Learning to Segment Actions from Observation and Narration

Daniel Fried, Jean-Baptiste Alayrac, Phil Blunsom,
Chris Dyer, Stephen Clark, and Aida Nematzadeh

Action Segmentation in Video

Task: `make pancakes: {add egg, add flour, …, pour batter, remove pancake}`

Actions: `background` | `pour batter` | `background` | `remove pancake`

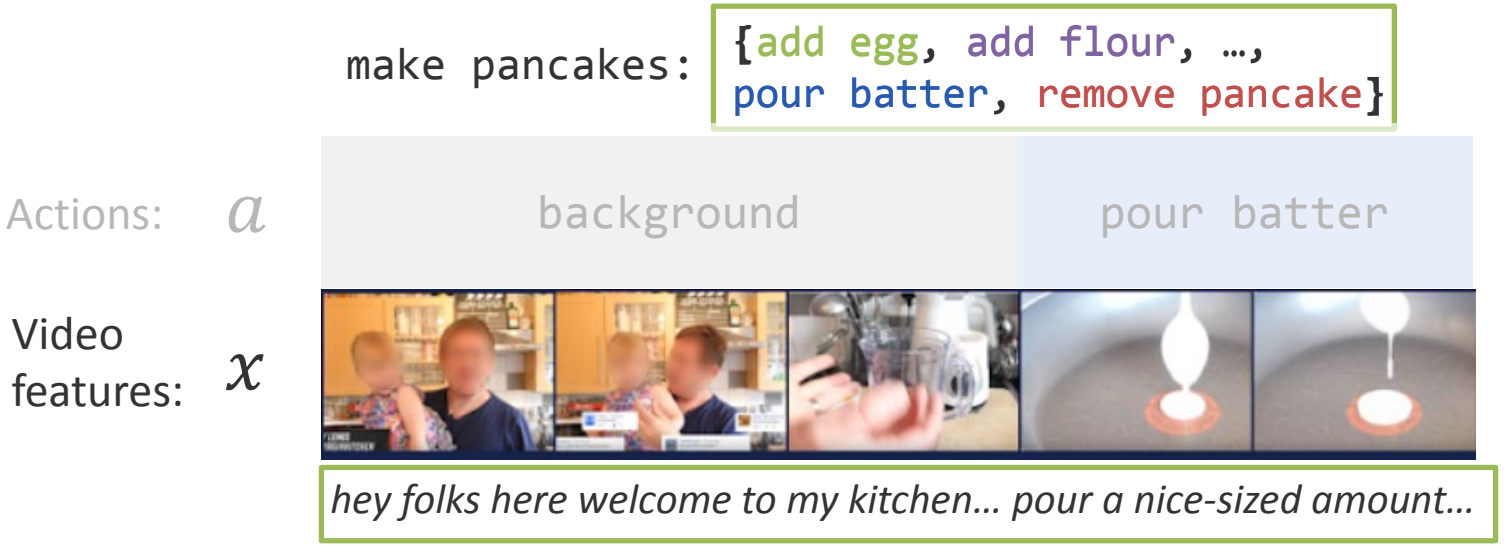
Video:

Narration: *hey folks here welcome to my kitchen … pour a nice-sized amount … change the angle to show … and take it out*

Challenges: visual diversity, noisy narration, varied task structure

How little supervision can we get away with?

Training Without Segment Labels



Generative: $\max_{\theta} \sum_{a} p_\theta(a, x)$ [Richard et al. 2018, Sener and Yao 2018]

Discriminative: $\max_{\theta, a} p_\theta(a|x)$ [Alayrac et al. 2016, Zhukov et al. 2019]

Weak-supervision for a:

- Likely ordering of the actions
- Time-aligned narration

How little supervision can we get away with?

Define a model that allows flexible training.

Semi-Markov Model

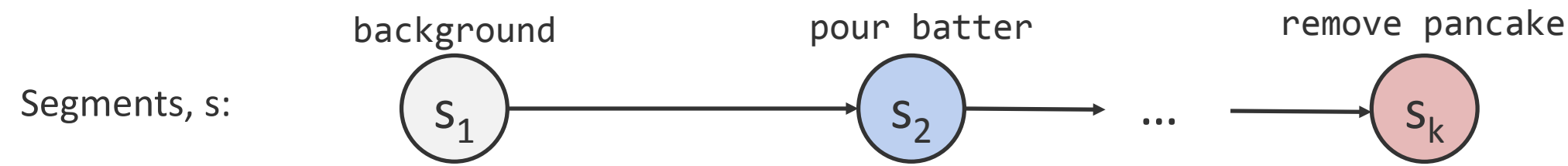
Actions

Video

Semi-Markov Model

tabular

$$\prod_k p(s_k|s_{k-1})$$

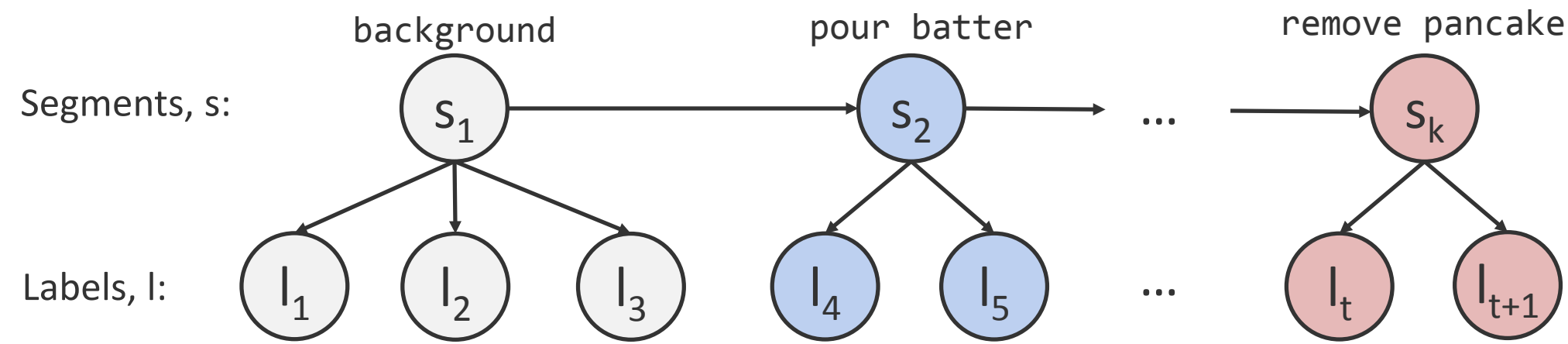


Actions

Video

Semi-Markov Model

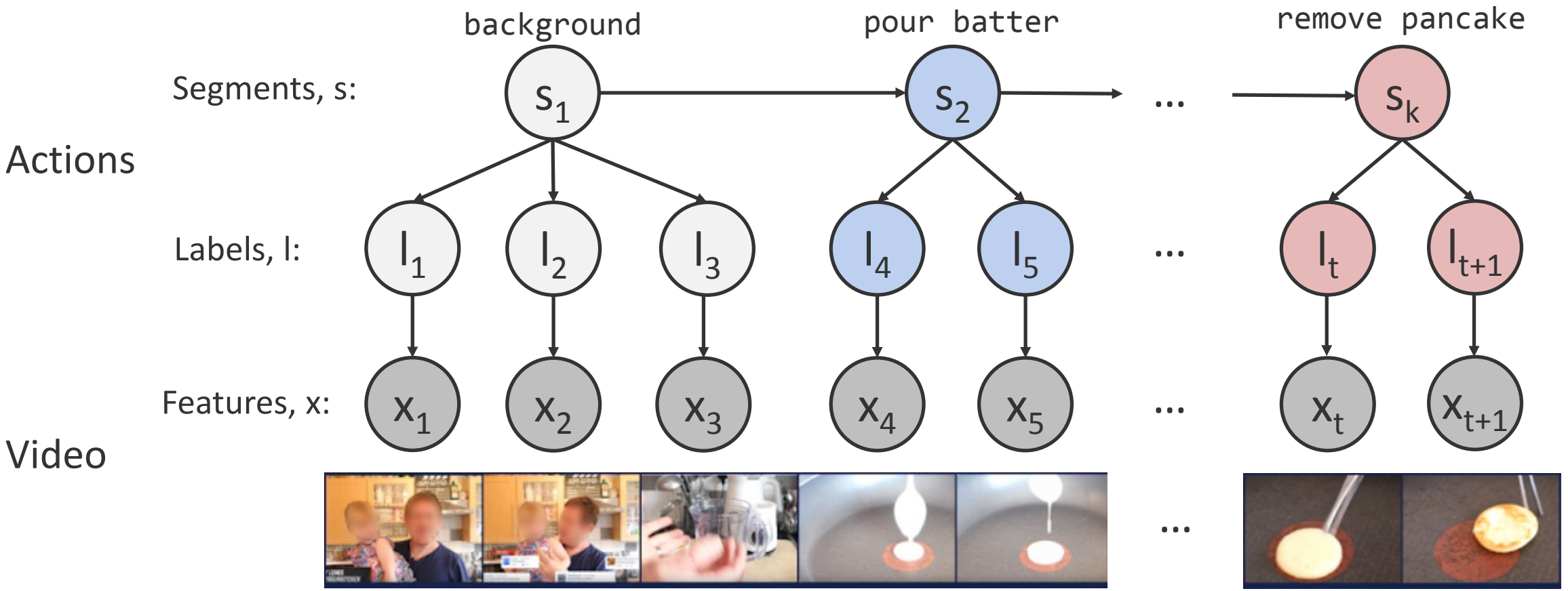
$$\prod_k \underbrace{p(s_k|s_{k-1})}_{\text{tabular}} \underbrace{p(\text{len}(s_k)|s_k)}_{\text{Poisson}}$$



Video

Semi-Markov Model

$$p(s,l,x) = \prod_k \overbrace{p(s_k|s_{k-1})}^{\text{tabular}} \overbrace{p(\text{len}(s_k)|s_k)}^{\text{Poisson}} \prod_t \overbrace{p(x_t|l_t)}^{\text{Gaussian}}$$



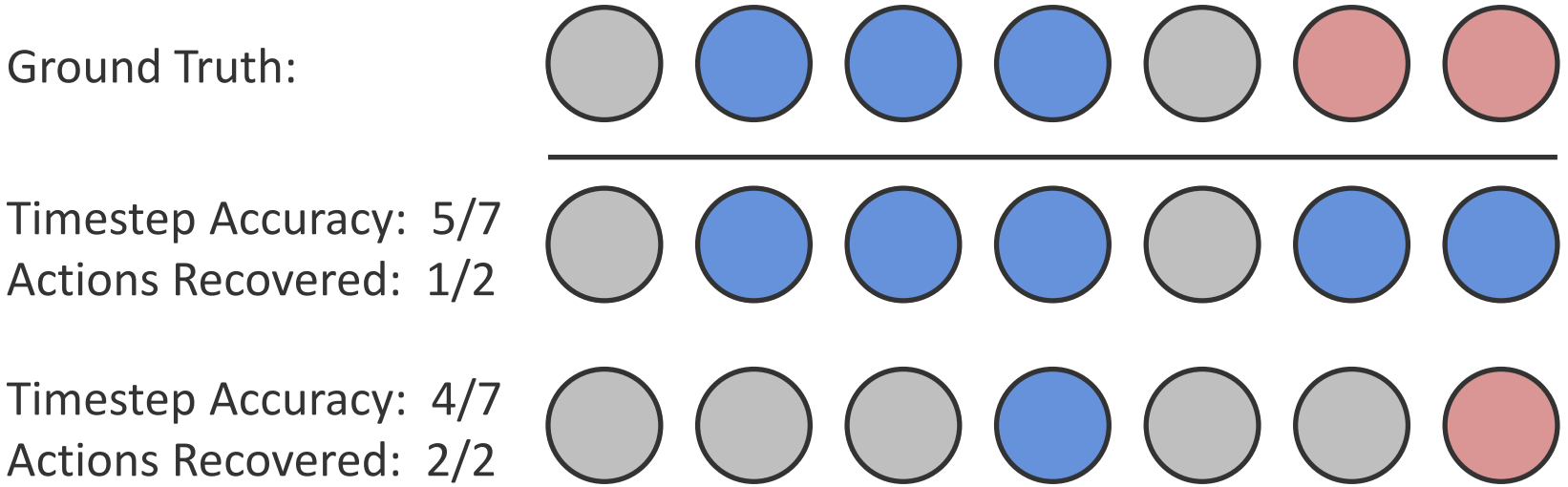
CrossTask Dataset [Zhukov et al. 2019]

- 2,700 instructional YouTube videos, with transcribed narration
- 18 household tasks, *e.g.* cooking, changing a tire, assembling furniture
- Features from ConvNets trained on other related tasks
  - Action recognition [Carreira and Zisserman 2017; Kay et al. 2017]
  - Object classification [He et al. 2016; Russakovsky et al. 2015]
  - Audio classification [Simonyan and Zisserman 2015; Abu-El-Haija et al. 2016]

Evaluation

Two main metrics from past work:

- Timestep accuracy (1-second intervals) [Sener and Yao 2018, Richard et al. 2018, inter alia]
- Action recovery (with one timestep per action) [Alayrac et al. 2016, Zhukov et al. 2019]



How little supervision can we get away with?

First, compare models in a supervised setting.

Supervised Training

Generative:

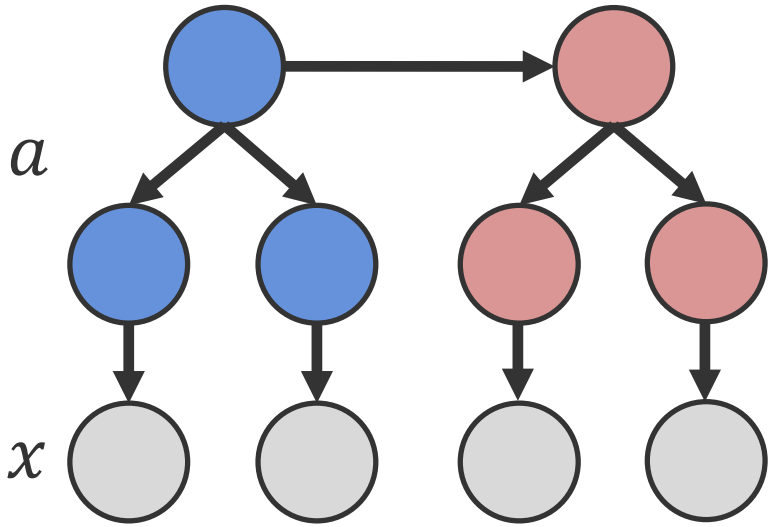
$p(a, x)$

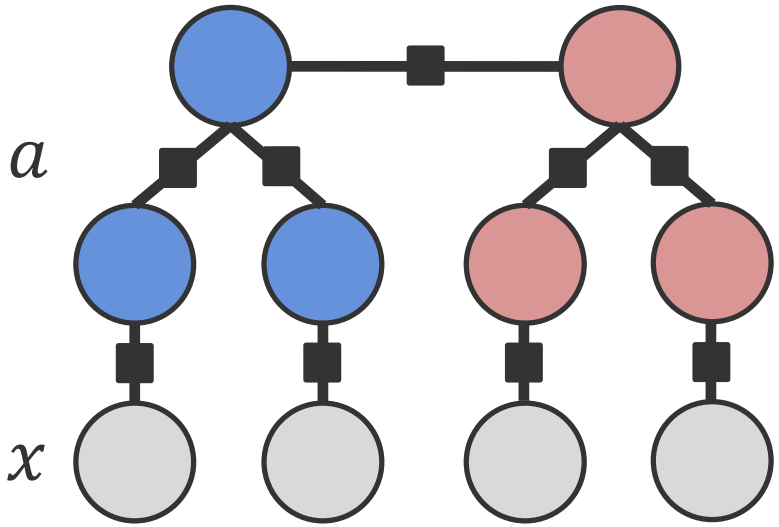
Discriminative:

$p(a|x)$

Structured:

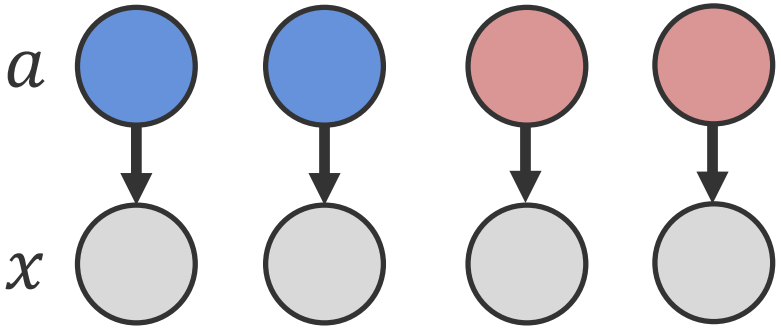
Semi-Markov model

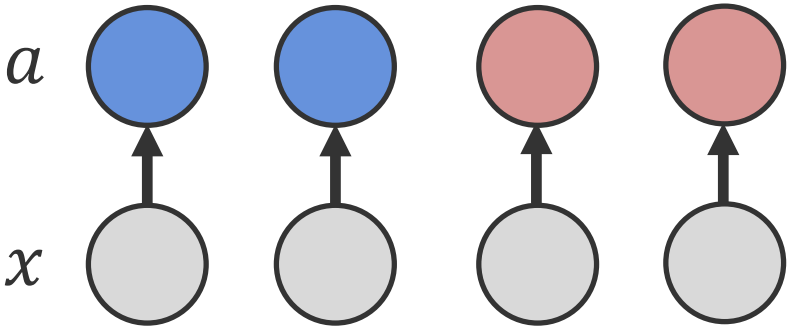




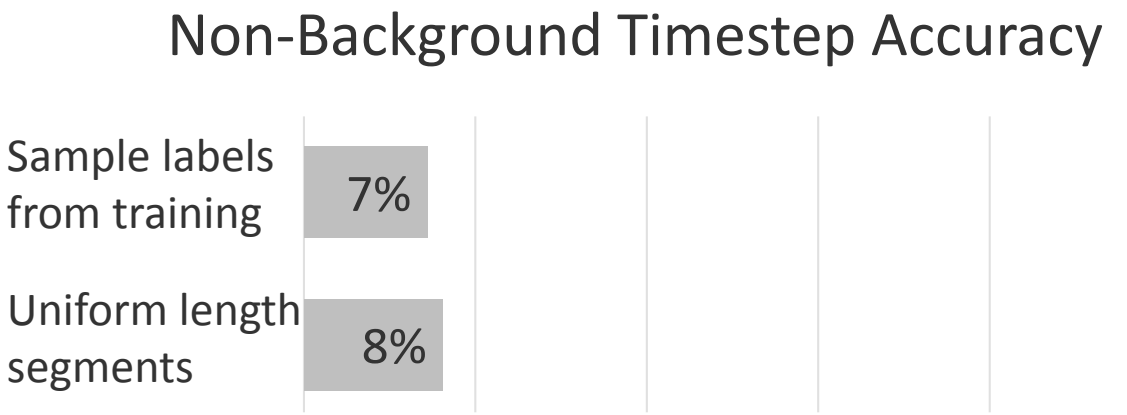
Unstructured:

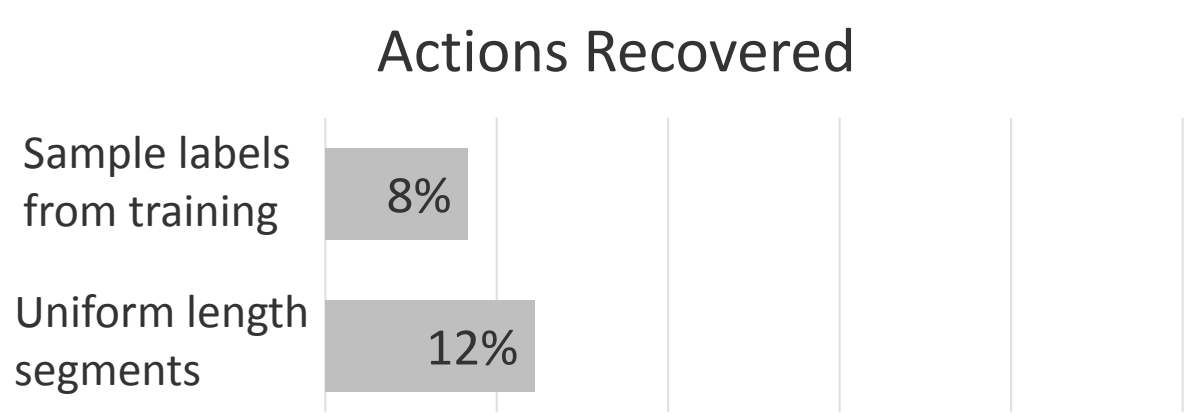
Independent classifier at each time-step



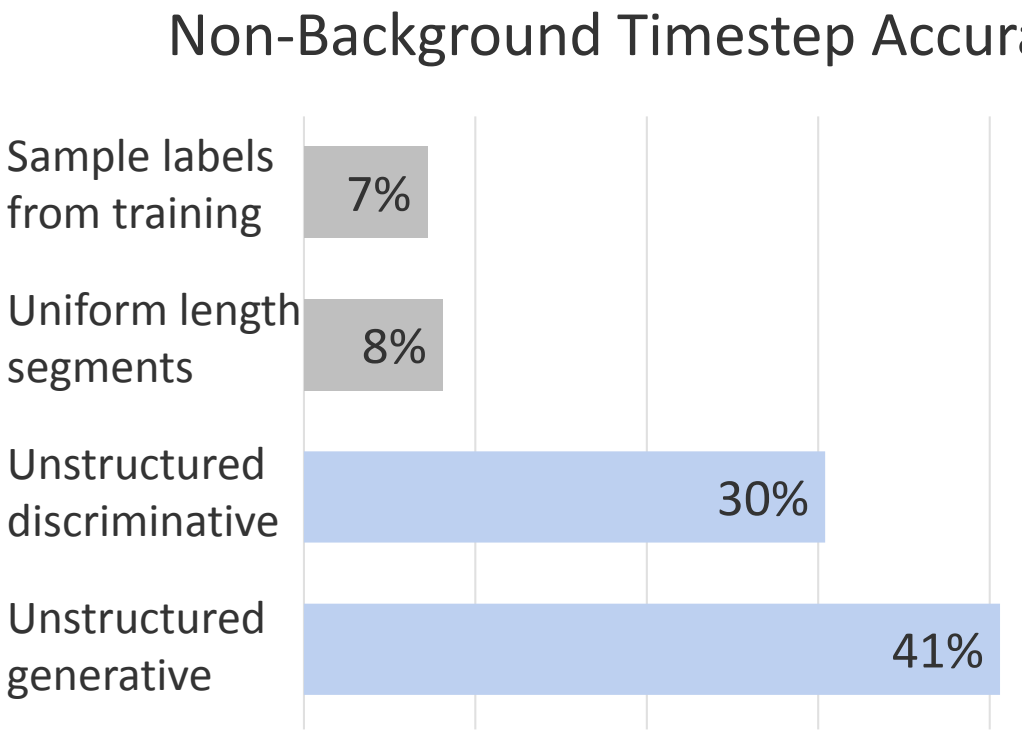


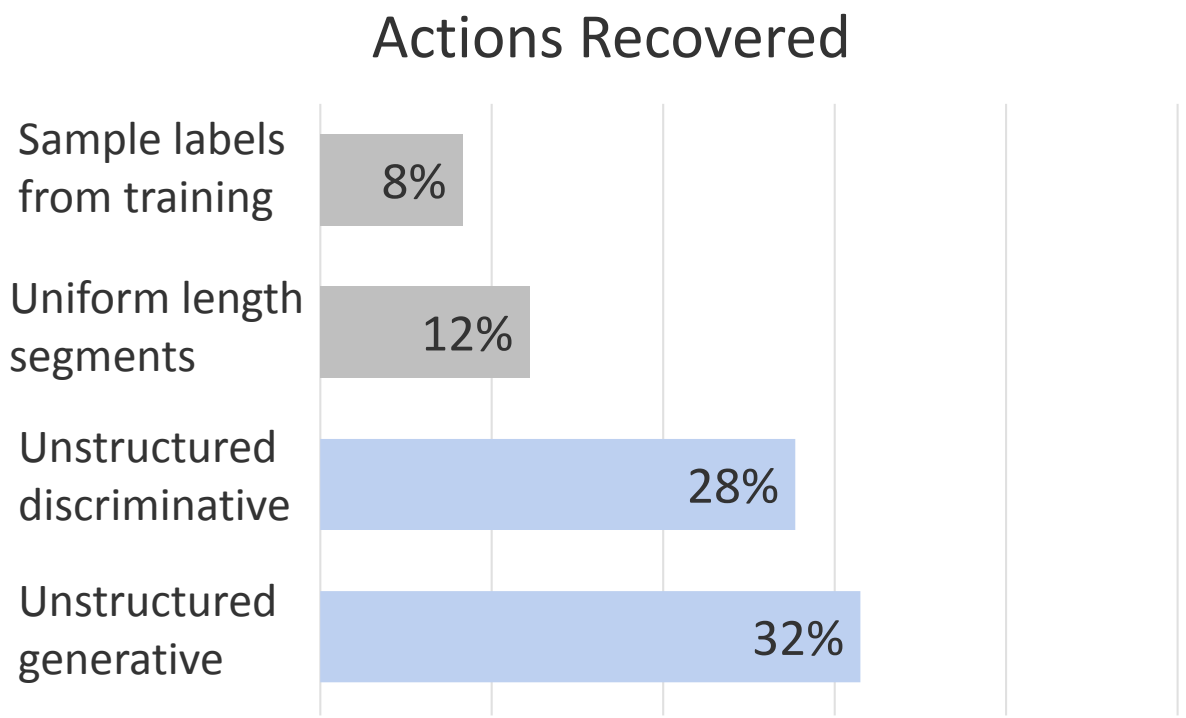
Supervised Results



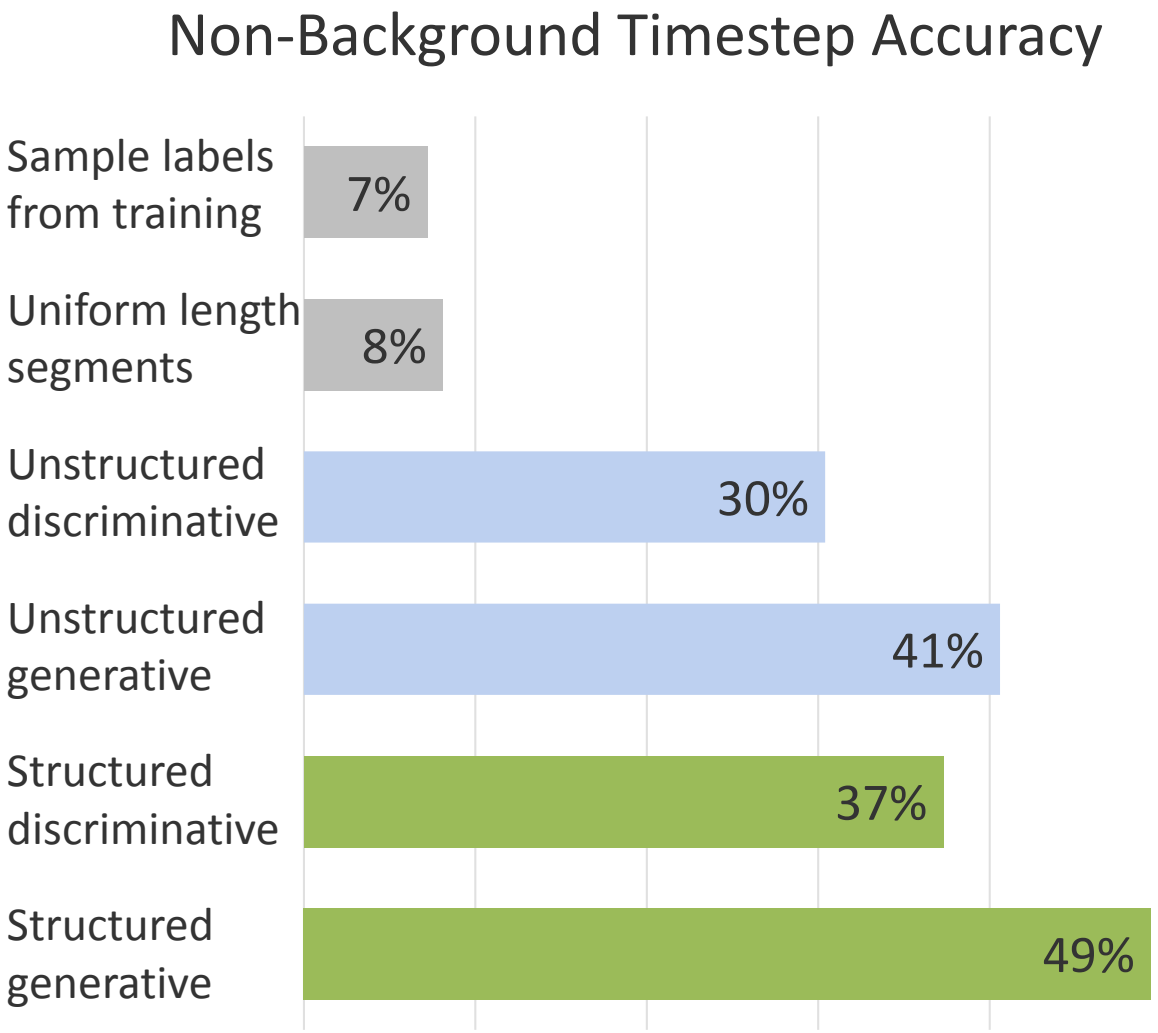


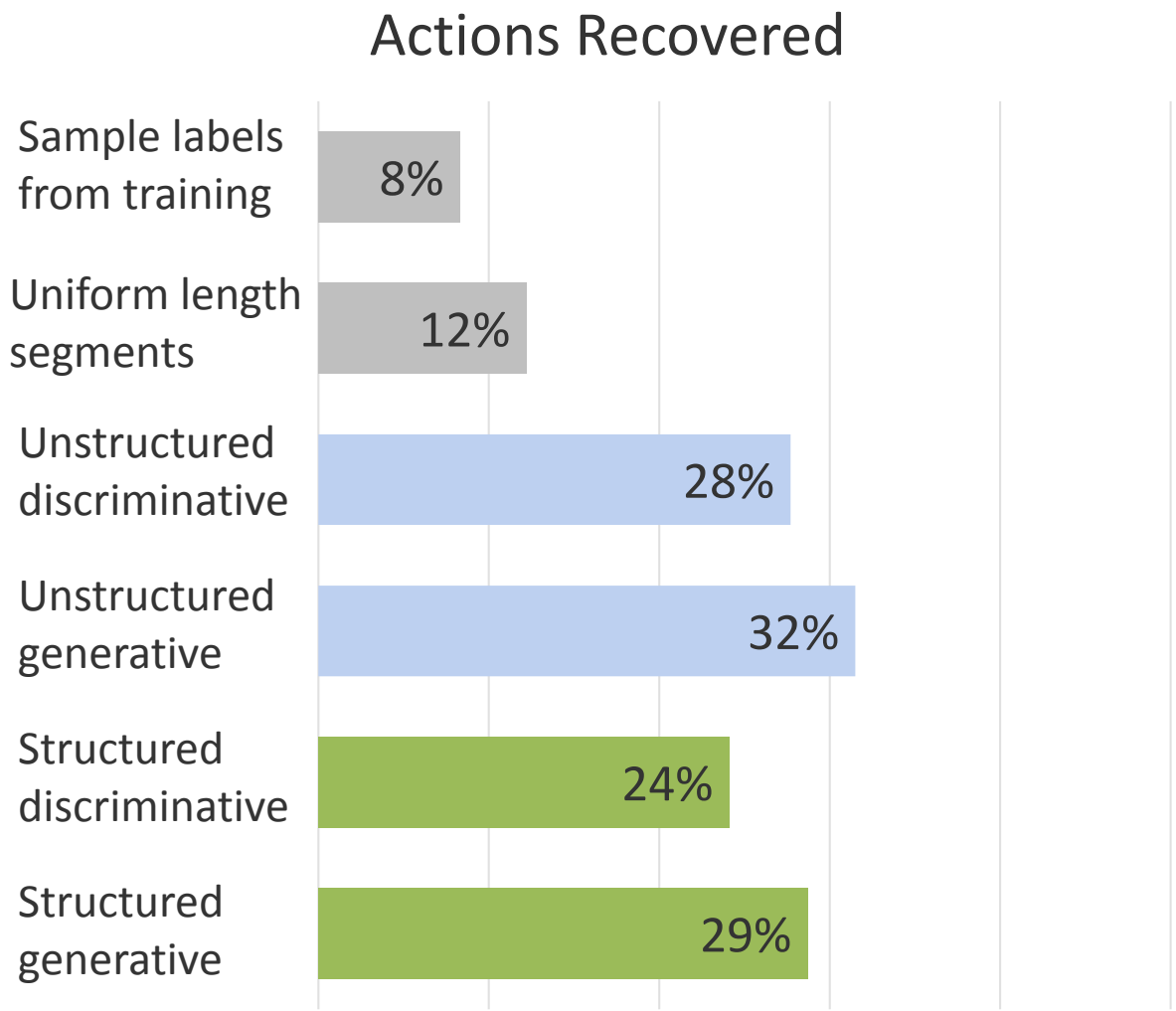
Supervised Results





Supervised Results

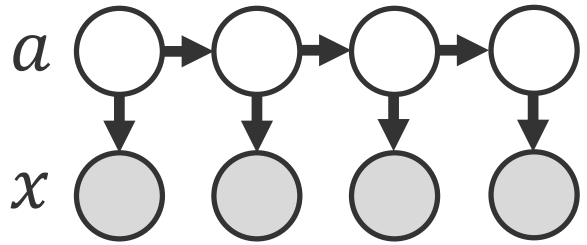




How little supervision can we get away with?

Train the structured, generative model without labels.

Training without Segment Labels



$$p(x) = \sum_{a_{1:T}} \prod_{t} p(a_t|a_{t-1})\, p(x_t|a_t)$$

Maximize $\log p(x)$ (gradient-based soft EM [Eisner 2016])

- Ordering Supervision

  Use a typical ordering of steps for each task, *e.g.* `add flour`, `add sugar`, ... [Zhukov et al.]

  Constrain $p(a_t|a_{t-1})$ to enforce this ordering over segments in all videos for the task

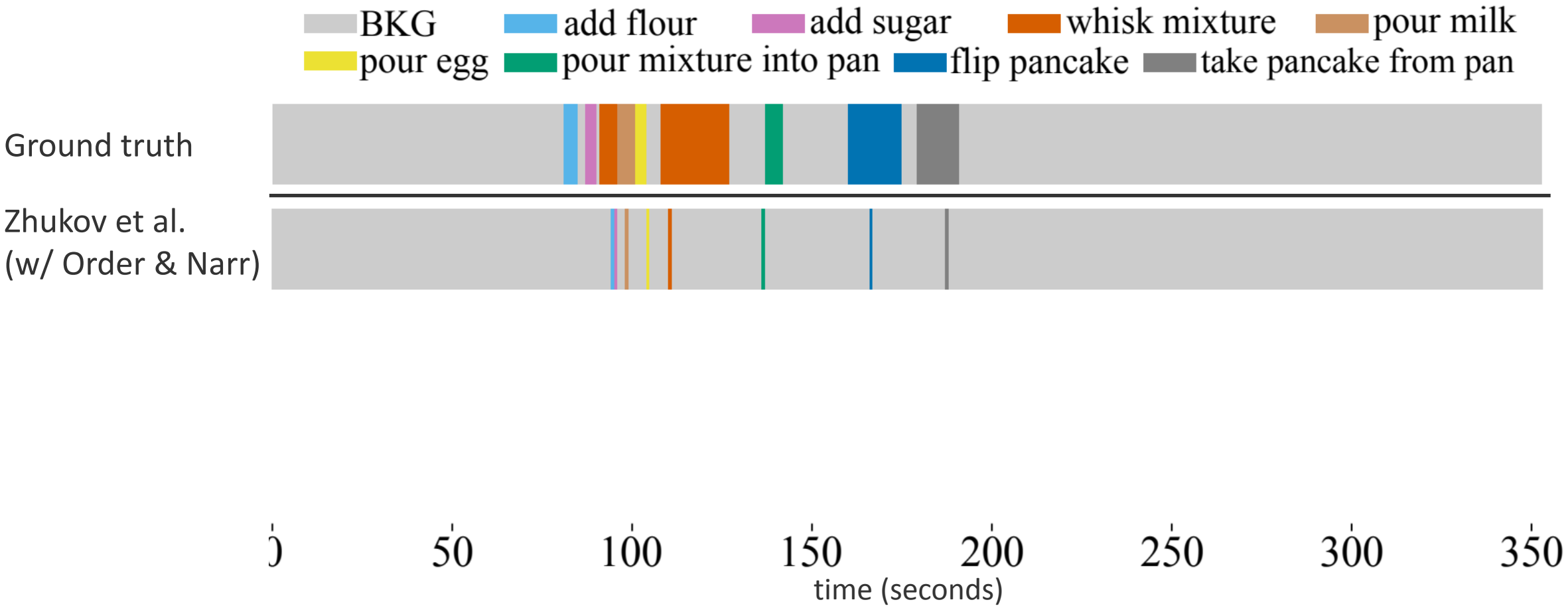
- Narration Supervision

  Use label—narration similarity and time alignment to constrain labels [Zhukov et al.]

  In training, constrain the sum over label assignments $\sum_{a_{1:T}}$

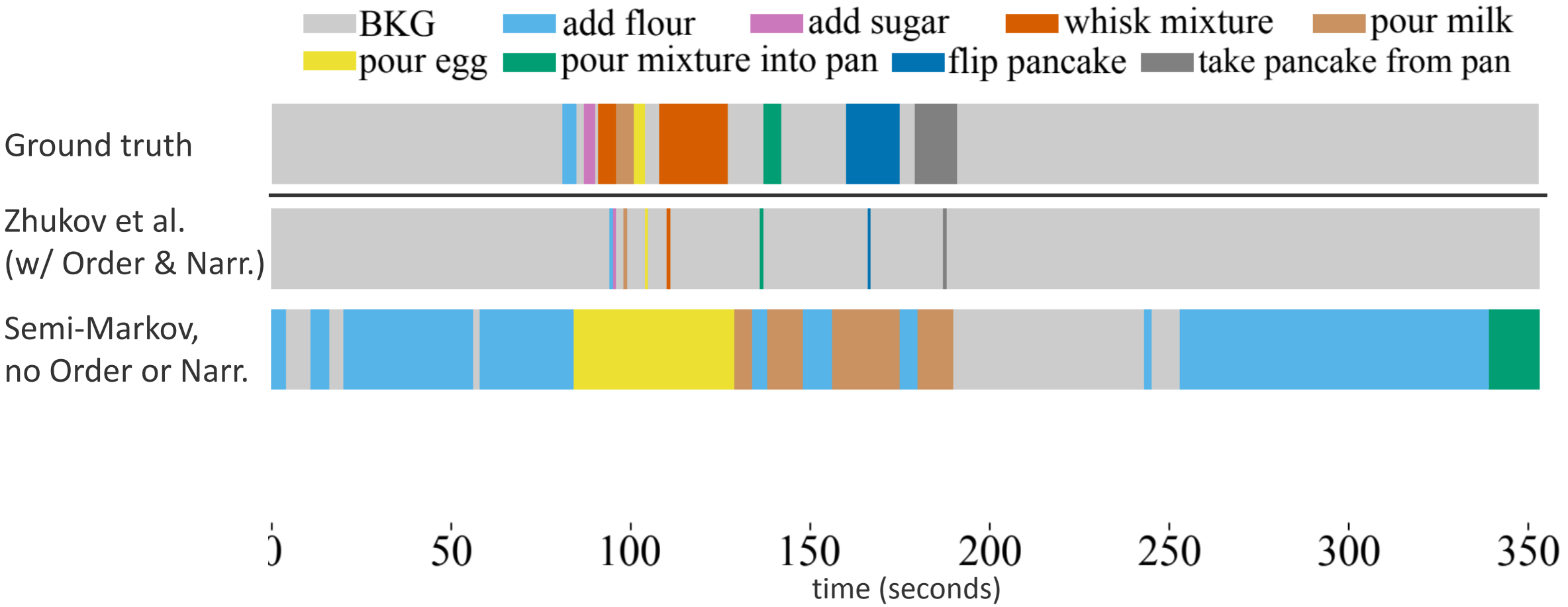
Training without Segment Labels

Task: *make pancakes*



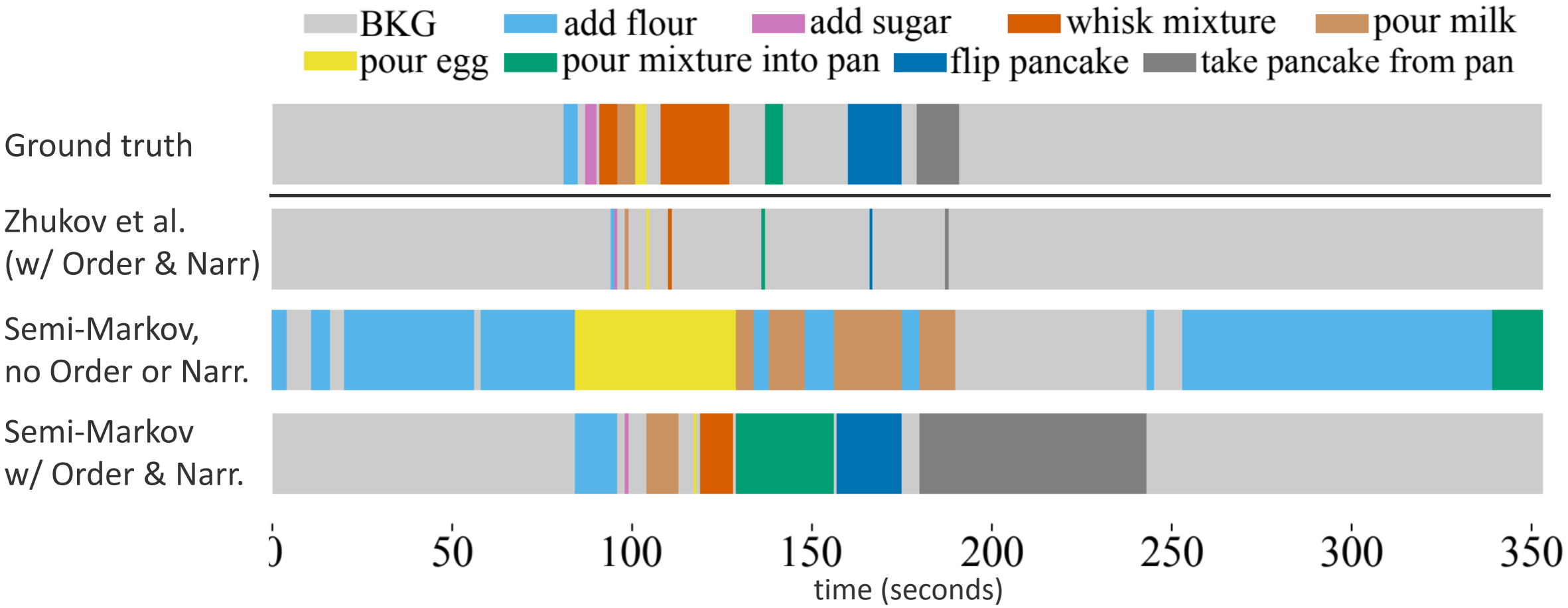
Training without Segment Labels

Task: *make pancakes*

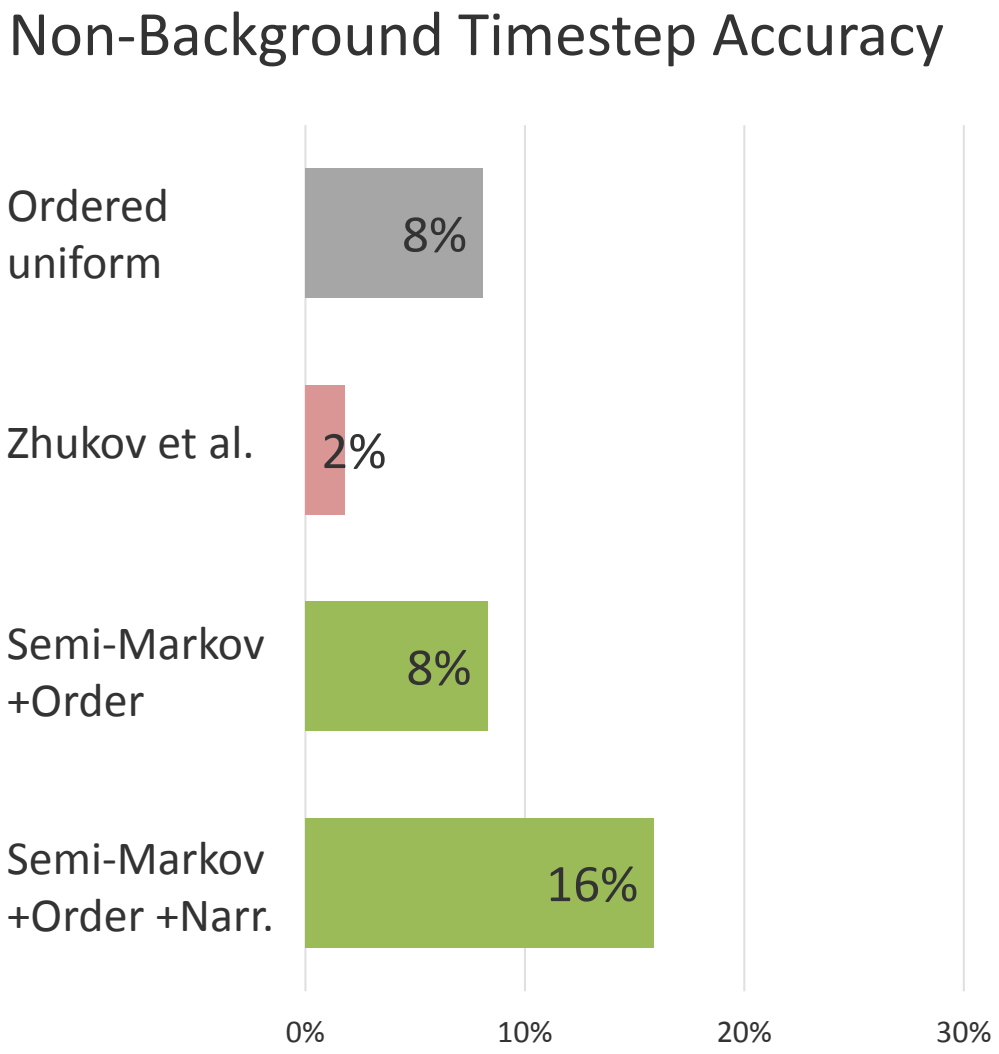


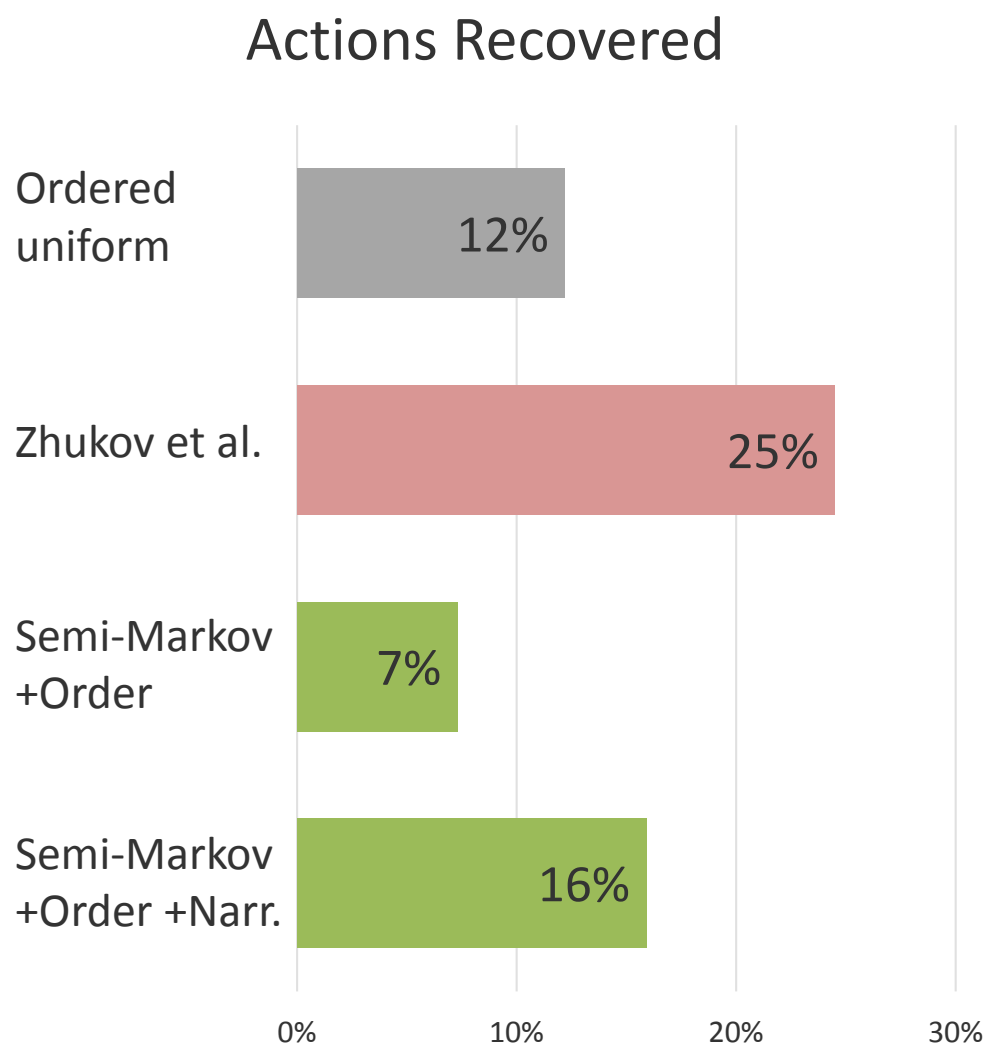
Training without Segment Labels

Task: *make pancakes*

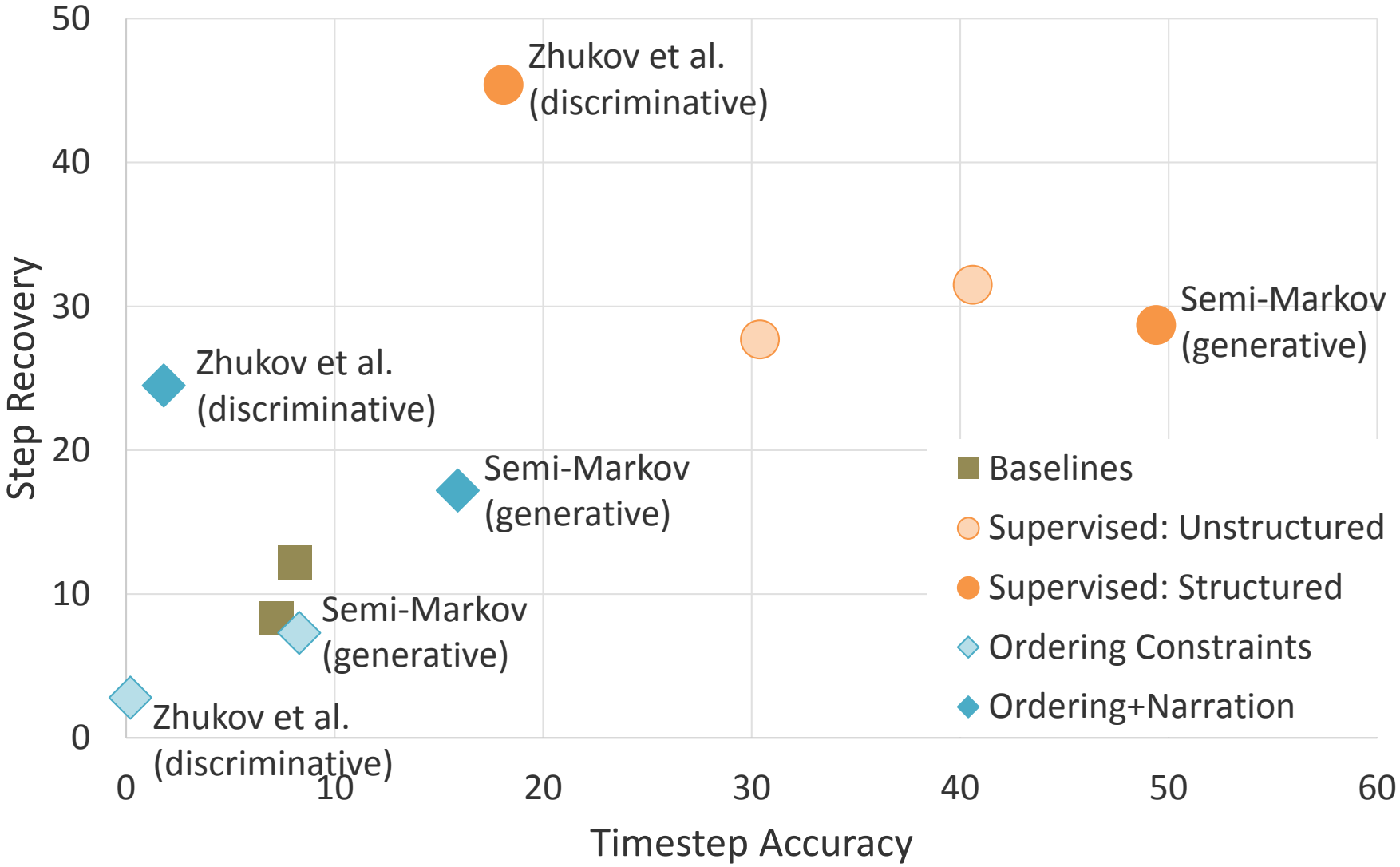


Training without Segment Labels



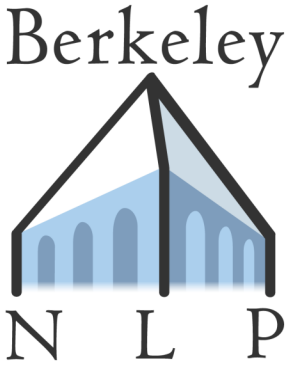


Effects of Supervision



How little supervision can we get away with?

Weak supervision from narration helps substantially!



QA Sessions:

Monday, July 6. 4B: Language Grounding-1. 18:00-19:00 UTC+0

Monday, July 6. 5B: Language Grounding-2. 21:00-22:00 UTC+0

`github.com/dpfried/action-segmentation`

Thank you!