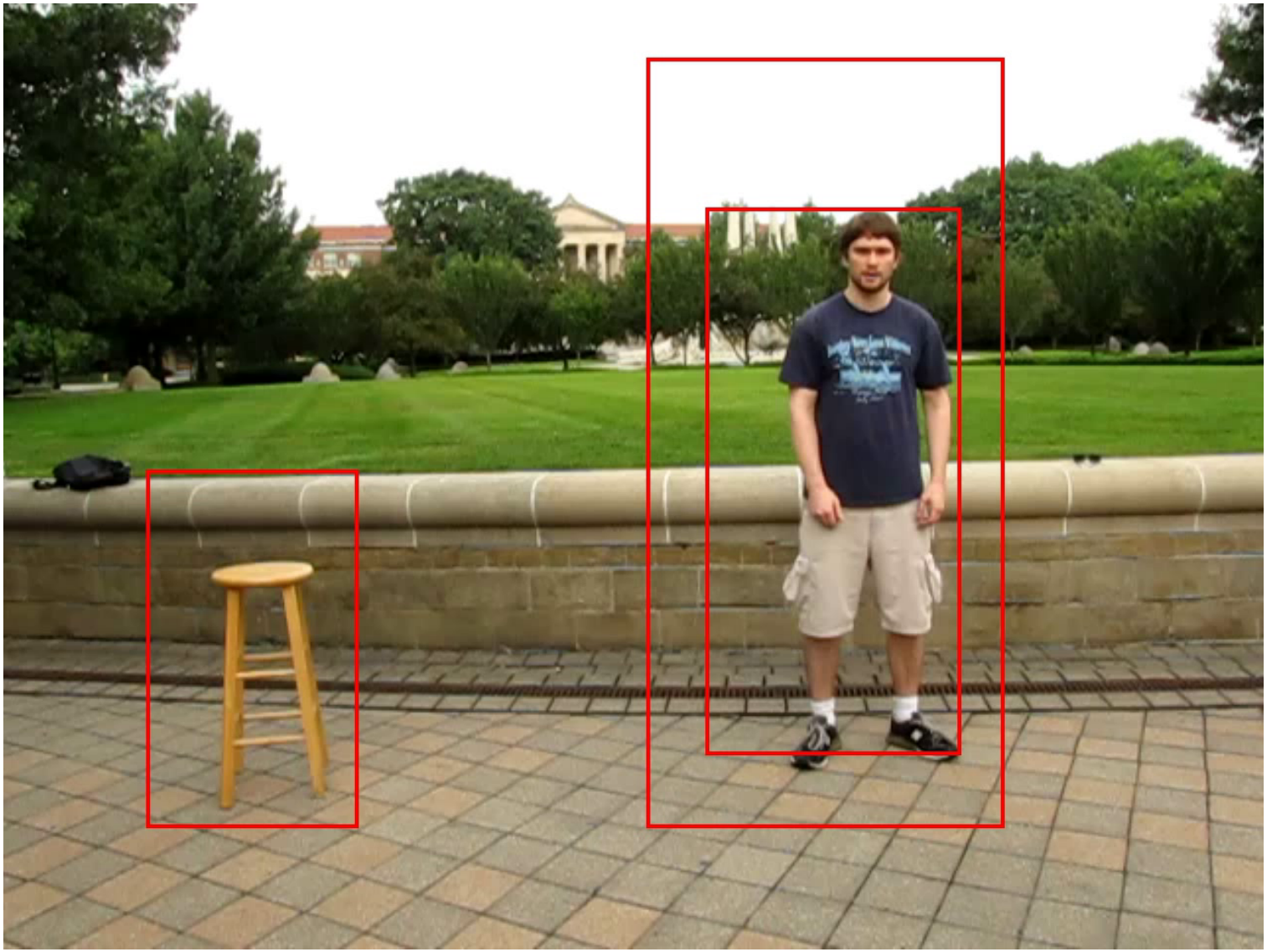


Trade more false positives for fewer false negatives.

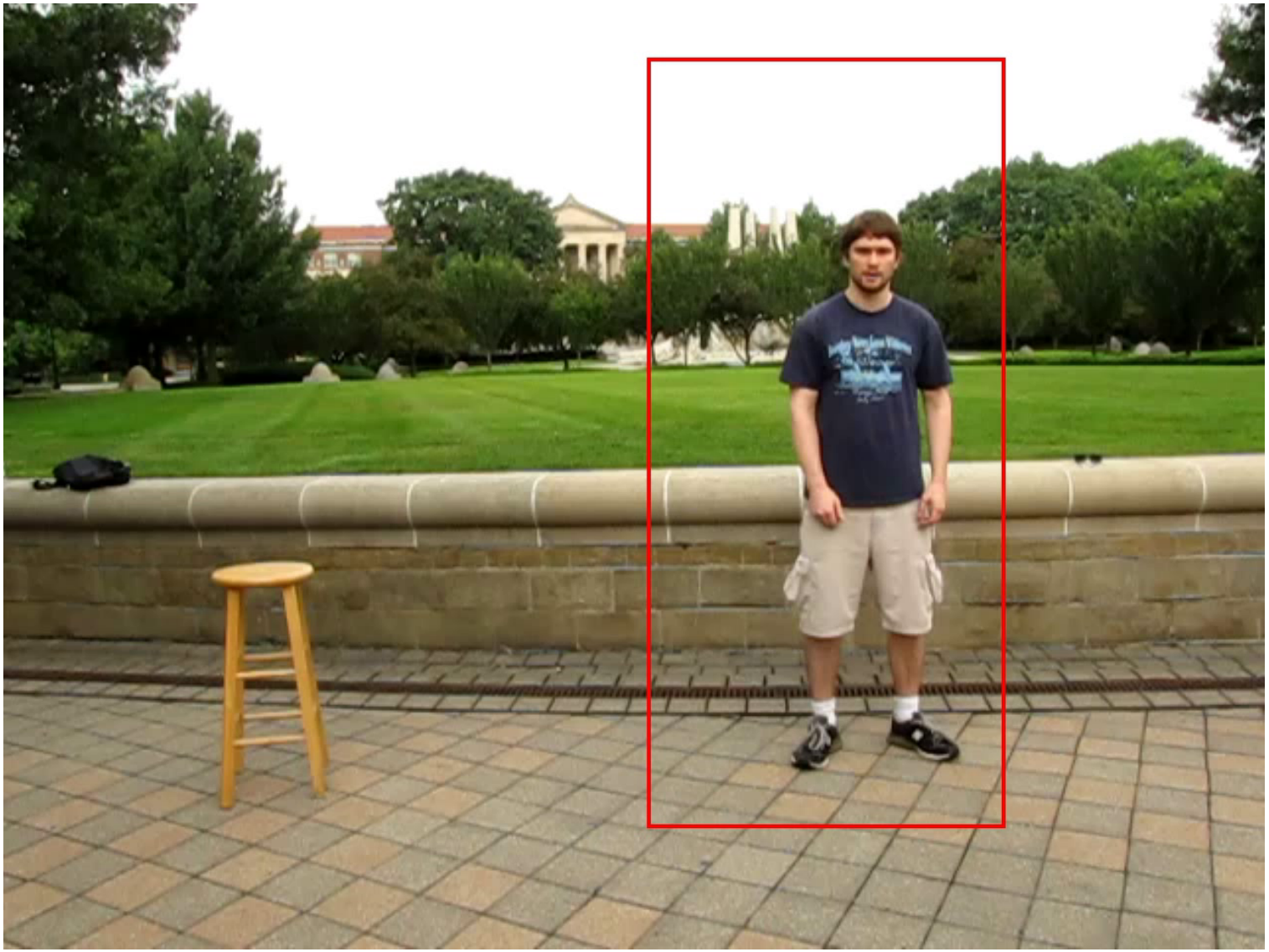
Object Detection *and Overgeneration*



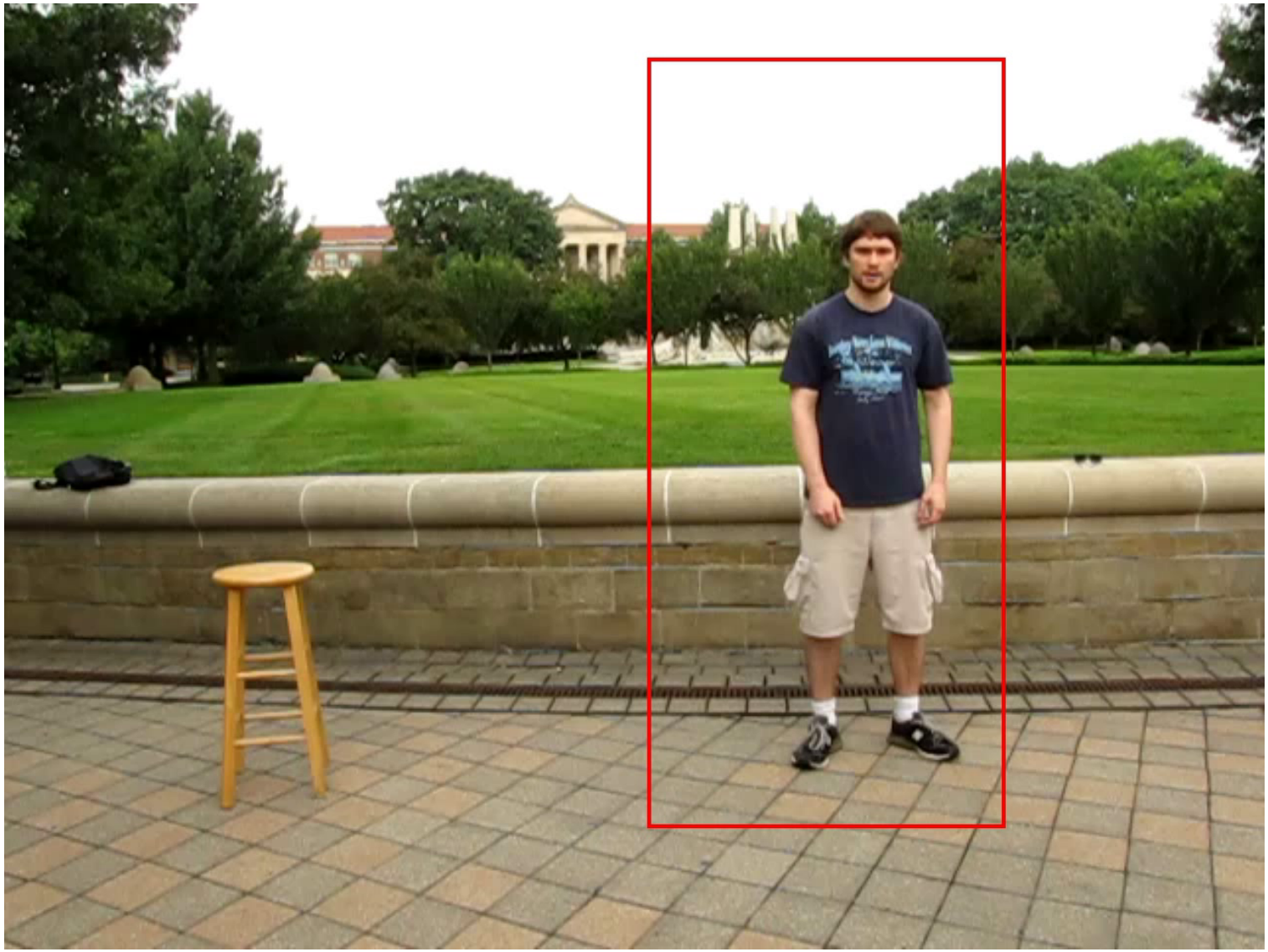
Object Detection *and Overgeneration*



Viterbi Tracker



Viterbi Tracker

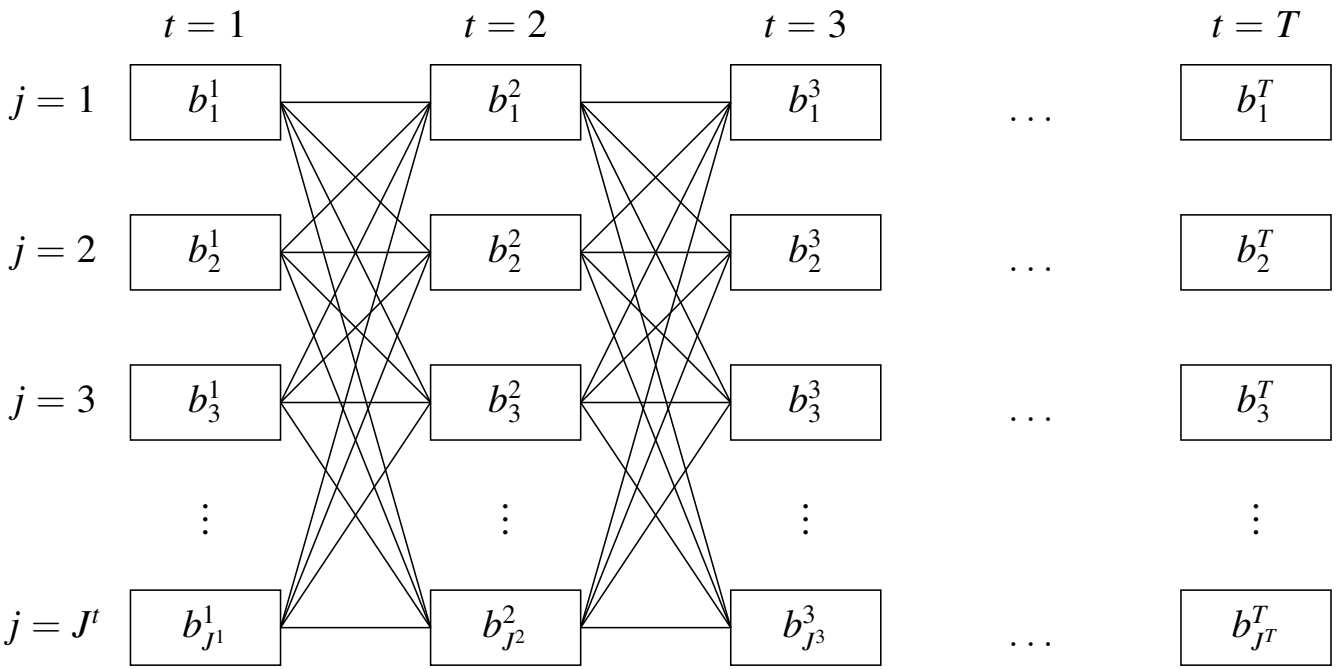


Viterbi Tracker

Optimal state sequence for HMM (Viterbi 1967)

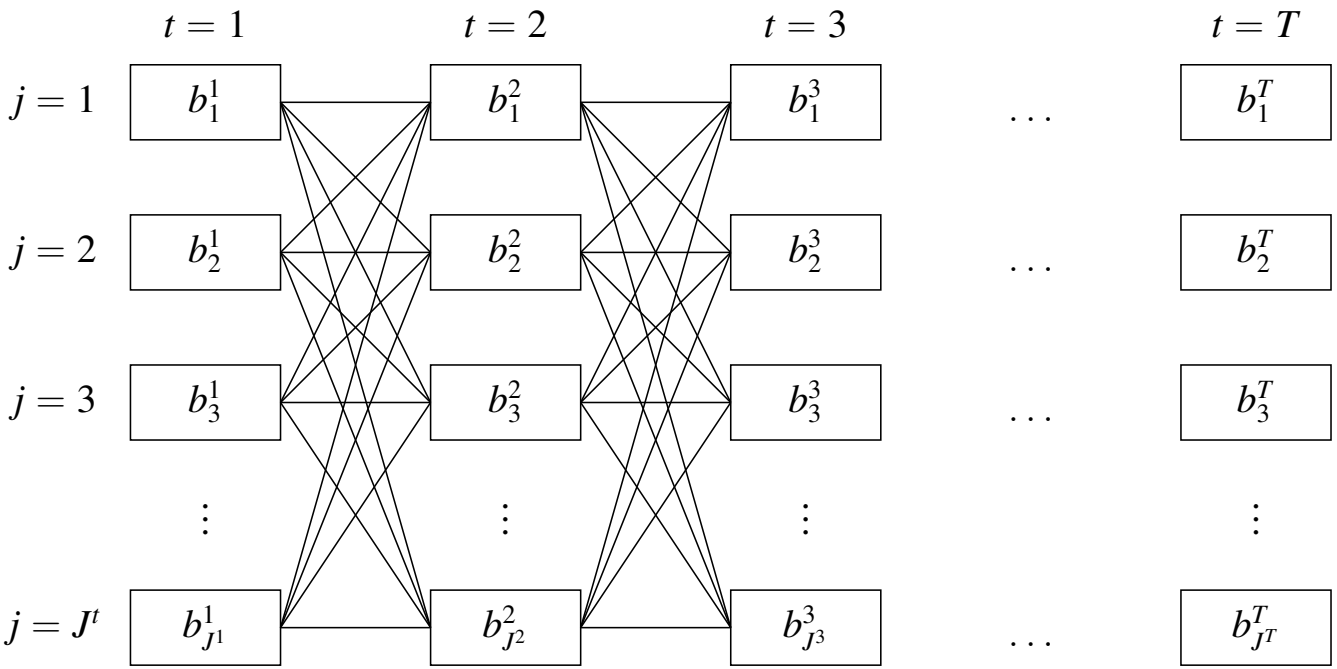
Viterbi Tracker

Optimal state sequence for HMM (Viterbi 1967)



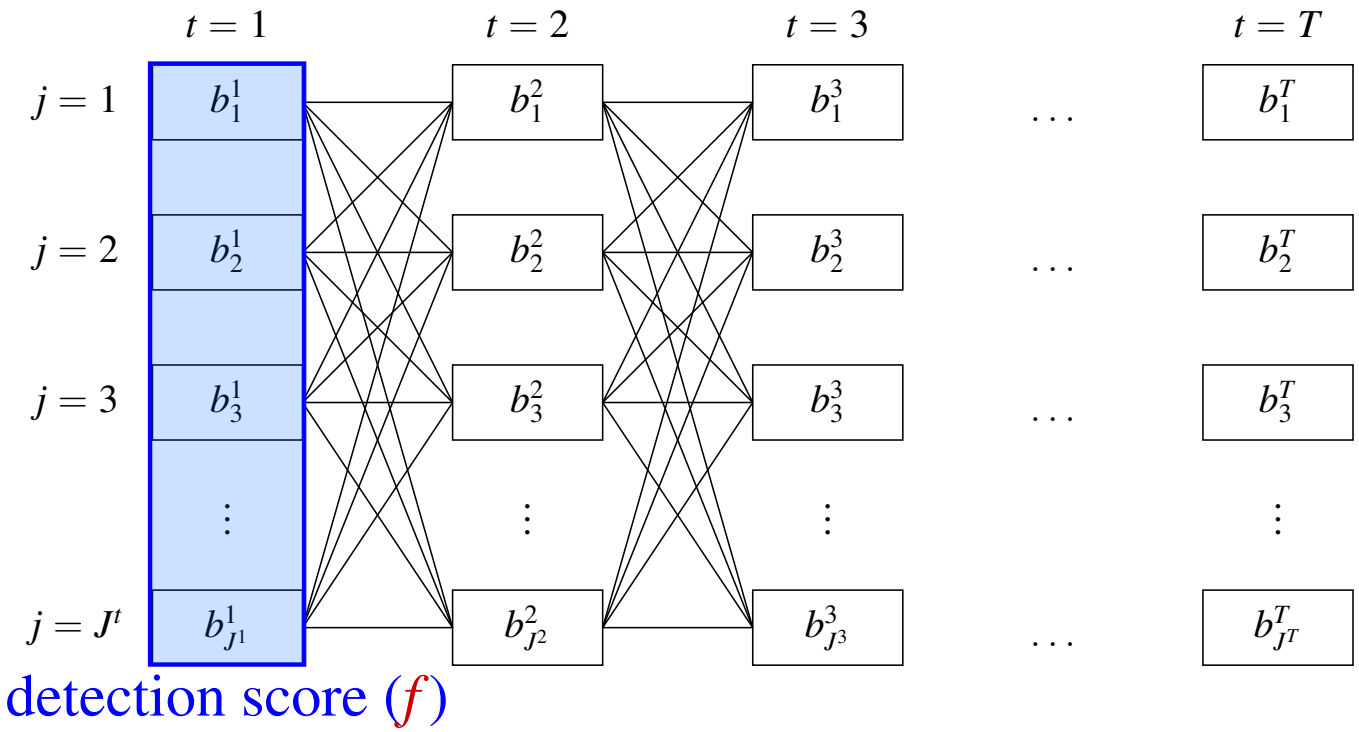
Viterbi Tracker

Optimal state sequence for HMM (Viterbi 1967)
Detection-based tracking (Viterbi 1971)



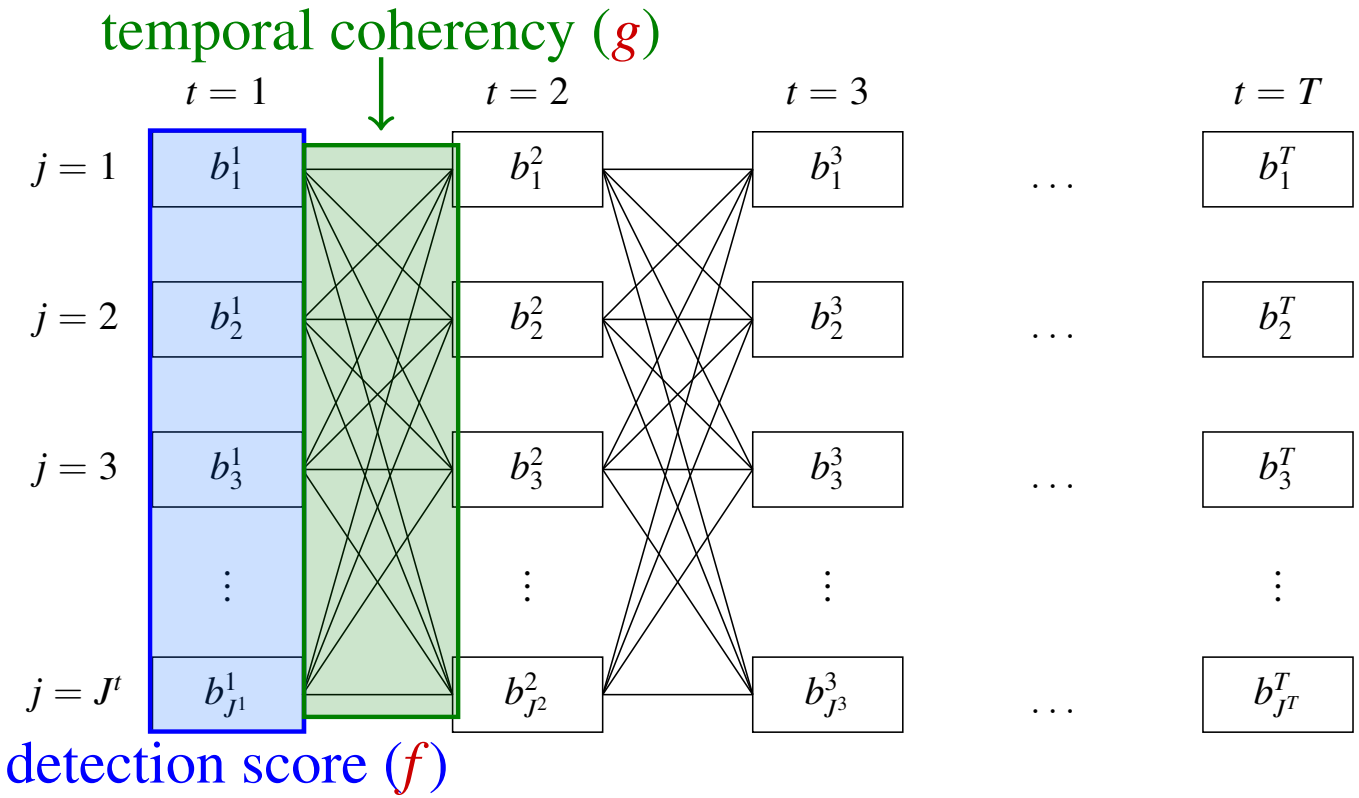
Viterbi Tracker

Optimal state sequence for HMM (Viterbi 1967)
Detection-based tracking (Viterbi 1971)



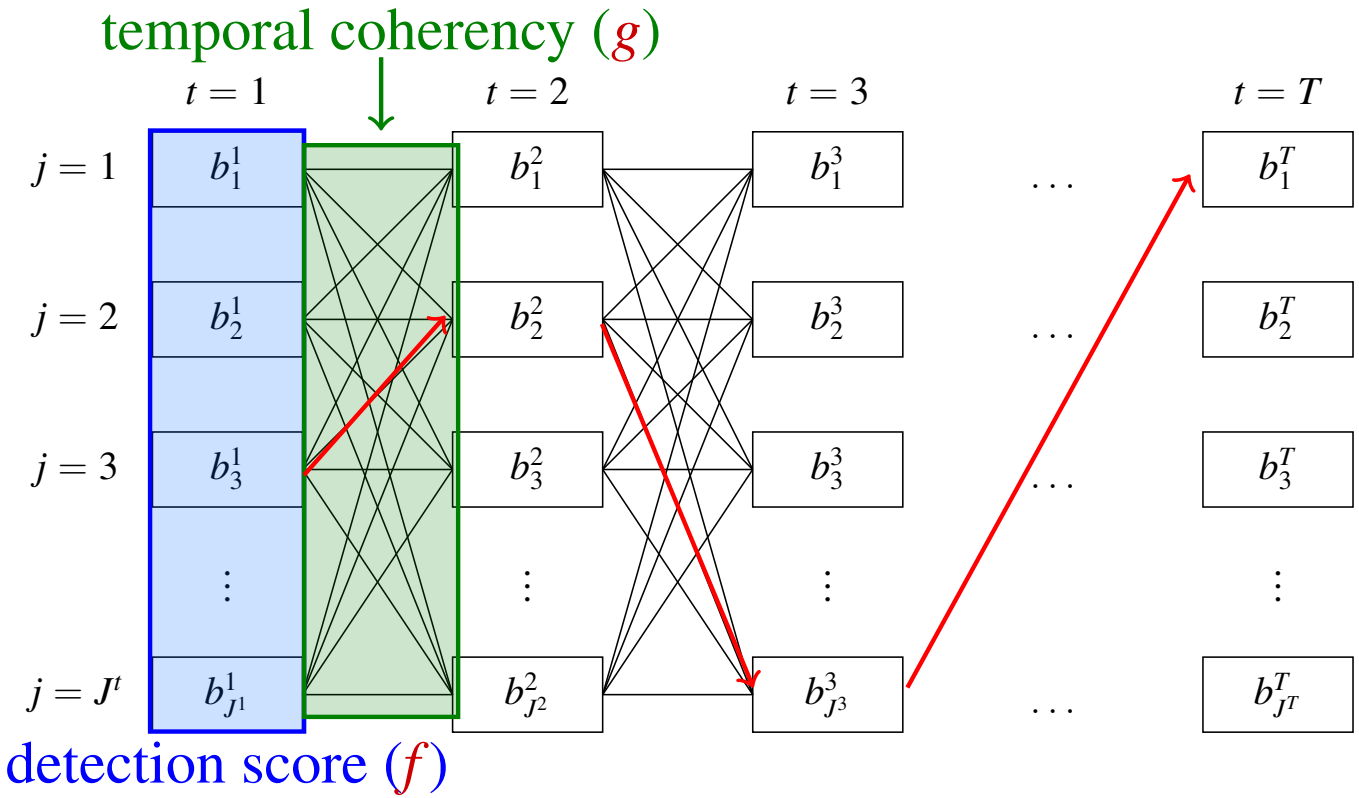
Viterbi Tracker

Optimal state sequence for HMM (Viterbi 1967)
Detection-based tracking (Viterbi 1971)



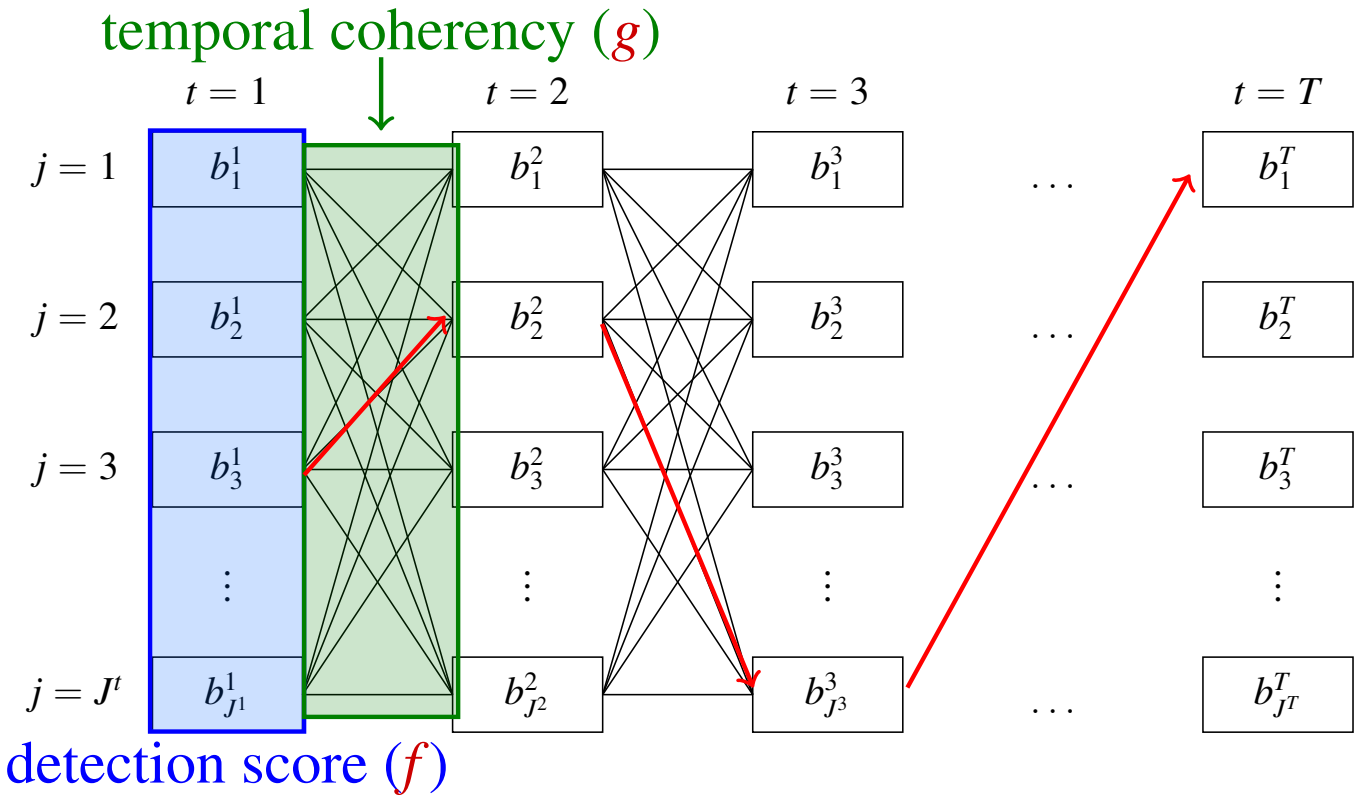
Viterbi Tracker

Optimal state sequence for HMM (Viterbi 1967)
Detection-based tracking (Viterbi 1971)



Viterbi Tracker

Optimal state sequence for HMM (Viterbi 1967)
Detection-based tracking (Viterbi 1971)



$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b_{j^t}^t) + \sum_{t=2}^{T} g(b_{j^{t-1}}^{t-1}, b_{j^t}^t)$$

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP} \\
\text{D} &\rightarrow \textit{the}
\end{aligned}
$$

$$
\begin{aligned}
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \text{emphaway from}
\end{aligned}
$$

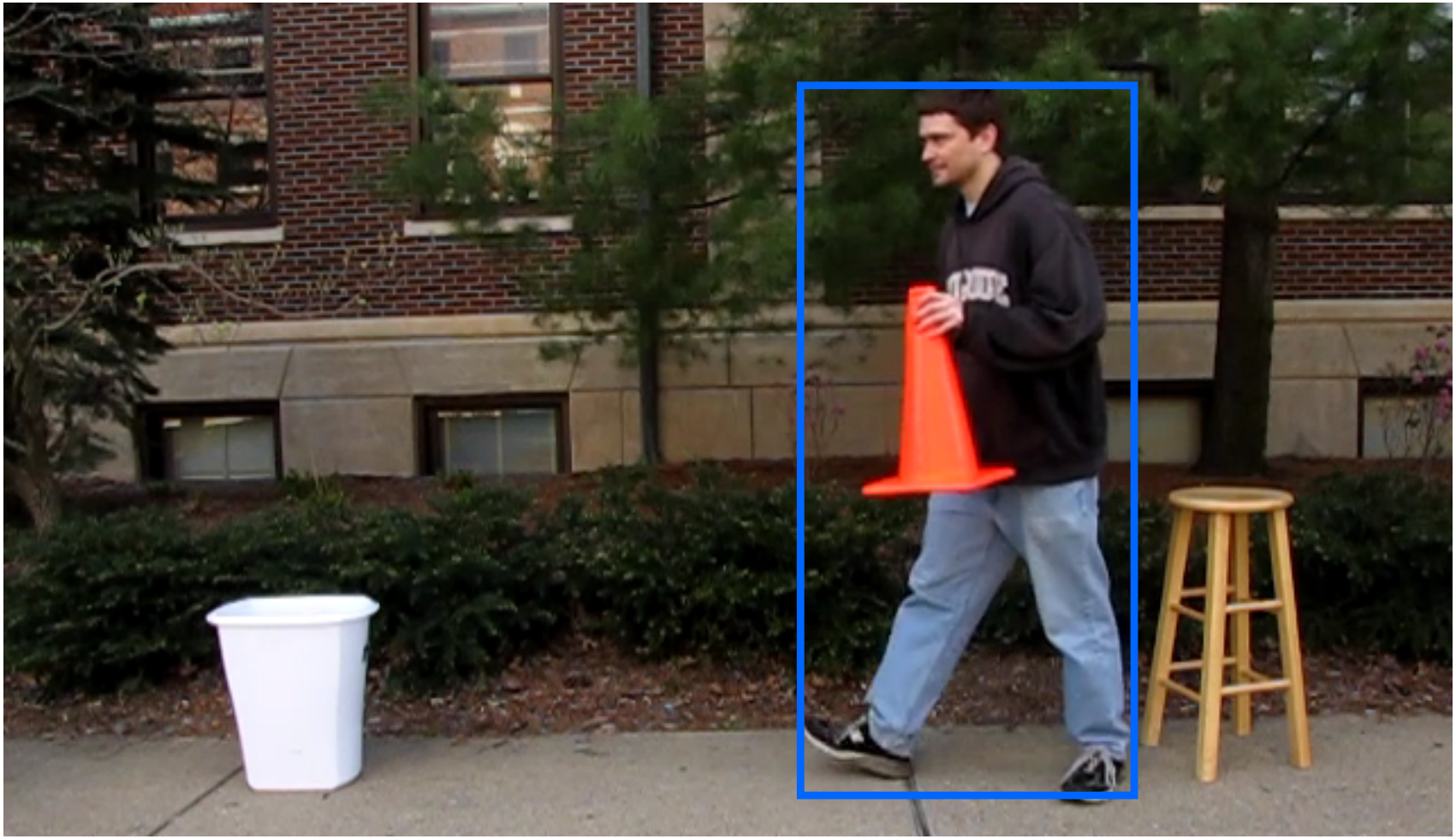
Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP} \\
\text{D} &\rightarrow \textit{the}
\end{aligned}
$$

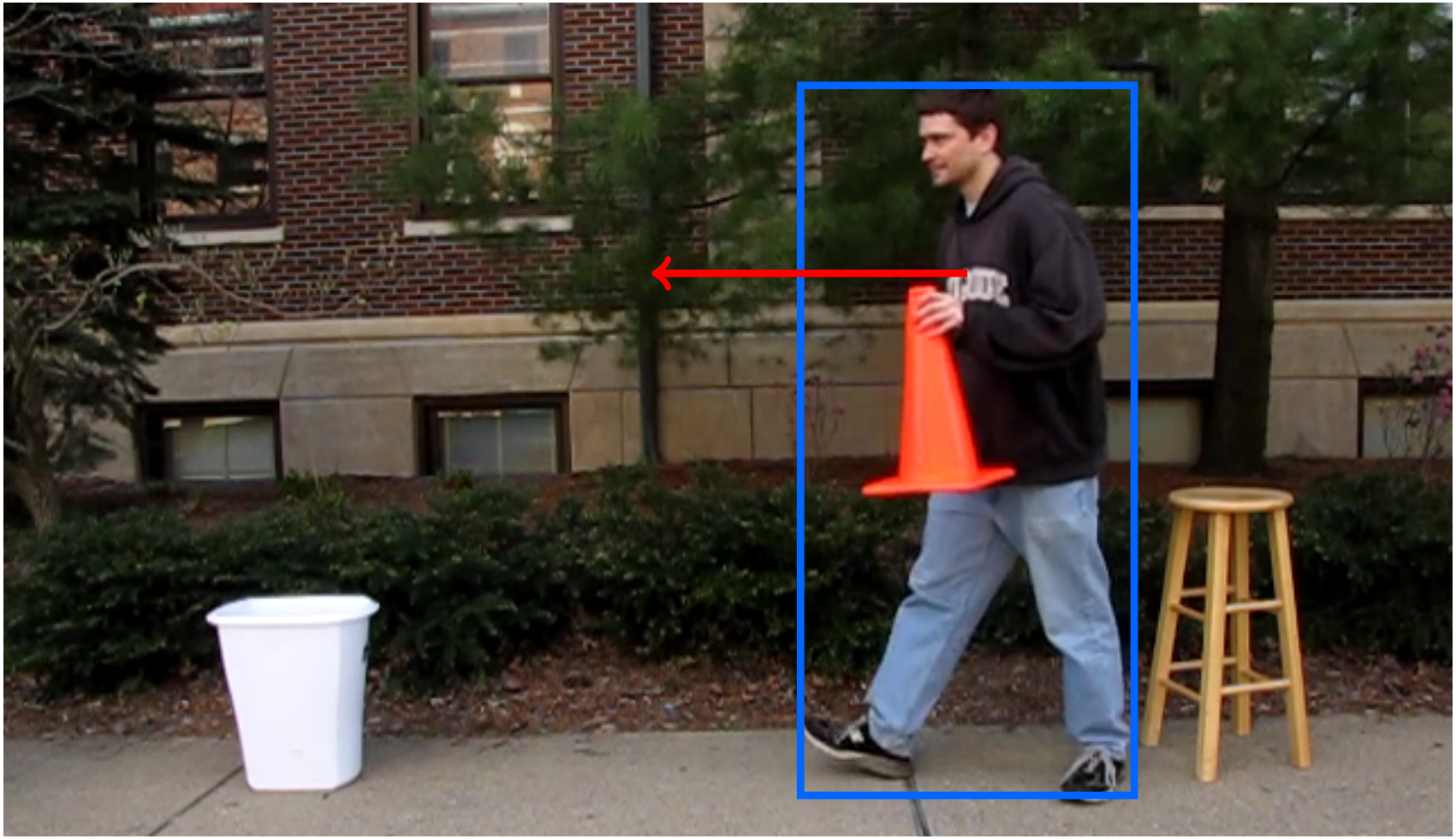
$$
\begin{aligned}
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \text{emphaway from}
\end{aligned}
$$

Visual Features

Visual Features

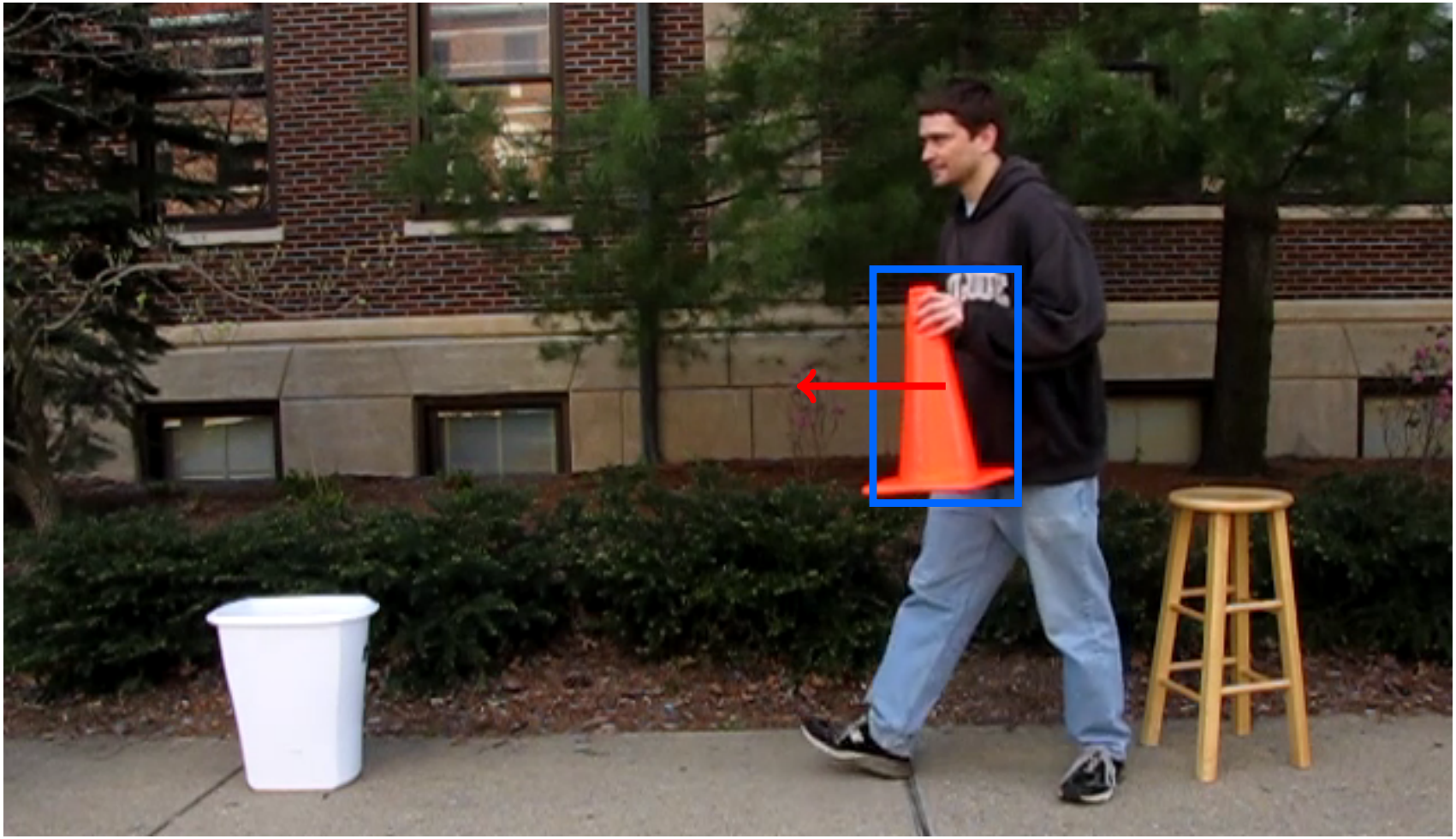


Visual Features



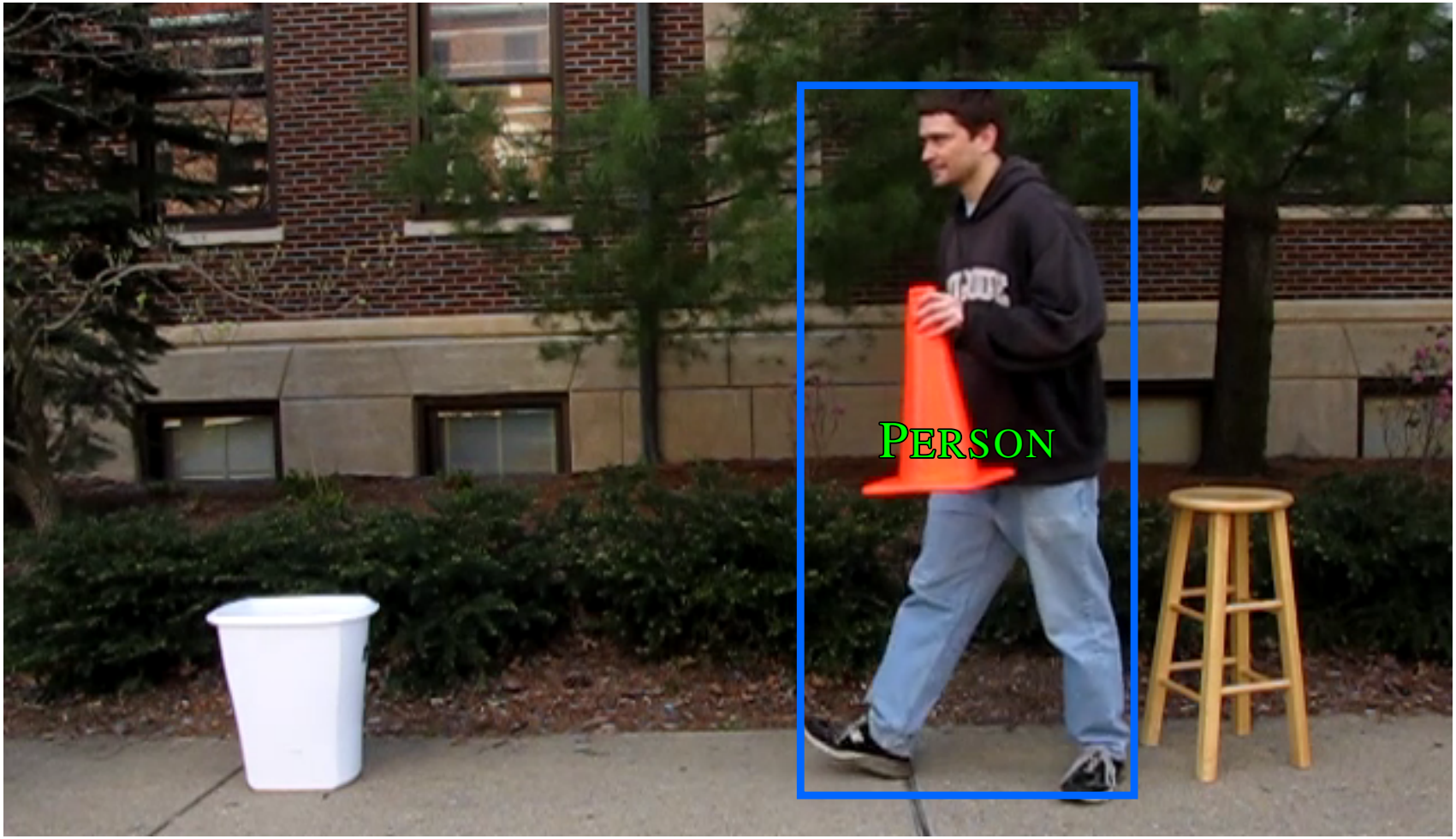
velocity object class distance orientation

Visual Features



velocity object class distance orientation

Visual Features

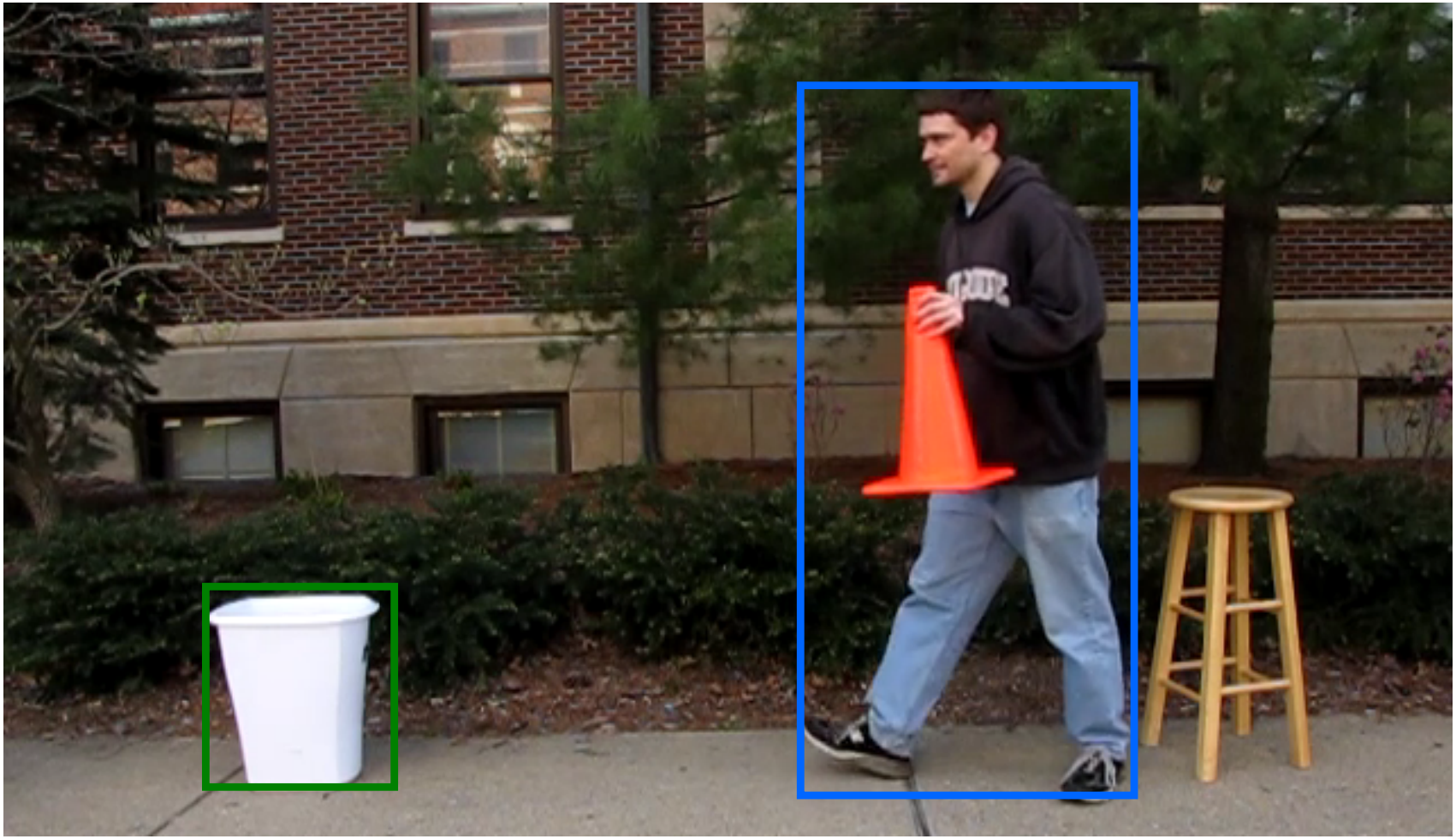


velocity object class distance orientation

Visual Features

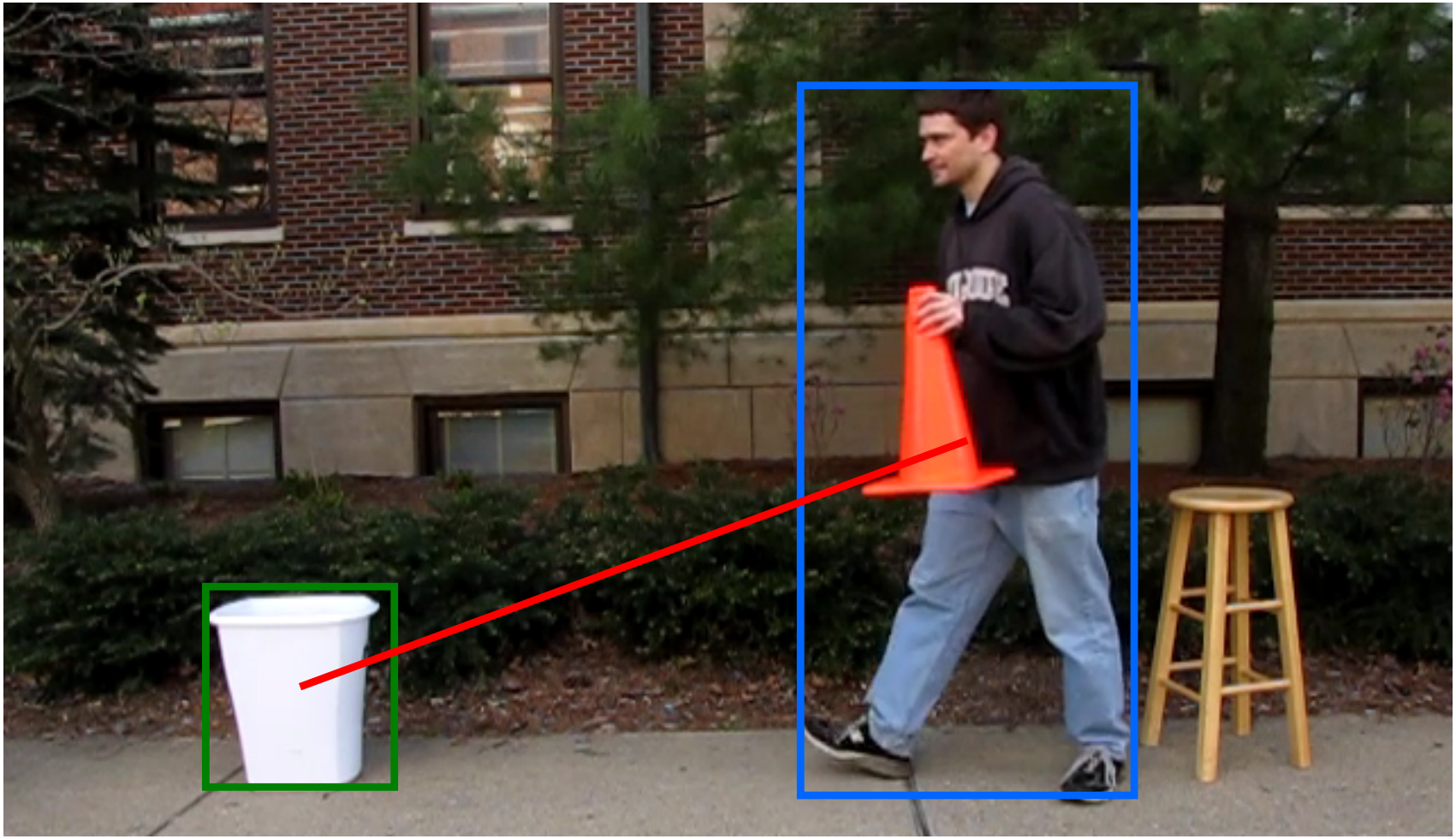
velocity object class distance orientation

Visual Features



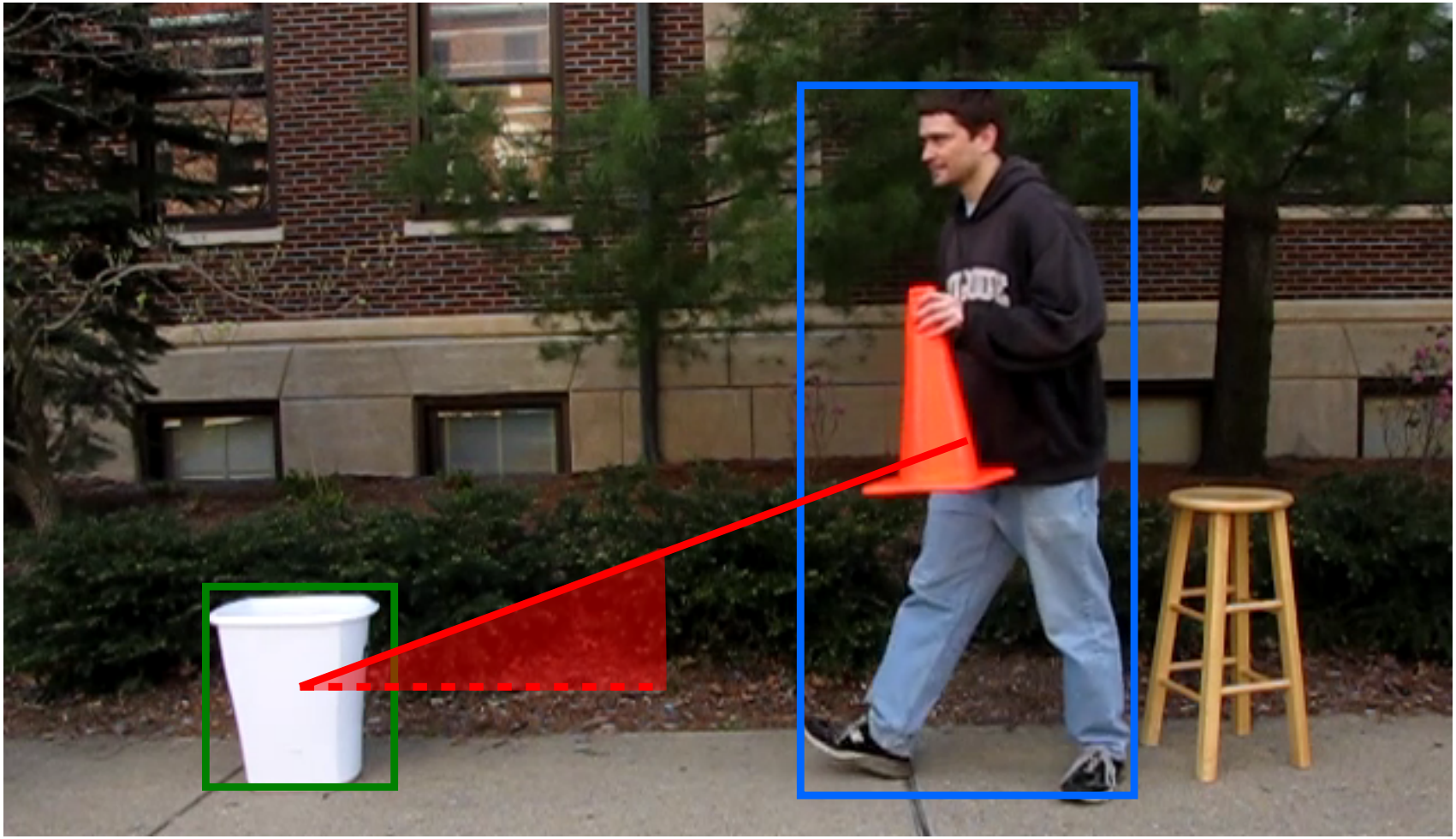
velocity object class distance orientation

Visual Features



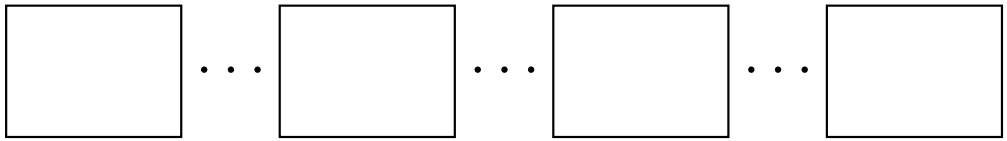
velocity object class distance orientation

Visual Features



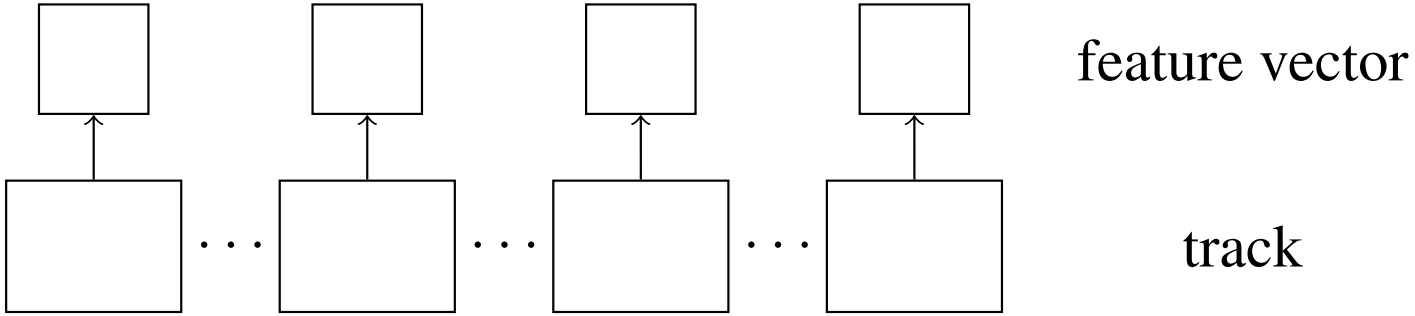
velocity object class distance orientation

Event Recognition with Hidden Markov Models

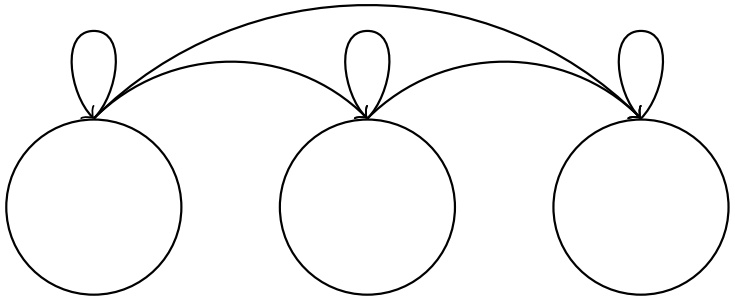


track

Event Recognition with Hidden Markov Models

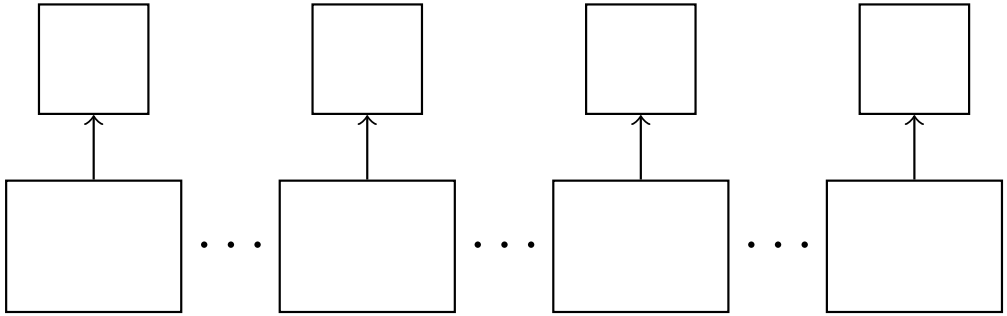


Event Recognition with Hidden Markov Models



a

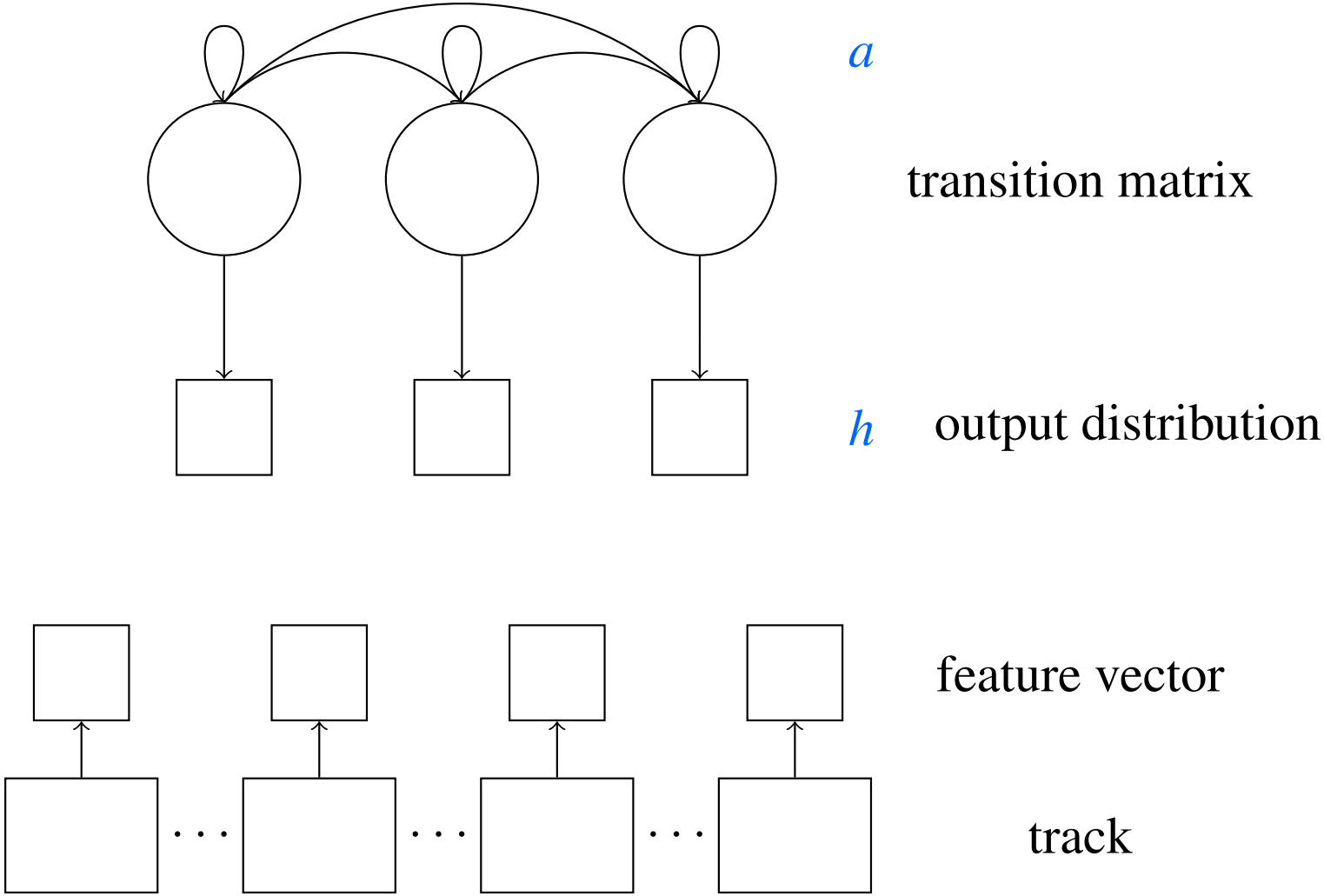
transition matrix



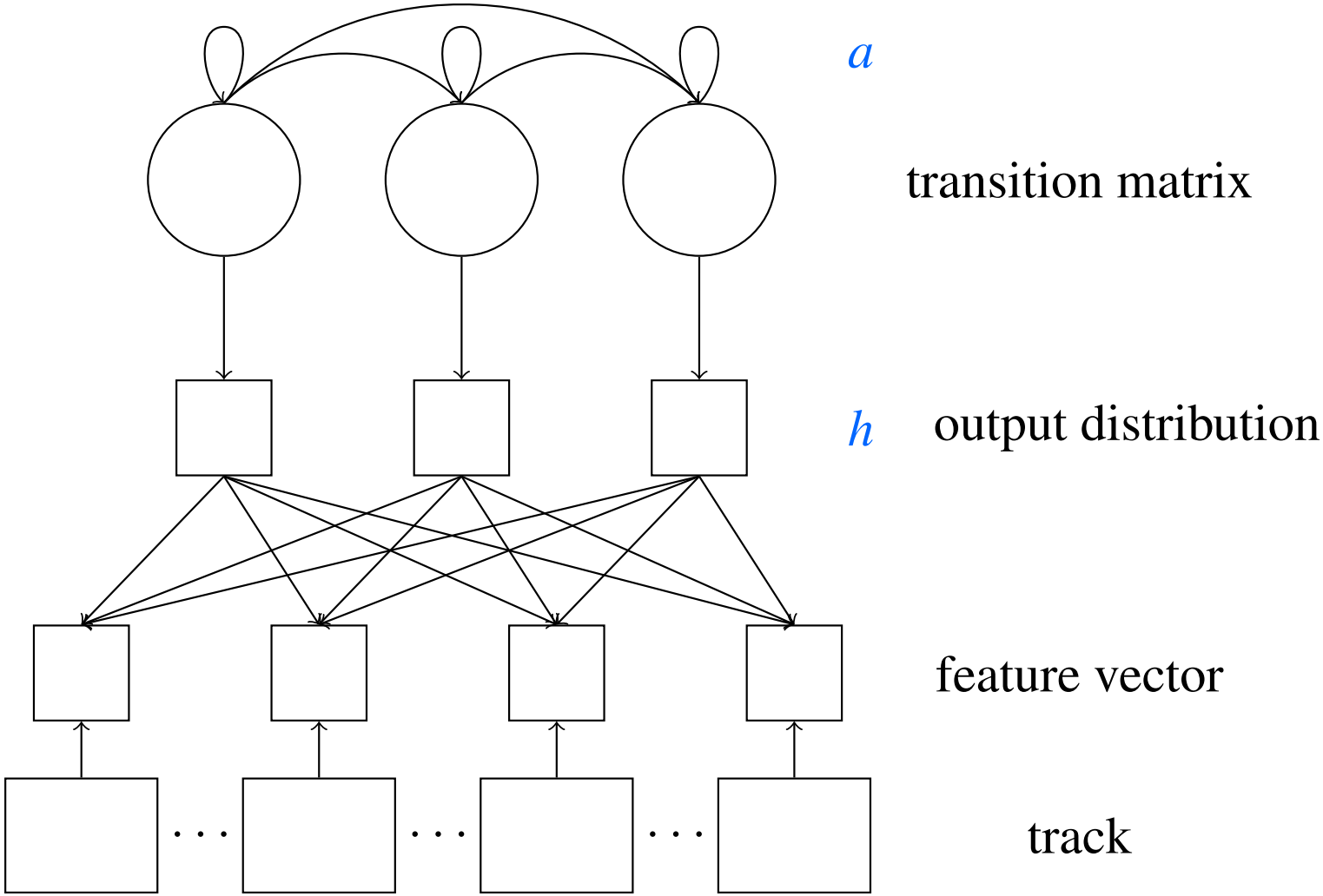
feature vector

track

Event Recognition with Hidden Markov Models

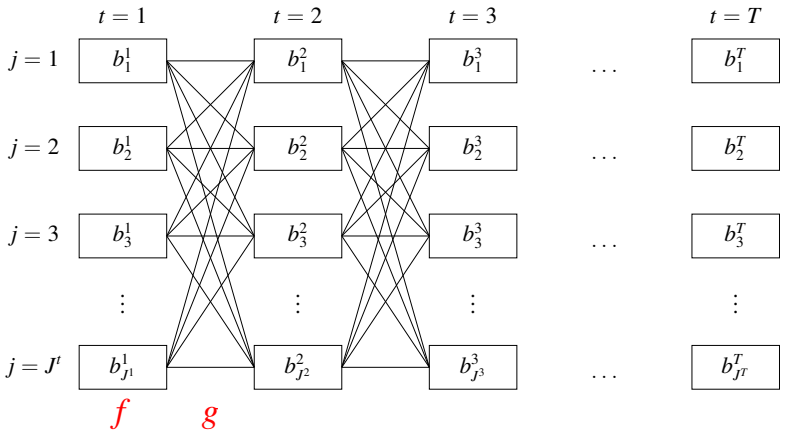


Event Recognition with Hidden Markov Models



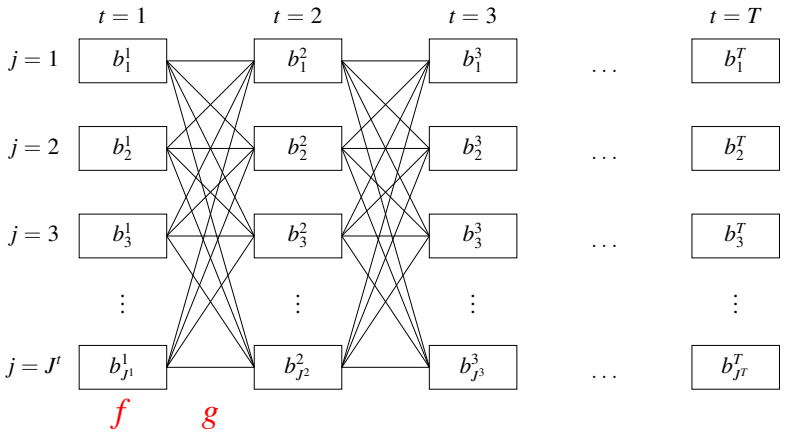
Event Tracker

Event Tracker

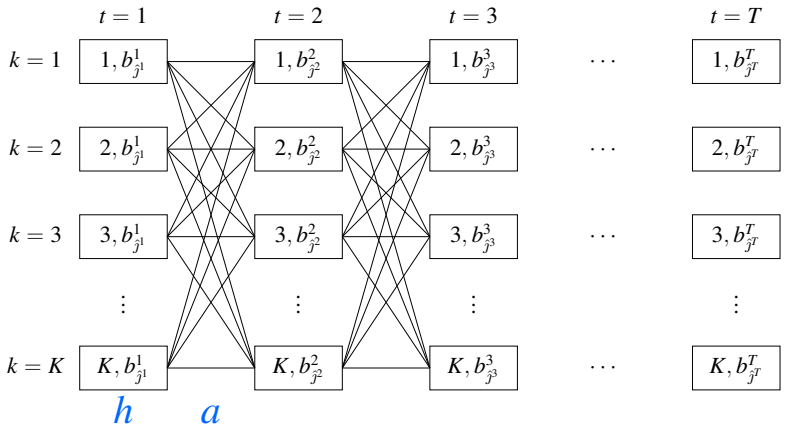


tracker lattice

Event Tracker

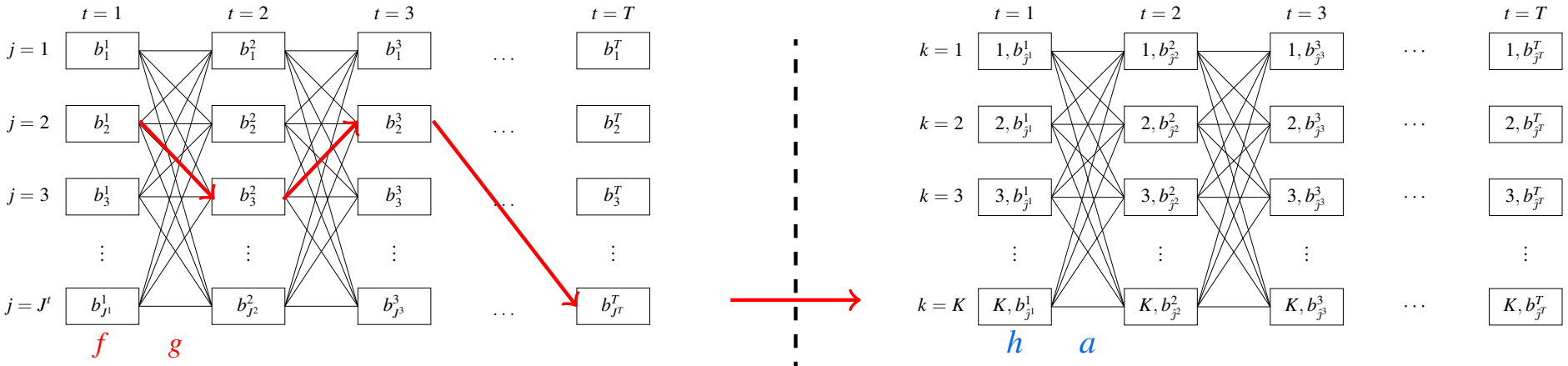


tracker lattice

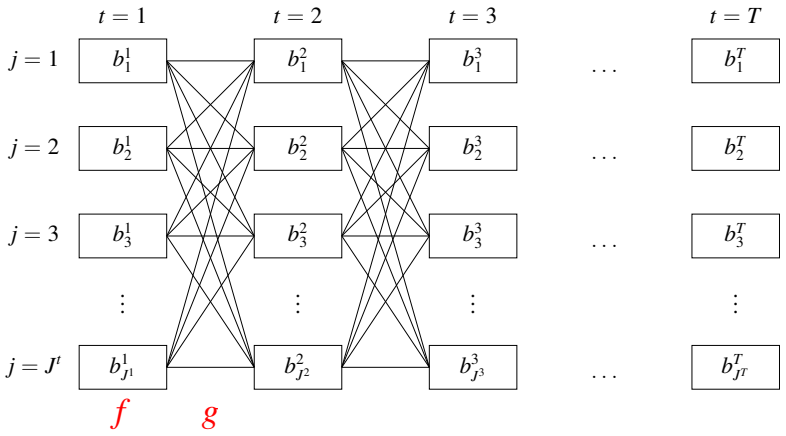


verb lattice

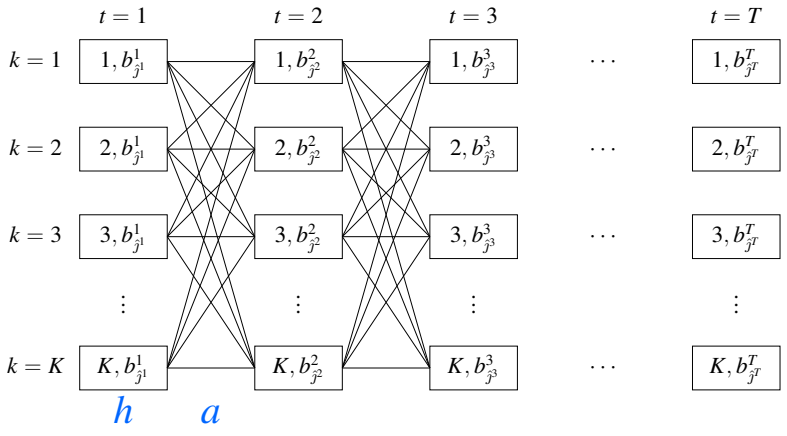
Event Tracker



Event Tracker



tracker lattice

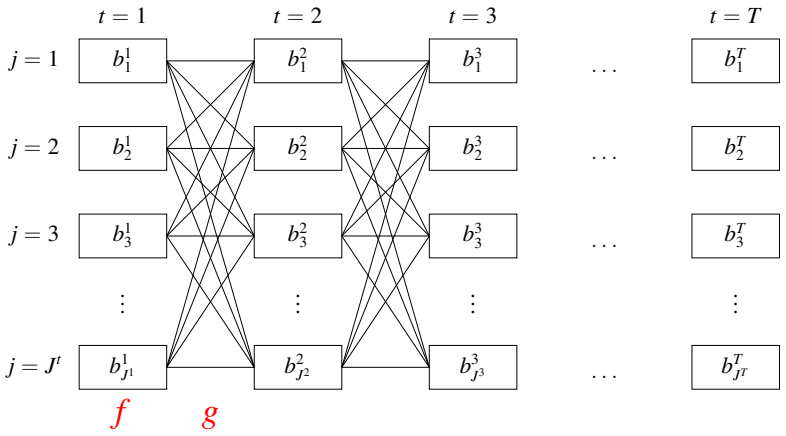


verb lattice

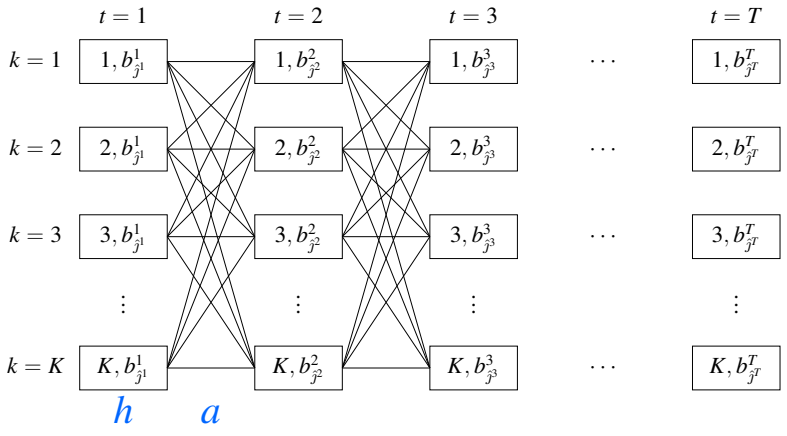
$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b_{j^t}^t) + \sum_{t=2}^{T} g(b_{j^{t-1}}^{t-1}, b_{j^t}^t)$$

$$\max_{k^1,\ldots,k^T} \sum_{t=1}^{T} h(k^t, b^t) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Event Tracker



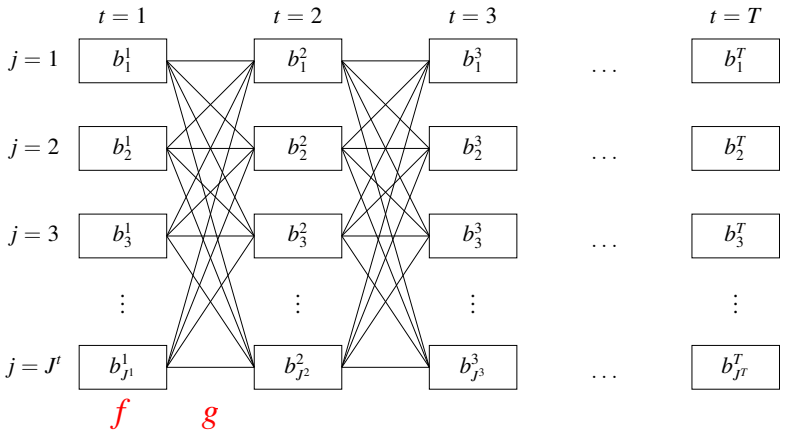
tracker lattice



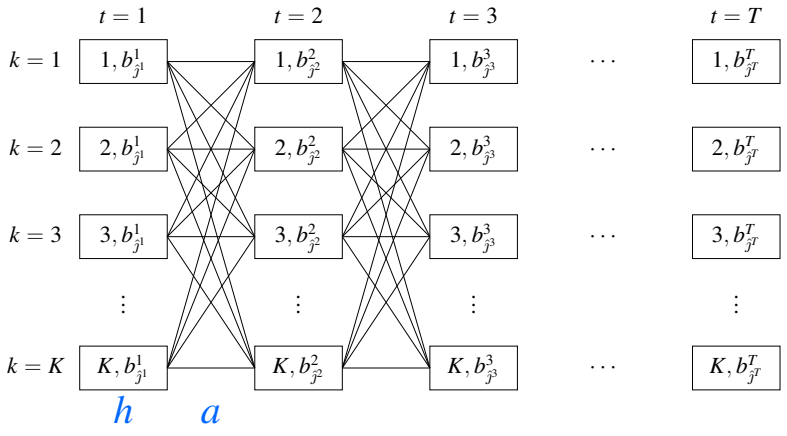
verb lattice

$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b_{j^t}^t) + \sum_{t=2}^{T} g(b_{j^{t-1}}^{t-1}, b_{j^t}^t) \quad + \quad \max_{k^1,\ldots,k^T} \sum_{t=1}^{T} h(k^t, b^t) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Event Tracker



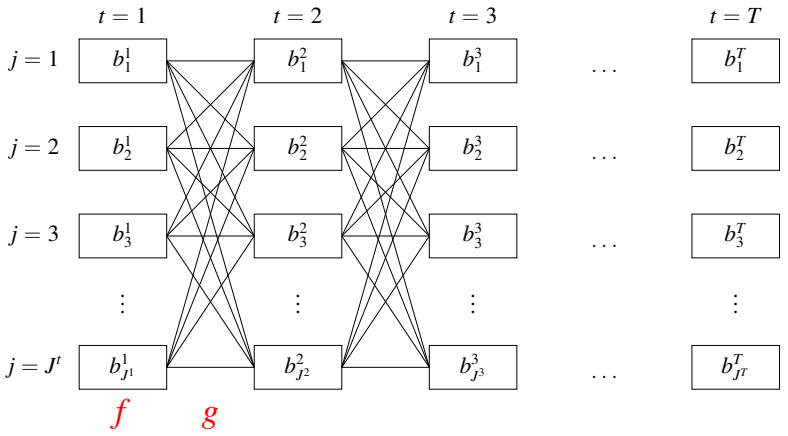
tracker lattice



verb lattice

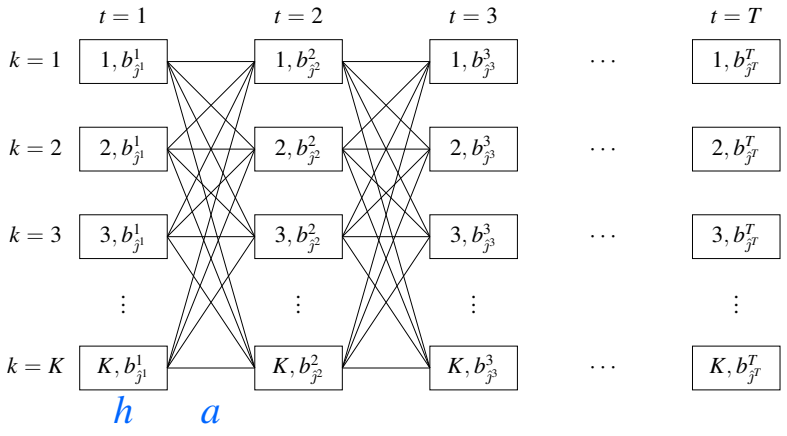
$$\max_{j^1,\ldots,j^T} \max_{k^1,\ldots,k^T} \sum_{t=1}^{T} f(b_{j^t}^t) + \sum_{t=2}^{T} g(b_{j^{t-1}}^{t-1}, b_{j^t}^t) + \sum_{t=1}^{T} h(k^t, b_{j^t}^t) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Event Tracker



tracker lattice

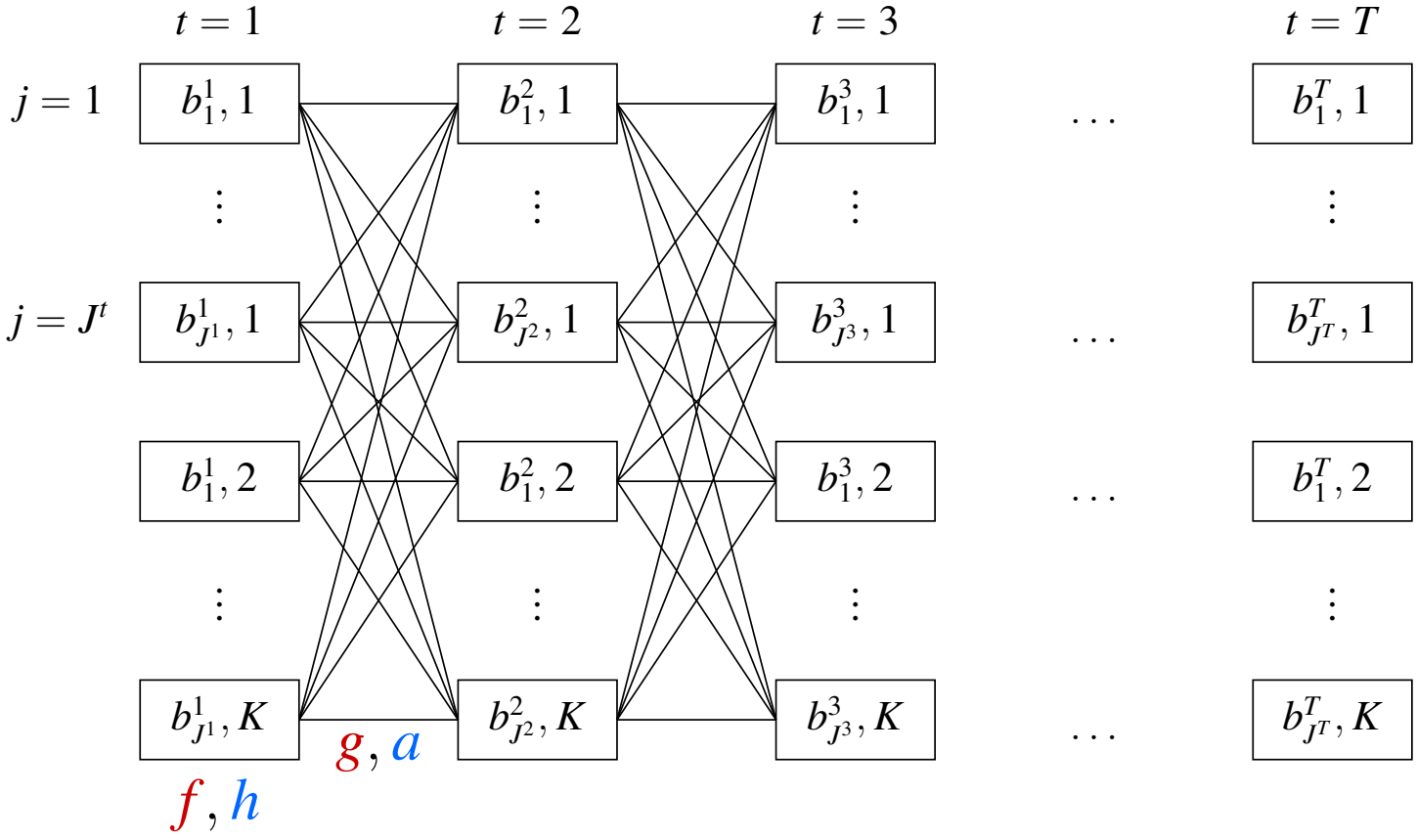
$\times$



verb lattice

$$\max_{j^1,\ldots,j^T} \max_{k^1,\ldots,k^T} \sum_{t=1}^{T} f(b_{j^t}^t) + \sum_{t=2}^{T} g(b_{j^{t-1}}^{t-1}, b_{j^t}^t) + \sum_{t=1}^{T} h(k^t, b_{j^t}^t) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Event Tracker



$$\max_{j^1,\ldots,j^T} \max_{k^1,\ldots,k^T} \sum_{t=1}^{T} f(b_{j^t}^t) + \sum_{t=2}^{T} g(b_{j^{t-1}}^{t-1}, b_{j^t}^t) + \sum_{t=1}^{T} h(k^t, b_{j^t}^t) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP} \\
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP} \\
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP} \\
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

- model other words also as HMMs

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP} \\
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

- model other words also as HMMs

  *The **jump** was fast.* (Some nouns are dynamic.)

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP} \\
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

- model other words also as HMMs

  *The **jump** was fast.* (Some nouns are dynamic.)

  *The person **held** the backpack.* (Some verbs are static.)

Background Knowledge

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP} \\
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

- model other words also as HMMs

  *The **jump** was fast.* (Some nouns are dynamic.)

  *The person **held** the backpack.* (Some verbs are static.)

- unified learning algorithm

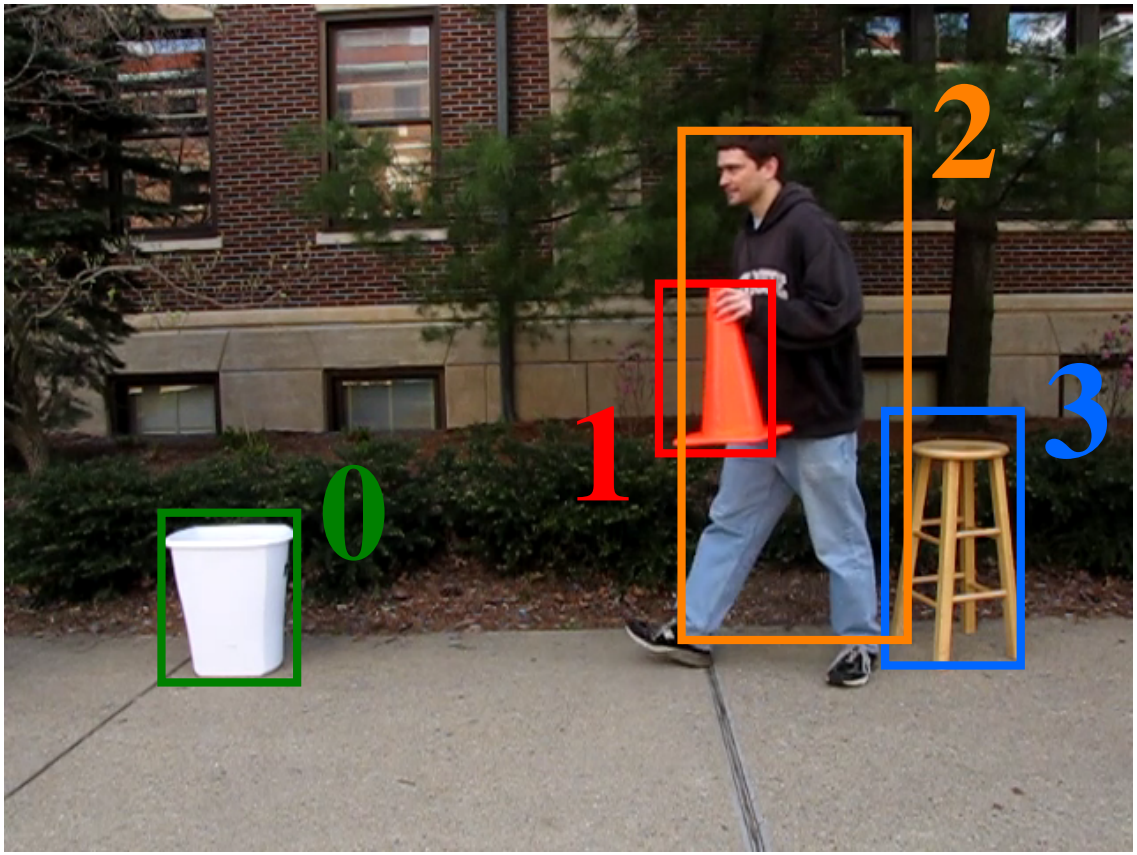
Sentence Tracker

The person to the left of the stool carried the traffic-cone towards the trash-can.

Sentence Tracker

The person to the left of the stool carried the traffic-cone towards the trash-can.

object 0 **object 1** **object 2** **object 3**

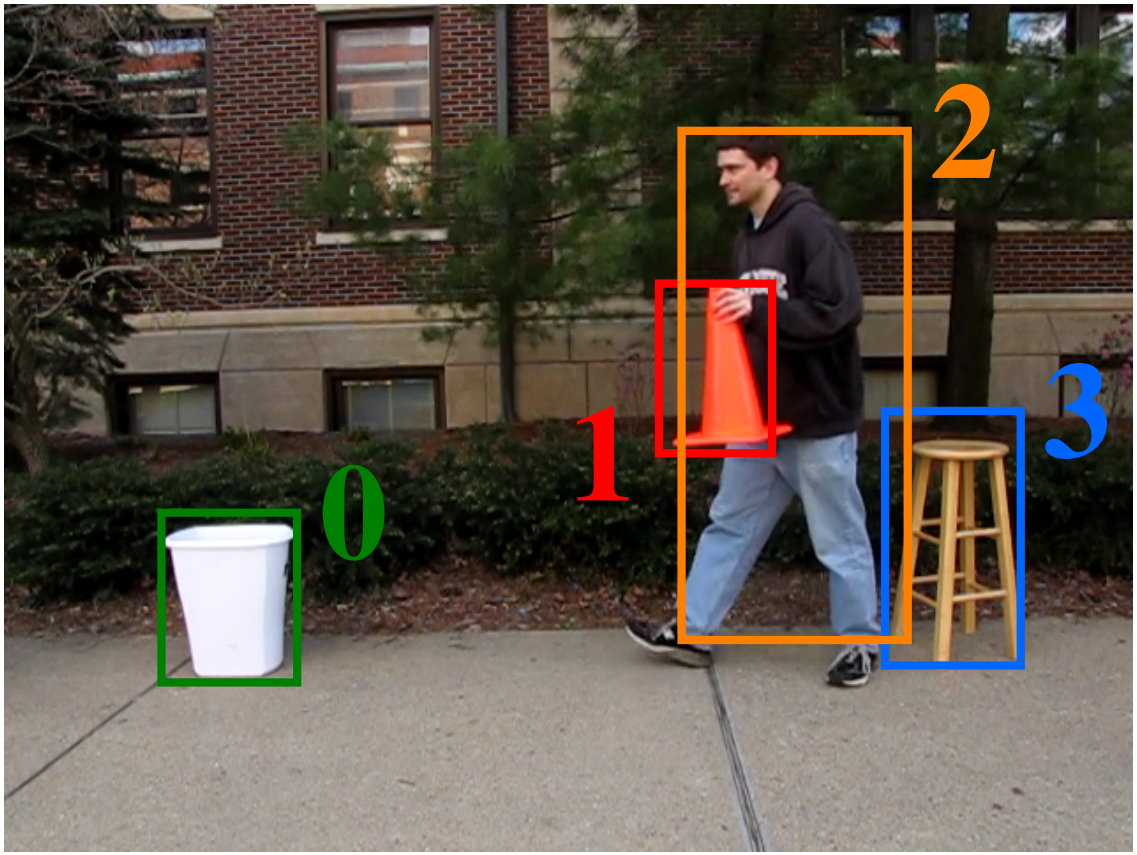


Sentence Tracker

The person to the left of the stool carried the traffic-cone towards the trash-can.

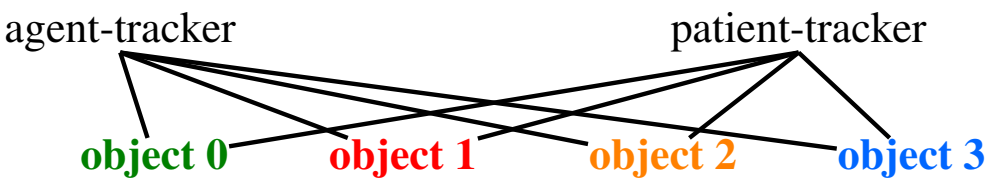
agent-tracker

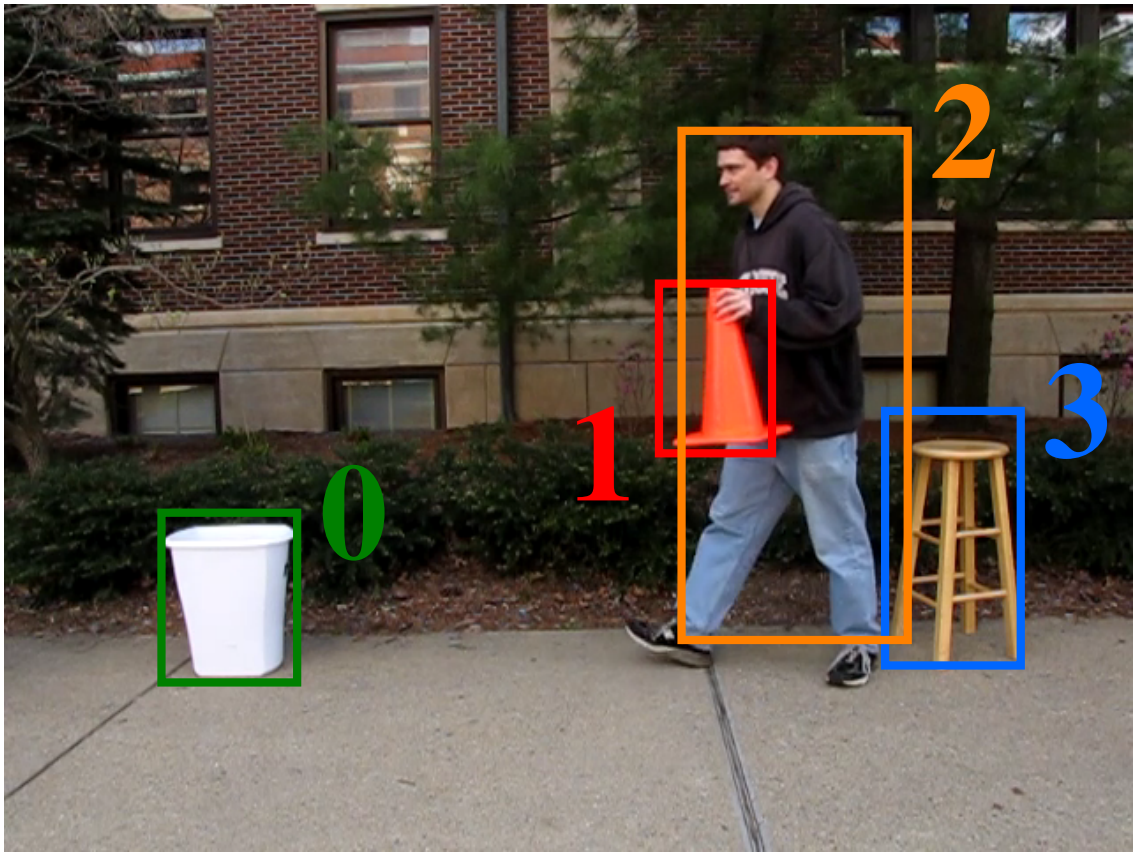
object 0 **object 1** **object 2** **object 3**



Sentence Tracker

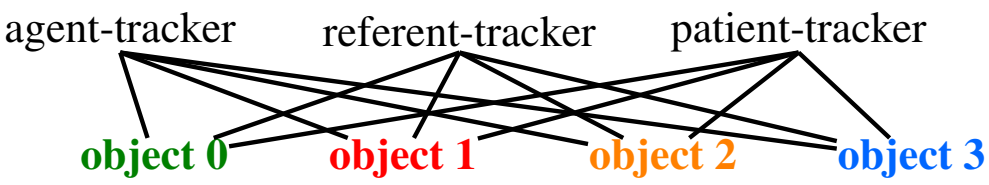
The person to the left of the stool carried the traffic-cone towards the trash-can.

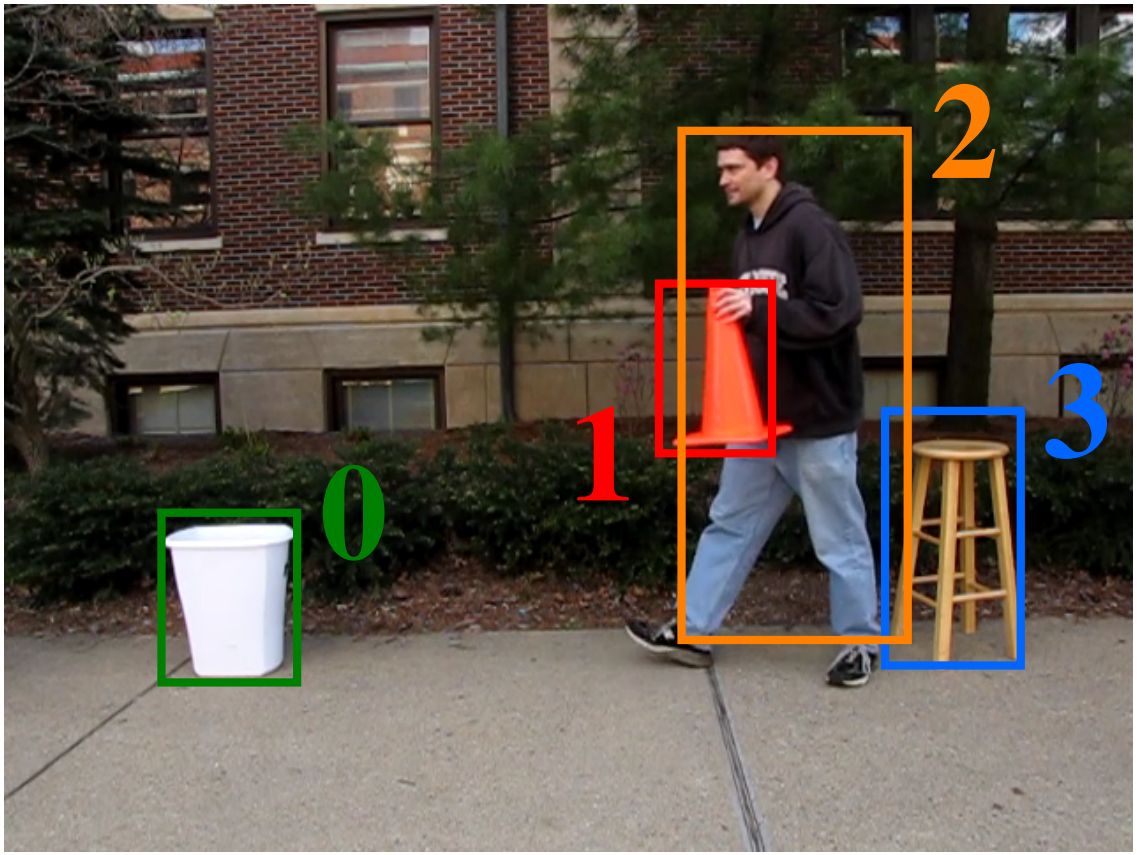




Sentence Tracker

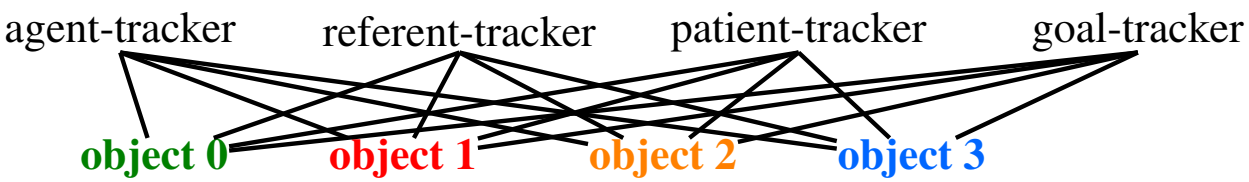
The person to the left of the stool carried the traffic-cone towards the trash-can.

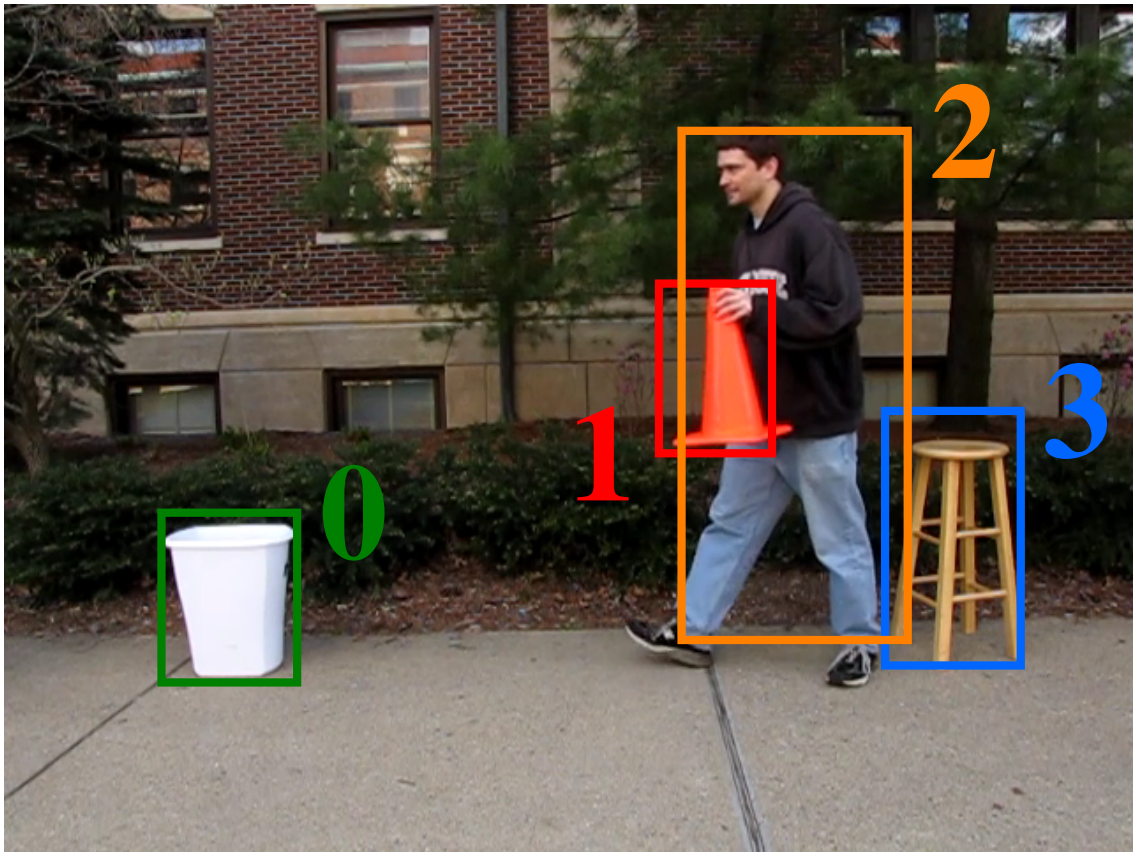




Sentence Tracker

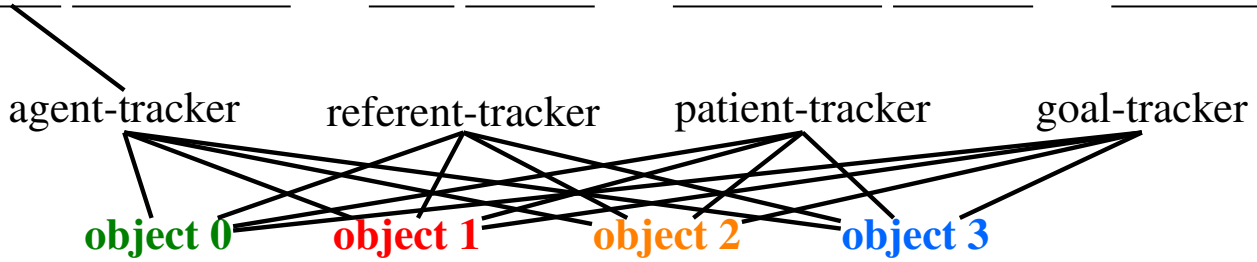
The person to the left of the stool carried the traffic-cone towards the trash-can.

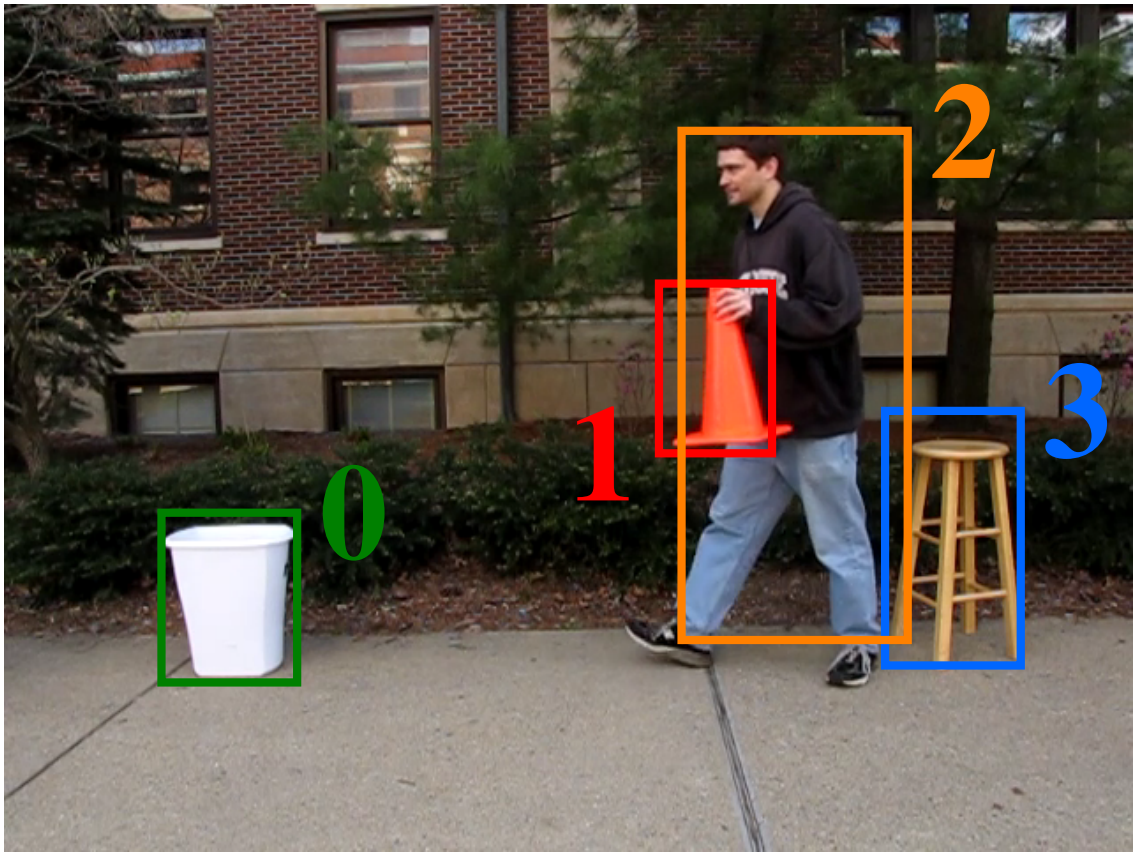




Sentence Tracker

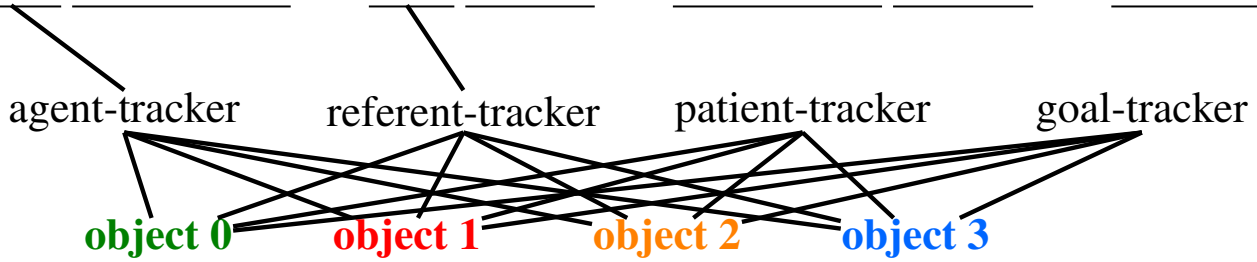
The person to the left of the stool carried the traffic-cone towards the trash-can.

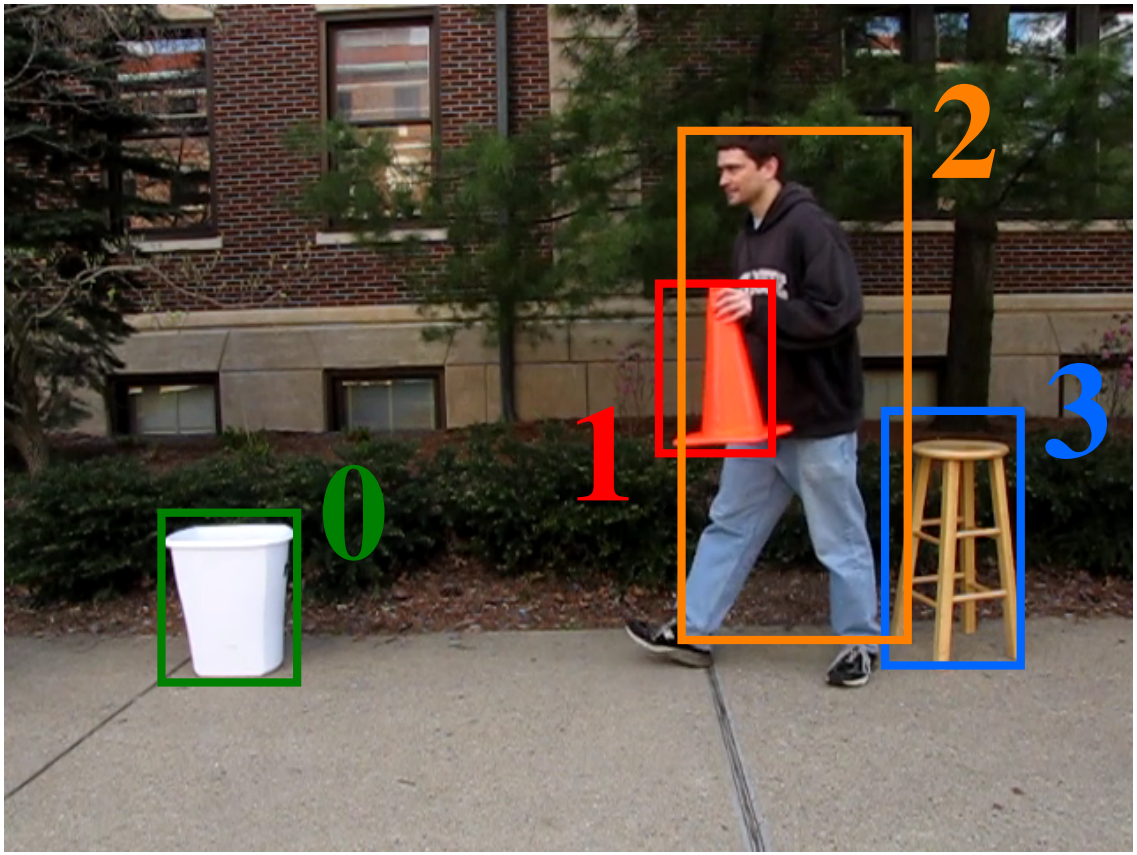




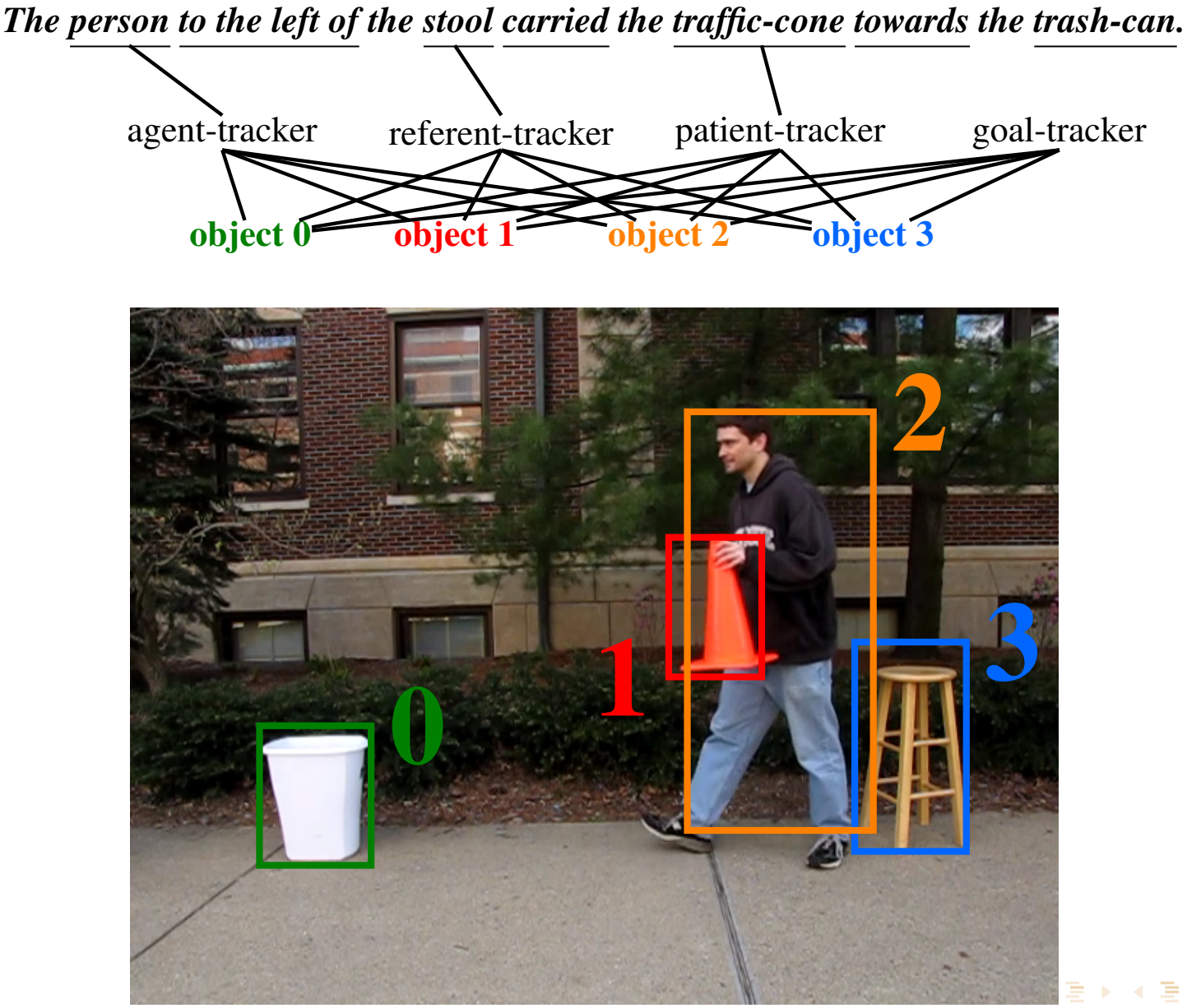
Sentence Tracker

The person to the left of the stool carried the traffic-cone towards the trash-can.

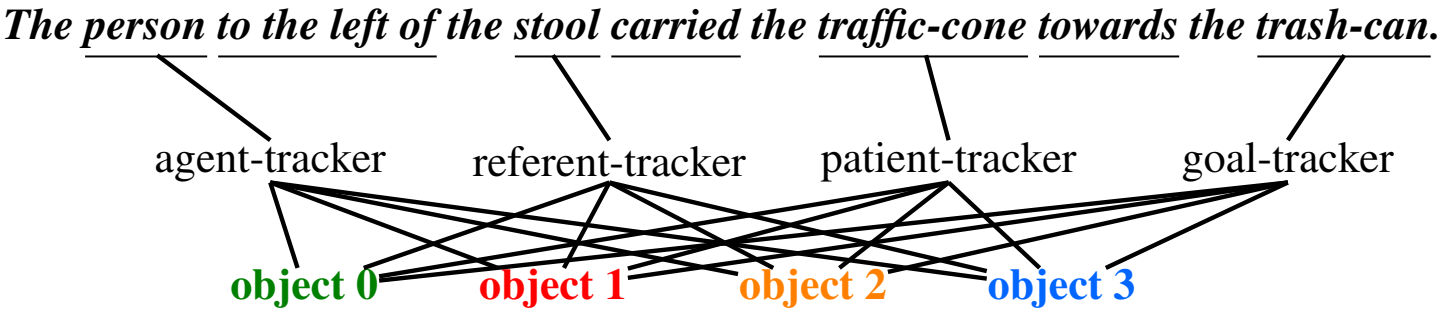


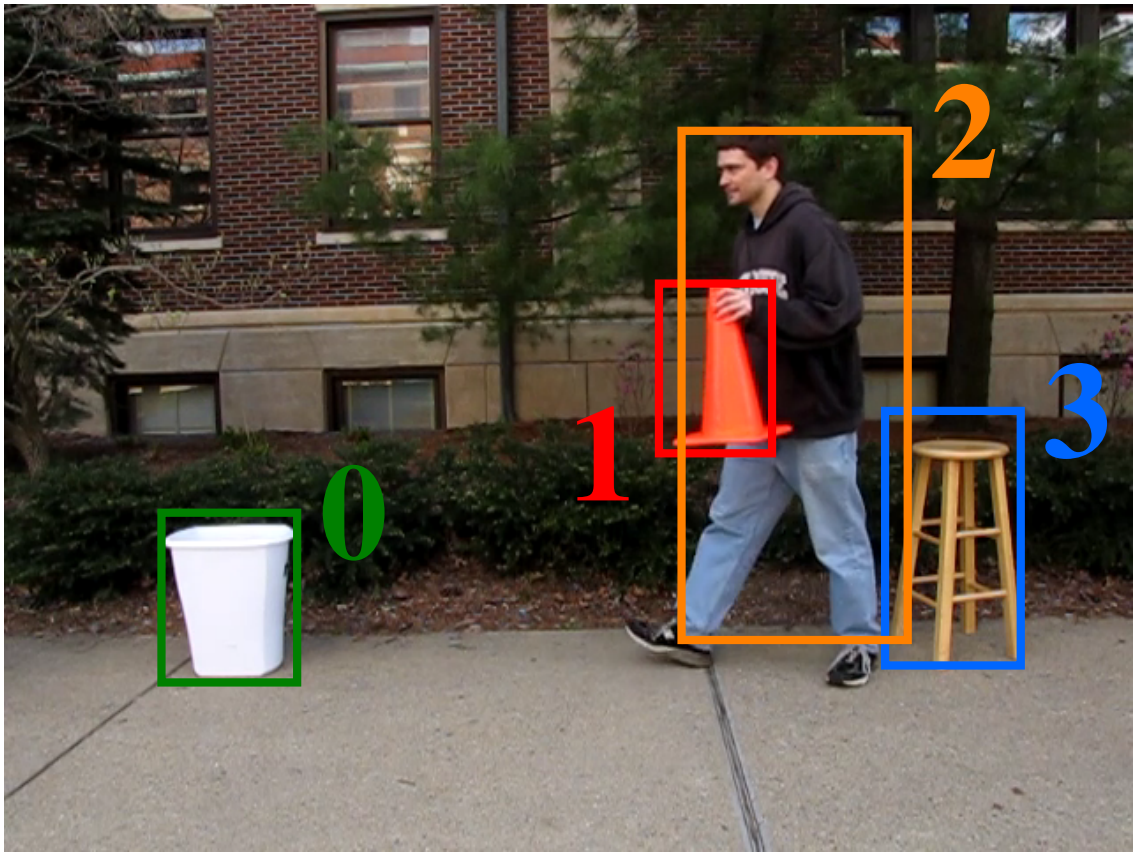


Sentence Tracker

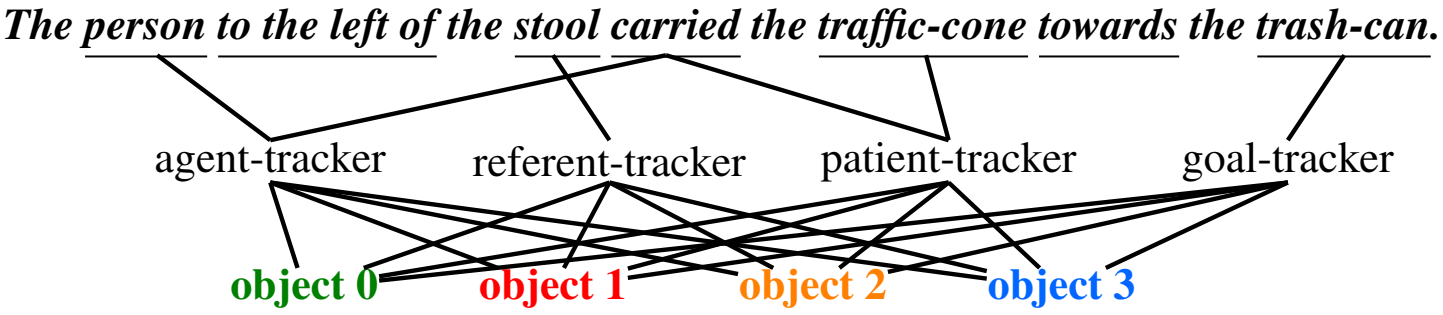


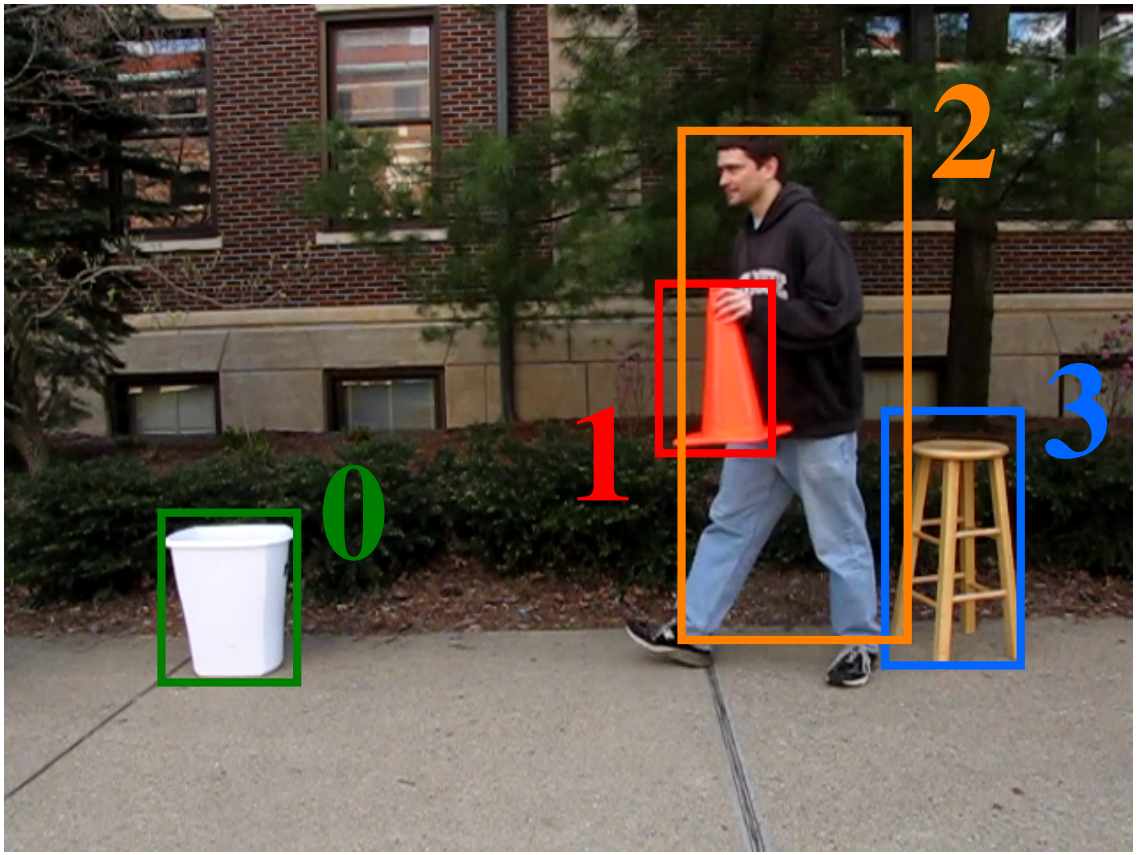
Sentence Tracker



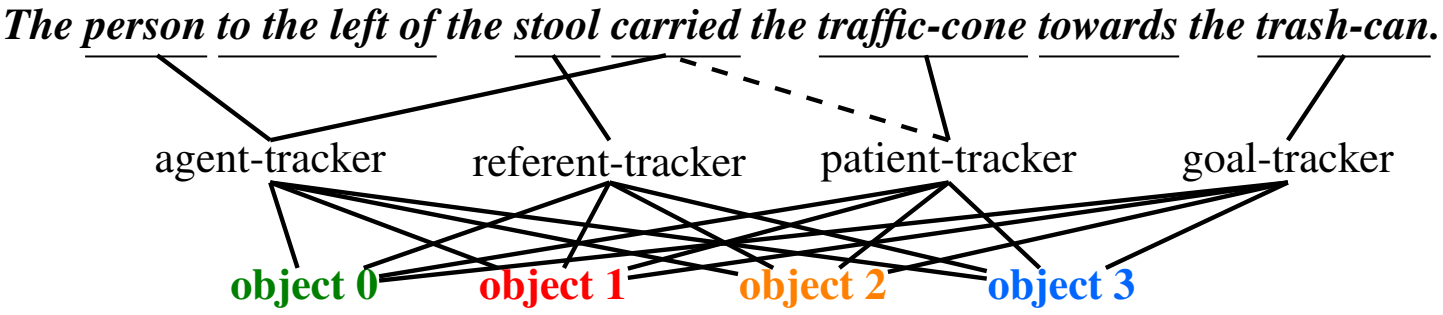


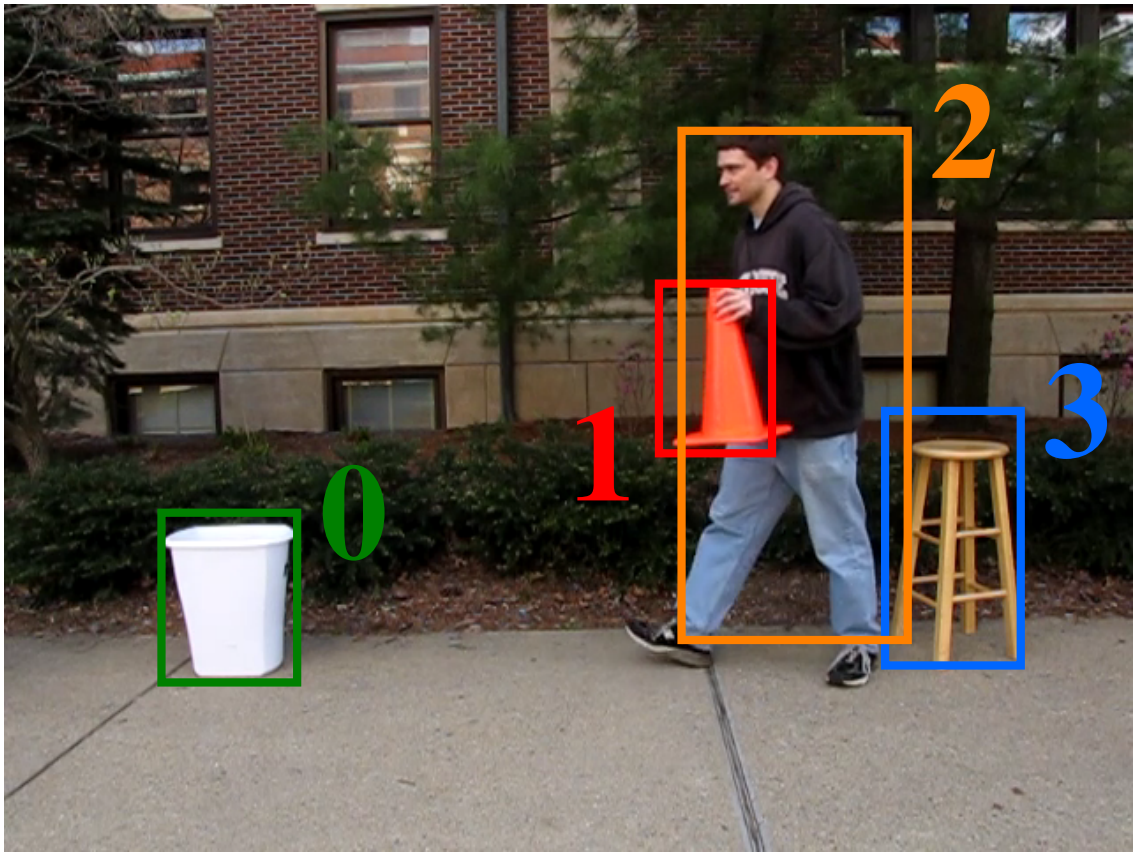
Sentence Tracker



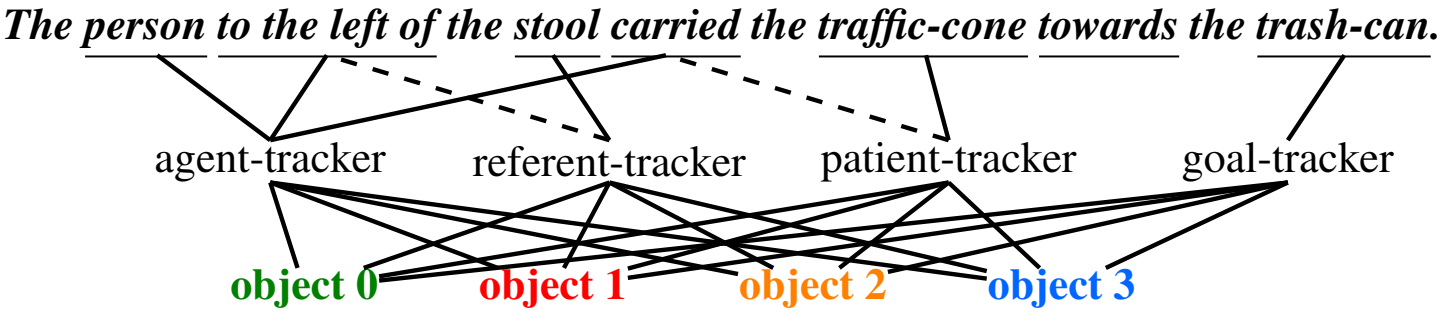


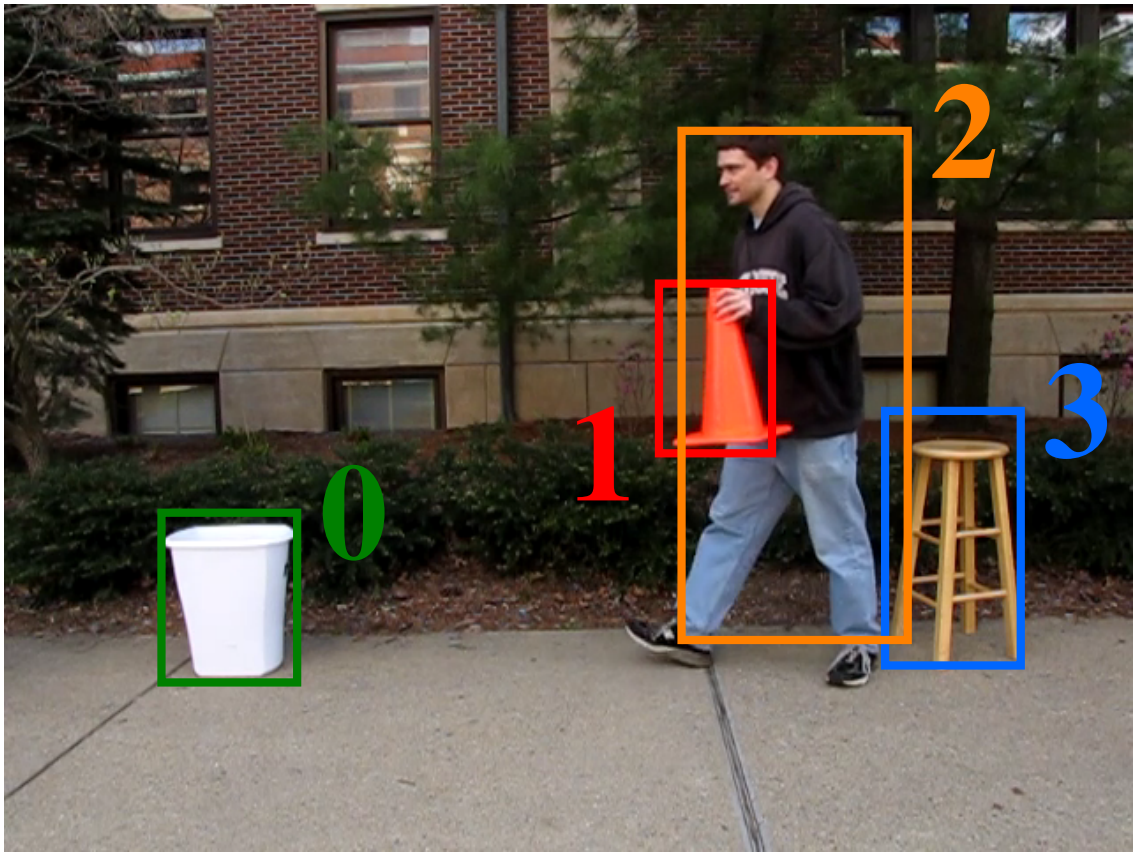
Sentence Tracker



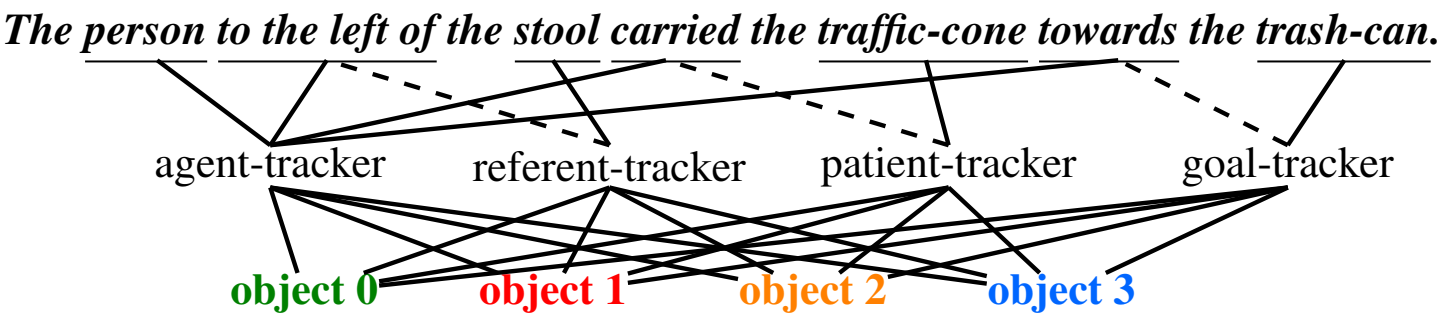


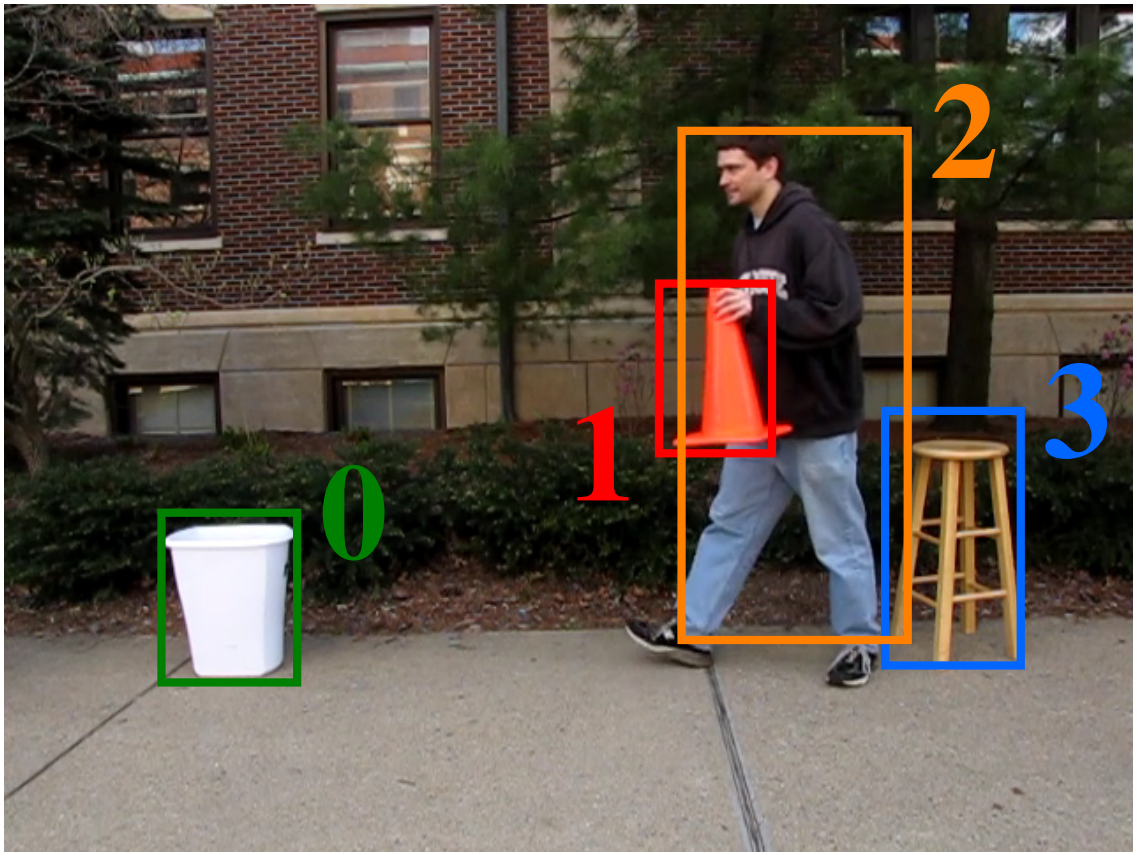
Sentence Tracker



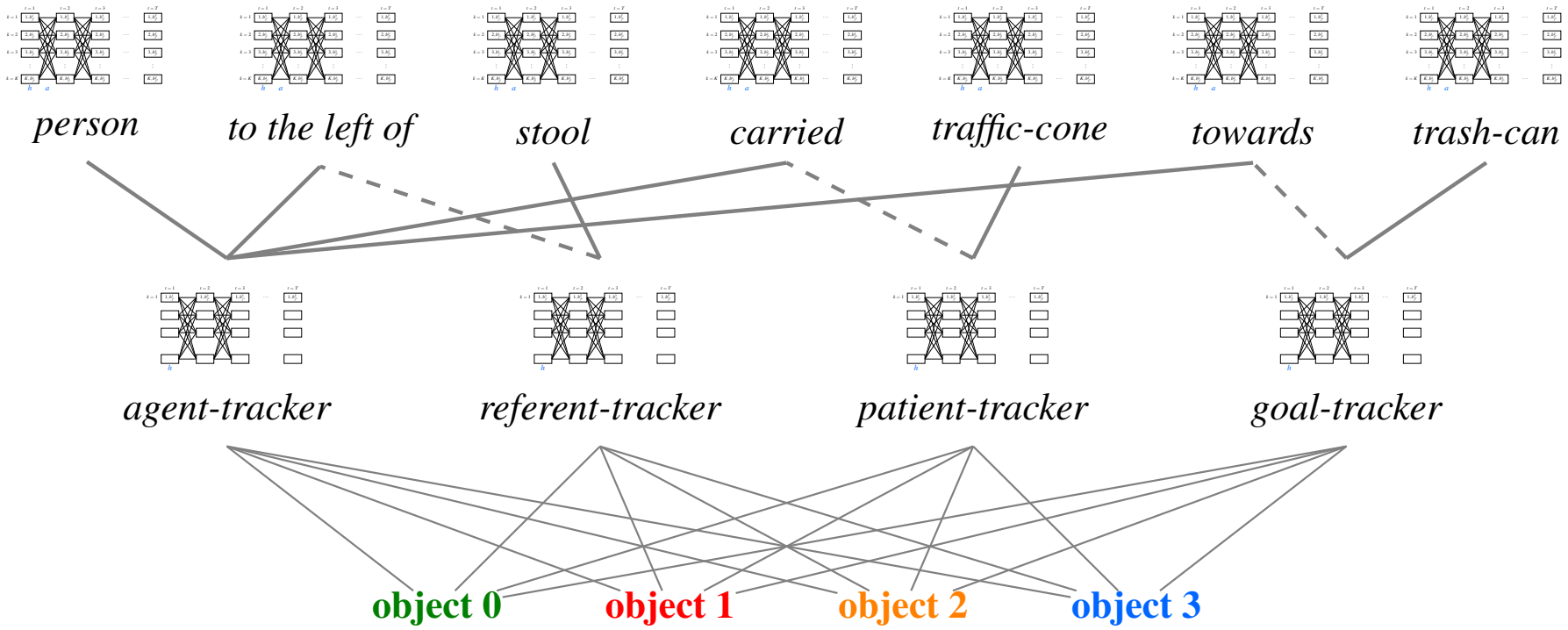


Sentence Tracker

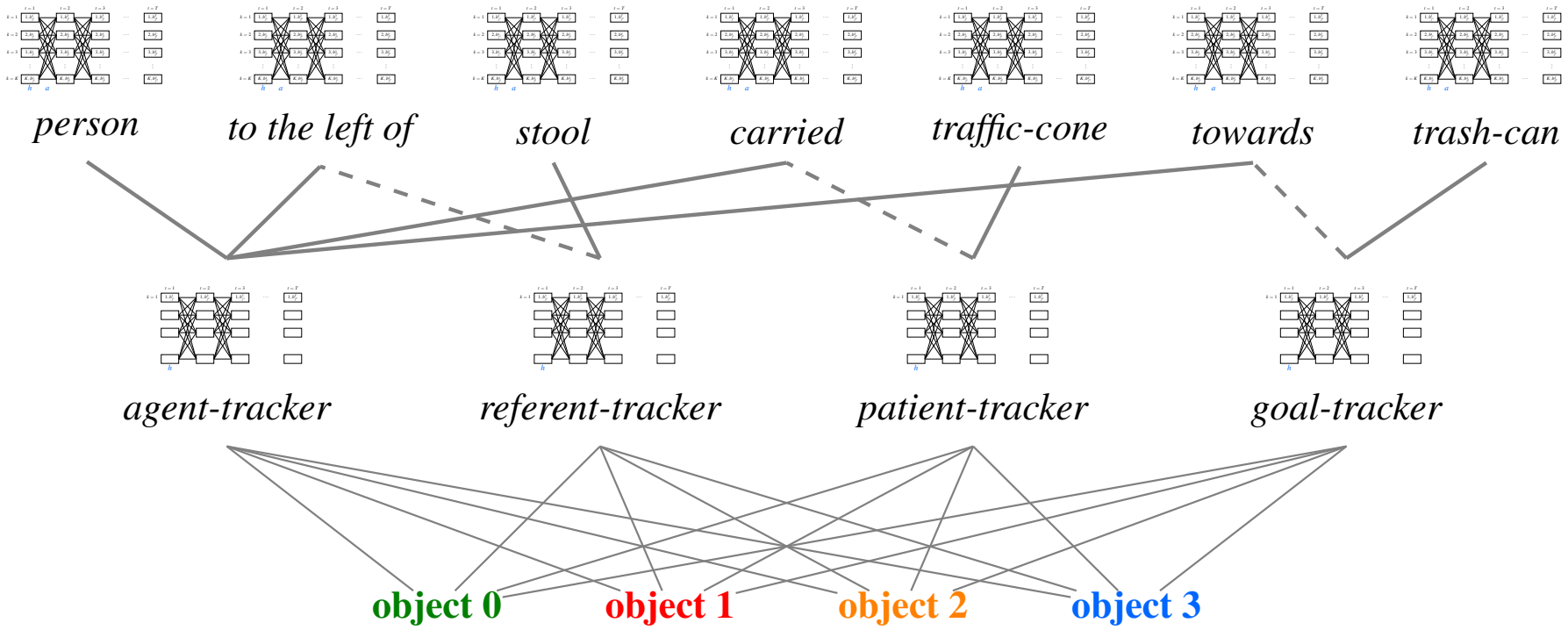




Sentence Tracker

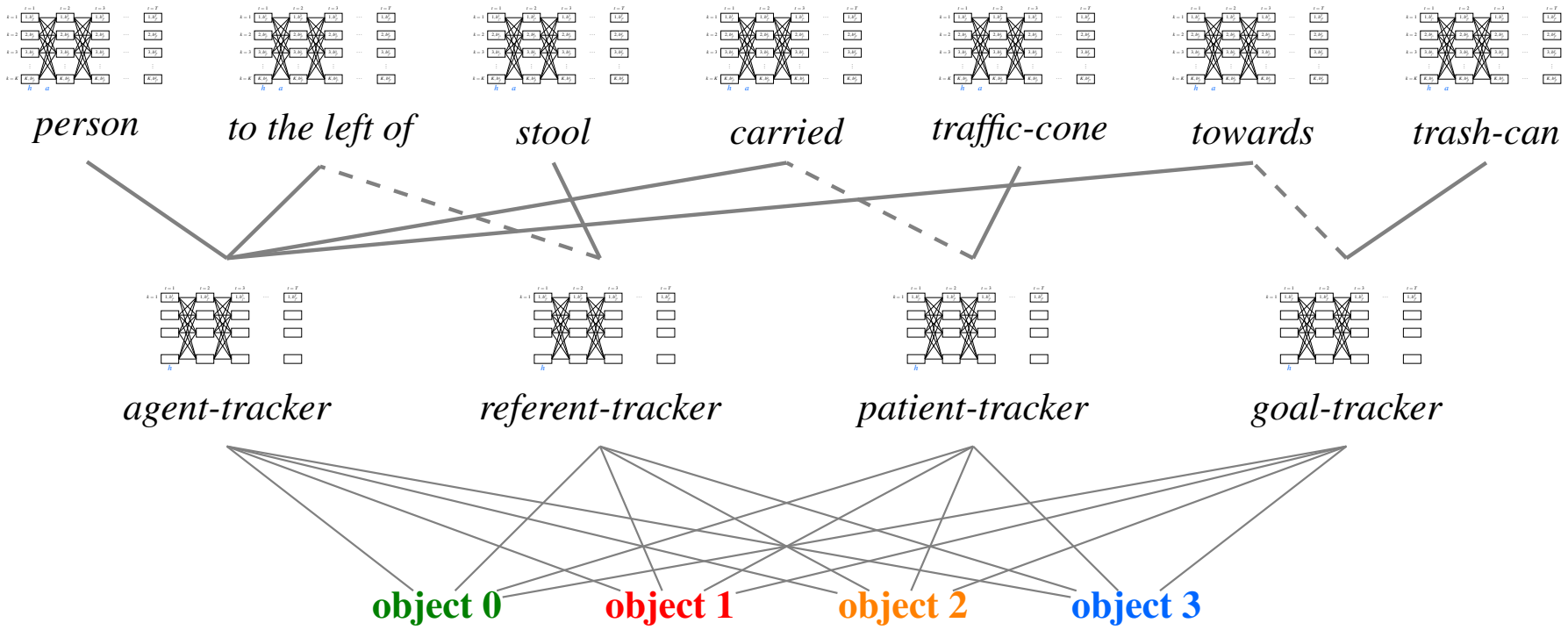


Sentence Tracker



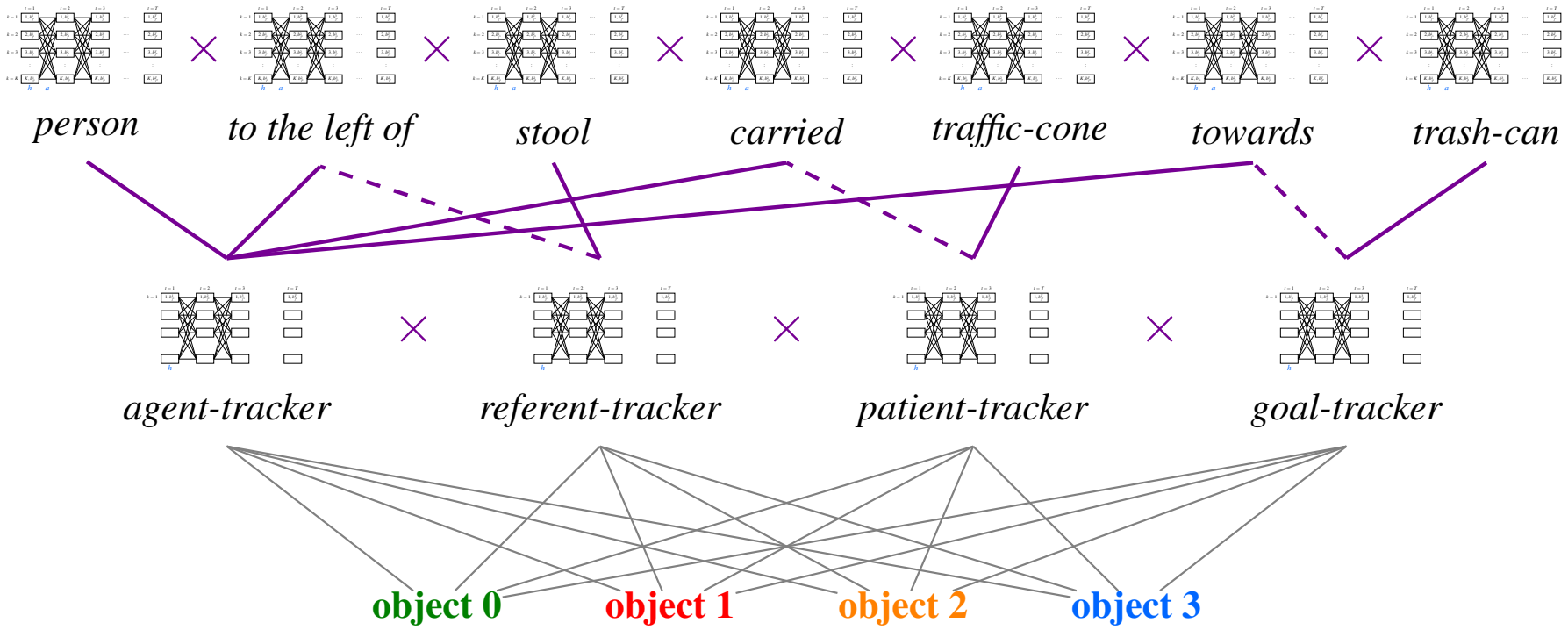
$$\max_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \left(\sum_{l=1}^{L} \sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t) \right) + \max_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \left(\sum_{w=1}^{W} \sum_{t=1}^{T} h_w(k_w^t, b_{w,1}^t, b_{w,2}^t) + \sum_{t=2}^{T} a_w(k_w^{t-1}, k_w^t) \right)$$

Sentence Tracker



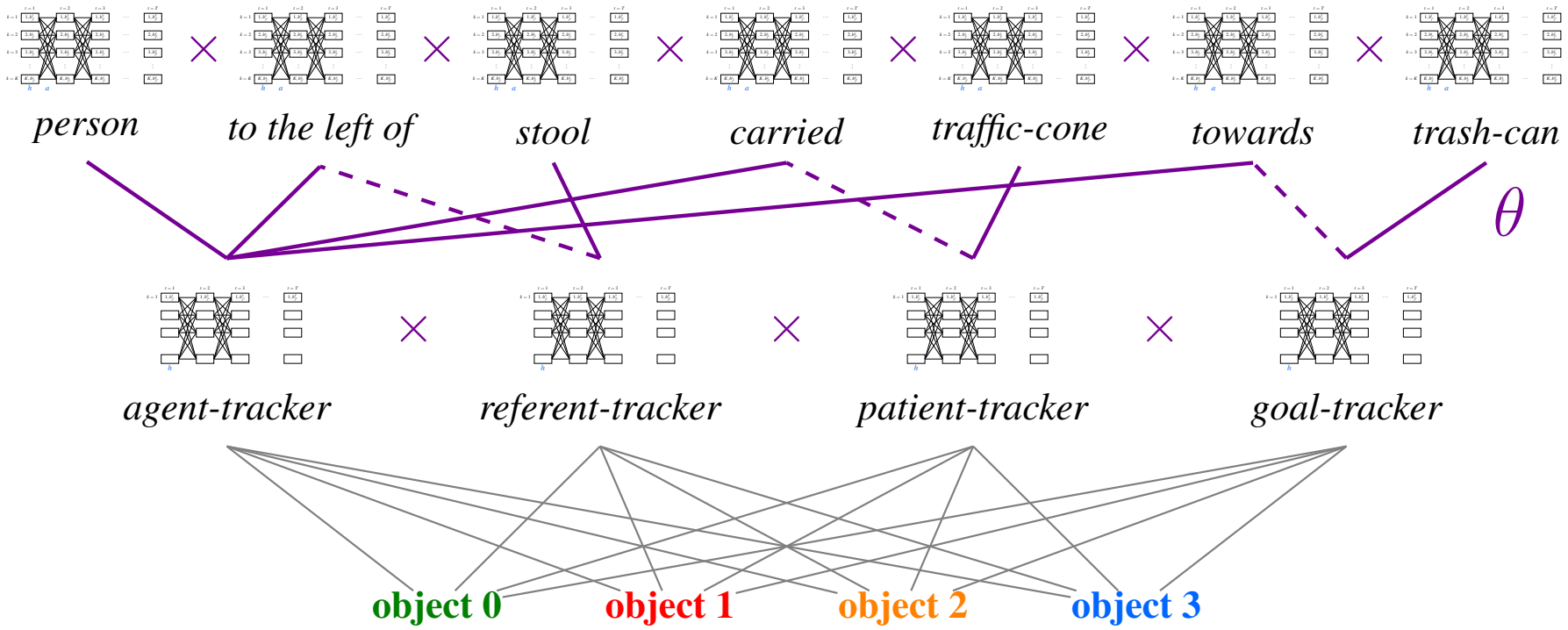
$$\max_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \max_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \sum_{l=1}^{L}\sum_{t=1}^{T} f(b^t_{j^t_l}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}_l}, b^t_{j^t_l}) + \sum_{w=1}^{W}\sum_{t=1}^{T} h_w(k^t_w, b^t_{j^t_{\theta^1_w}}, b^t_{j^t_{\theta^2_w}}) + \sum_{t=2}^{T} a_w(k^{t-1}_w, k^t_w)$$

Sentence Tracker



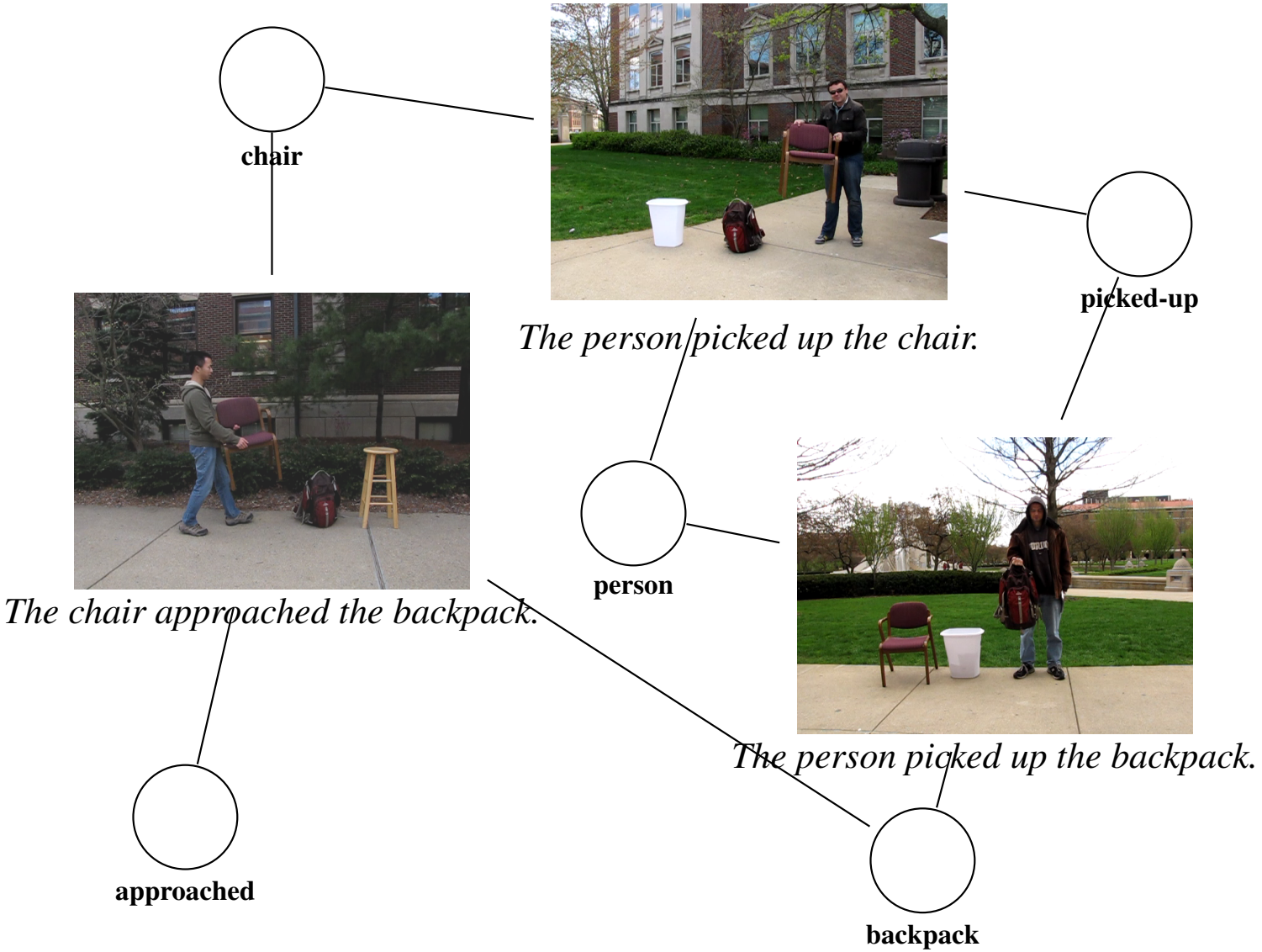
$$\max_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \max_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \sum_{l=1}^{L}\sum_{t=1}^{T} f(b^t_{j^t_l}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}_l}, b^t_{j^t_l}) + \sum_{w=1}^{W}\sum_{t=1}^{T} h_w(k^t_w, b^t_{j^t_{\theta^1_w}}, b^t_{j^t_{\theta^2_w}}) + \sum_{t=2}^{T} a_w(k^{t-1}_w, k^t_w)$$

Sentence Tracker

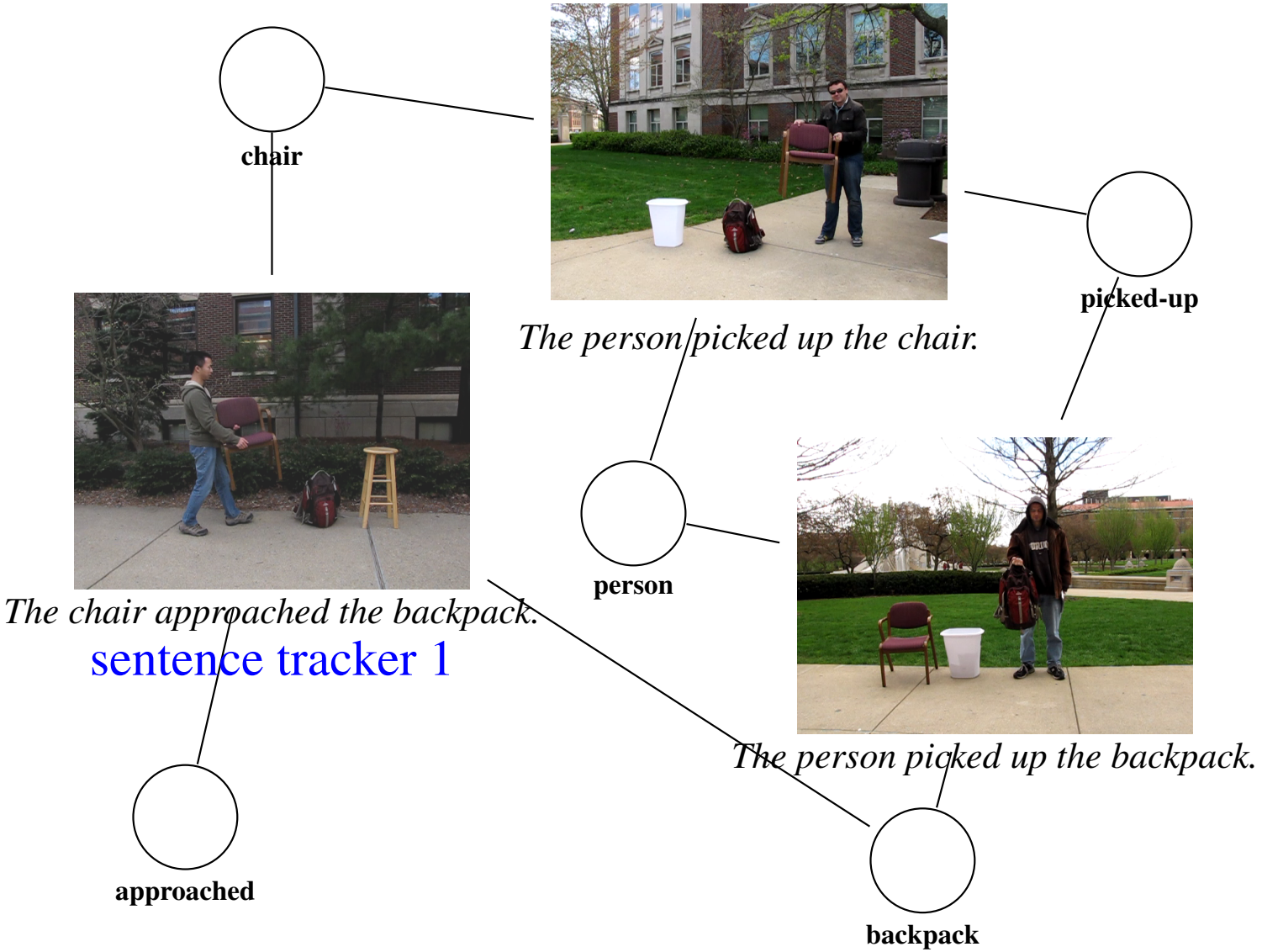


$$\max_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \max_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \sum_{l=1}^{L}\sum_{t=1}^{T} f(b^t_{j^t_l}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}_l}, b^t_{j^t_l}) + \sum_{w=1}^{W}\sum_{t=1}^{T} h_w(k^t_w, b^t_{j^t_{\theta^1_w}}, b^t_{j^t_{\theta^2_w}}) + \sum_{t=2}^{T} a_w(k^{t-1}_w, k^t_w)$$

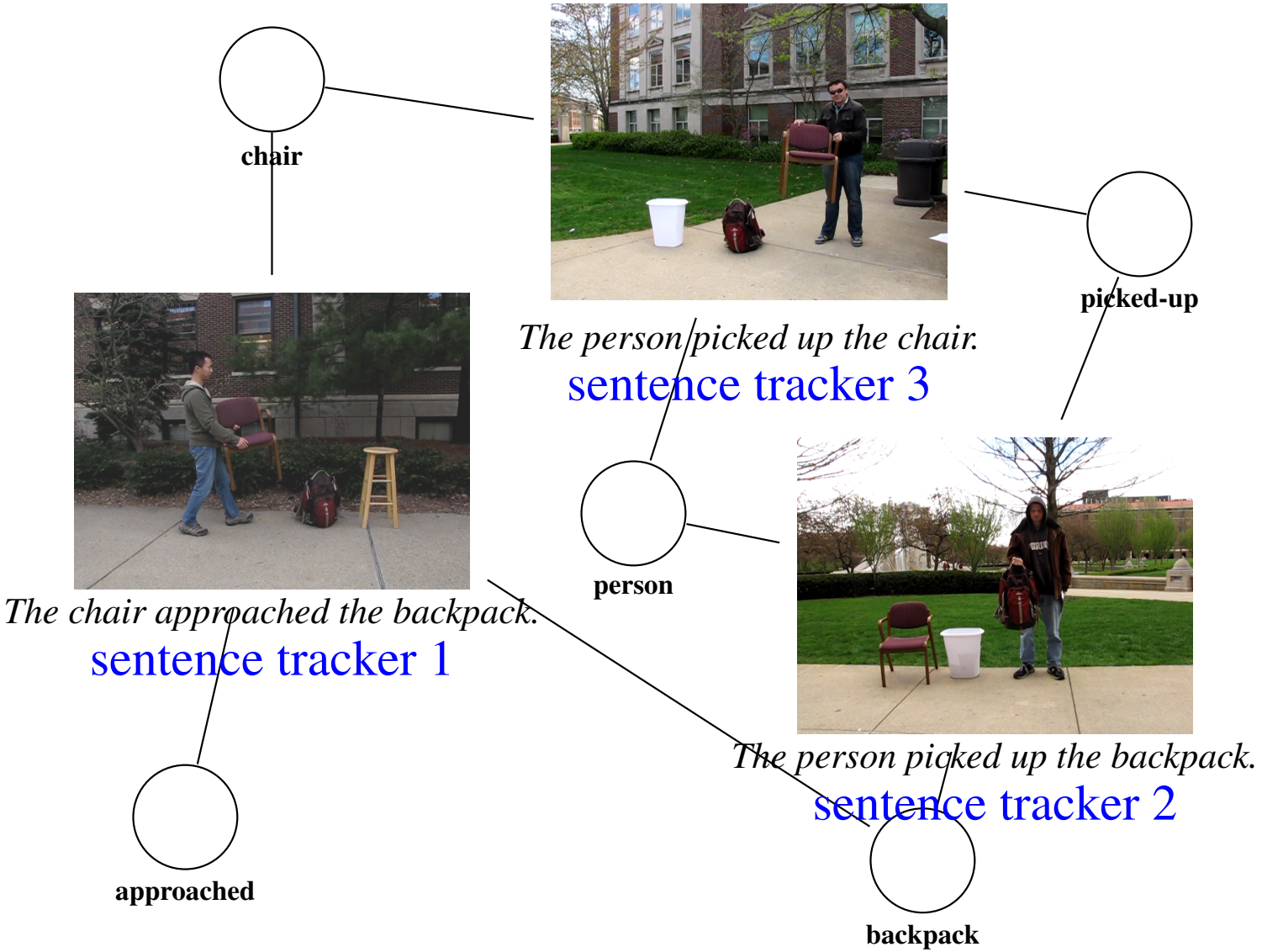
Sentence Tracker in EM



Sentence Tracker in EM



Sentence Tracker in EM



Sentence Tracker in EM

$$\max_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \max_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \begin{array}{l} \sum_{l=1}^{L} \sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t) + \\ \sum_{w=1}^{W} \sum_{t=1}^{T} h_w(k_w^t, b_{j_{\theta_w^1}^t}^t, b_{j_{\theta_w^2}^t}^t) + \sum_{t=2}^{T} a_w(k_w^{t-1}, k_w^t) \end{array}$$

Sentence Tracker in EM

$$\log \sum_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \sum_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \exp\left[\begin{array}{l} \sum_{l=1}^{L} \sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t) + \\ \sum_{w=1}^{W} \sum_{t=1}^{T} h_w(k_w^t, b_{j^t_{\theta_w^1}}^t, b_{j^t_{\theta_w^2}}^t) + \sum_{t=2}^{T} a_w(k_w^{t-1}, k_w^t) \end{array} \right]$$

Sentence Tracker in EM

$$\log \sum_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \sum_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \exp\left[\begin{array}{l} \sum_{l=1}^{L} \sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t) + \\ \sum_{w=1}^{W} \sum_{t=1}^{T} h_w(k_w^t, b_{j_{\theta_w^1}^t}^t, b_{j_{\theta_w^2}^t}^t) + \sum_{t=2}^{T} a_w(k_w^{t-1}, k_w^t) \end{array} \right]$$

- Wrap the sum of log likelihoods of all video-sentence pairs in EM.

Sentence Tracker in EM

$$\log \sum_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \sum_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \exp\left[\begin{array}{l} \sum_{l=1}^{L} \sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t) + \\ \sum_{w=1}^{W} \sum_{t=1}^{T} h_w(k_w^t, b_{j_{\theta_w^1}^t}^t, b_{j_{\theta_w^2}^t}^t) + \sum_{t=2}^{T} a_w(k_w^{t-1}, k_w^t) \end{array} \right]$$

- Wrap the sum of log likelihoods of all video-sentence pairs in EM.
- In the E-step, compute probability for tracks, HMM states, and outputs.

Sentence Tracker in EM

$$\log \sum_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \sum_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \exp\left[\begin{array}{l} \sum_{l=1}^{L} \sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t) + \\ \sum_{w=1}^{W} \sum_{t=1}^{T} h_w(k_w^t, b_{j_{\theta_w^1}^t}^t, b_{j_{\theta_w^2}^t}^t) + \sum_{t=2}^{T} a_w(k_w^{t-1}, k_w^t) \end{array} \right]$$

- Wrap the sum of log likelihoods of all video-sentence pairs in EM.
- In the E-step, compute probability for tracks, HMM states, and outputs.
- In the M-step, the transition matrix $a_w(k_w^{t-1}, k_w^t)$ and output distribution $h_w(k_w^t, b_{j_{\theta_w^1}^t}^t, b_{j_{\theta_w^2}^t}^t)$ are re-estimated.

Experiment Results

Experiment Results

- learn all content words in the lexicon

Experiment Results

$$
\begin{aligned}
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

Experiment Results

- learn all content words in the lexicon
- 95 video clips, each video clip contains 1 person + 2 or 3 objects

Experiment Results

- learn all content words in the lexicon
- 95 video clips, each video clip contains 1 person + 2 or 3 objects
- about 200 training video-sentence pairs + 240 test video-sentence pairs

Experiment Results

- learn all content words in the lexicon
- 95 video clips, each video clip contains 1 person + 2 or 3 objects
- about 200 training video-sentence pairs + 240 test video-sentence pairs
- test on videos/sentences *never seen* in training set

Experiment Results

- learn all content words in the lexicon
- 95 video clips, each video clip contains 1 person + 2 or 3 objects
- about 200 training video-sentence pairs + 240 test video-sentence pairs
- test on videos/sentences *never seen* in training set

The person to the left of the stool picked up the chair.

Experiment Results

- learn all content words in the lexicon
- 95 video clips, each video clip contains 1 person + 2 or 3 objects
- about 200 training video-sentence pairs + 240 test video-sentence pairs
- test on videos/sentences *never seen* in training set

The person to the left of the stool picked up the chair.

The person carried the backpack towards the stool.

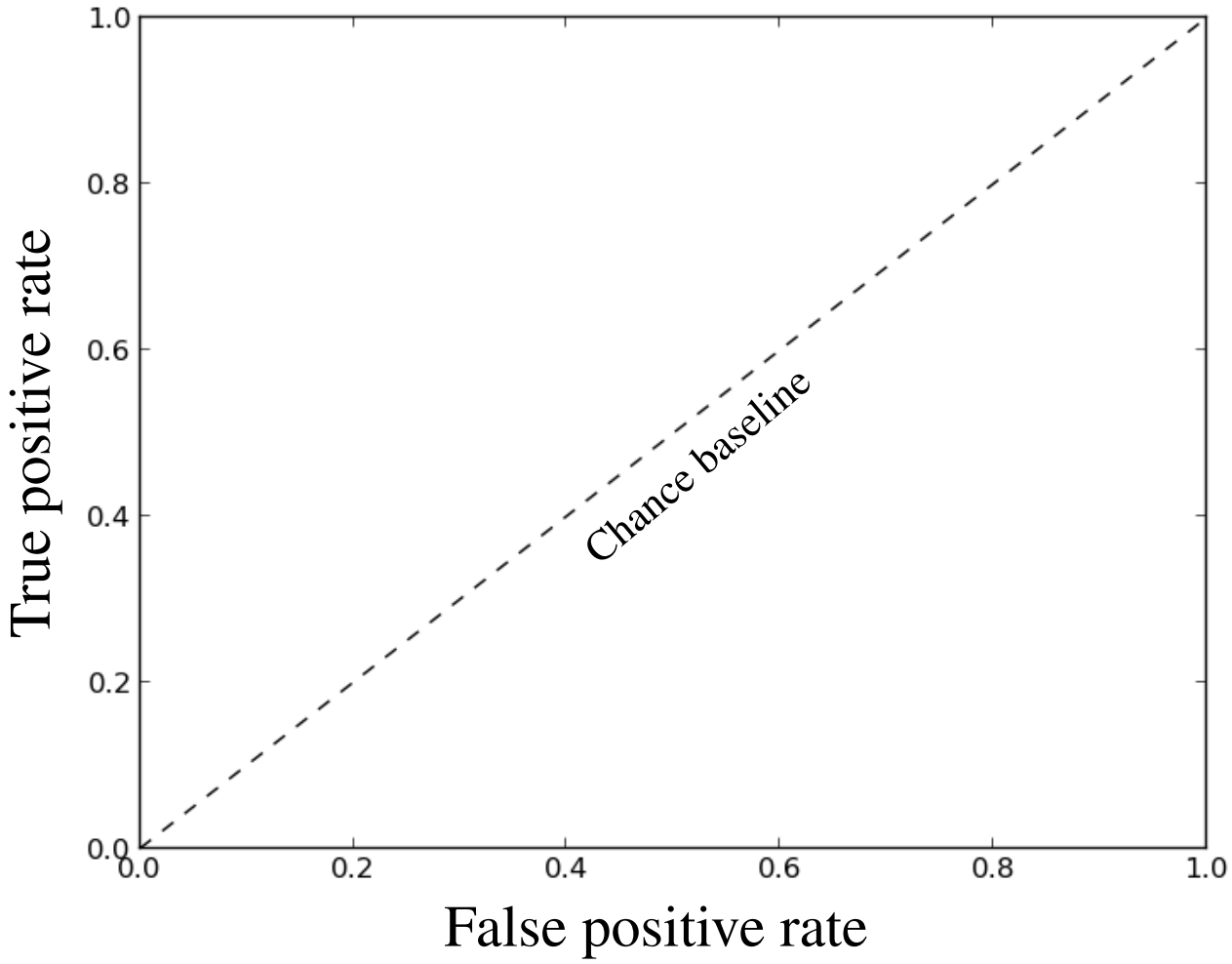
Experiment Results

- learn all content words in the lexicon
- 95 video clips, each video clip contains 1 person + 2 or 3 objects
- about 200 training video-sentence pairs + 240 test video-sentence pairs
- test on videos/sentences *never seen* in training set

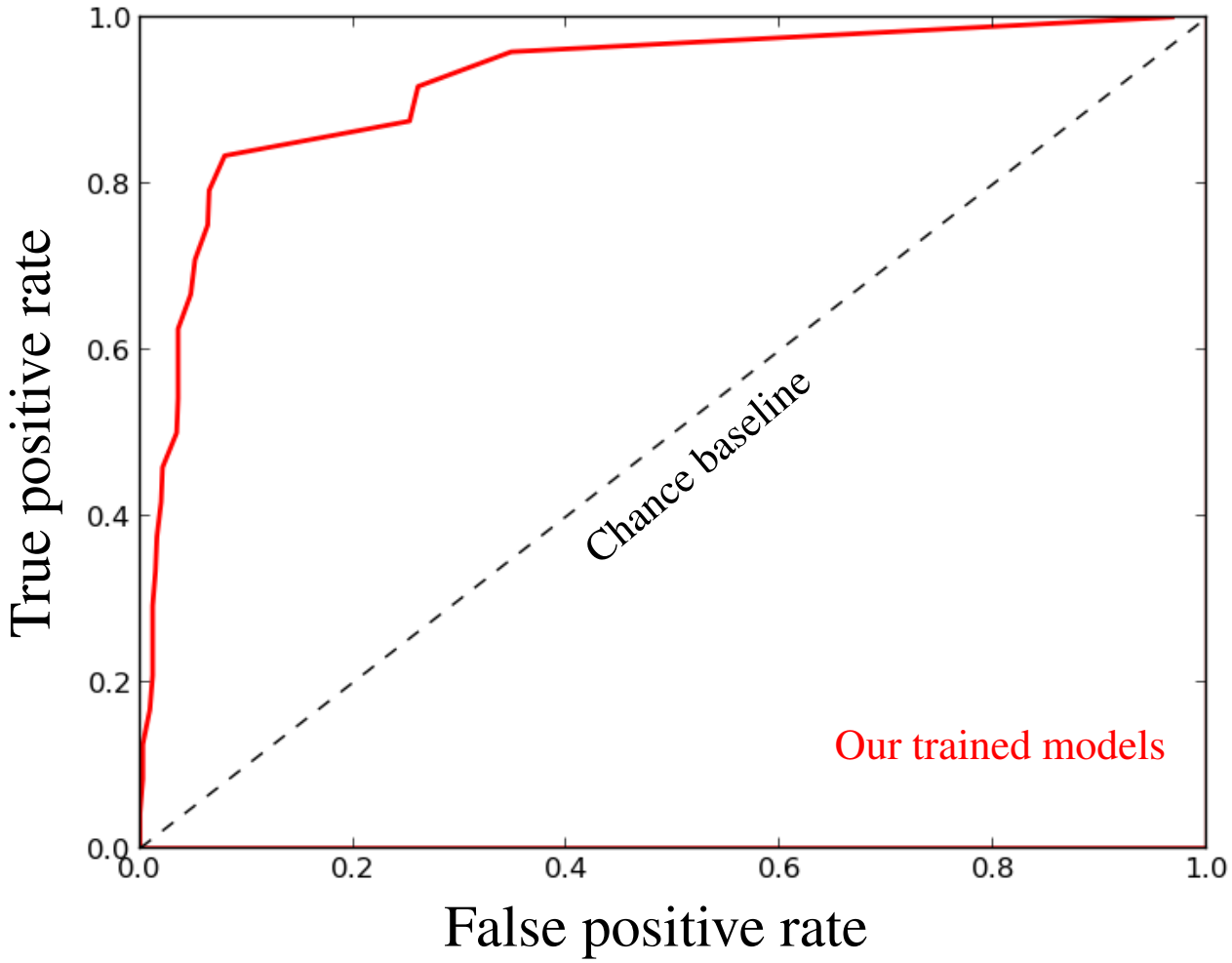
The person to the left of the stool picked up the chair.

The person carried the backpack towards the stool.

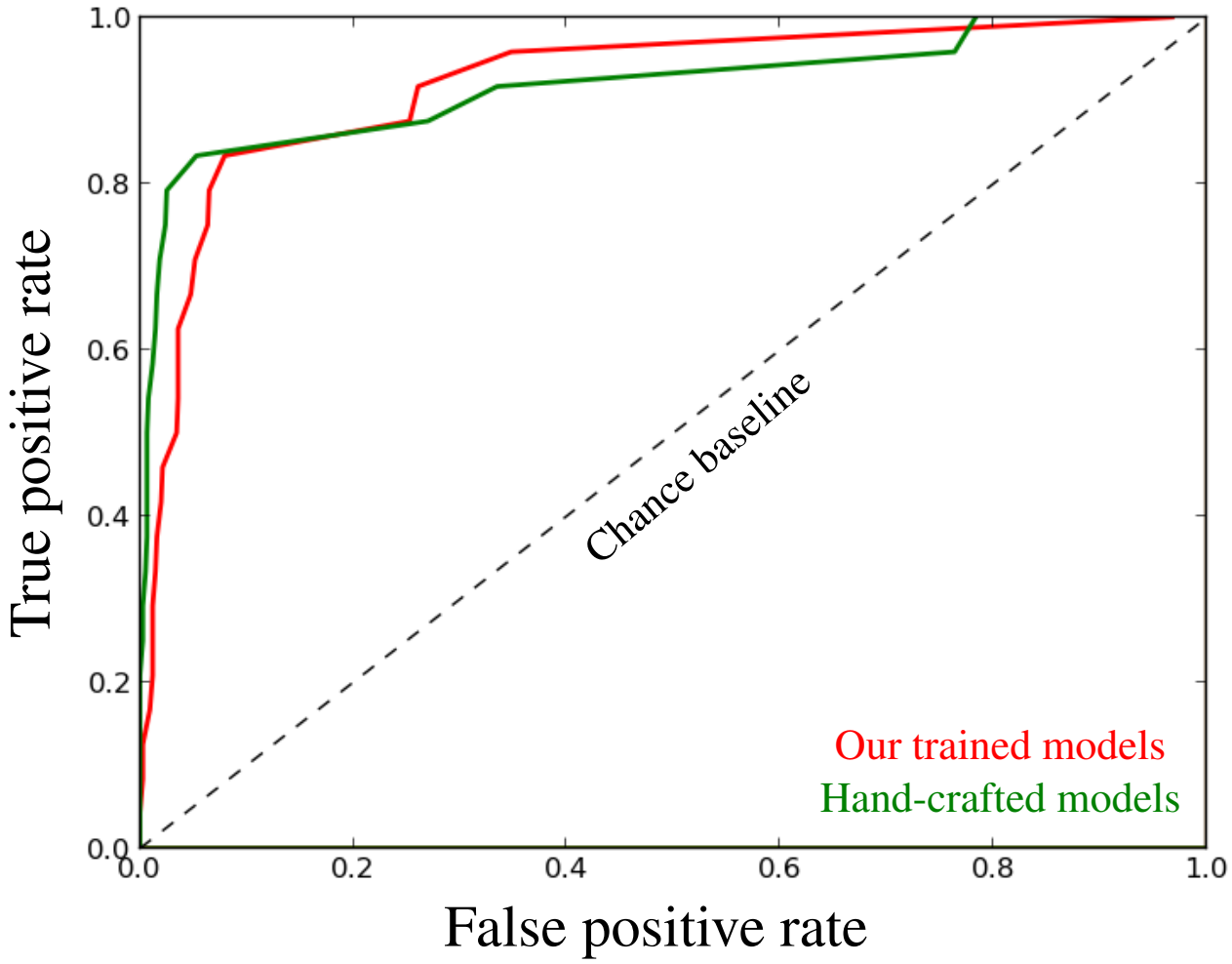
ROC curves


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
True positive rate
False positive rate
Chance baseline

ROC curves


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
True positive rate
False positive rate
Chance baseline
Our trained models

ROC curves


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
True positive rate
False positive rate
Chance baseline
Our trained models
Hand-crafted models

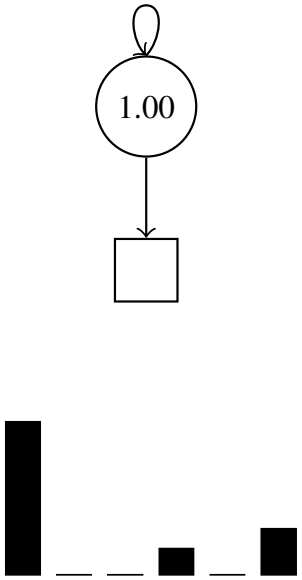
Examples of Trained Models

nouns

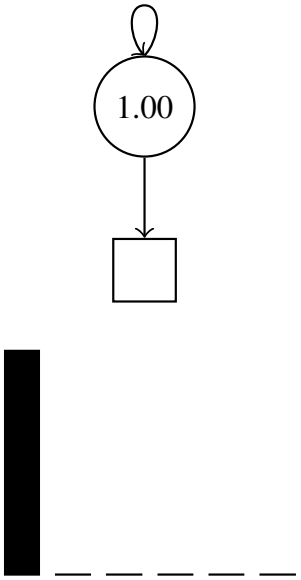
Examples of Trained Models

nouns

person



trained



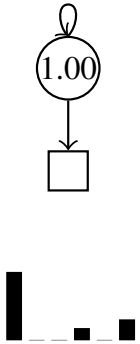
hand-crafted

OBJECT CLASS

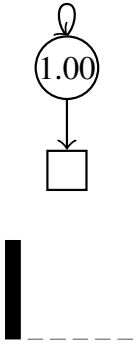
Examples of Trained Models

nouns

person



trained

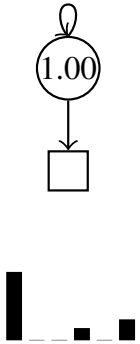


hand-crafted

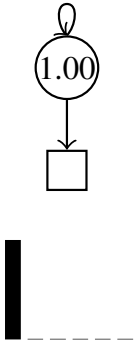
Examples of Trained Models

nouns

person

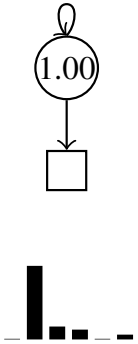


trained

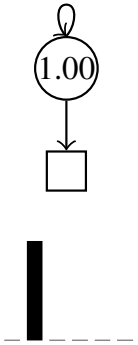


hand-crafted

backpack

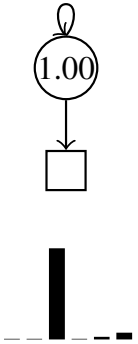


trained

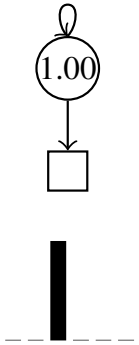


hand-crafted

chair

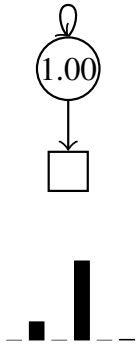


trained

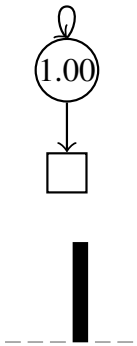


hand-crafted

trash-can

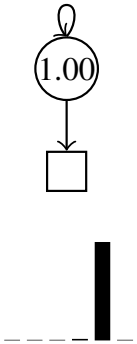


trained

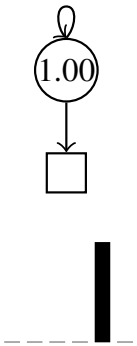


hand-crafted

stool

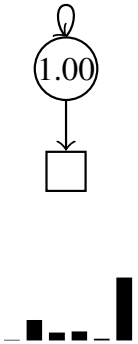


trained

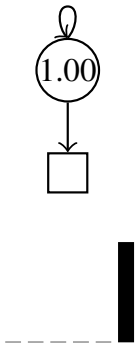


hand-crafted

traffic-cone



trained

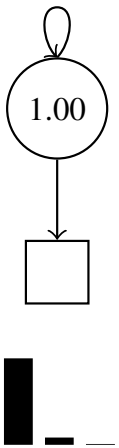


hand-crafted

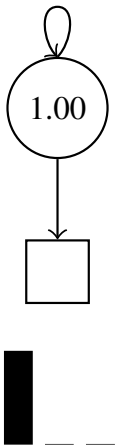
Examples of Trained Models

spatial-relation prepositions

to-the-left-of



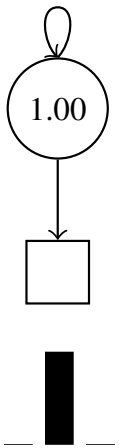
trained



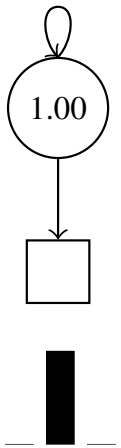
hand-crafted

RELATIVE X-POSITION

to-the-right-of



trained



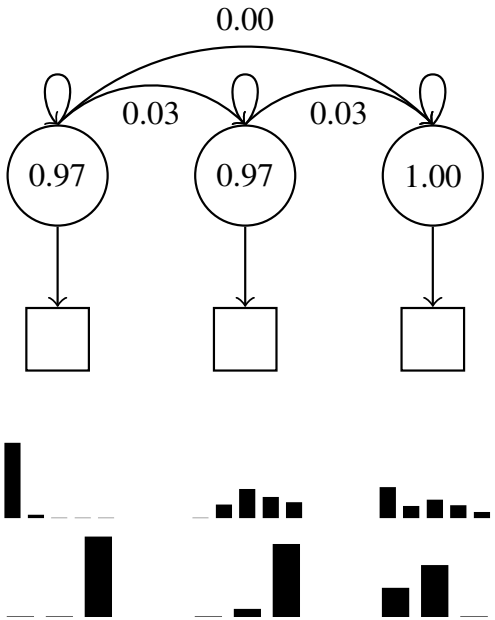
hand-crafted

Examples of Trained Models

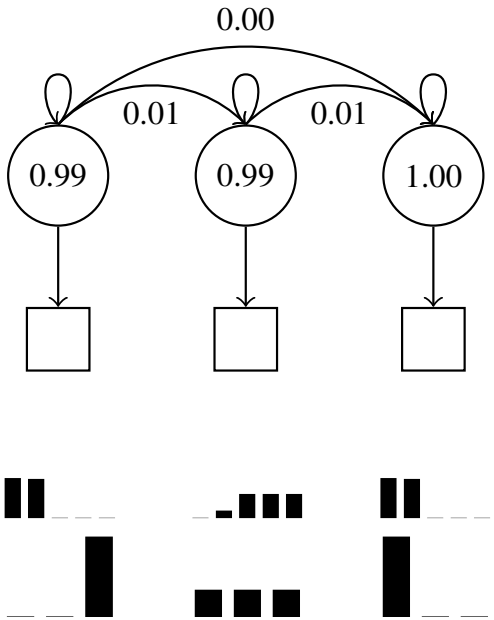
motion prepositions

towards

trained



hand-crafted



VELOCITY

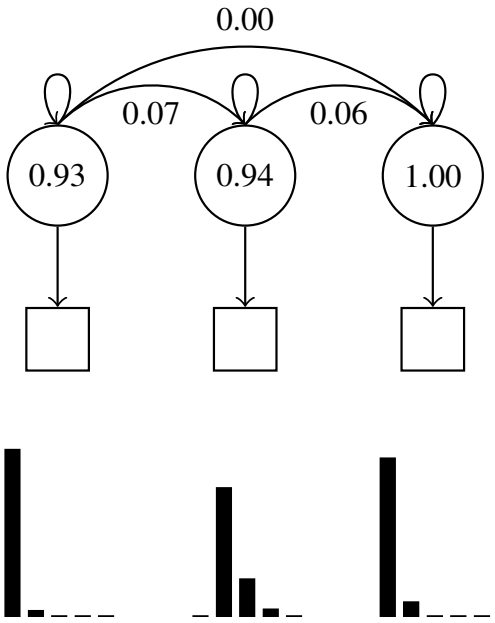
DISTANCE

Examples of Trained Models

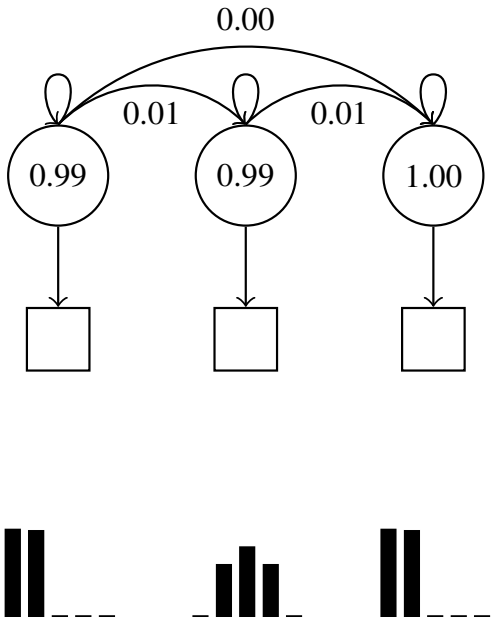
adverbs

slowly

trained



hand-crafted



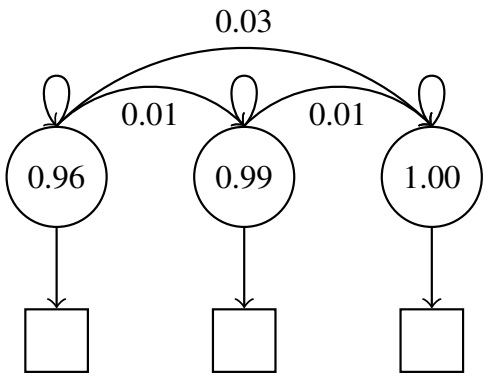
VELOCITY

Examples of Trained Models

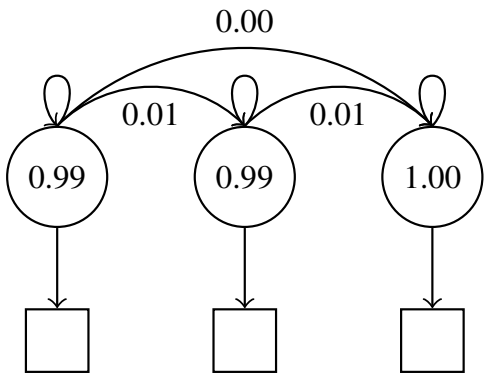
verbs

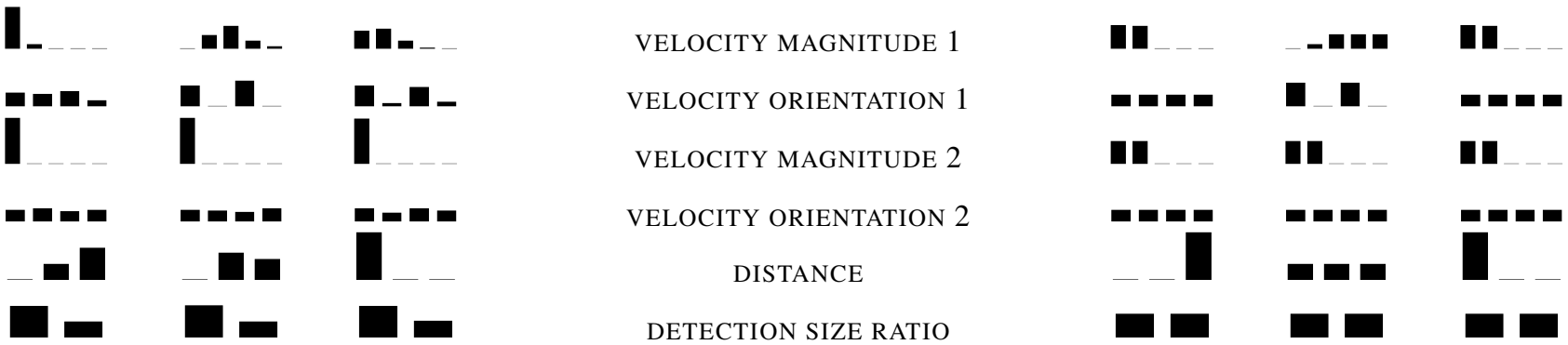
approached

trained

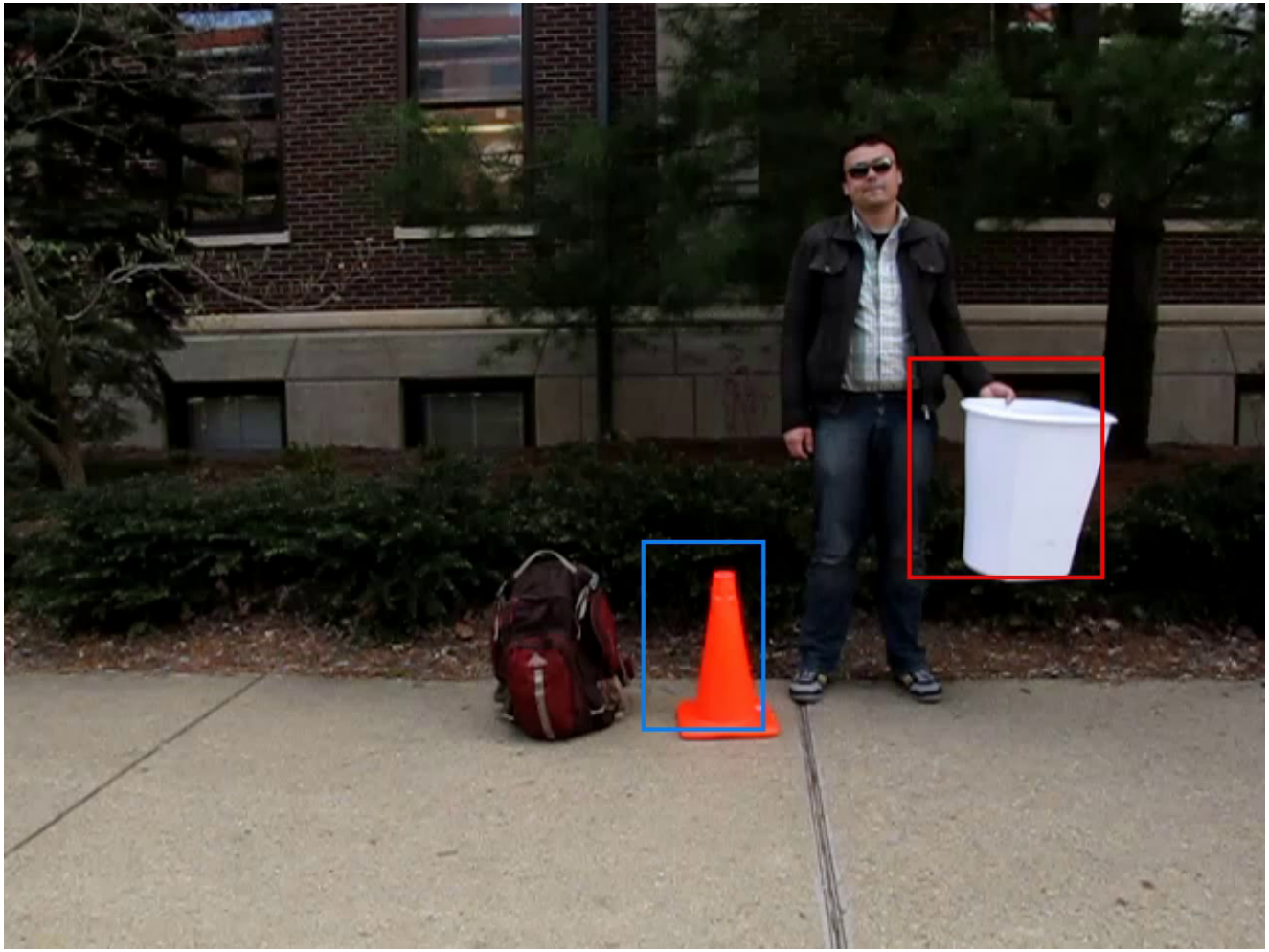


hand-crafted





New Sentence Generation with Trained Models

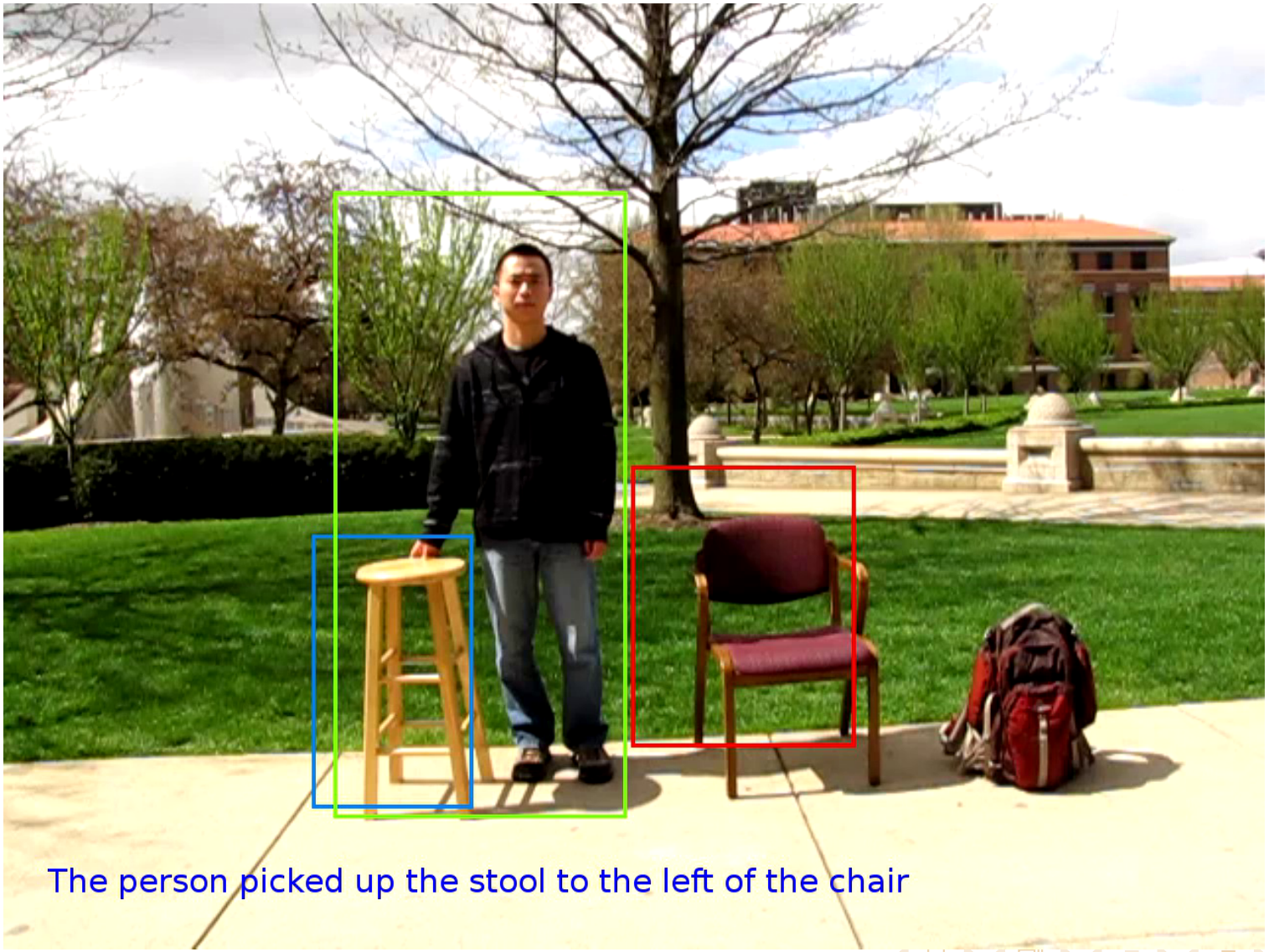
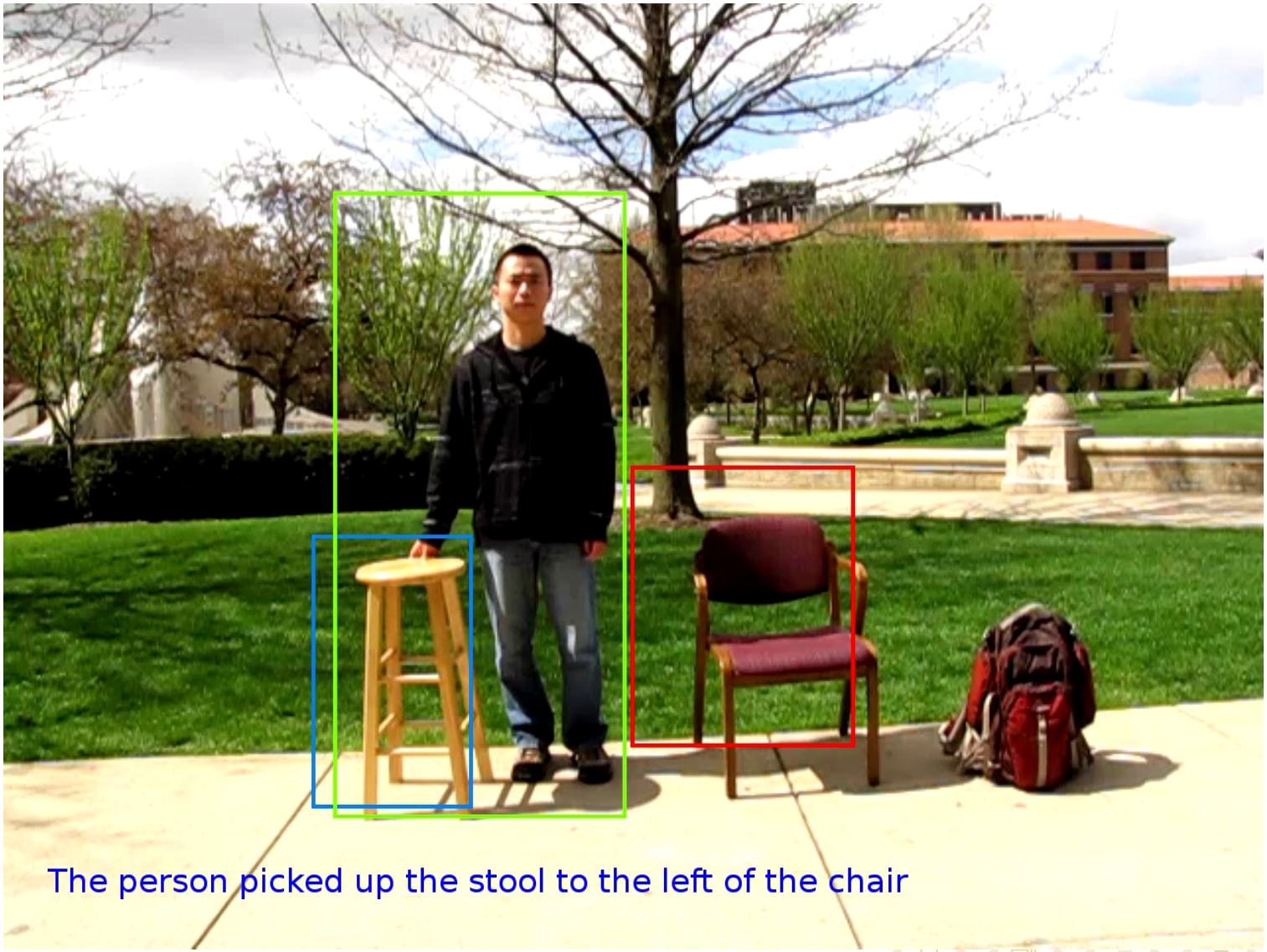


The trash-can approached the traffic-cone.

New Sentence Generation with Trained Models

The trash-can approached the traffic-cone.

New Sentence Generation with Trained Models



New Sentence Generation with Trained Models