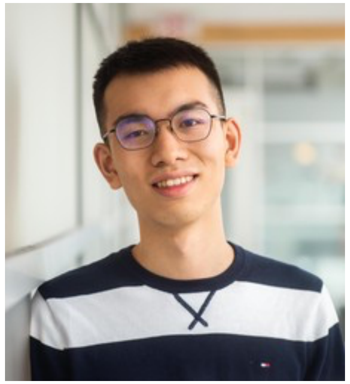
CVPR
VIRTUAL JUNE 19-25

Learning to Segment Actions from Visual and Language Instructions via Differentiable Weak Sequence Alignment



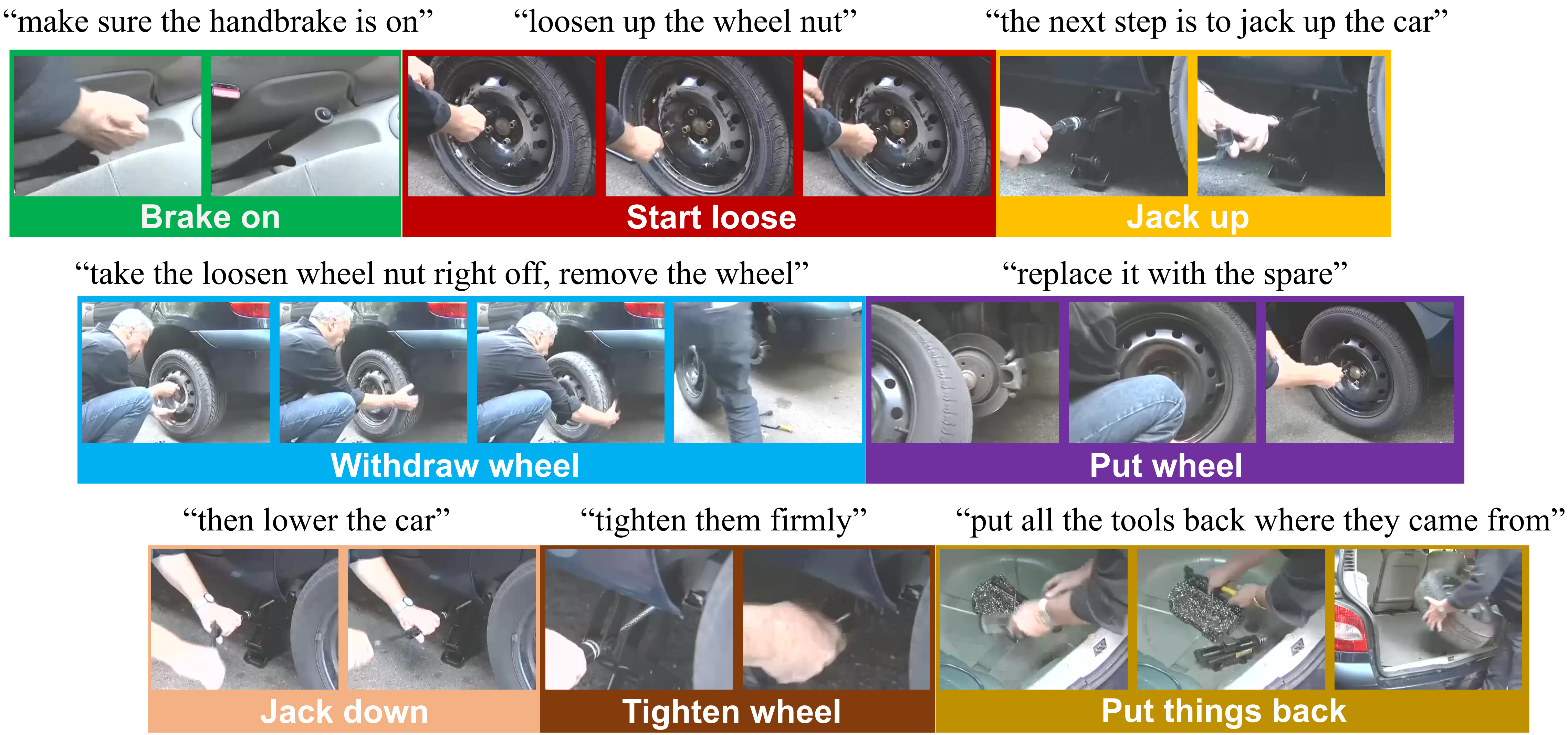
Yuhan Shen
Khoury College of Computer Sciences
Northeastern University

Lu Wang
Computer Science and Engineering
University of Michigan

Ehsan Elhamifar
Khoury College of Computer Sciences
Northeastern University

Action Segmentation

Goal: unsupervised action segmentation in instructional (procedural) videos



Prior Work

- **Visual-Only** [Sener-Yao'18, Elhamifar-Naing'19, GoelBrunskill'19, Kukleva et al'19, Elhamifar-Huynh'20] → **Cannot use narrations**

"So now we are going to slowly **lower the car back down**."

"To begin with, **find the jacking point** closest to the wheel."

"And then we can **jack the car up**."

Correct label: start loose

Not aligned!

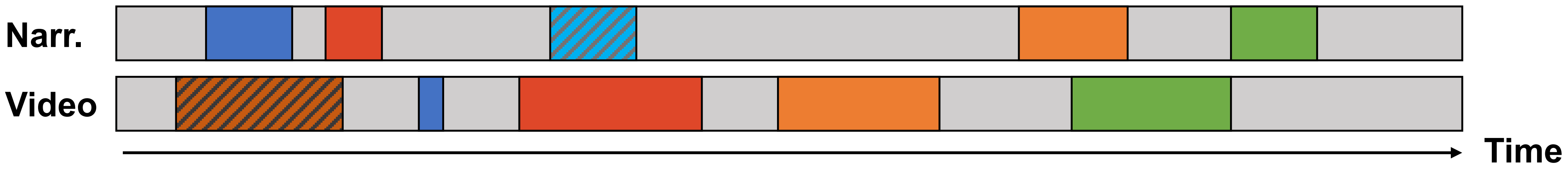
- **Visual+Narration** [Malmaud et al'15, Alayrac et al'16, Fried et al'20]
  - Assume temporal alignment → **Often violated**
  - Use precomputed features → **Cannot perform feature learning**

Contributions

CVPR VIRTUAL JUNE 19-25

- **Unsupervised action segmentation using visual data and narrations**
  - Soft ordered prototype learning: **extract key-steps**
  - Differentiable weak sequence alignment: **weakly align** videos

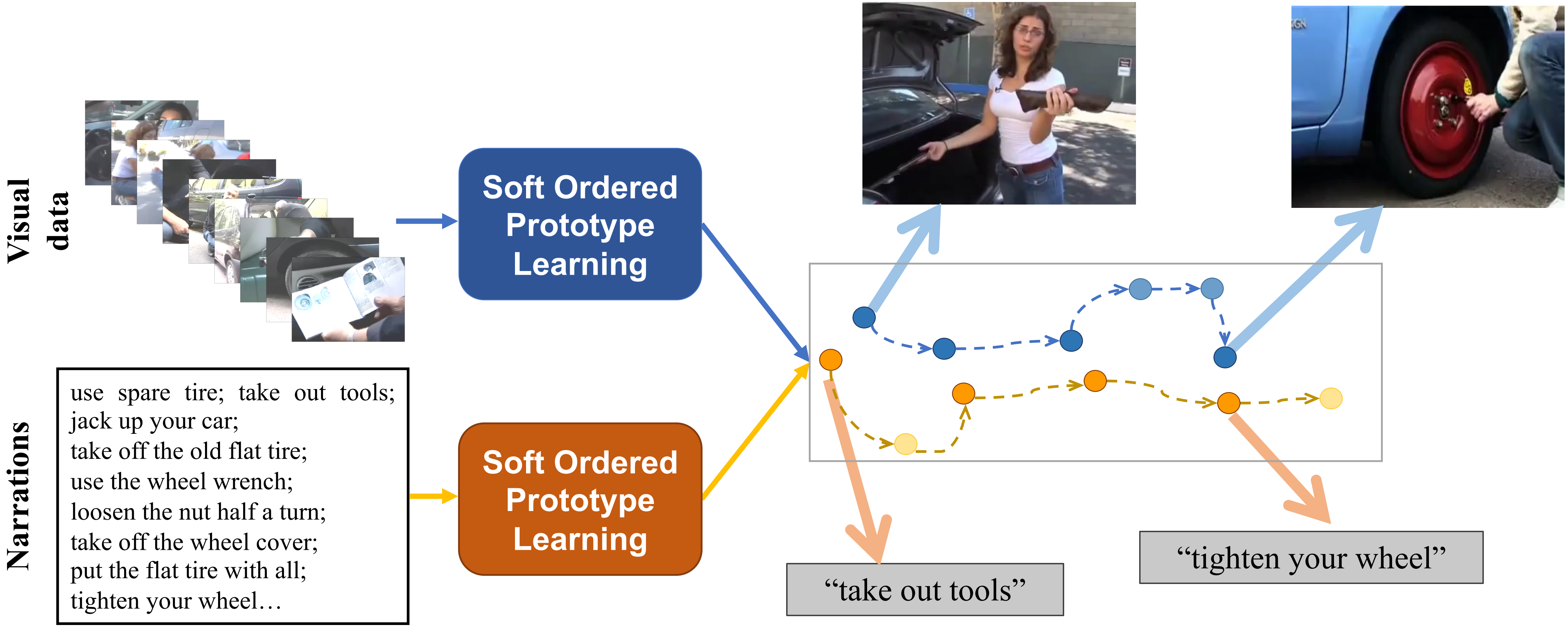
- **Observation:** Sequences of visual and linguistic key-steps are **weakly-aligned**



- **Self-supervised multi-modal feature learning**

Proposed Approach

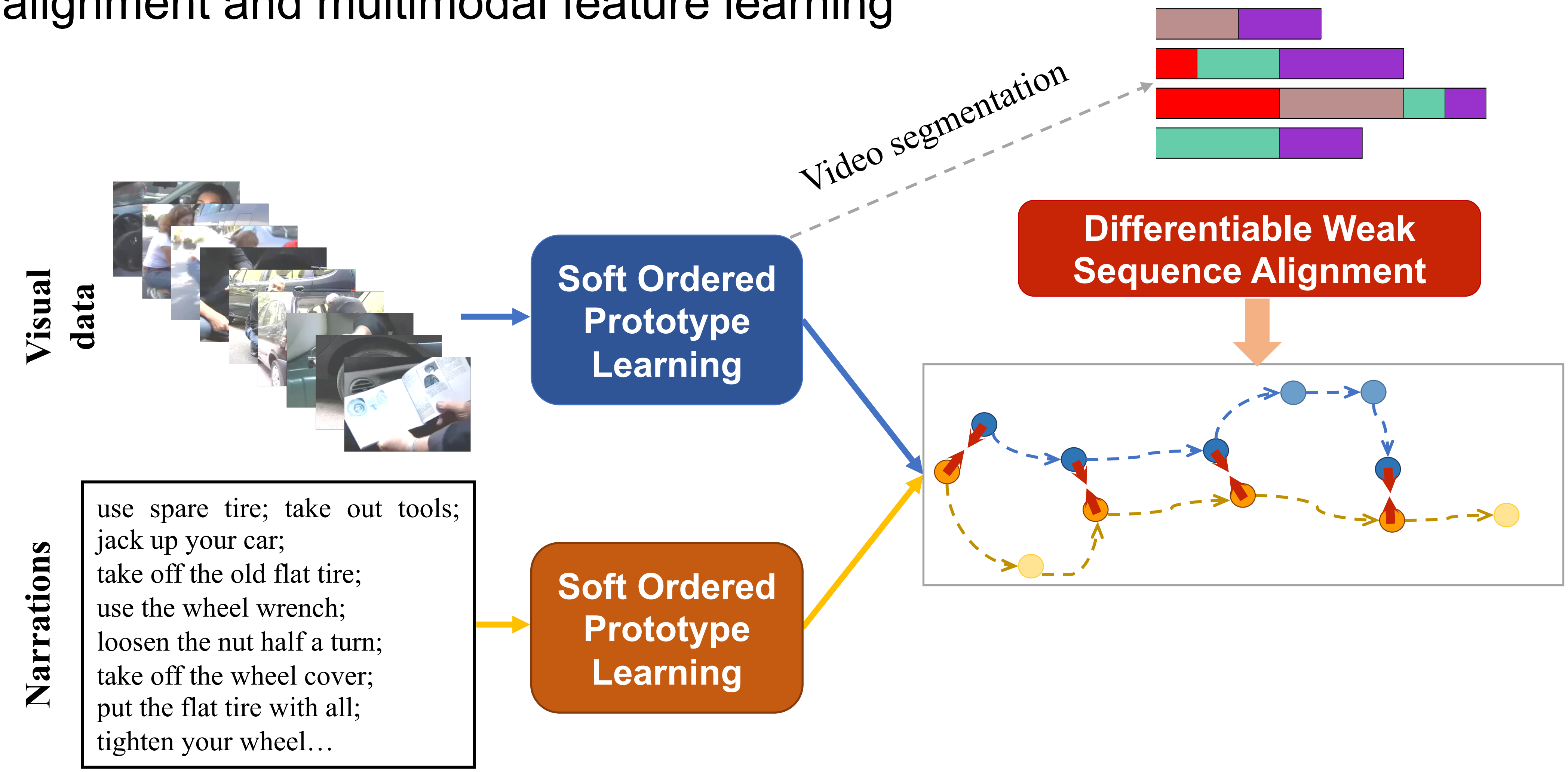
- **Soft Ordered Prototype Learning (SOPL):** recover visual and linguistic prototype sequences representing key-steps



Proposed Approach

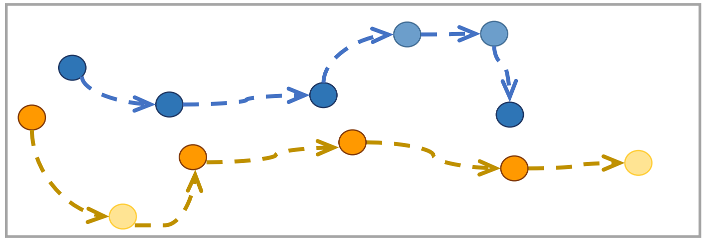
CVPR VIRTUAL JUNE 19-25

- **Differentiable Weak Sequence Alignment (DWSA):** allow weak sequence alignment and multimodal feature learning



Differentiable Weak Sequence Alignment

- **Consider two symbolic sequences:**



$a \rightarrow b \rightarrow c \rightarrow d \rightarrow f \rightarrow g$

$a \rightarrow c \rightarrow d \rightarrow e \rightarrow h \rightarrow f$

Step 1: Insert empty slots

Step 2: Compute pairwise alignment cost

	$\emptyset$	a	$\emptyset$	b	$\emptyset$	c	$\emptyset$	d	$\emptyset$	f	$\emptyset$	g	$\emptyset$
a	0	-1	0	1	0	1	0	1	0	1	0	1	0
c	0	1	0	1	0	-1	0	1	0	1	0	1	0
d	0	1	0	1	0	1	0	-1	0	1	0	1	0
e	0	1	0	1	0	1	0	1	0	1	0	1	0
h	0	1	0	1	0	1	0	1	0	1	0	1	0
f	0	1	0	1	0	1	0	1	0	-1	0	1	0

Pairwise Cost $\{\delta_{i,j}\}$

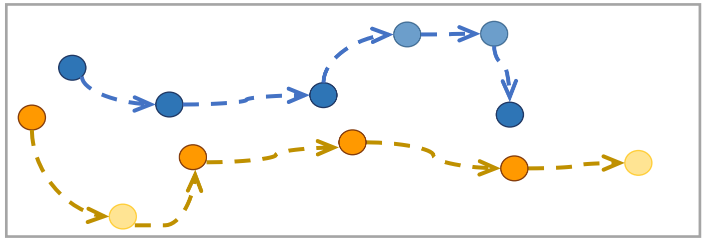
Alignment Cost Function: $\begin{cases} \delta(x,x) = -1 \\ \delta(x,y) = 1 \\ \delta(x,\phi) = 0 \end{cases}$

Differentiable Weak Sequence Alignment

- **Consider two symbolic sequences:**

$a \rightarrow b \rightarrow c \rightarrow d \rightarrow f \rightarrow g$

$a \rightarrow c \rightarrow d \rightarrow e \rightarrow h \rightarrow f$



Step 3: Dynamic program to update cumulative cost matrix

$$\textbf{Update rule: } d_{i,j} = \begin{cases} \delta_{i,j} + \min_{\beta}\{d_{i-1,1}, \cdots, d_{i-1,j}\}, j \text{ is odd} \\ \delta_{i,j} + \min_{\beta}\{d_{i-1,1}, \cdots, d_{i-1,j-1}\}, j \text{ is even} \end{cases}$$

$$\boldsymbol{\min_{\beta}(a_1, a_2, \cdots) = -\beta \log \sum_k e^{-\frac{a_k}{\beta}}}$$

$\boldsymbol{\delta(o_i^v, o_j^l)}$

	∅	a	∅	b	∅	c	∅	d	∅	f	∅	g	∅
a	0	-1	0	1	0	1	0	1	0	1	0	1	0
c	0	1	0	1	0	-1	0	1	0	1	0	1	0
d	0	1	0	1	0	1	0	-1	0	1	0	1	0
e	0	1	0	1	0	1	0	1	0	1	0	1	0
h	0	1	0	1	0	1	0	1	0	1	0	1	0
f	0	1	0	1	0	1	0	1	0	-1	0	1	0

Pairwise Cost $\{\boldsymbol{\delta_{i,j}}\}$

	∅	a	∅	b	∅	c	∅	d	∅	f	∅	g	∅
a	0	-1	0	1	0	1	0	1	0	1	0	1	0
c	0	1	-1	0	-1	-2	-1	0	-1	0	-1	0	-1
d	0	1	-1	0	-1	0	-2	-3	-2	-1	-2	-1	-2
e	0	1	-1	0	-1	0	-2	-1	-3	-2	-3	-2	-3
h	0	1	-1	0	-1	0	-2	-1	-3	-2	-3	-2	-3
f	0	1	-1	0	-1	0	-2	-1	-3	-4	-3	-2	-3

Cumulative Cost $\{\boldsymbol{d_{i,j}}\}$

a	a
∅	b
c	c
d	d
e	∅
h	∅
f	f
∅	g

Total Cost: -4

Experiments

- **Datasets:** ProceL (Elhamifar et al. ICCV'19), CrossTask (Zhukov et al. CVPR'19)
- **Baselines:**

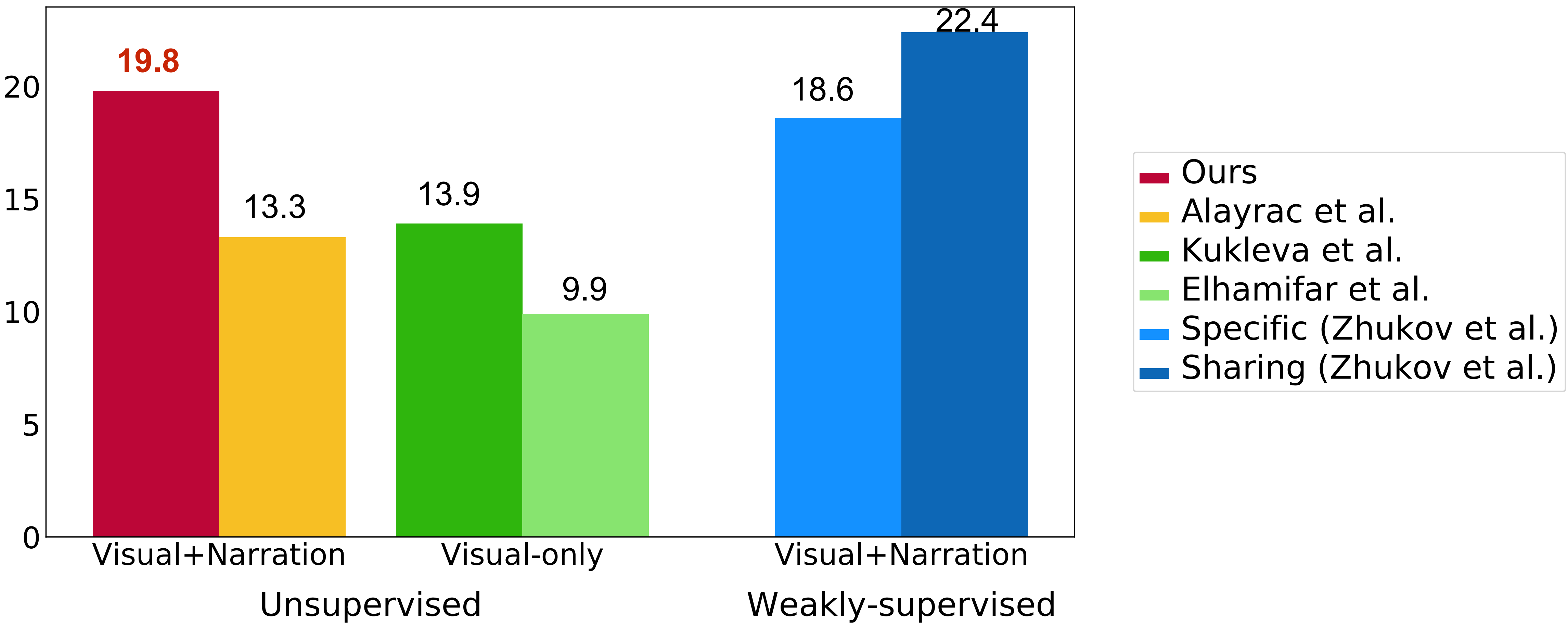
  Visual+Narration: Alayrac et al. CVPR'16

  Visual-only: Kukleva et al. CVPR'19, Elhamifar et al. ECCV'20
- **We improve SOTA by ~4.7% on F1 score on both datasets**

	ProceL		CrossTask	
	Precision (%)	F1 (%)	Precision (%)	F1 (%)
Alayrac et al.	12.25	5.54	6.80	4.46
Kukleva et al.	11.69	16.39	9.82	15.27
Elhamifar et al.	9.49	14.00	10.14	16.30
SOPL+Soft-DTW	14.29	18.41	14.36	19.83
SOPL+DWSA (Ours)	**16.51**	**21.07**	**15.21**	**21.00**

Experiments

- **Step Detection (Recall)** on CrossTask: detect one frame per key-step in each video.
- Outperform all unsupervised baselines; similar performance to weakly-supervised methods



Experiments

- **Qualitative results: more correct alignments after feature learning via DWSA**

Alignment between the prototypes in two modalities before and after feature learning using DWSA.

Thanks!