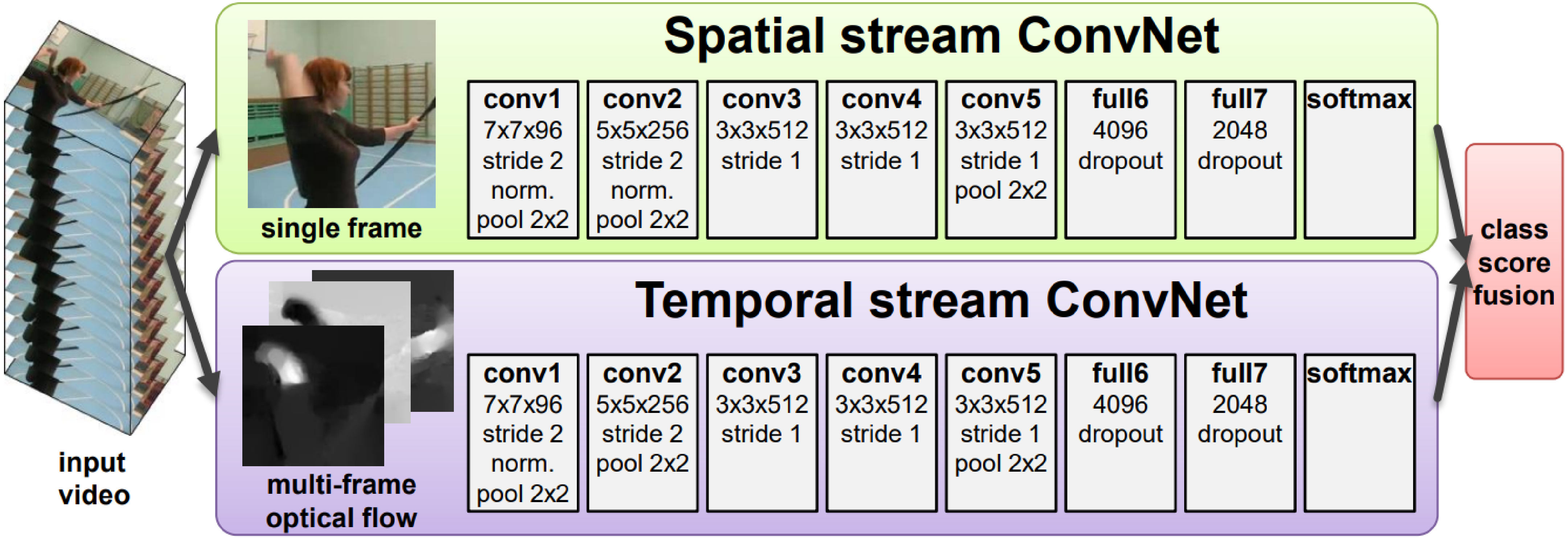
Understanding Actions in Video

Hazel Doughty

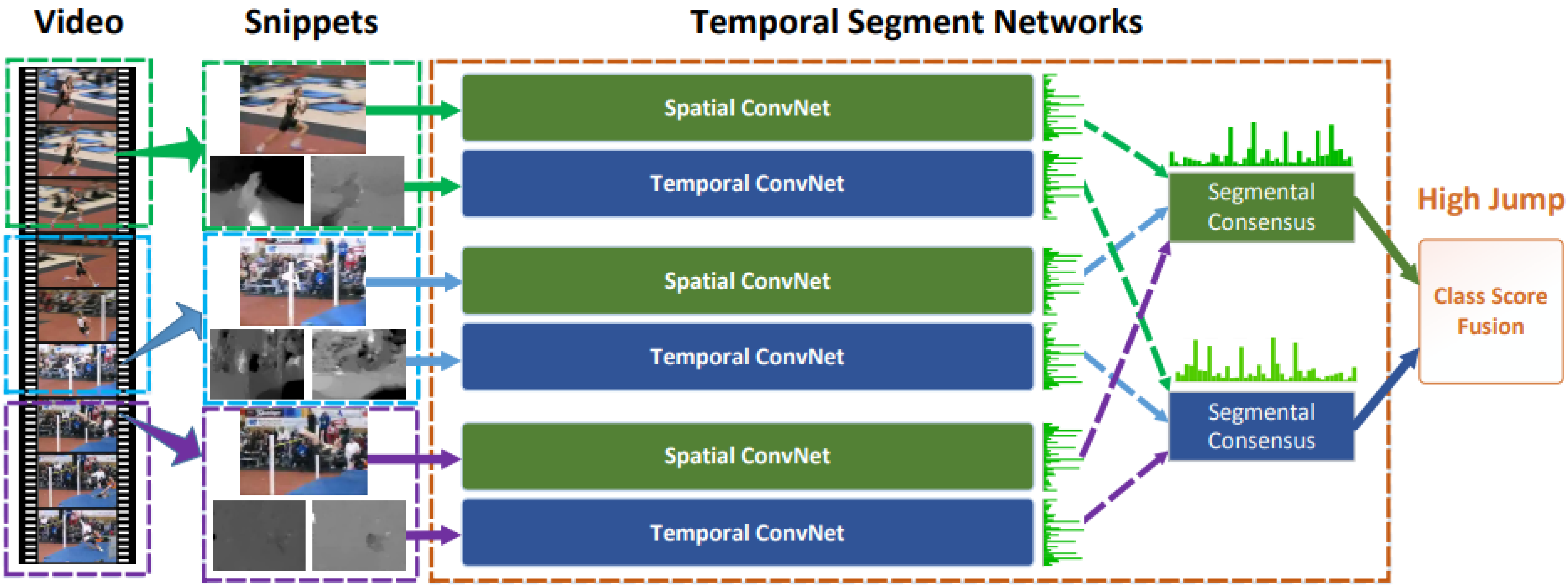
Action Recognition

Deep Learning for Action Recognition



Karen Simonyan and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *NeurIPS* (2014).

Deep Learning for Action Recognition



Limin Wang et al. "Temporal segment networks: Towards good practices for deep action recognition." *ECCV,* 2016

Fine-Grained Action Recognition

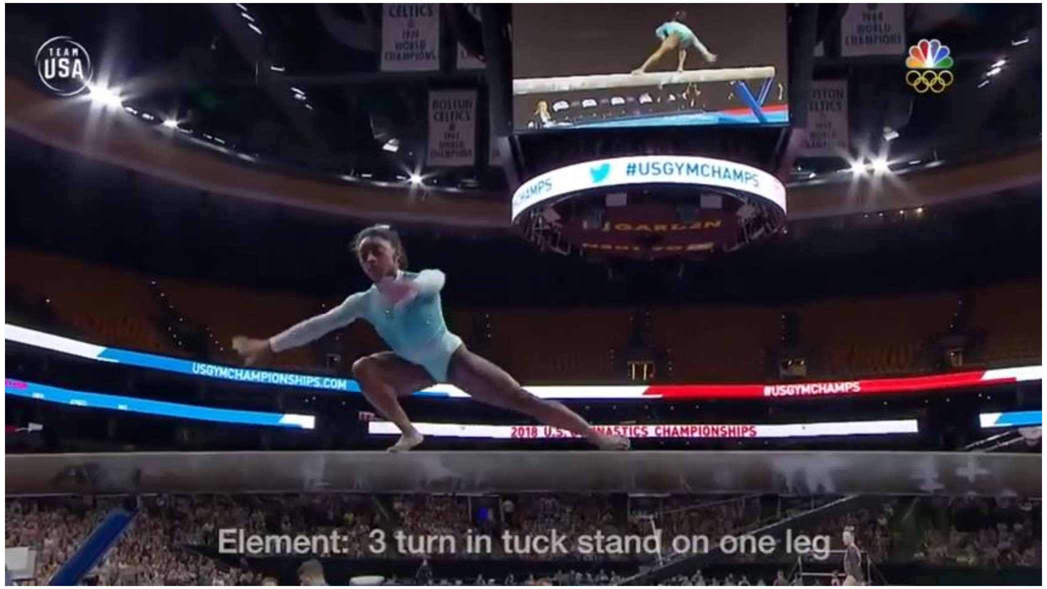
What is happening?

Coarse-grained: cooking

Fine-grained: cutting bell pepper

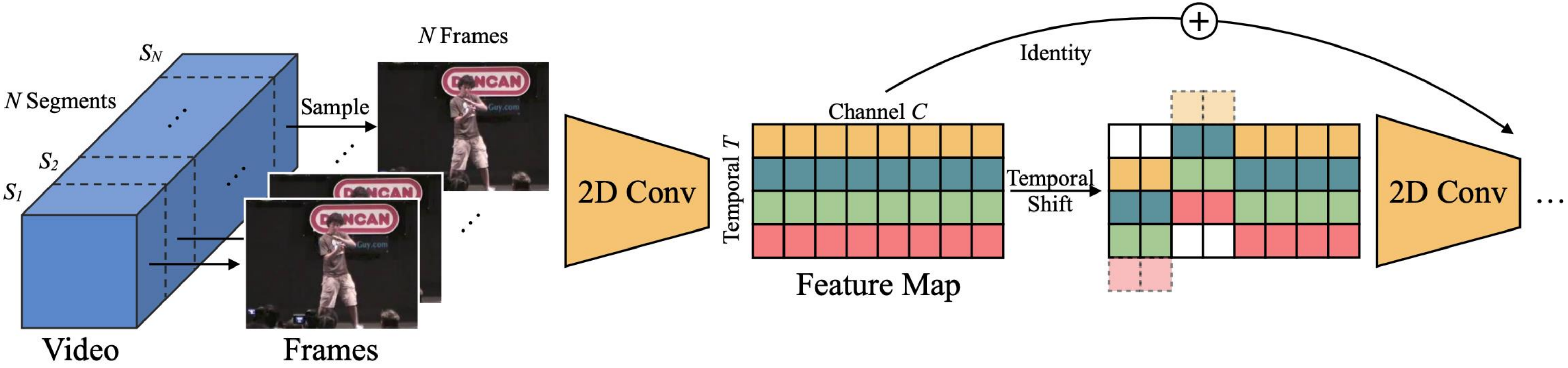
Fine-Grained Action Recognition

EPIC Kitchens, Damen et al. ECCV 2020



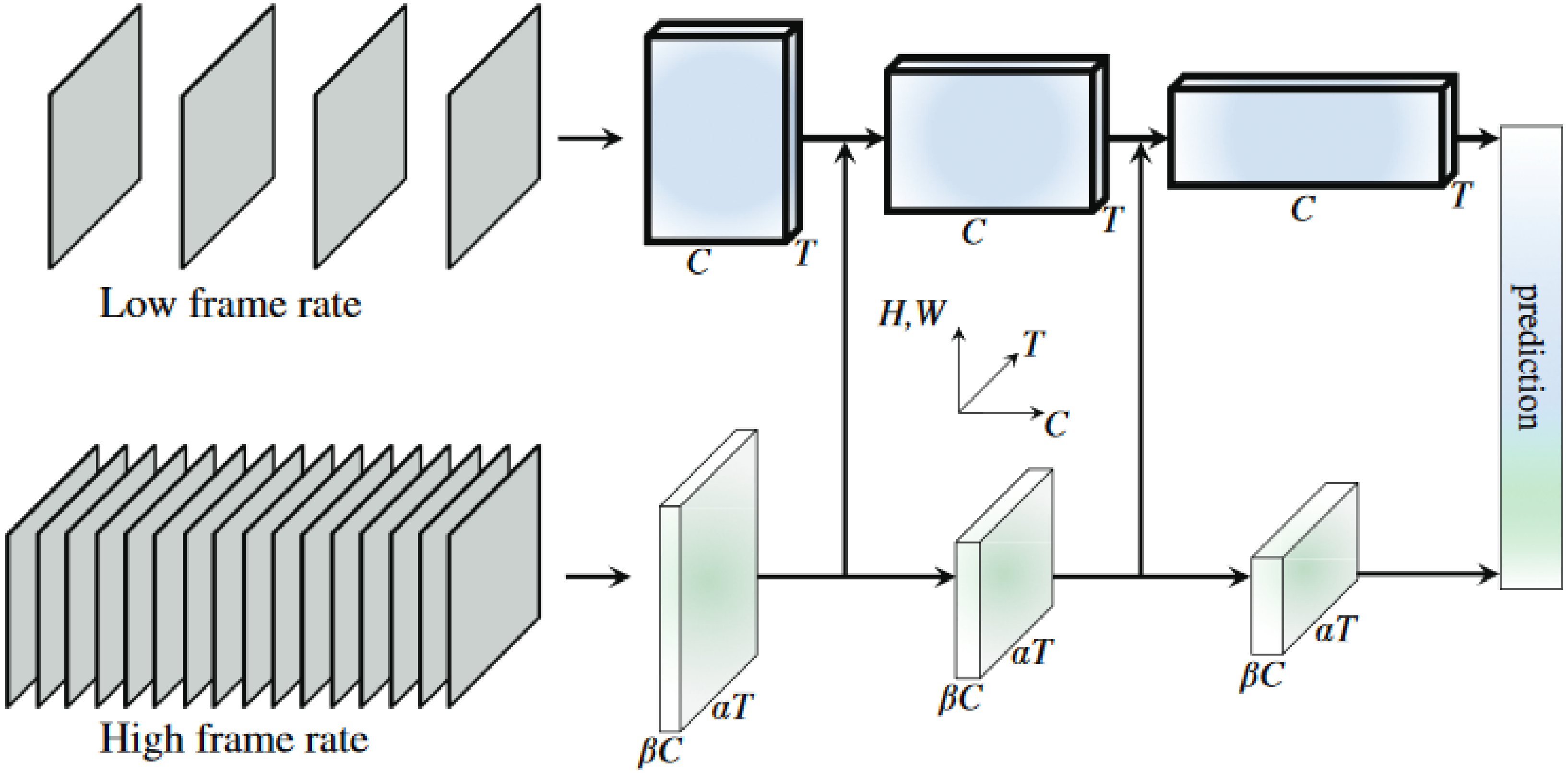
FineGym, Shao et al. CVPR 2020

Deep Learning for Action Recognition



Ji Lin, Chuang Gan, and Song Han. "Tsm: Temporal shift module for efficient video understanding." CVPR. 2019.

Deep Learning for Action Recognition



Christoph Feichtenhofer, et al. "Slowfast networks for video recognition." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

Issue with Supervision

Action Modifiers: Learning from Adverbs in Instructional Videos

CVPR 2020

Hazel Doughty

University of Bristol

Ivan Laptev

INRIA
École Normale Supérieure

Walterio Mayol-Cuevas

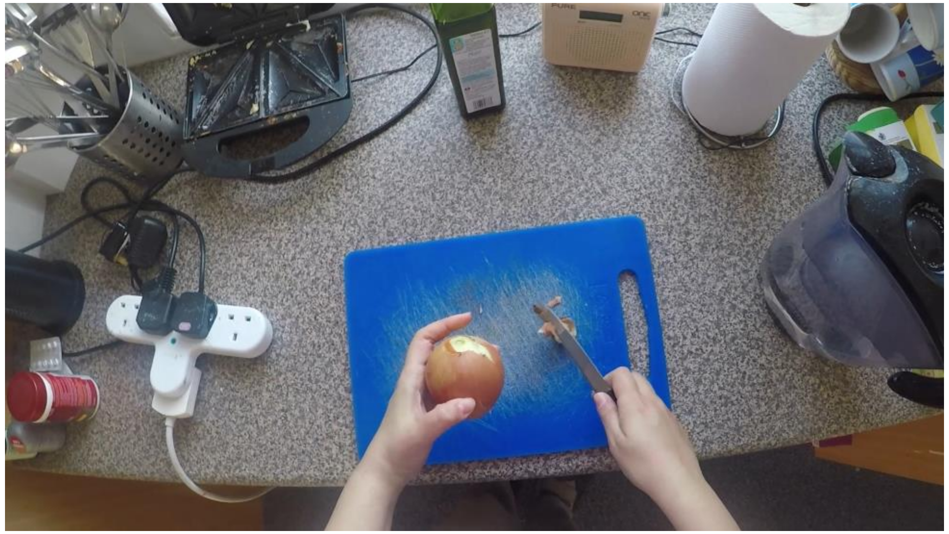
University of Bristol
Amazon

Dima Damen

University of Bristol

More info: https://hazeldoughty.github.io/Papers/ActionModifiers/

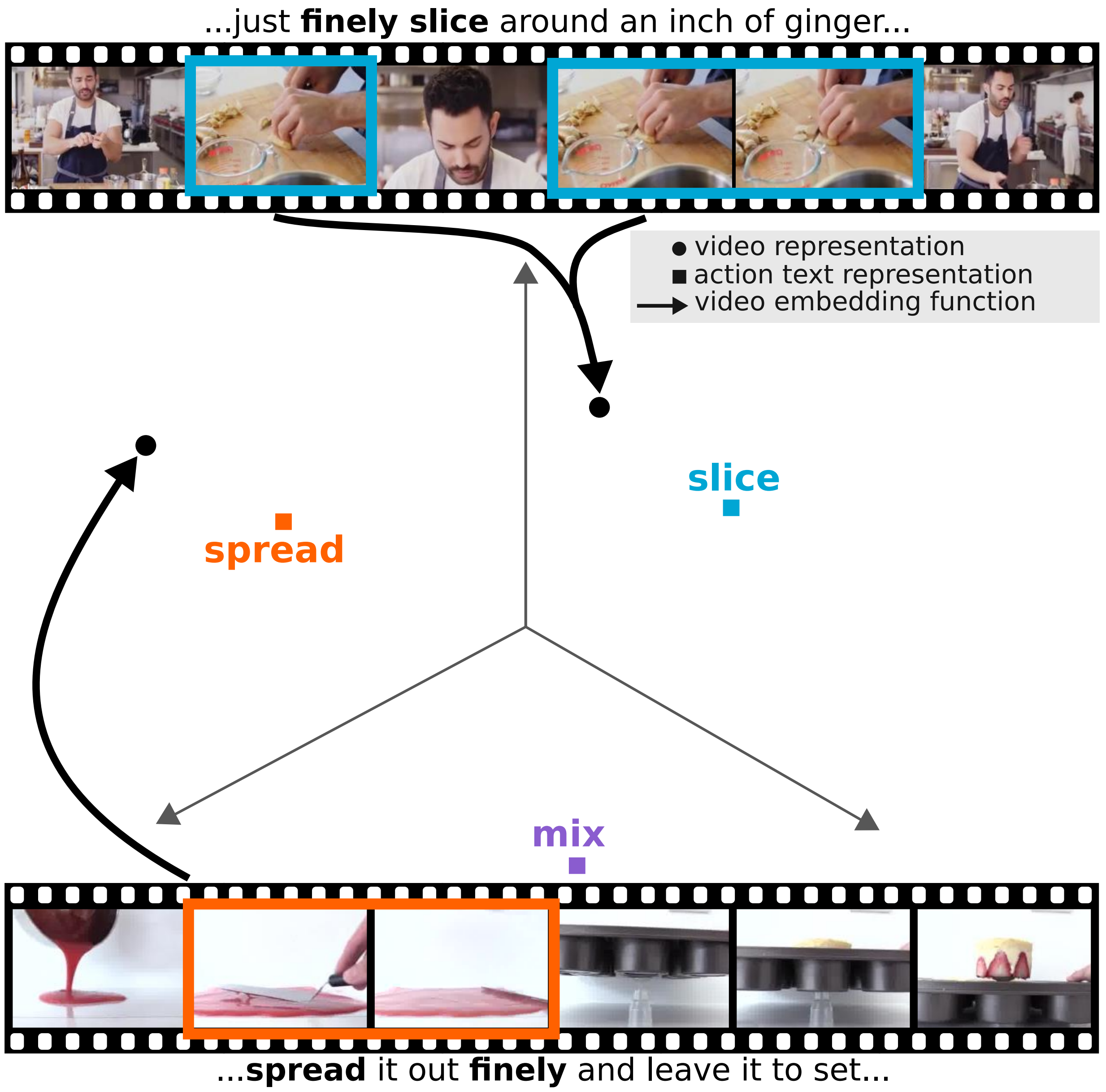
Beyond ‘What’ is Happening

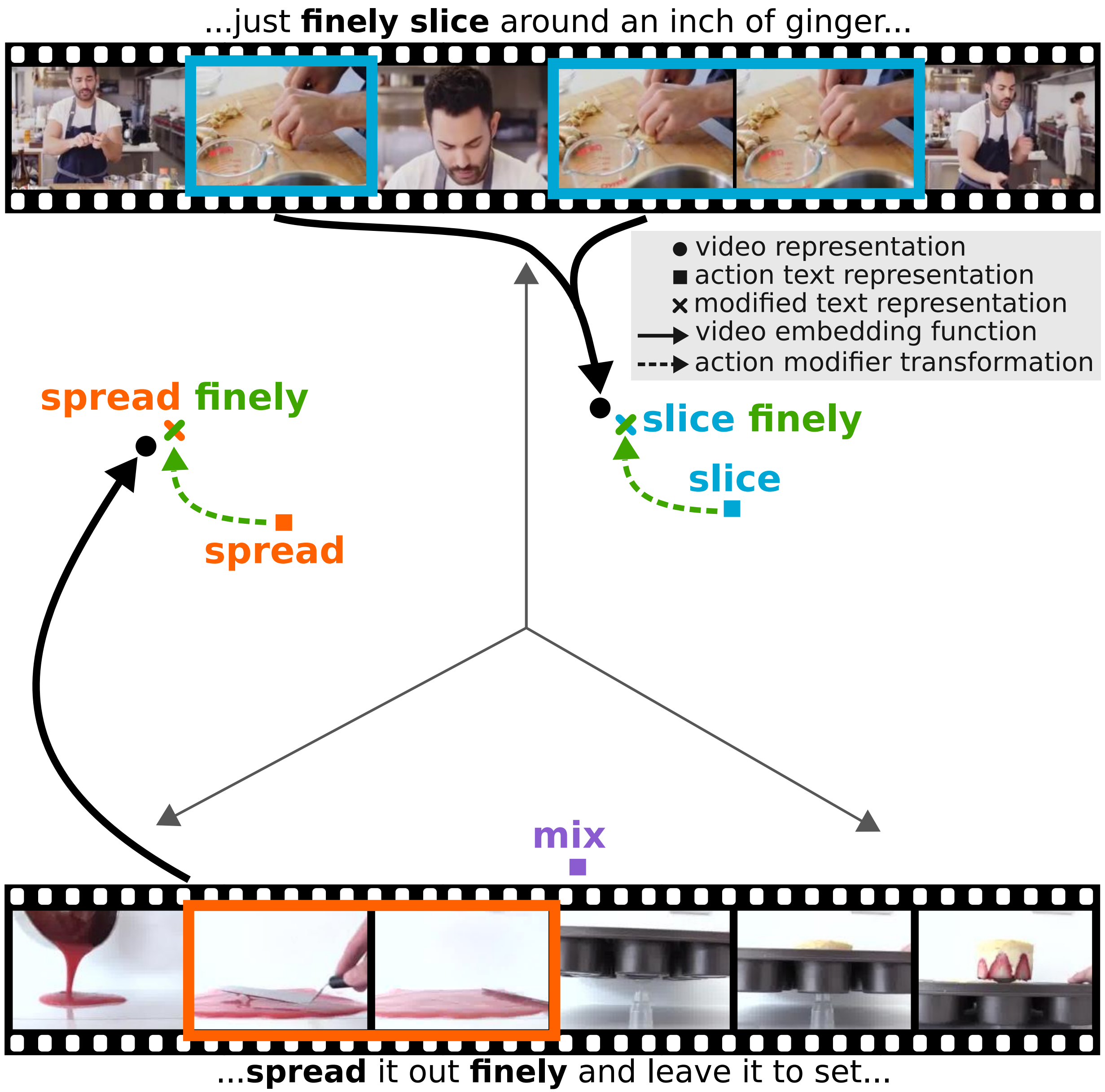


worse	peel onion	better
slowly	put onion peel in bin	quickly
messily	slice onion	neatly

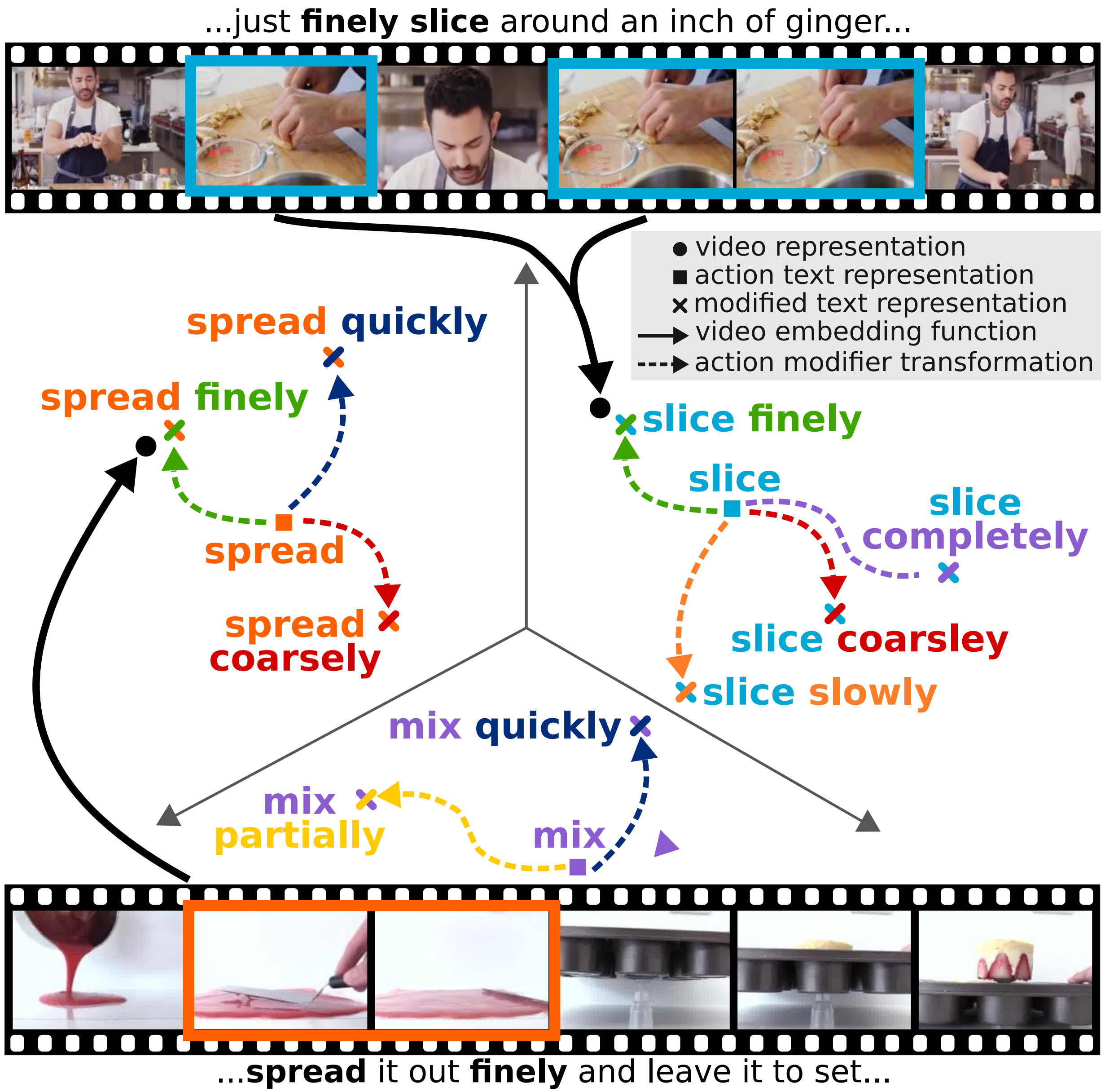
Adverbs


... if you **turn** the bowl upside down **slowly** they won't come out ...
... mix it well until it is **completely dissolved** ...
... you want to make sure you **fill** it up **partially** ...
... you want to **dice** it **finely**...
-10 seconds
timestamp
+10 seconds


...just **finely slice** around an inch of ginger...
video representation
action text representation
video embedding function
slice
spread
mix
...**spread** it out **finely** and leave it to set...



...just finely slice around an inch of ginger...
video representation
action text representation
modified text representation
video embedding function
action modifier transformation
spread finely
spread
slice finely
slice
mix
...spread it out finely and leave it to set...



...just **finely slice** around an inch of ginger...
video representation
action text representation
modified text representation
video embedding function
action modifier transformation
spread quickly
spread finely
spread
spread coarsely
slice finely
slice
slice completely
slice coarsley
slice slowly
mix quickly
mix partially
mix
...**spread** it out **finely** and leave it to set...

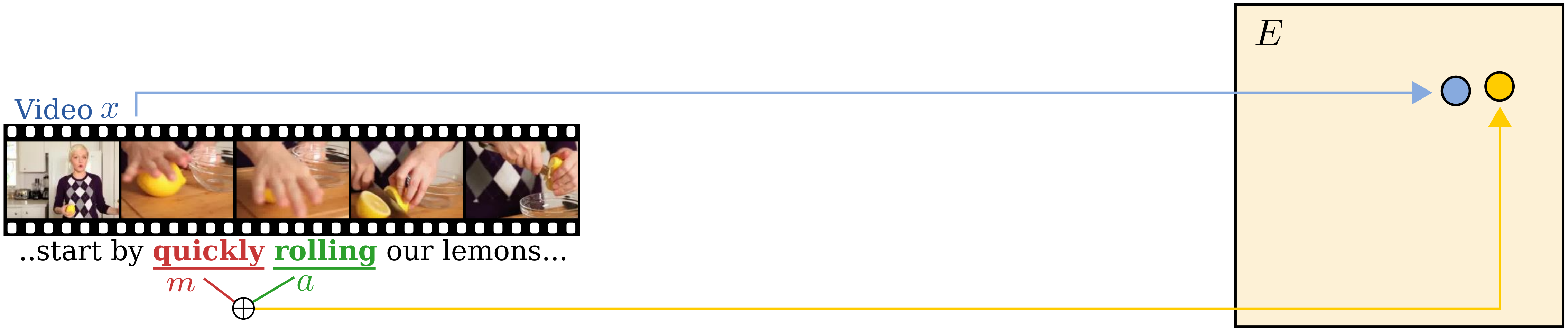
Adverbs – Action Modifiers

Video x

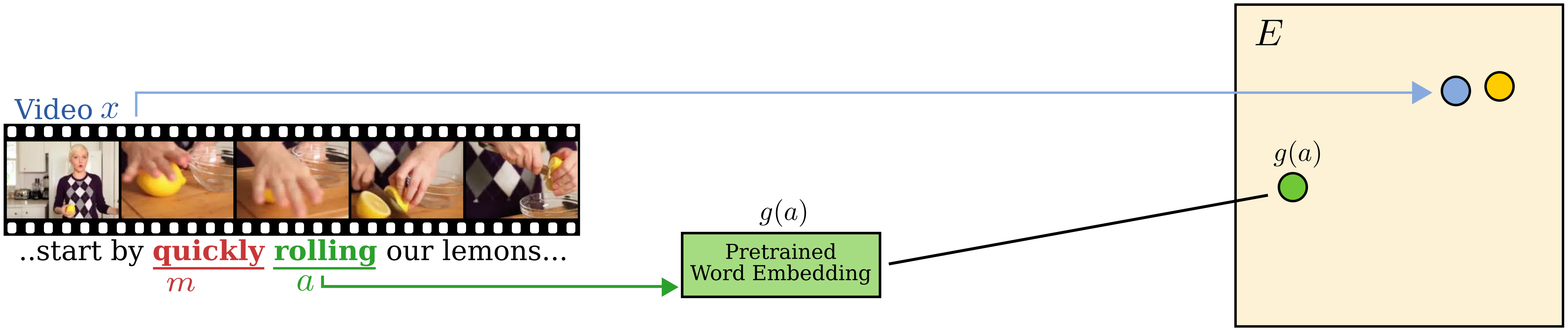
..start by **quickly** **rolling** our lemons...

m a

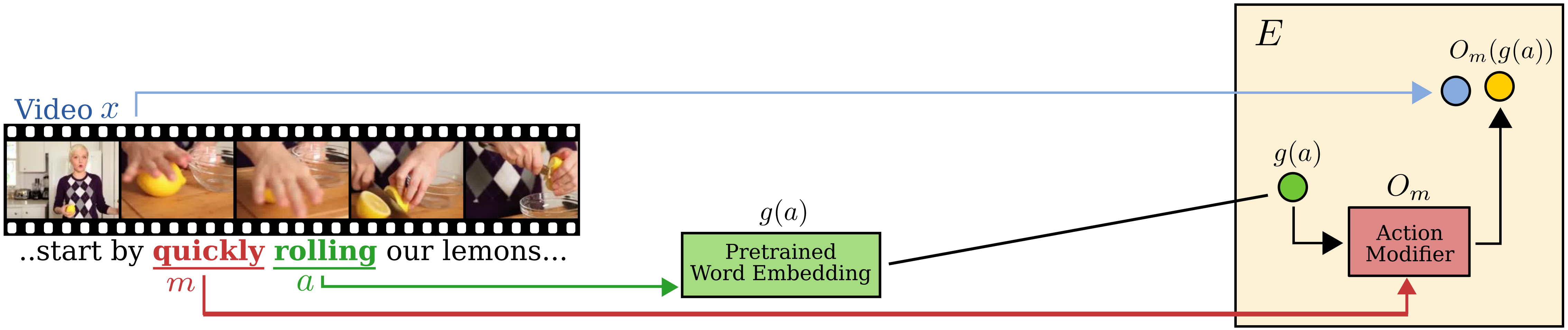
Adverbs – Action Modifiers



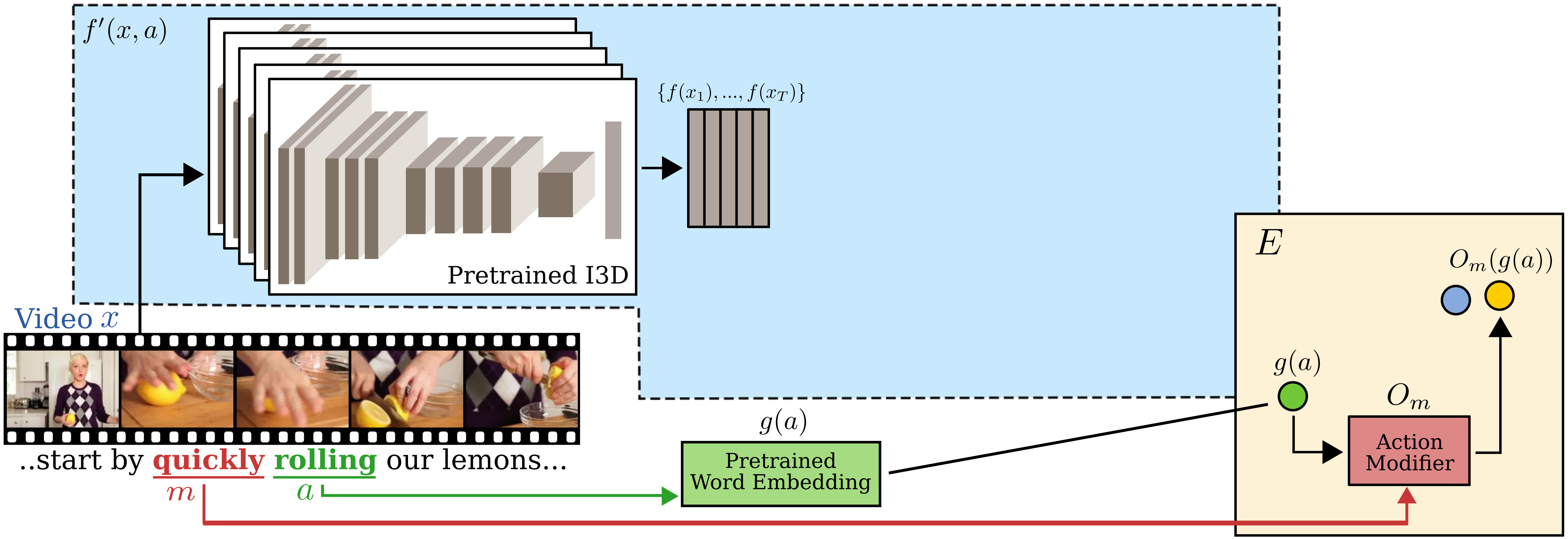
Adverbs – Action Modifiers



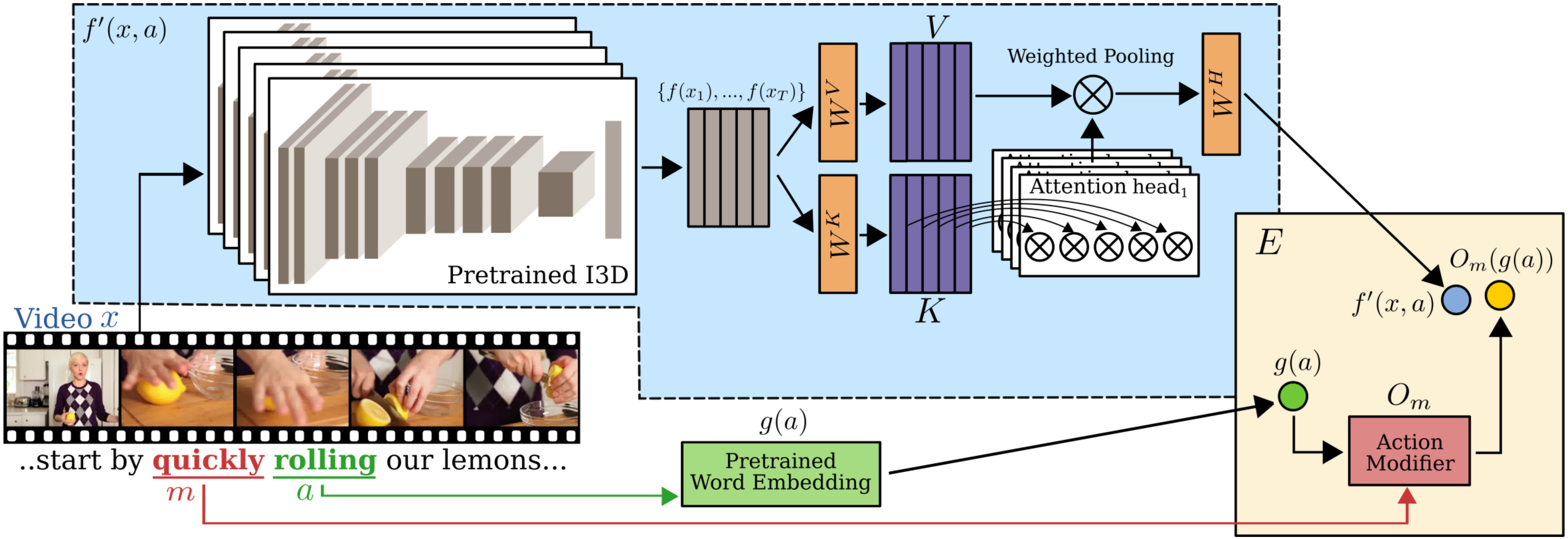
Adverbs – Action Modifiers



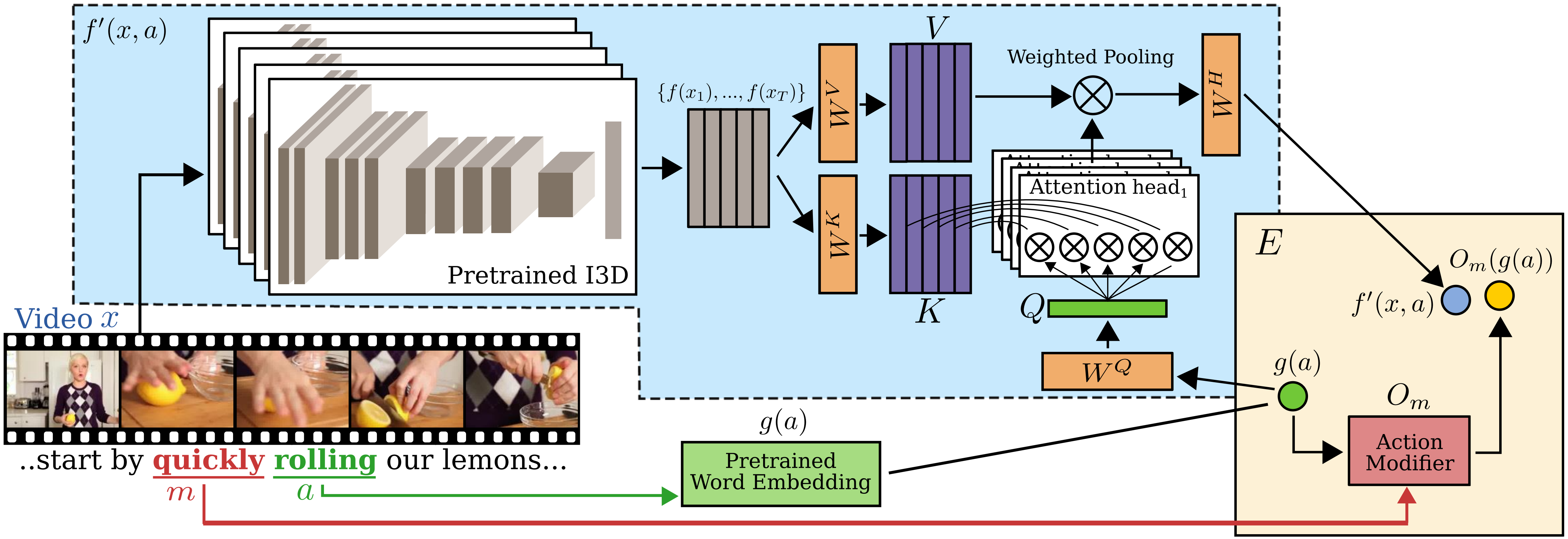
Adverbs – Action Modifiers



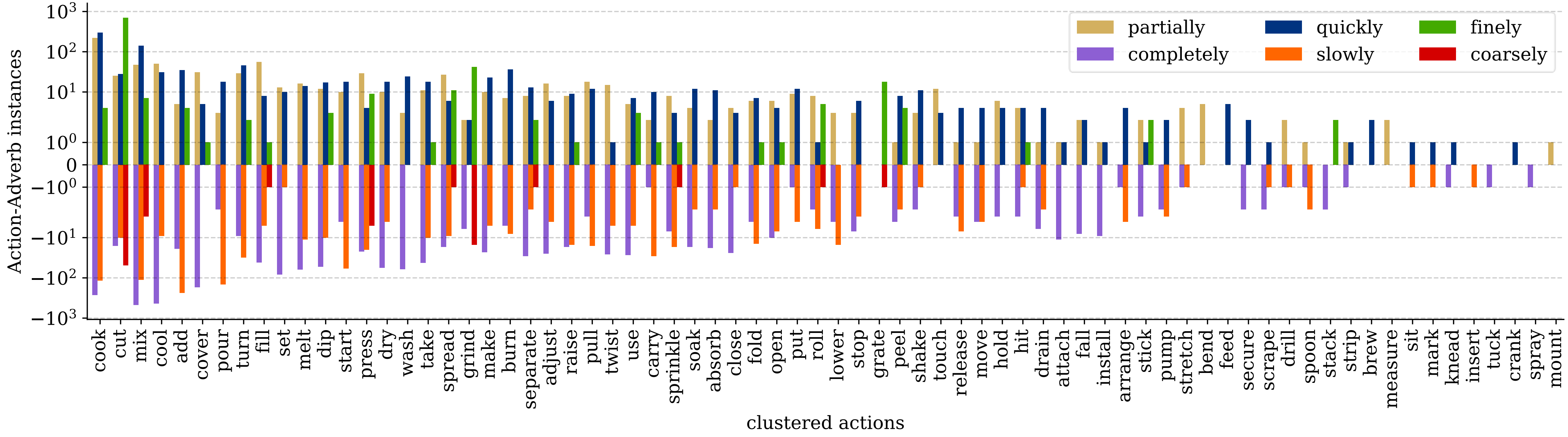
Adverbs – Action Modifiers



Adverbs – Action Modifiers



Adverbs - Dataset



Video: https://www.youtube.com/watch?v=rajo0x7WF-c&t=100s

... we're going to mix these up real **quick**...

... get under there then turn real **quick**...

...wash, roll up and spin it to **completely** dip it...

Conclusions

- The proposed method can learn how adverbs compose with different actions
- We can successfully learn adverb representations with weak supervision
- Open challenges:
    - *Representing more adverbs*
    - *Spatial disambiguation from weak supervision*
    - *Utilizing adverbs for other tasks*

How Do You Do It? Fine-Grained Action Understanding with Pseudo-Adverbs

CVPR 2022

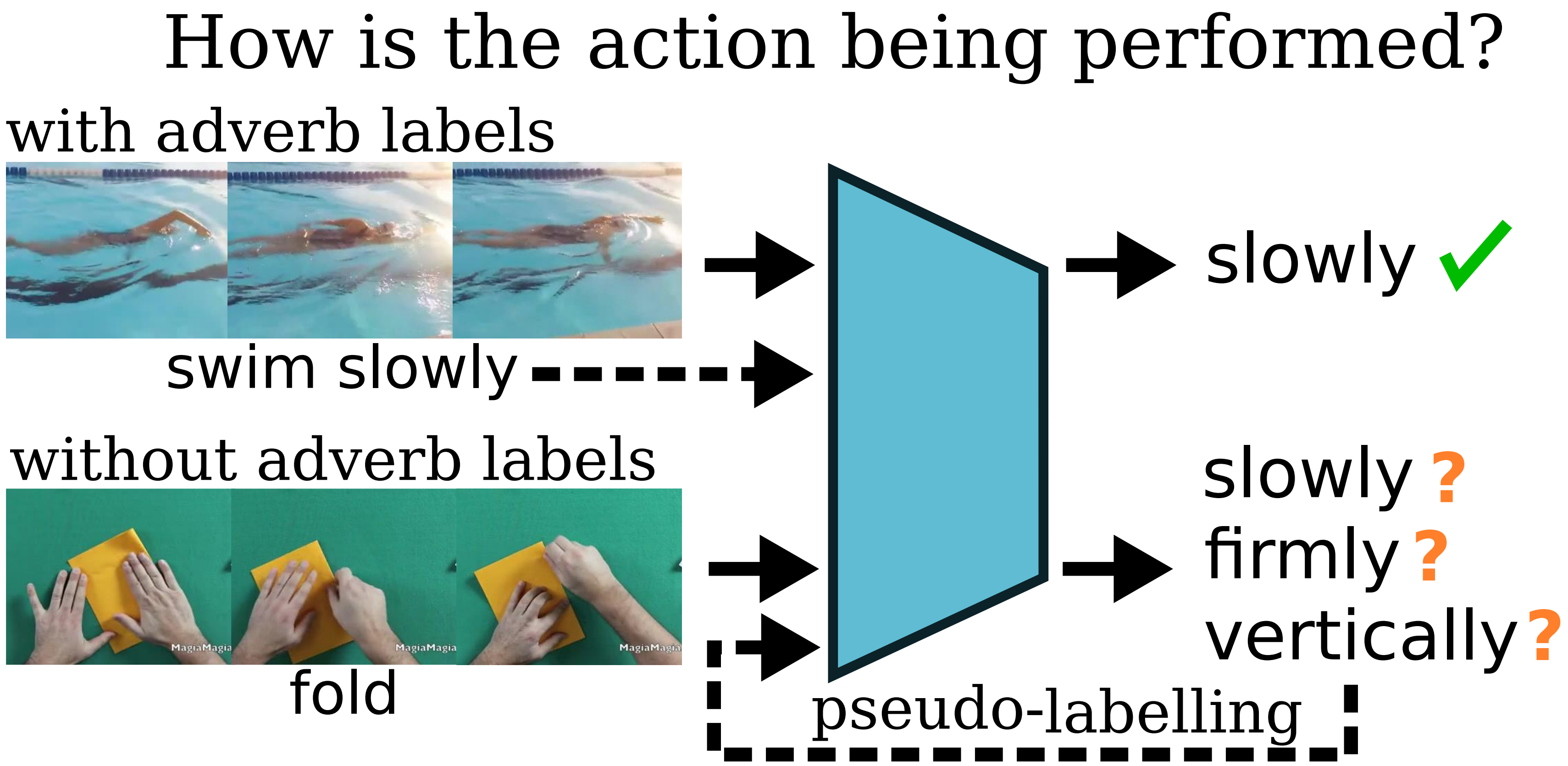
Hazel Doughty

Cees Snoek

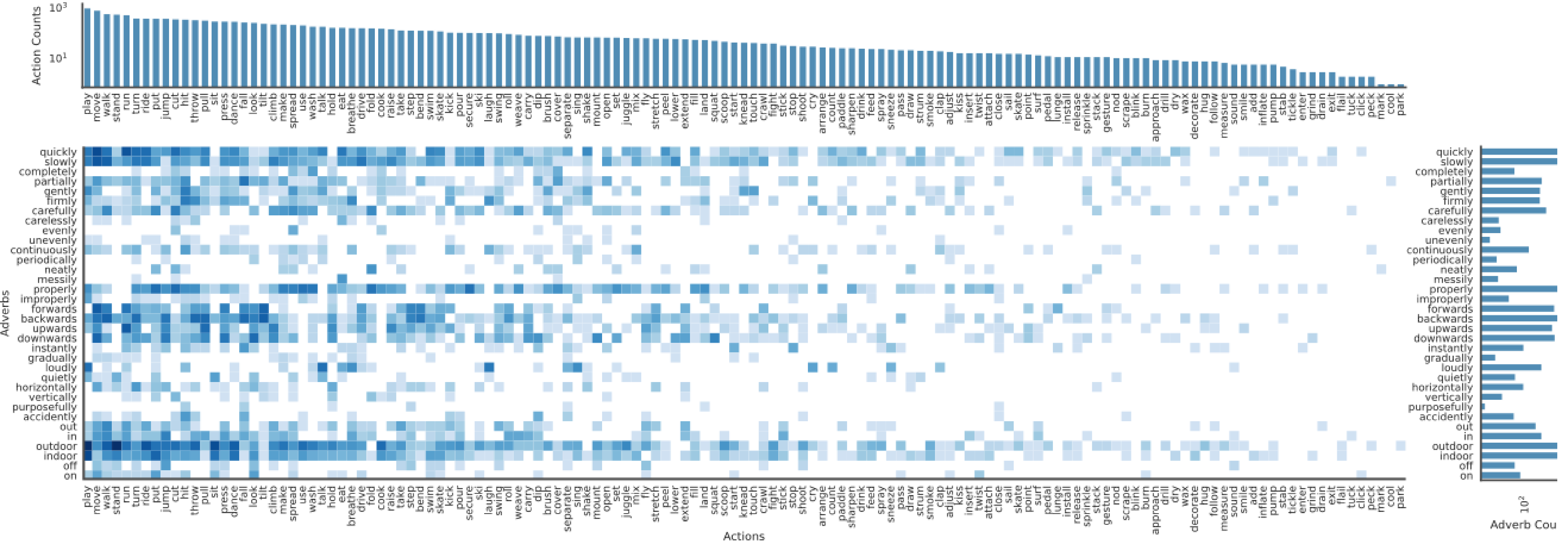
University of Amsterdam

More info: https://hazeldoughty.github.io/Papers/PseudoAdverbs/

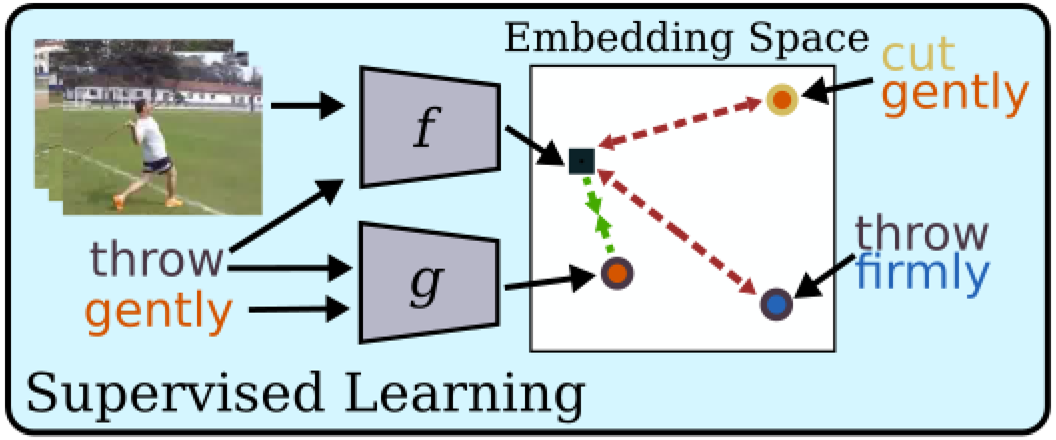
Idea


How is the action being performed?
with adverb labels
swim slowly
slowly
without adverb labels
fold
slowly ?
firmly ?
vertically ?
pseudo-labelling

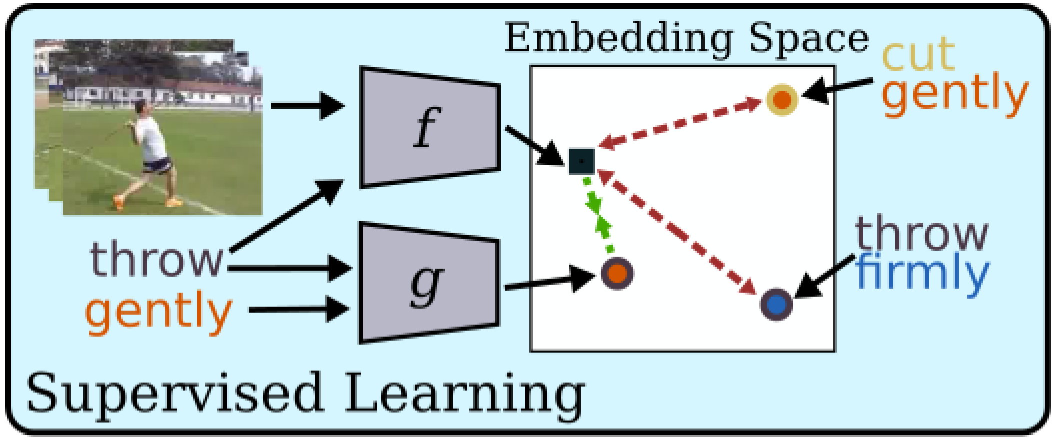
Adverb Datasets

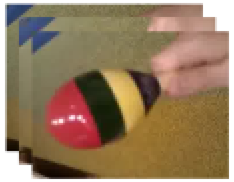


Semi Supervised Learning of Adverbs



Semi Supervised Learning of Adverbs

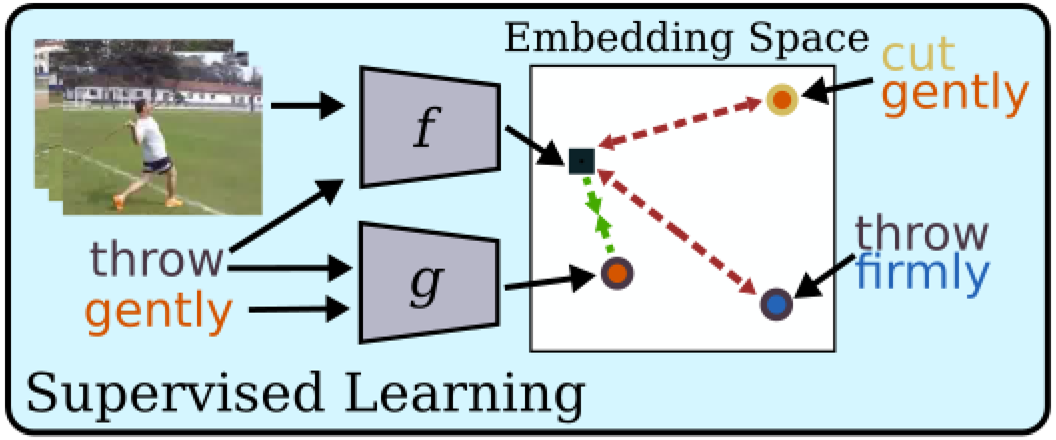


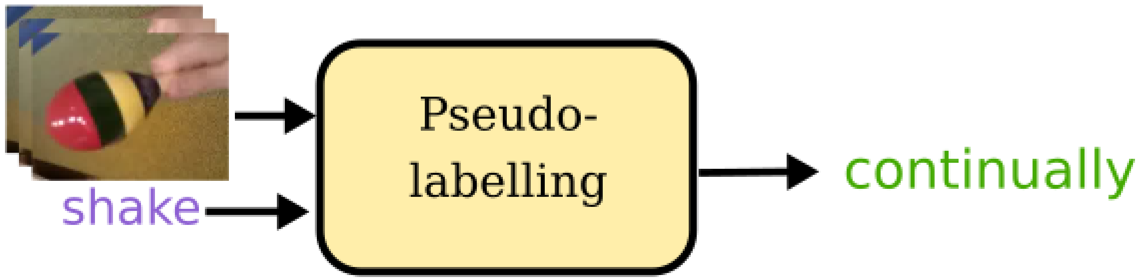


shake

Action-Only Labels

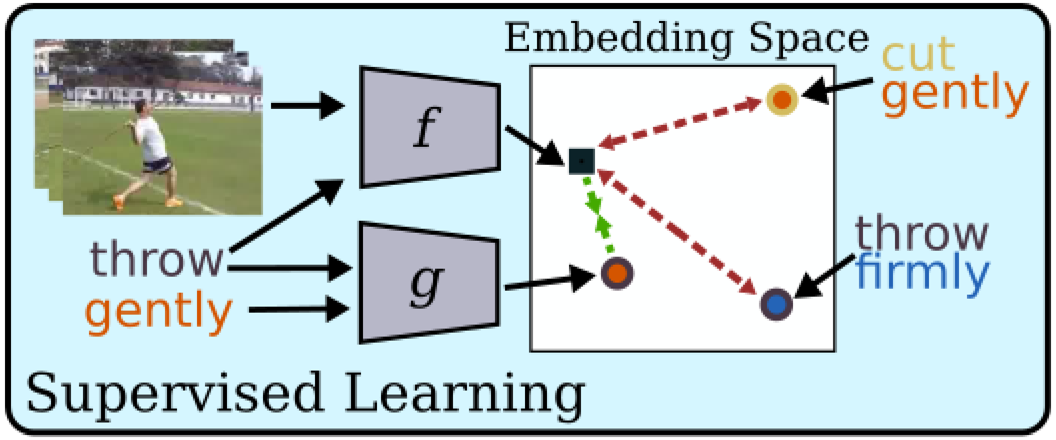
Semi Supervised Learning of Adverbs

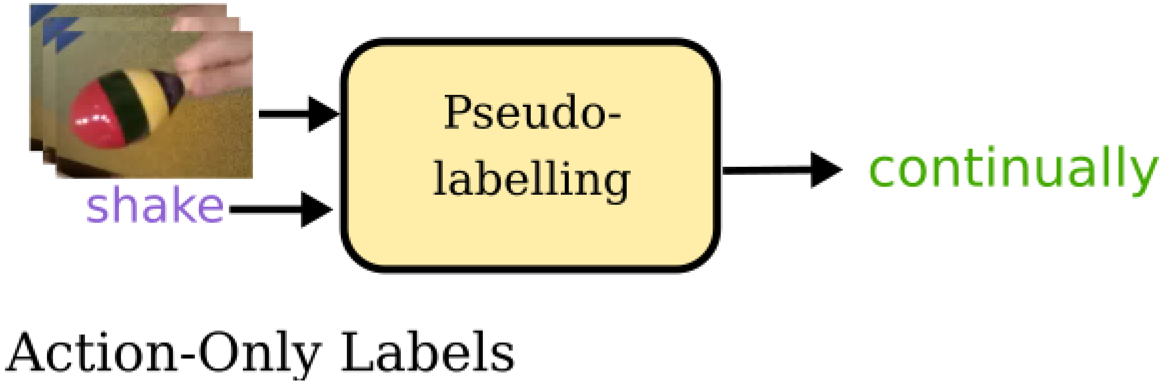


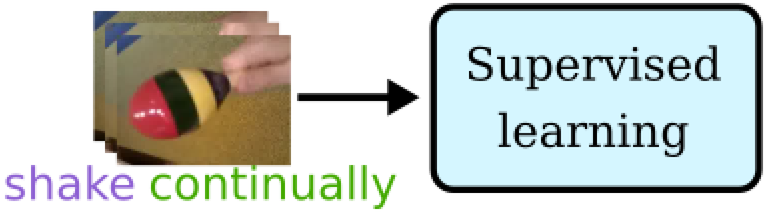


Action-Only Labels

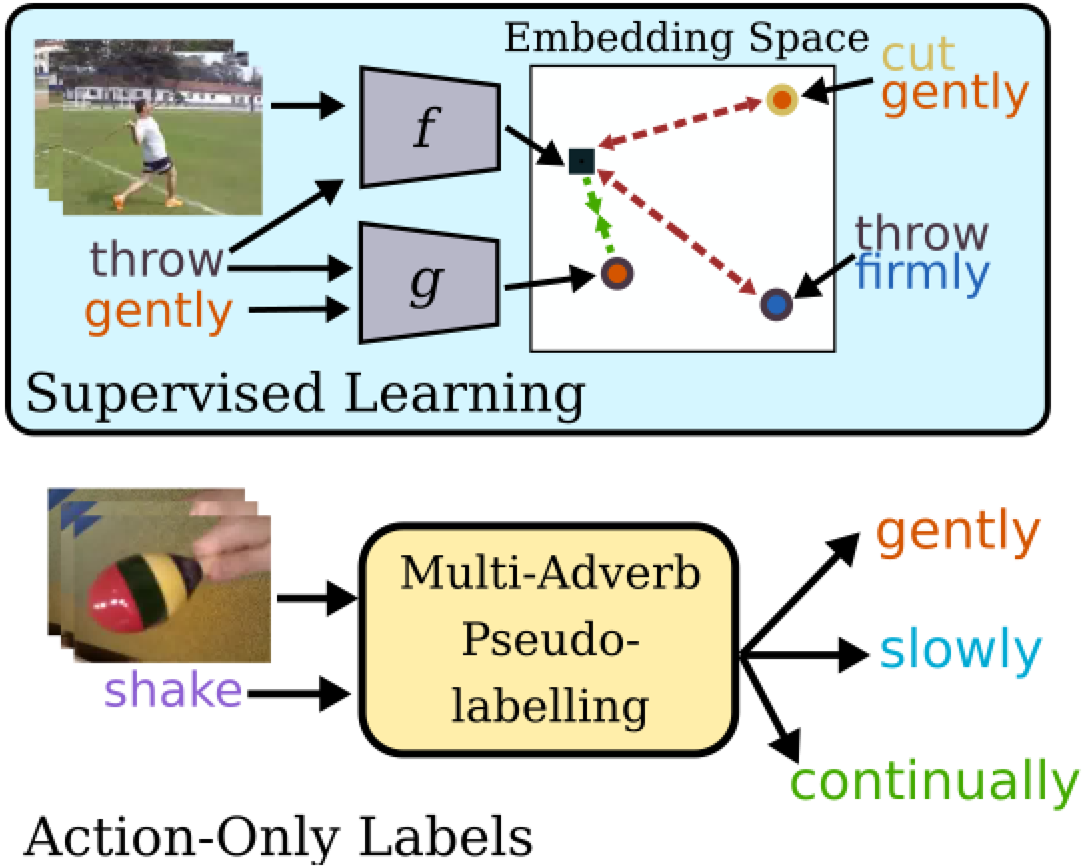
Semi Supervised Learning of Adverbs



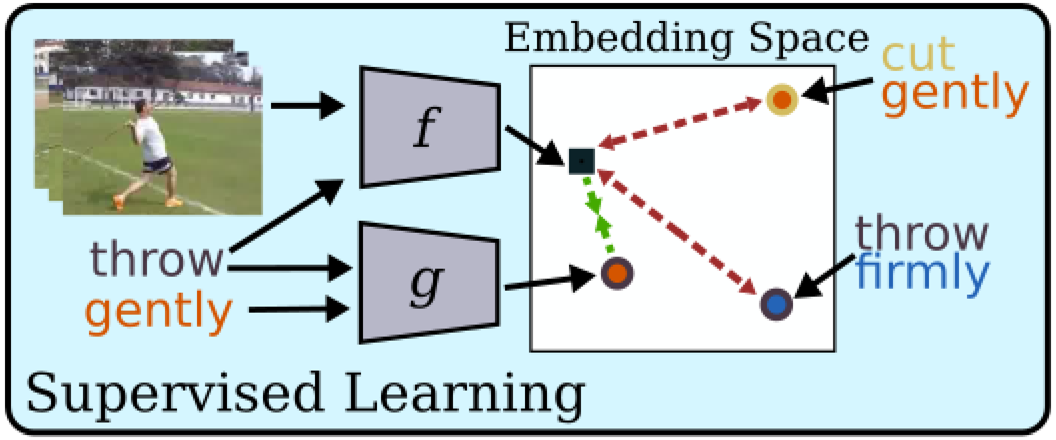


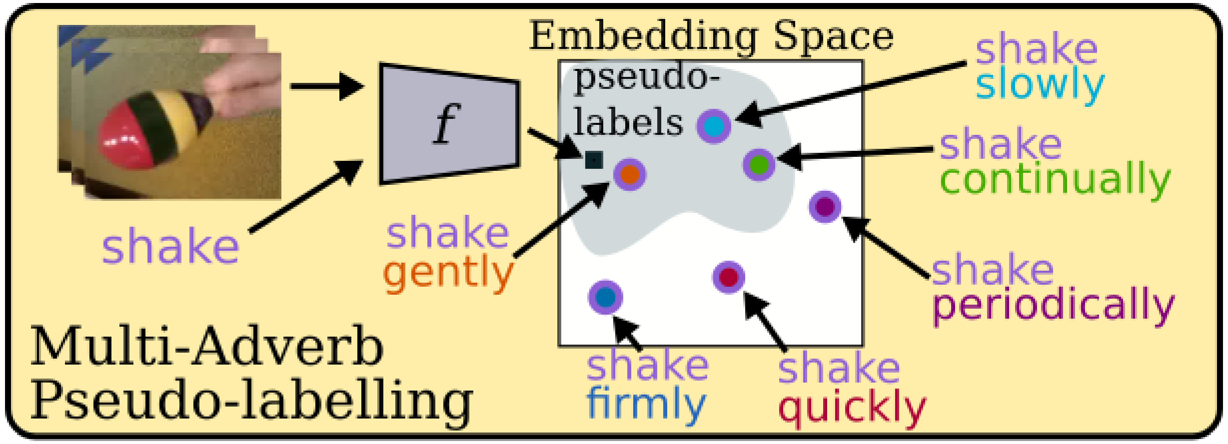


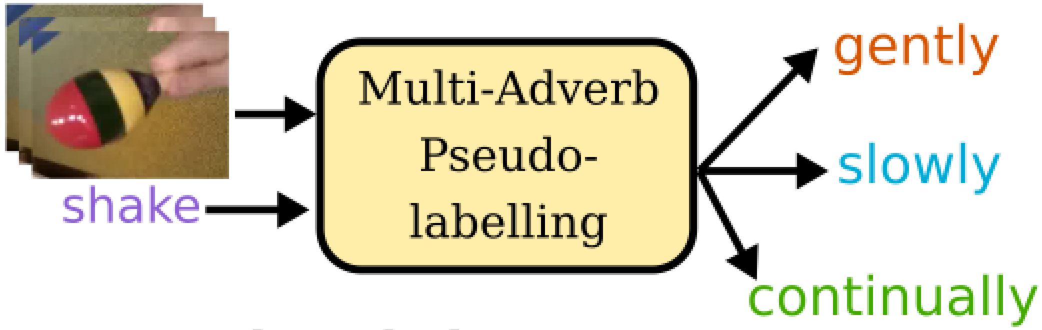
Semi Supervised Learning of Adverbs



Semi Supervised Learning of Adverbs

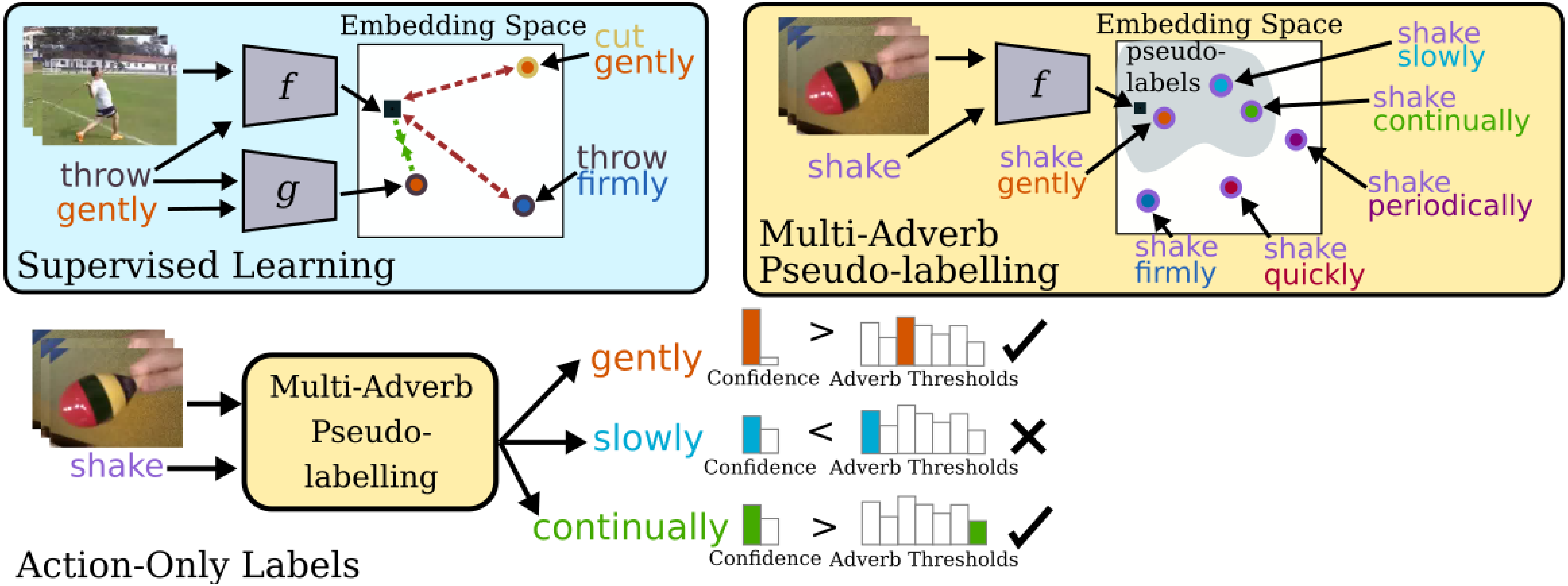




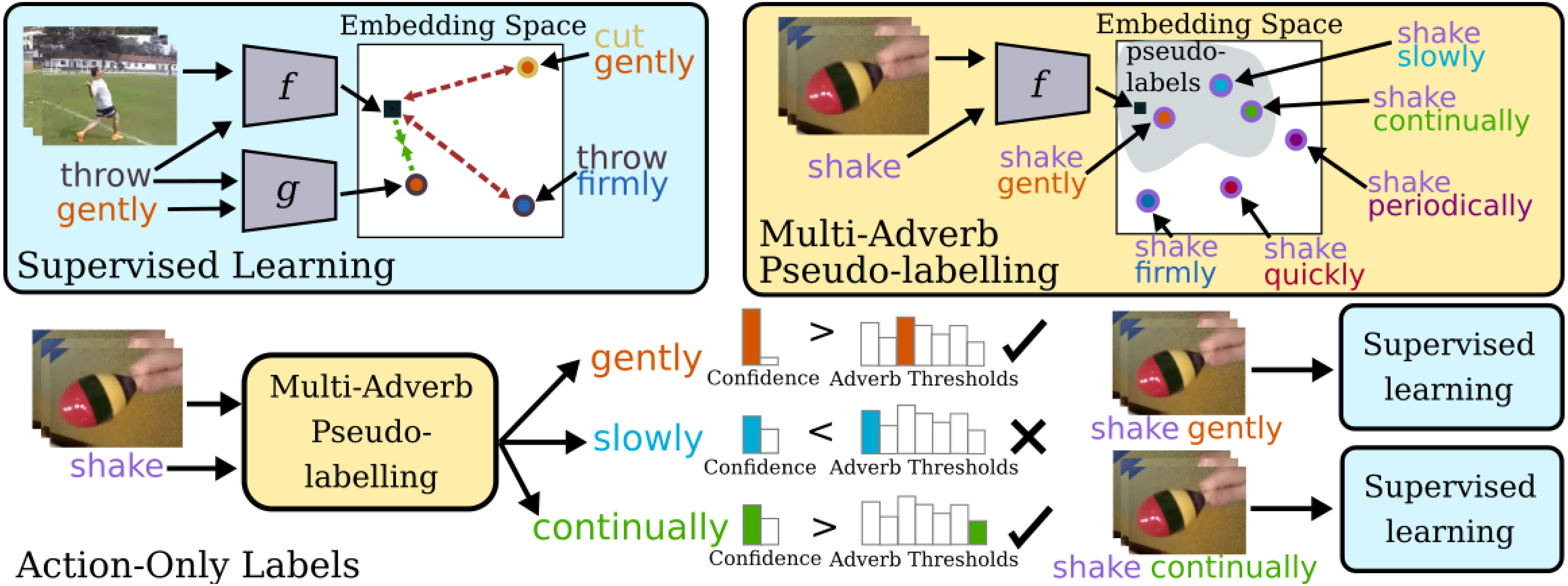


Action-Only Labels

Semi Supervised Learning of Adverbs



Semi Supervised Learning of Adverbs



Results – Unseen Compositions

Method	Accuracy
Supervised only	52.2
Ours	56.1
Training with full labels	65.1

Table 4. **Unseen compositions** in VATEX Adverbs. Our method improves generalization to unseen action-adverb compositions.

Results – New Domains

Train

fold gently

swim slowly

Test

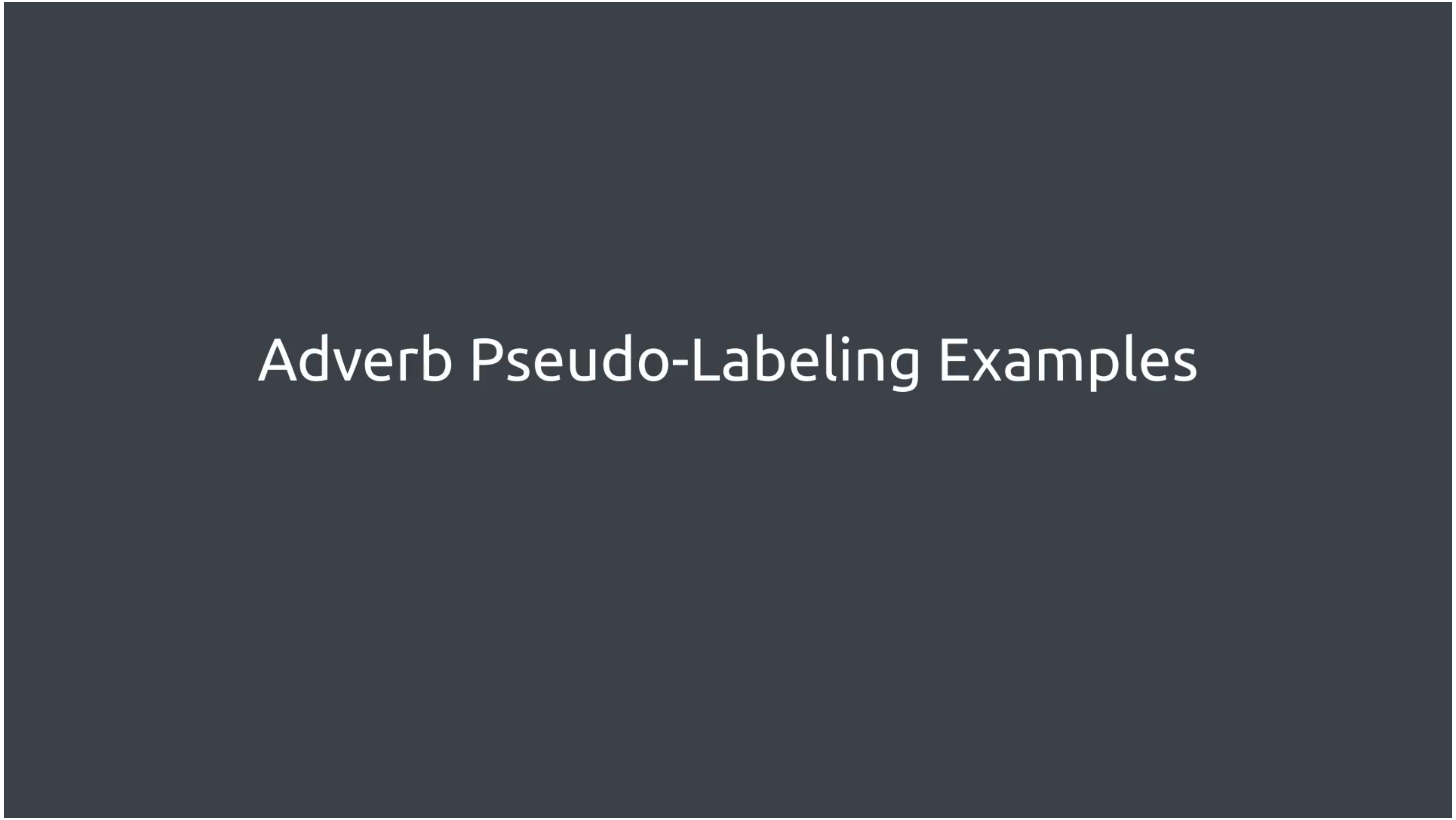
fold gently

swim slowly

Method	MSR-VTT Adverbs	ActivityNet Adverbs
Source only	62.9	67.2
Pseudo-Label	63.9	66.4
Ours	65.0	66.6
Source + Target	67.5	71.6
Target only	70.5	71.8

Table 5. Transfer to **unseen domains** from VATEX-Adverbs. Our method aids generalization to similar domains (MSR-VTT Adverbs), but struggles with larger shifts (ActivityNet Adverbs).

Video: https://hazeldoughty.github.io/Papers/PseudoAdverbs/



Conclusions

- Using multi-adverb pseudo-labelling allows us to use action labelled videos
- We can successfully learn adverbs in a long-tailed distribution
- Open challenges:
    - *Recognizing unseen action-adverb combinations*
    - *Infeasible combinations*
    - *Generalization from few contexts*
    - *Utilizing adverbs for other tasks*

How SEVERE is Benchmark Sensitivity in Video Self-Supervised Learning?


Fida Mohammad Thoker


Hazel Doughty


Piyush Bagad


Cees Snoek

University of Amsterdam

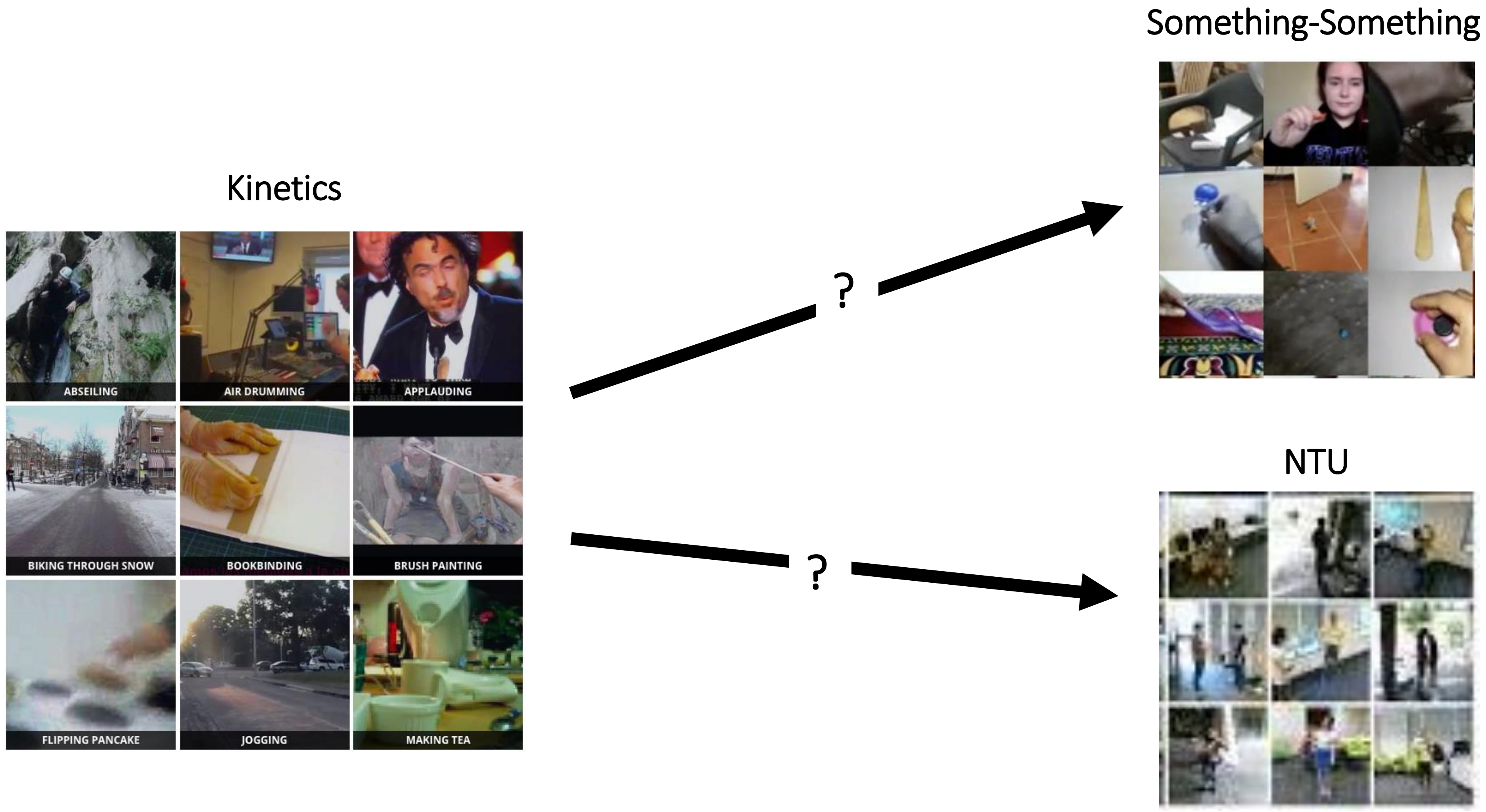
Current Evaluation

Kinetics

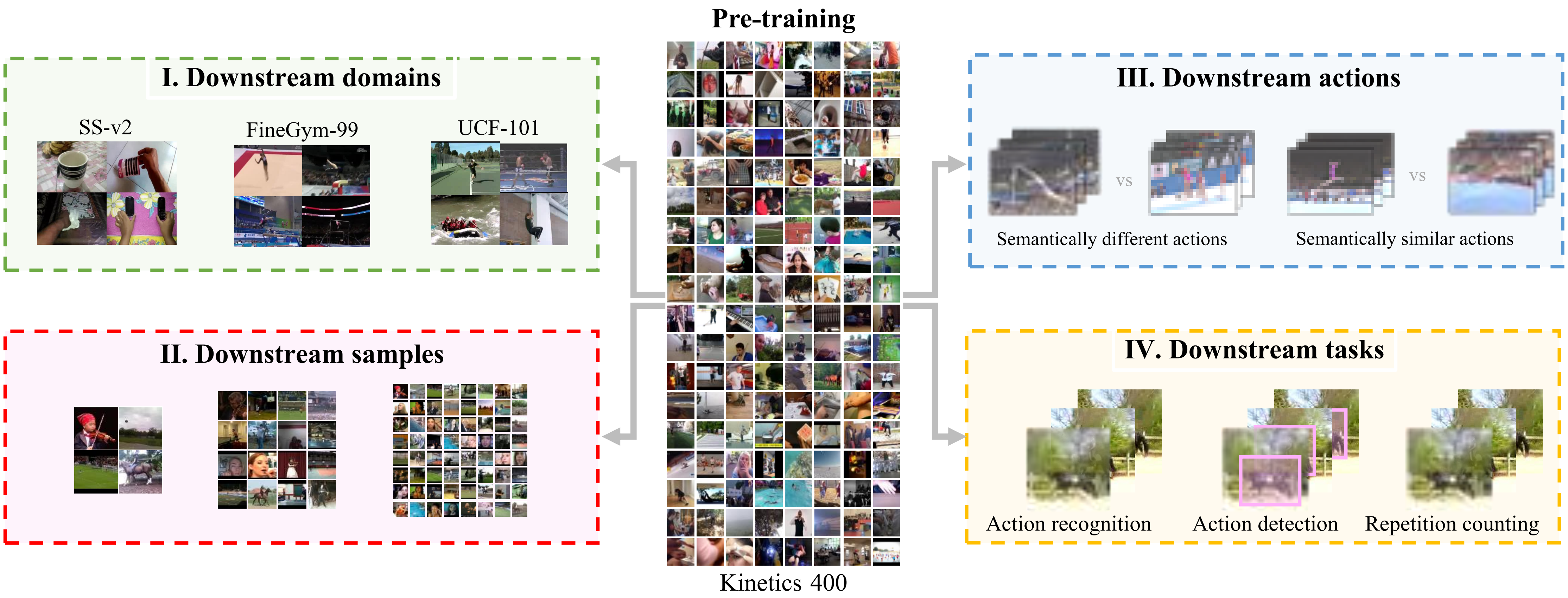
UCF-101

HMDB

Current Evaluation

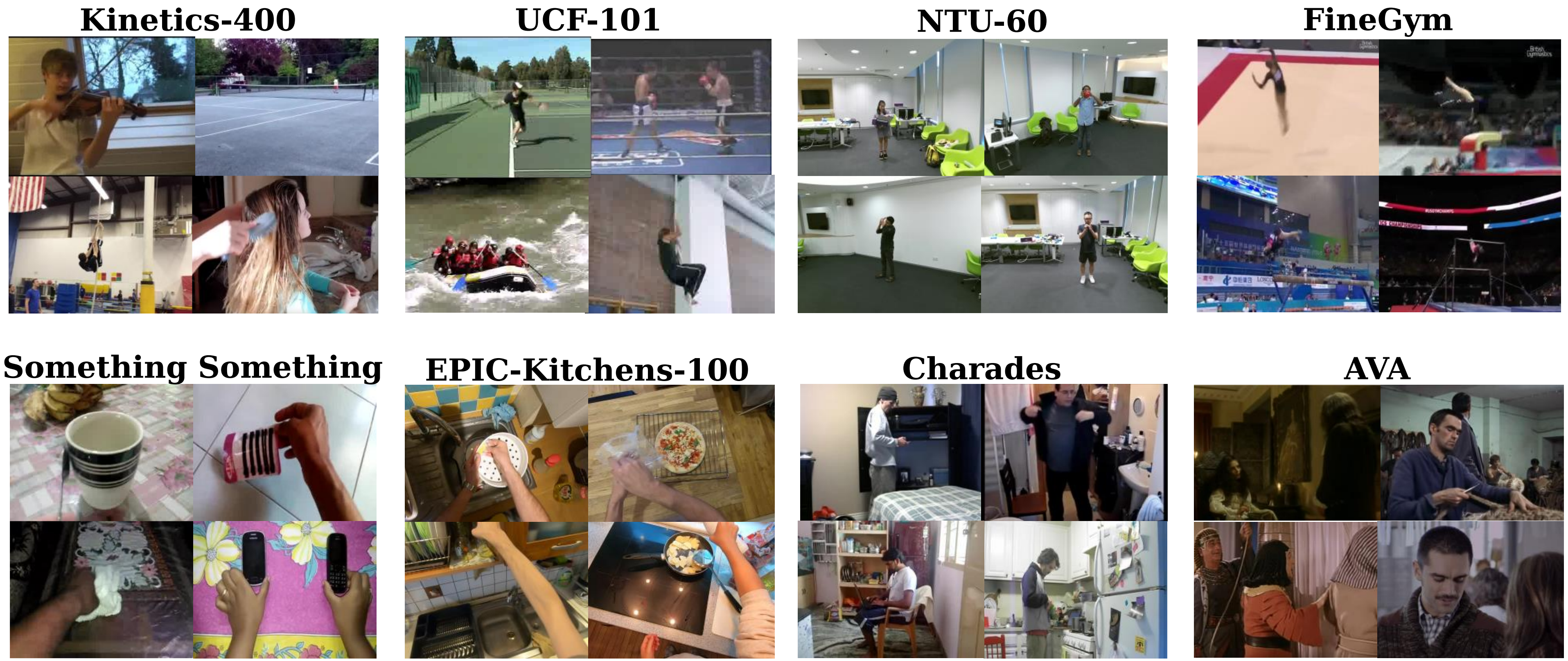


Factors We Investigate



Factors We Investigate

Downstream Datasets



Downstream Domain

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

More dissimilar to source

Downstream Domain

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

More dissimilar to source

Downstream Domain

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

More dissimilar to source

Downstream Domain

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

More dissimilar to source

Downstream Domain

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

More dissimilar to source

Downstream Domain

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

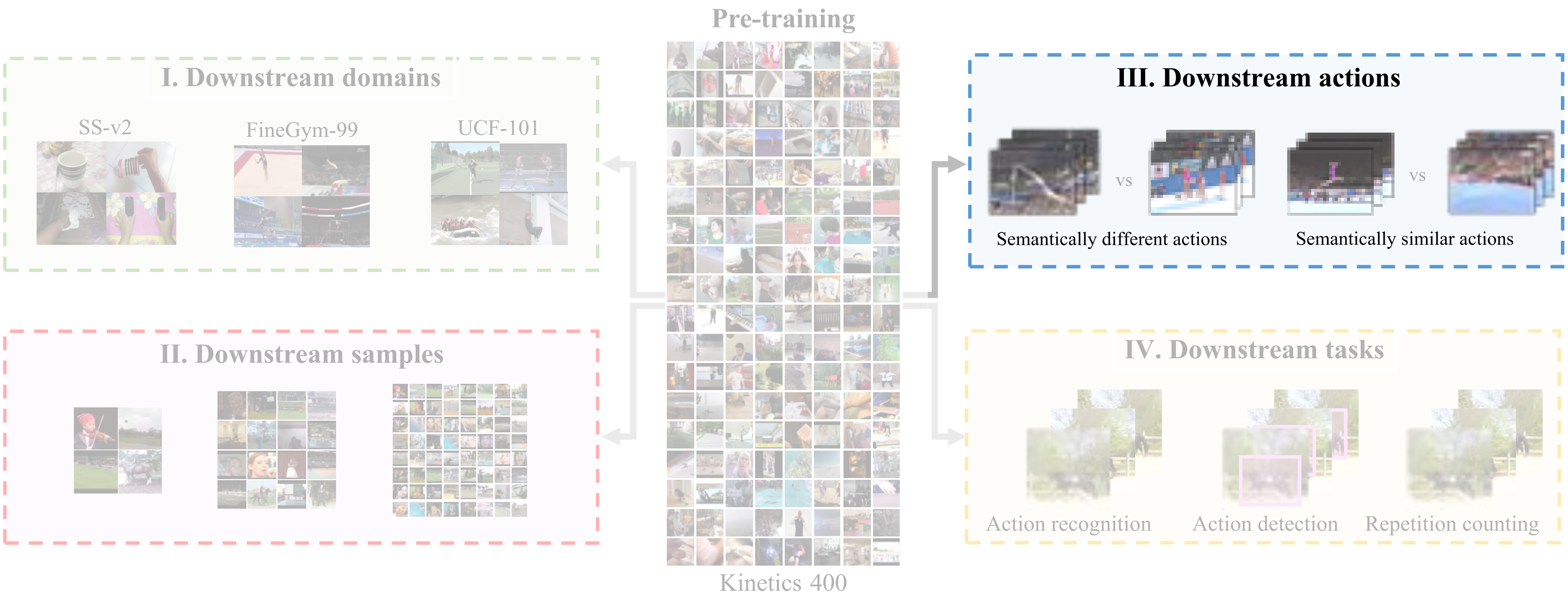
More dissimilar to source

Downstream Domain

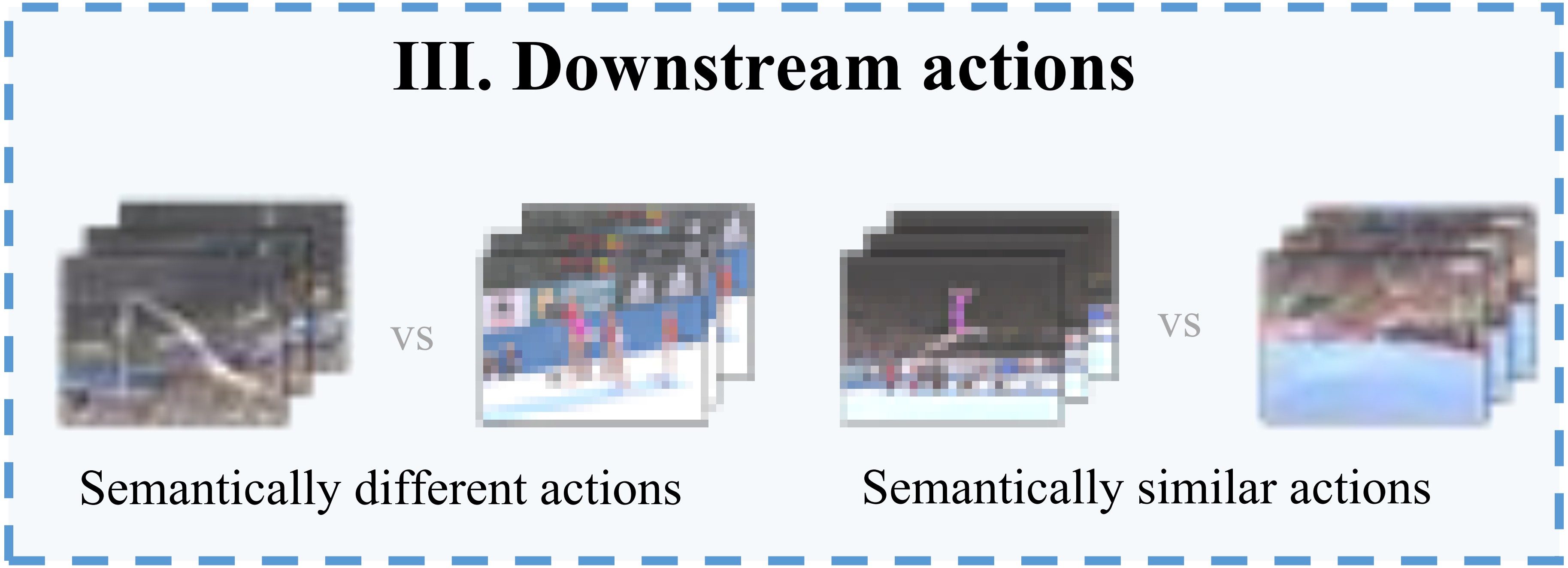
Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

More dissimilar to source

Factors We Investigate



Downstream Actions



Downstream Actions

Pre-training	Gym99				
	Across Events	Within Event		Within Set	
	All	Vault	Floor	FX-S1	UB-S1
None	84.4	24.7	75.9	45.0	84.0
SeLaVi	84.8	25.4	76.0	50.2	81.5
Pretext-contrast	85.7	28.5	81.4	65.8	86.2
AVID-CMA	85.8	30.4	82.7	67.2	88.4
MoCo	86.2	33.2	83.3	65.1	85.0
VideoMoCo	86.4	28.4	79.5	60.4	82.1
GDT	86.5	36.9	83.6	65.7	81.6
RSPNet	87.6	33.4	82.7	63.5	85.1
TCLR	88.0	29.8	84.3	61.0	85.3
CtP	88.3	26.8	86.2	79.7	88.4
Supervised	88.0	37.7	86.1	81.0	86.9

More fine-grained

Downstream Actions

Pre-training	**Gym99**				
	Across Events	Within Event		Within Set	
	All	Vault	Floor	FX-S1	UB-S1
None	84.4	24.7	75.9	45.0	84.0
SeLaVi	84.8	25.4	76.0	50.2	81.5
Pretext-contrast	85.7	28.5	81.4	65.8	86.2
AVID-CMA	85.8	30.4	82.7	67.2	88.4
MoCo	86.2	33.2	83.3	65.1	85.0
VideoMoCo	86.4	28.4	79.5	60.4	82.1
GDT	86.5	36.9	83.6	65.7	81.6
RSPNet	87.6	33.4	82.7	63.5	85.1
TCLR	88.0	29.8	84.3	61.0	85.3
CtP	88.3	26.8	86.2	79.7	88.4
Supervised	88.0	37.7	86.1	81.0	86.9

More fine-grained

Downstream Actions

Pre-training	**Gym99**				
	Across Events	Within Event		Within Set	
	All	Vault	Floor	FX-S1	UB-S1
None	84.4	24.7	75.9	45.0	84.0
SeLaVi	84.8	25.4	76.0	50.2	81.5
Pretext-contrast	85.7	28.5	81.4	65.8	86.2
AVID-CMA	85.8	30.4	82.7	67.2	88.4
MoCo	86.2	33.2	83.3	65.1	85.0
VideoMoCo	86.4	28.4	79.5	60.4	82.1
GDT	86.5	36.9	83.6	65.7	81.6
RSPNet	87.6	33.4	82.7	63.5	85.1
TCLR	88.0	29.8	84.3	61.0	85.3
CtP	88.3	26.8	86.2	79.7	88.4
Supervised	88.0	37.7	86.1	81.0	86.9

More fine-grained

Downstream Actions

Pre-training	Gym99				
	Across Events	Within Event		Within Set	
	All	Vault	Floor	FX-S1	UB-S1
None	84.4	24.7	75.9	45.0	84.0
SeLaVi	84.8	25.4	76.0	50.2	81.5
Pretext-contrast	85.7	28.5	81.4	65.8	86.2
AVID-CMA	85.8	30.4	82.7	67.2	88.4
MoCo	86.2	33.2	83.3	65.1	85.0
VideoMoCo	86.4	28.4	79.5	60.4	82.1
GDT	86.5	36.9	83.6	65.7	81.6
RSPNet	87.6	33.4	82.7	63.5	85.1
TCLR	88.0	29.8	84.3	61.0	85.3
CtP	88.3	26.8	86.2	79.7	88.4
Supervised	88.0	37.7	86.1	81.0	86.9

More fine-grained

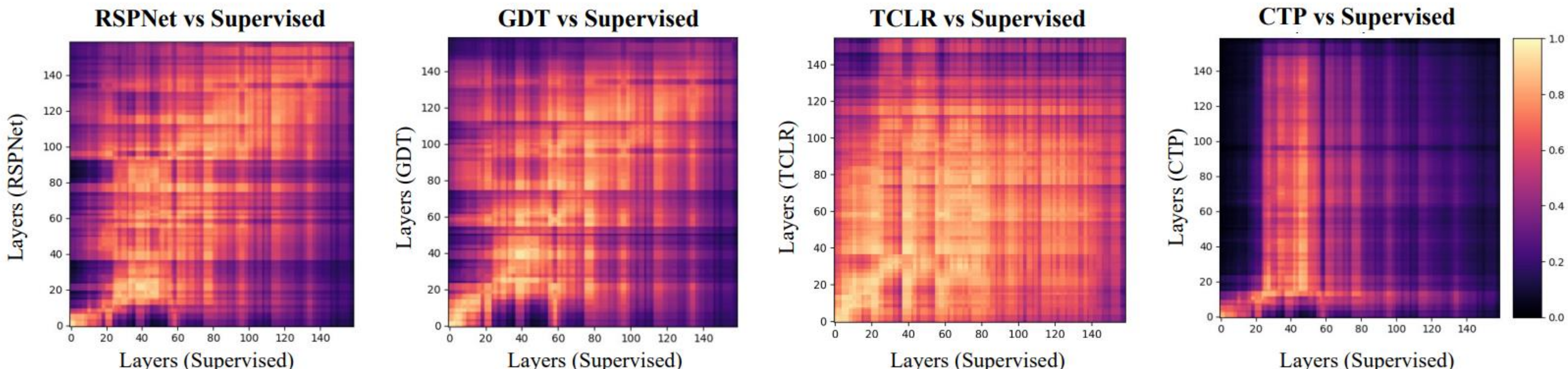
Downstream Actions

Pre-training	Gym99				
	Across Events	Within Event		Within Set	
	All	Vault	Floor	FX-S1	UB-S1
None	84.4	24.7	75.9	45.0	84.0
SeLaVi	84.8	25.4	76.0	50.2	81.5
Pretext-contrast	85.7	28.5	81.4	65.8	86.2
AVID-CMA	85.8	30.4	82.7	67.2	88.4
MoCo	86.2	33.2	83.3	65.1	85.0
VideoMoCo	86.4	28.4	79.5	60.4	82.1
GDT	86.5	36.9	83.6	65.7	81.6
RSPNet	87.6	33.4	82.7	63.5	85.1
TCLR	88.0	29.8	84.3	61.0	85.3
CtP	88.3	26.8	86.2	79.7	88.4
Supervised	88.0	37.7	86.1	81.0	86.9

More fine-grained

Overall Observations

- Different methods are better in different downstream settings
- Supervised pre-training dominates
- Contrasting parts of a video clip increases generalizability
- Too many augmentations can harm generalizability to fine-grained settings
- CtP generalizes well and doesn't use contrastive learning



On Semantic Similarity in Video Retrieval

CVPR 2021

Michael Wray

Hazel Doughty

Dima Damen

University of Bristol

More info: https://mwray.github.io/SSVR/

Video Retrieval

Which of these captions correspond to the following video?

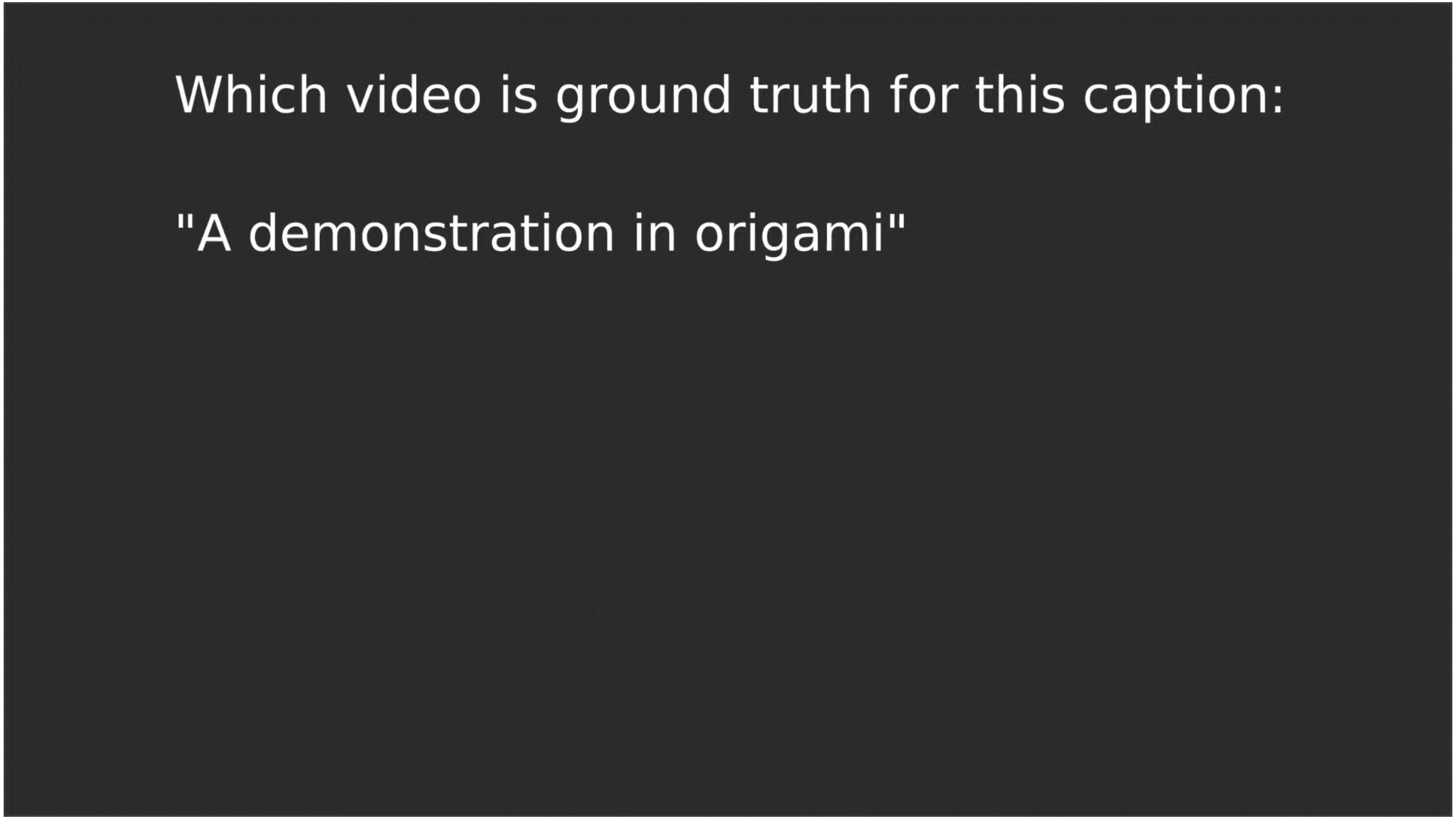
A band is performing for the crowd

A man is peeling fruit.

A girl is sitting in a chair

Add prawns to the pan and mix

Video: https://www.youtube.com/watch?v=AO7zUbxMn6o



Retrieval Assumption

Current methods make the following assumption

“There exists only one corresponding caption for a given video and vice versa”

Peel and chop the potatoes
Peel and cut up the potato
Peel the potatoes and cut them
Peel and cut the potatoes into chunks
Peel the potatoes and cut them into halves

YouCook2

Put fork and spoon in drying rack
Put spoons in drying rack
Put spoon in drying rack
Put bowl in drying rack
Put plate in drying rack

EPIC-KITCHENS

MSR-VTT

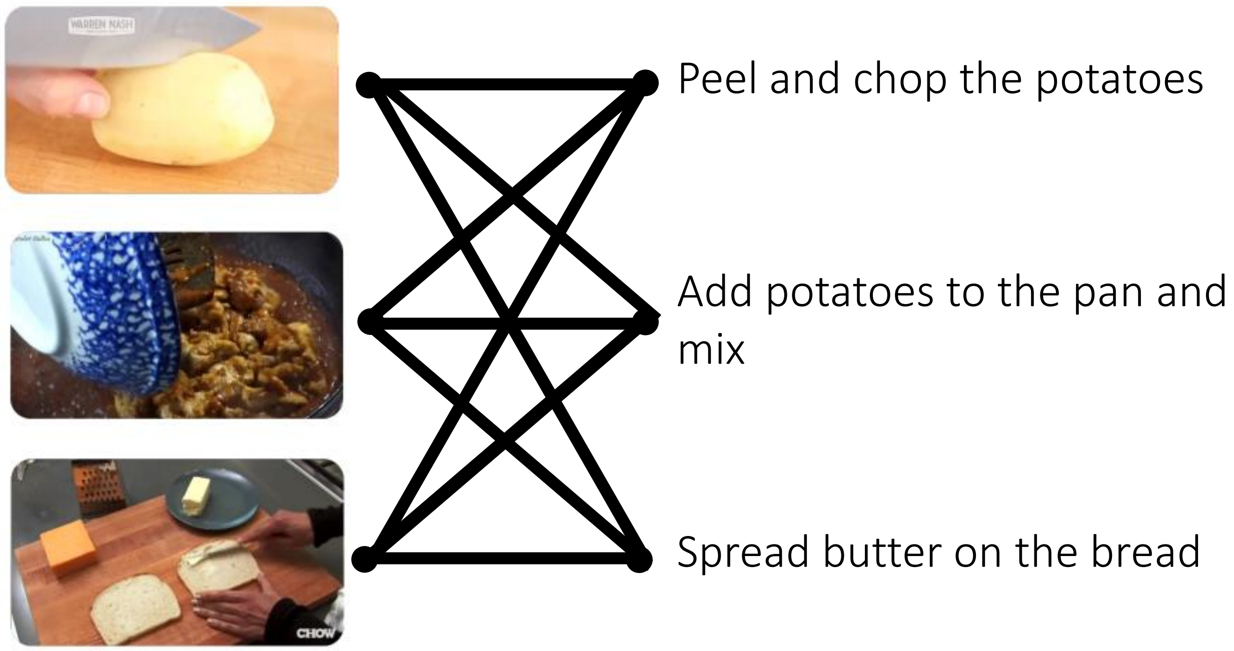
A band is performing for the crowd
A band is performing on a brightly lit stage
A band is playing a show
A band and singers perform
3 guys singing and playing instruments on a stage

Semantic Similarity

Two main goals for semantic similarity:

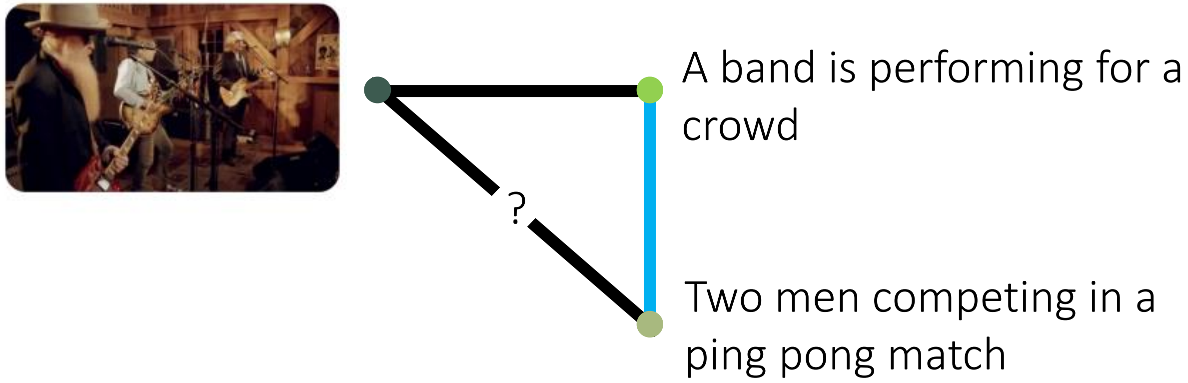
Move from a one-to-one relationship between videos and captions to many-to-many.

Allow for differing levels of similarity



Proxy Measures

Want to relate two items semantically.



Assume that a caption sufficiently describes a video.

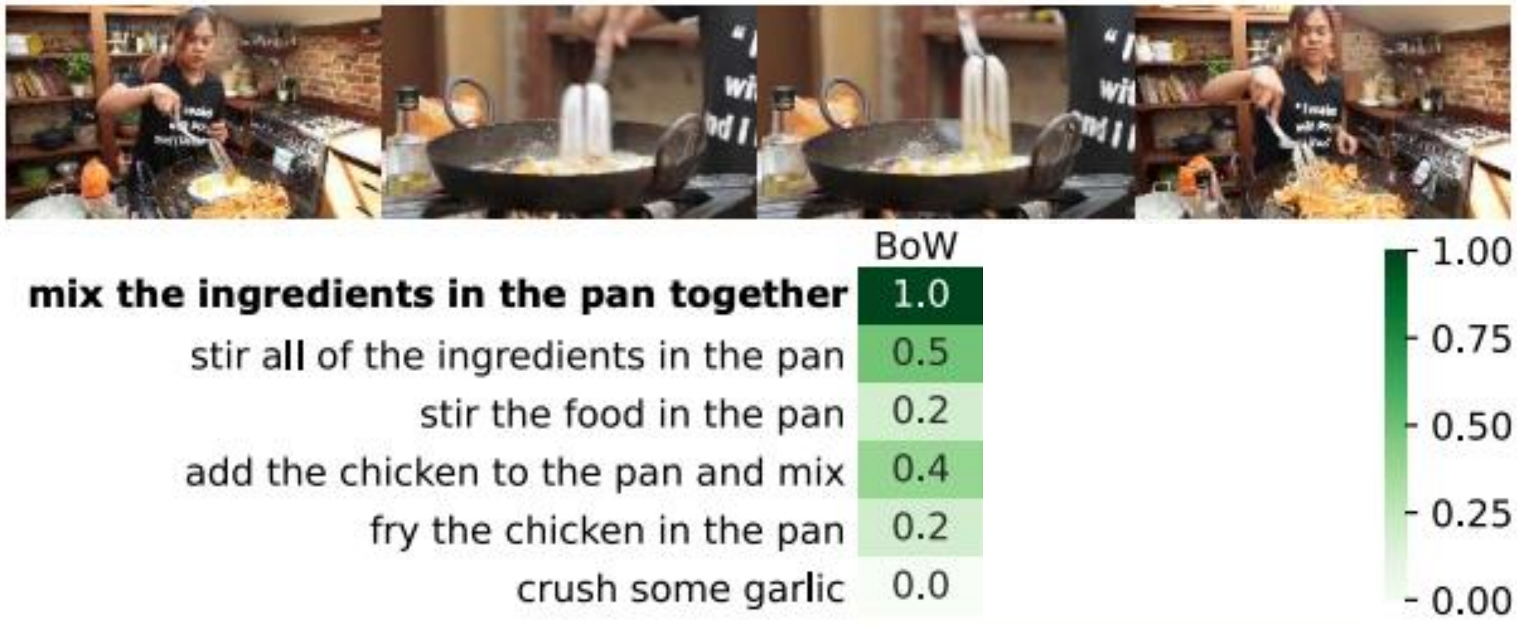
Define a proxy function that relates captions

$$S(x_i, y_j) = S'(y_i, y_j)$$

Example Proxy Measures

We introduce three other metrics based on:

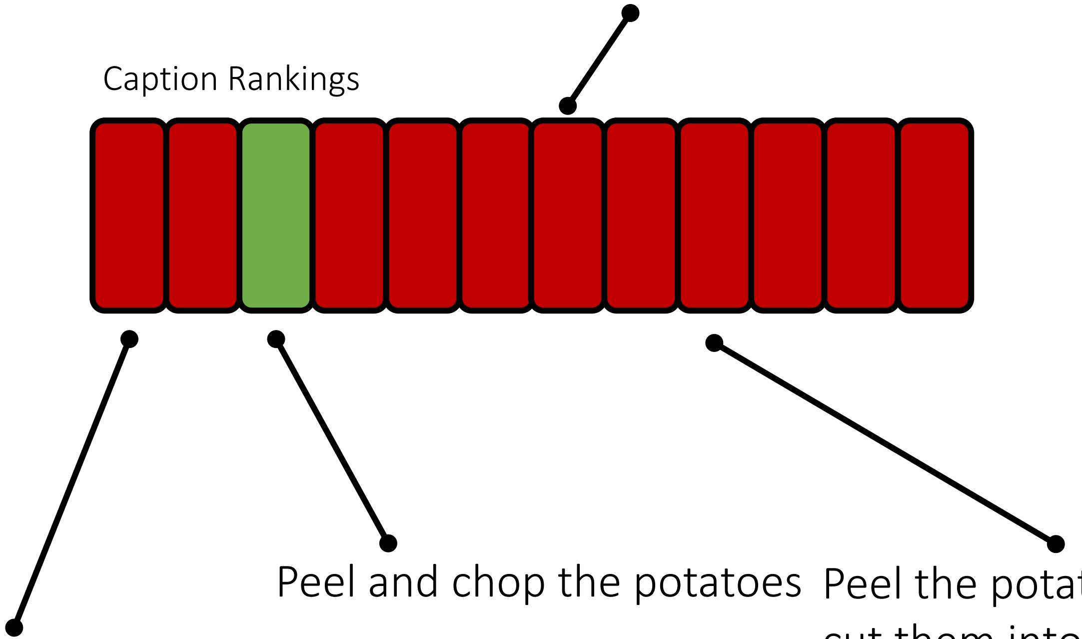
- Parts of Speech
- Synsets
- METEOR



Problems with Instance Retrieval

When evaluating with a single caption, the correct caption can be arbitrary.

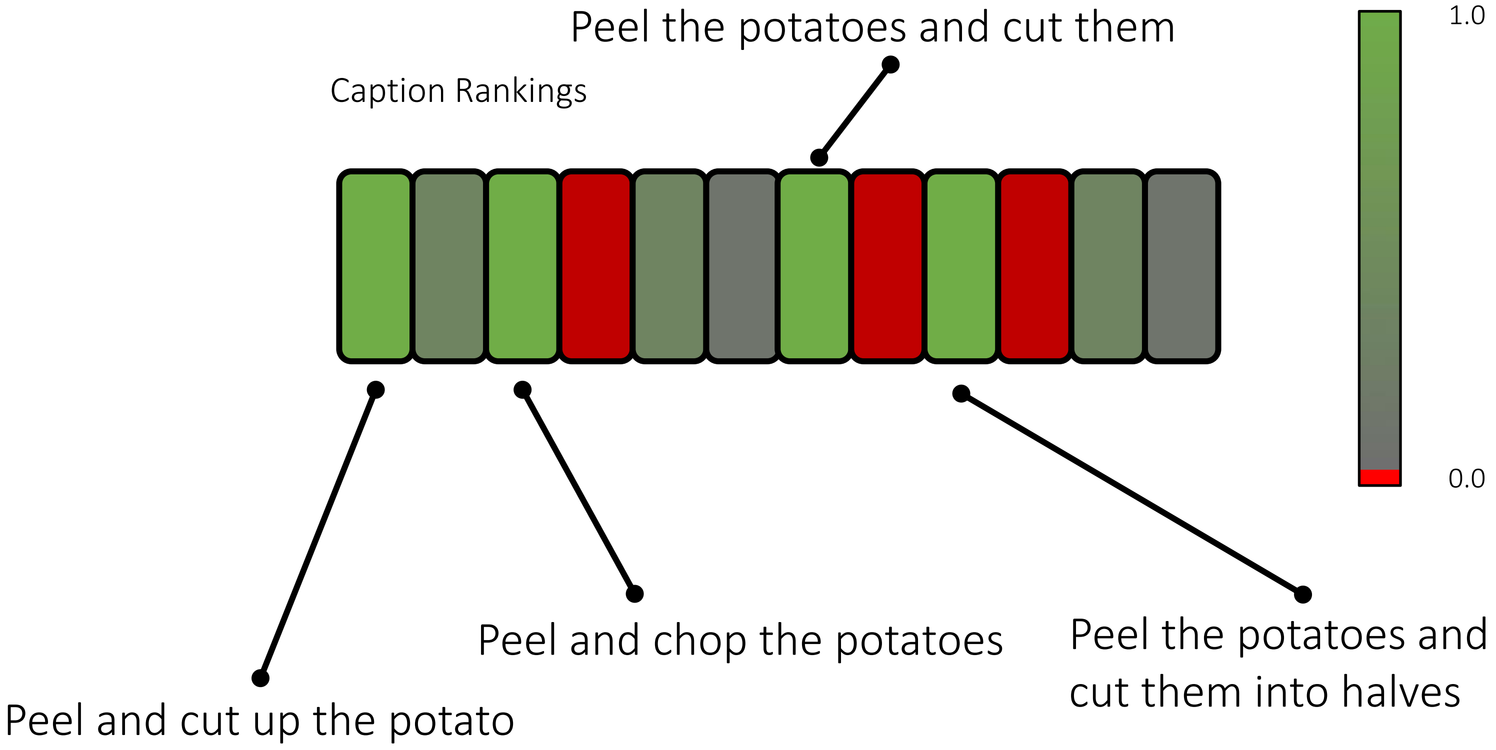
Query Video



Evaluating Semantic Retrieval

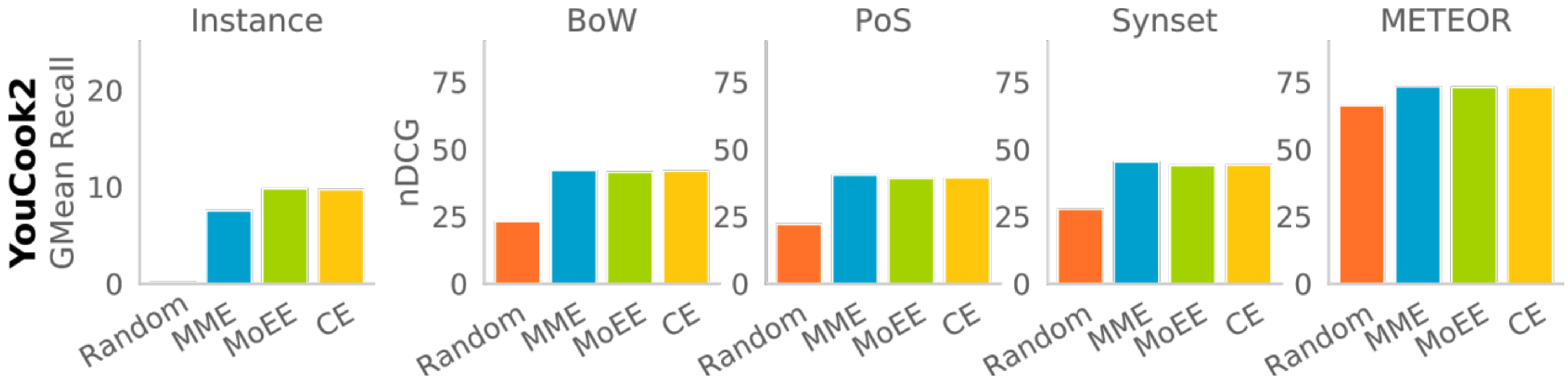
We use normalised Discounted Cumulative Gain to evaluate multiple items with differing relevance.

Query Video



Evaluating with Semantic Similarity

Whilst models outperform the MLP baseline (MME) for Instance Video Retrieval, this isn't the case when Semantic Similarity is used.

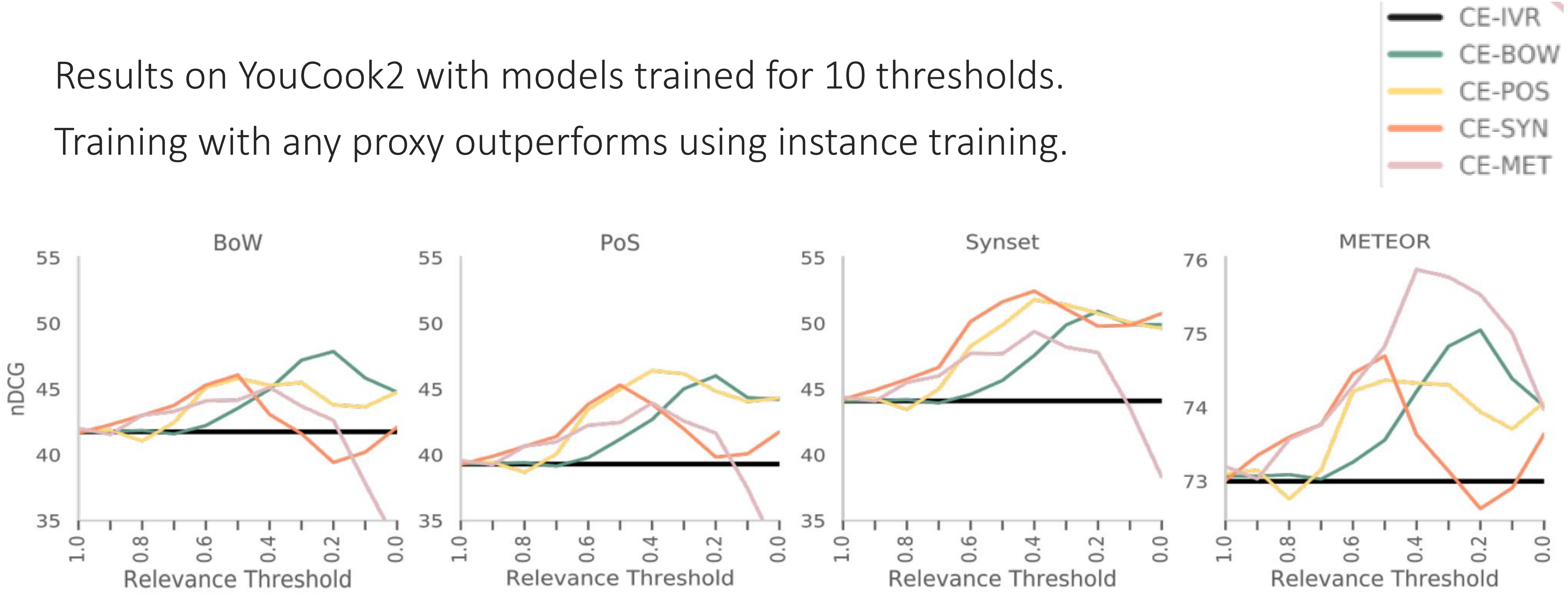


MoEE: Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. CoRR, abs/1804.02516, 2018
CE: Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In BMVC, 2019
JPoSE: Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple partsof-speech embeddings. In ICCV, 2019

Training with Semantic Similarity

Results on YouCook2 with models trained for 10 thresholds.

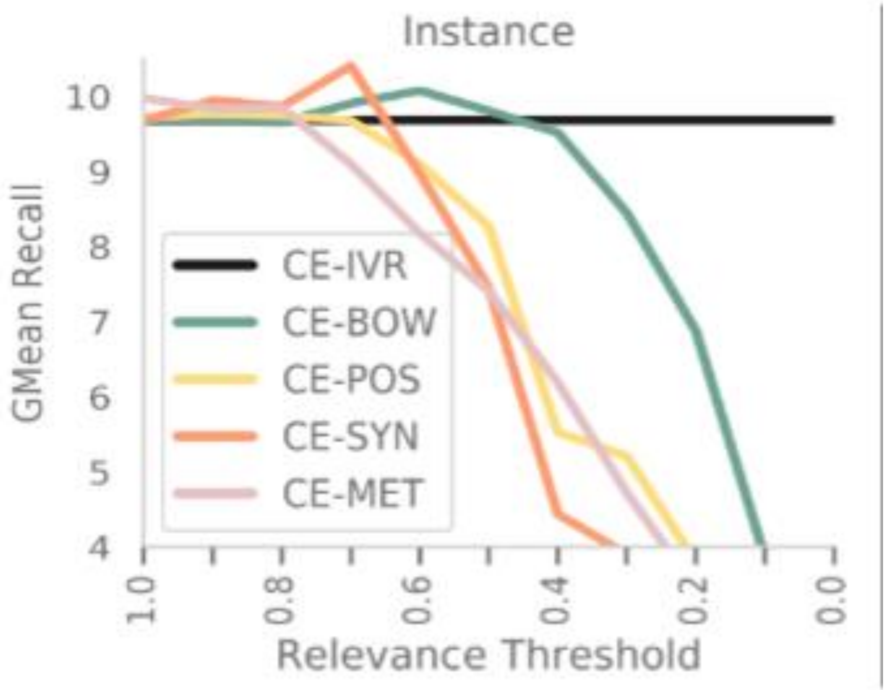
Training with any proxy outperforms using instance training.



Training with Semantic Similarity II

Results on YouCook2 with models trained for 10 thresholds.

Training with any proxy outperforms using instance training.



Conclusions

- There is an issue with the current instance-based metrics in video retrieval
- We propose a new metric which allows many-to-many relevancy and non-binary similarity
- These relevancies can be calculated via our proxies
- Considering multiple relevant captions can improve video retrieval results