Detailed Human Action Understanding from Unlabeled Videos

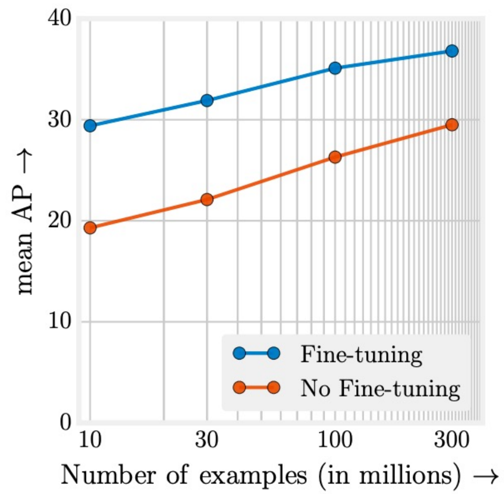
Chen Sun

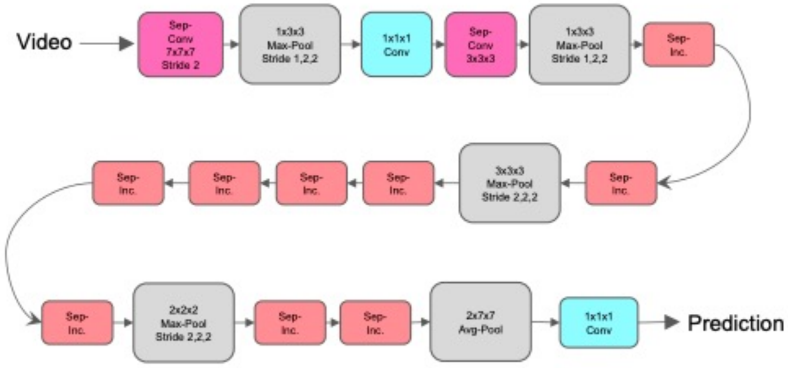
BROWN

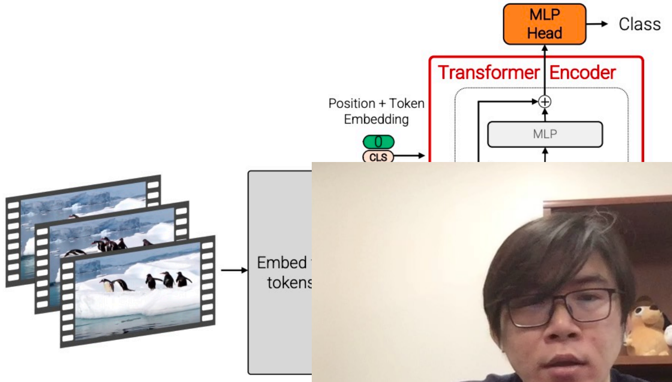
Google Research

My Research at Google: Large-scale Visual Understanding

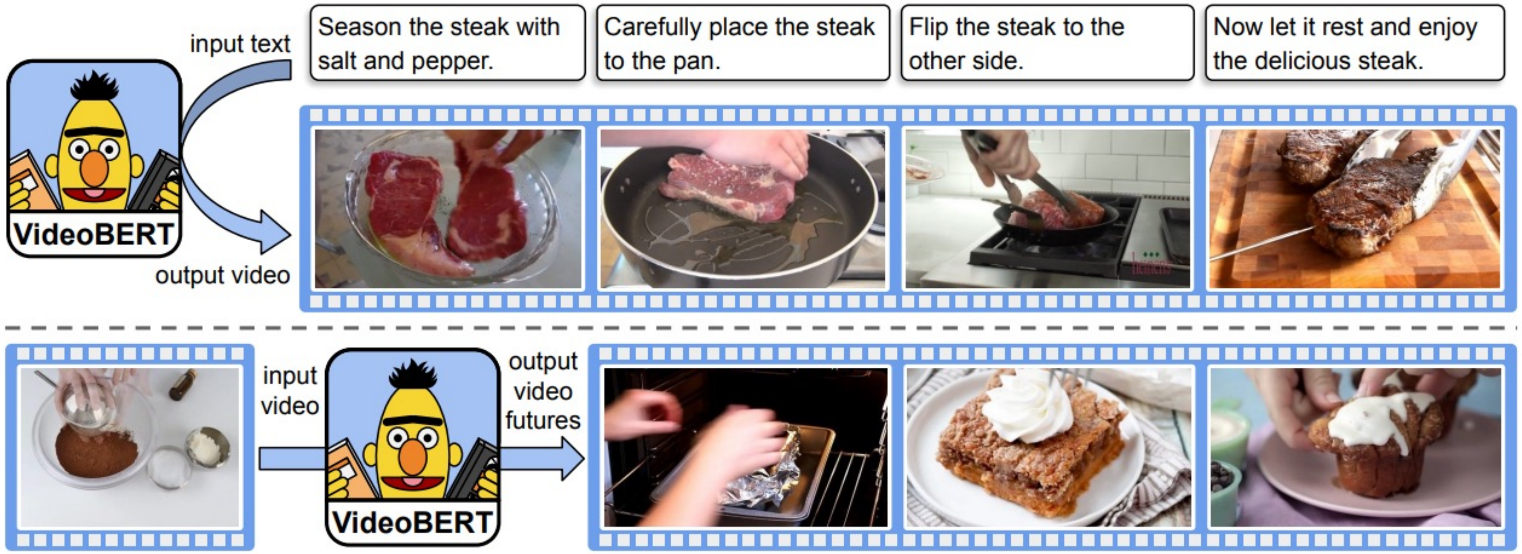
Left: Stand, Watch; Middle: Stand, Play instrument; Right: Sit, Play instrument

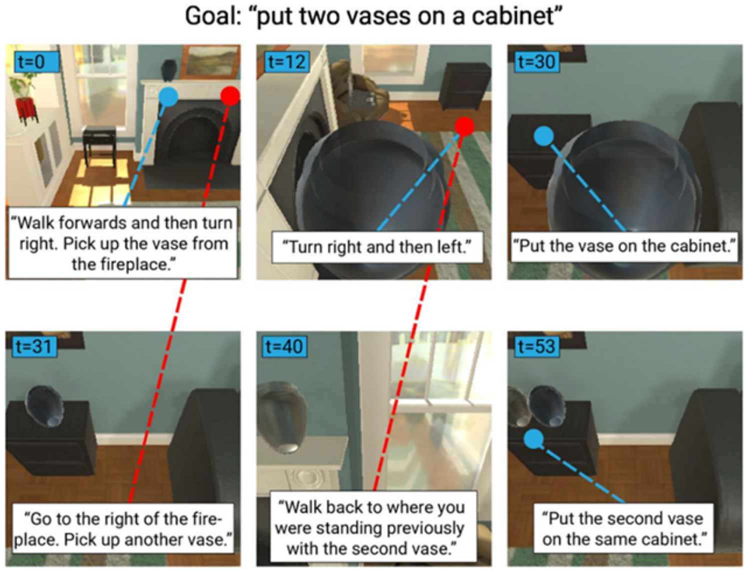


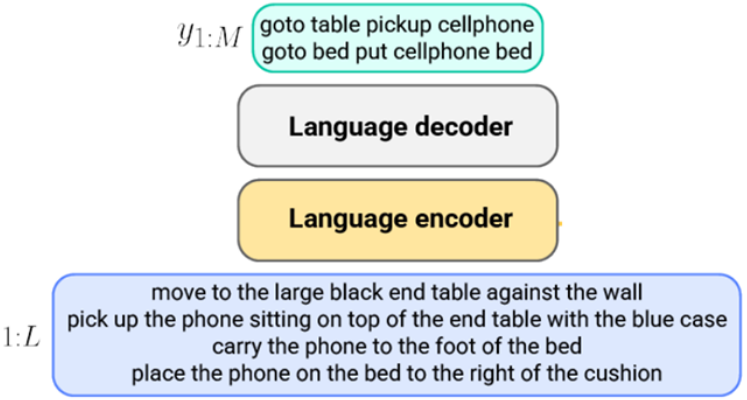




My Research at Brown: Structured Video Understanding

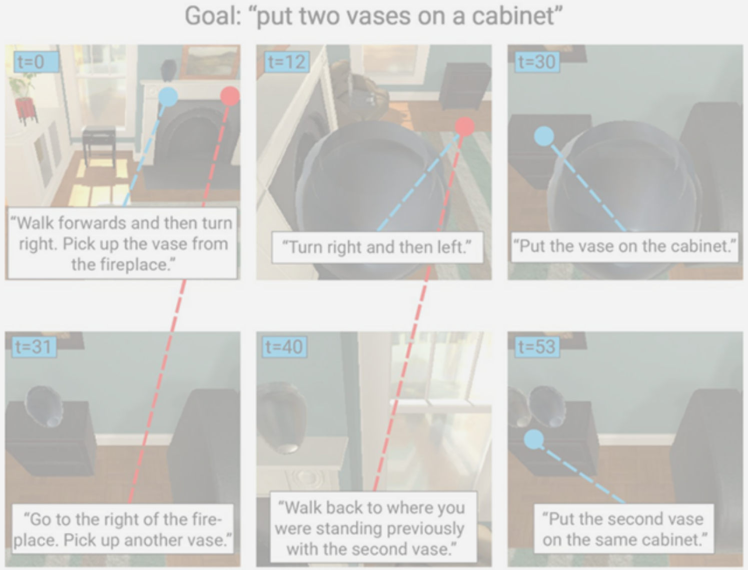






My Research at Brown: Structured Video Understanding

We are hiring PhD students!



What can we learn from videos?

A frame from the Atomic Visual Actions (AVA) dataset

What can we learn from videos?

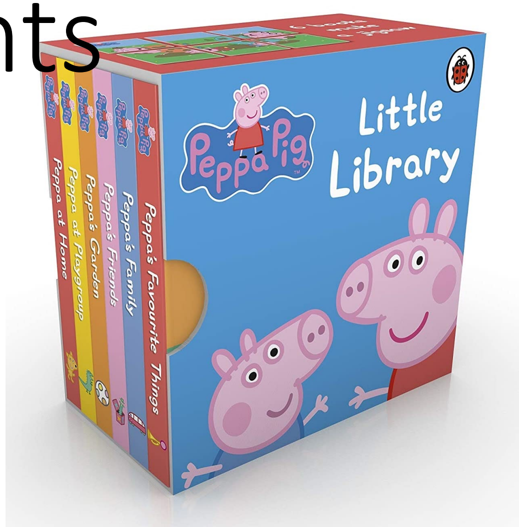
A frame from the Atomic Visual Actions (AVA) dataset

Object detection:
Person, silverware, food
Action detection:
Sit, eat, talk
Human-object interaction:
Person hold fork / eat food
Near-future prediction*:*
Stand

Encyclopedia of Multimedia Contents

Place the ingredients onto a bowl of hot steamed rice.

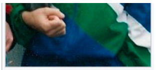
YouTube



Ferguson years (1986–2013)

Main article: History of Manchester United F.C. (1986–2013)

Alex Ferguson managed the team between 1986 and 2013.

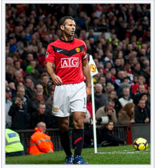


Bryan Robson was the captain of Manchester United for 12 years, longer than any other player.[36]

Alex Ferguson and his assistant Archie Knox arrived from Aberdeen on the day of Atkinson's dismissal,[41] and guided the club to an 11th-place finish in the league.[42] Despite a second-place finish in 1987–88, the club was back in 11th place the following season.[43] Reportedly on the verge of being dismissed, victory over Crystal Palace in the 1990 FA Cup Final replay (after a 3–3 draw) saved Ferguson's career.[44][45] The following season, Manchester United claimed their first UEFA Cup Winners' Cup title. That triumph allowed the club to compete in the European Super Cup for the very first time, where United beat European Cup holders Red Star Belgrade 1–0 in the final at Old Trafford. A second consecutive League Cup final appearance in 1992 saw the club win that competition for the first time as well, following a 1–0 win against Nottingham Forest at Wembley Stadium.[40] In 1993, the club won its first league title since 1967, and a year later, for the first time since 1957, it won a second consecutive title – alongside the FA Cup – to complete the first "Double" in the club's history.[40] United then became the first English club to do the Double twice when they won both competitions again in 1995–96,[46] before retaining the league title once more in 1996–97 with a game to spare.[47]

In the 1998–99 season, Manchester United became the first team to win the Premier League, FA Cup and UEFA Champions League – "The Treble" – in the same season.[48] Losing 1–0 going into injury time in the 1999 UEFA Champions League Final, Teddy Sheringham and Ole Gunnar Solskjær scored late goals to claim a dramatic victory over Bayern Munich, in what is considered one of the greatest comebacks of all time.[49] The club then became the only British team to ever win the Intercontinental Cup after beating Palmeiras 1–0 in Tokyo.[50] Ferguson was subsequently knighted for his services to football.[51]

Front three: Manchester United's treble medals of the



Ryan Giggs is the most decorated player in English football history.[52]

Manchester United won the league again in the 1999–2000 and 2000–01 seasons, becoming only the fourth club to win the English title three times in a row. The team finished third in 2001–02, before regaining the title in 2002–03.[53] They won the 2003–04 FA Cup, beating Millwall 3–0 in the final at the Millennium Stadium in Cardiff to lift the trophy for a record 11th time.[54] In the 2005–06 season, Manchester United failed to qualify for the knockout phase of the UEFA Champions League for the first time in over a decade,[55] but recovered to secure a second-place league finish and victory over Wigan Athletic in the 2006 Football League Cup Final. The club regained the Premier League in the 2006–07 season, before completing the European double in 2007–08 with a 6–5 penalty shoot-out victory over Chelsea in the 2008 UEFA Champions League Final in Moscow to go with their 17th English league title. Ryan Giggs made a record 759th appearance for the club in that game, overtaking previous record holder Bobby Charlton.[56] In December 2008, the club became the first British team to win the FIFA Club World Cup and followed this with the 2008–09 Football League Cup, and its third successive Premier League t forward Cristiano Ronaldo was sold to Real Madrid for a world record £80 million.[59] In 2010, Manchester United defeated Aston Villa 2–1 at Wembley to retain t successful defence of a knockout cup competition.[60]

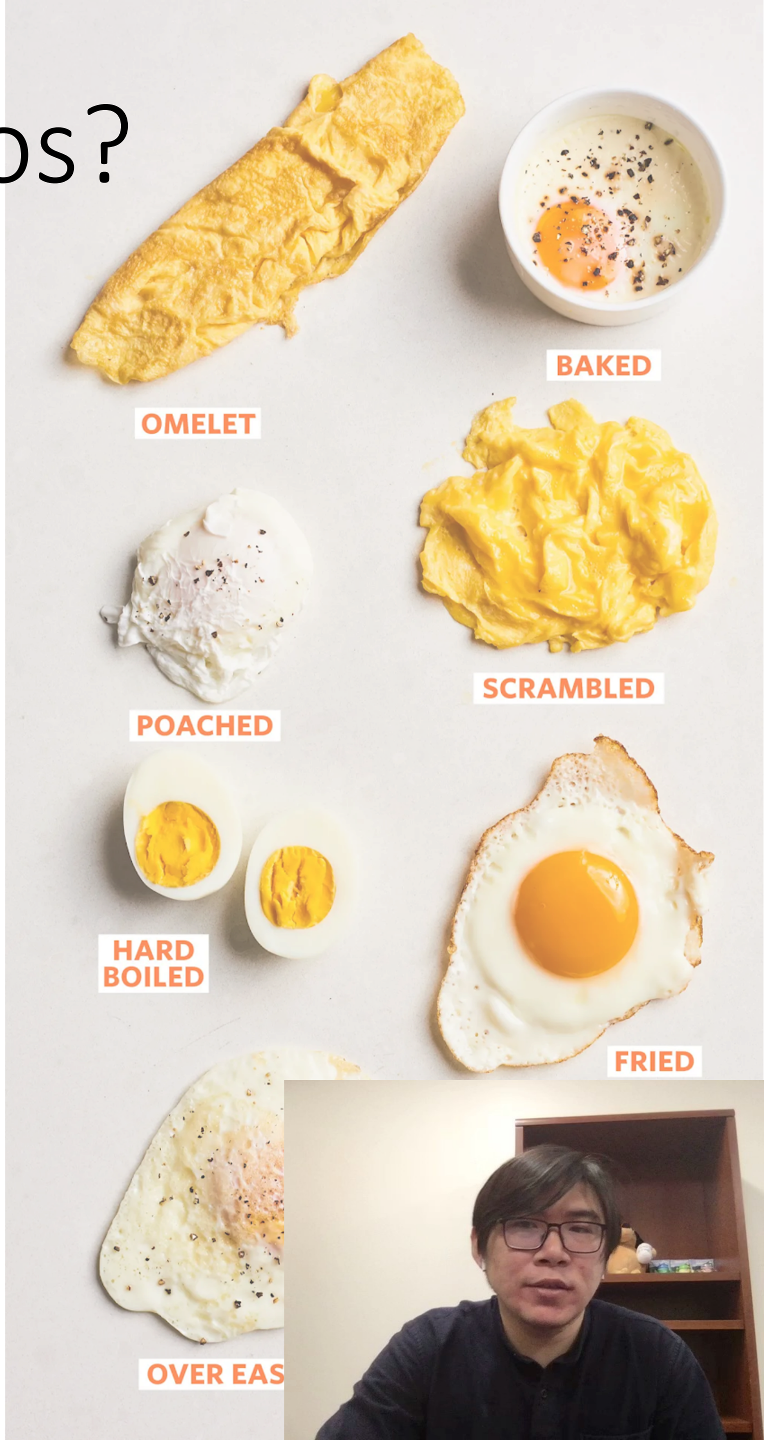
After finishing as runner-up to Chelsea in the 2009–10 season, United achieved a record 19th league title in 2010–11, securing the championship with a 1–1 awa Rovers on 14 May 2011.[61] This was extended to 20 league titles in 2012–13, securing the championship with a 3–0 home win against Aston Villa on 22 April 20

2013–present

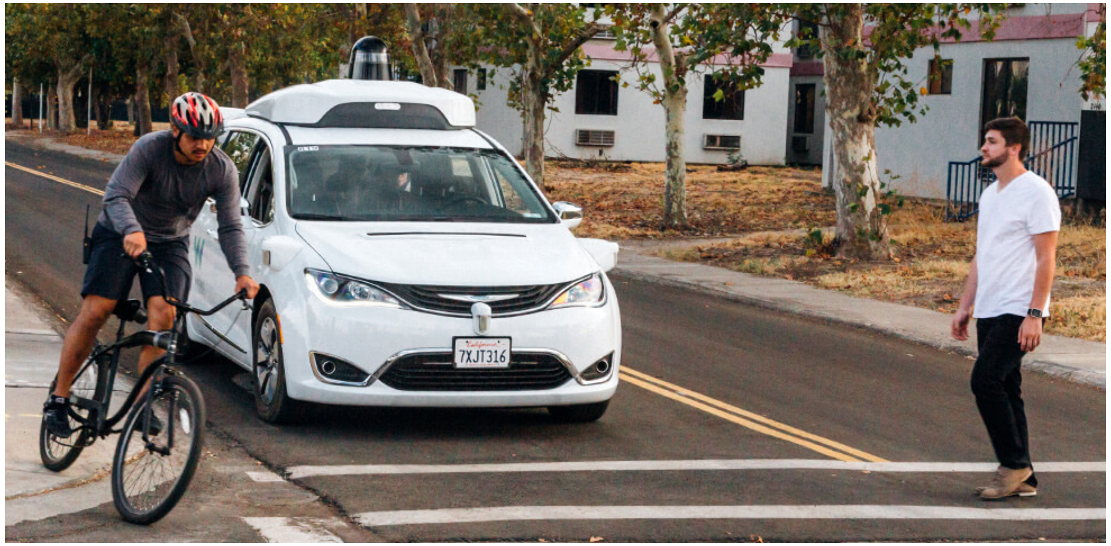
On 8 May 2013, Ferguson announced that he was to retire as manager at the end of the football season, but would remain at the club as a director and club ambassador.[63][64] He retired as the in football history.[65][66] The club announced the next day that Everton manager David Moyes would replace him from 1 July, having signed a six-year contract.[67][68][69] Ryan Giggs took over a months later, on 22 April 2014, when Moyes was sacked after a poor season in which the club failed to defend their Premier League title and failed to qualify for the UEFA Champions League fo 96.[70] They also failed to qualify for the Europa League, meaning that it was the first time Manchester United had not qualified for a European competition since 1990.[71] On 19 May 2014, it wa

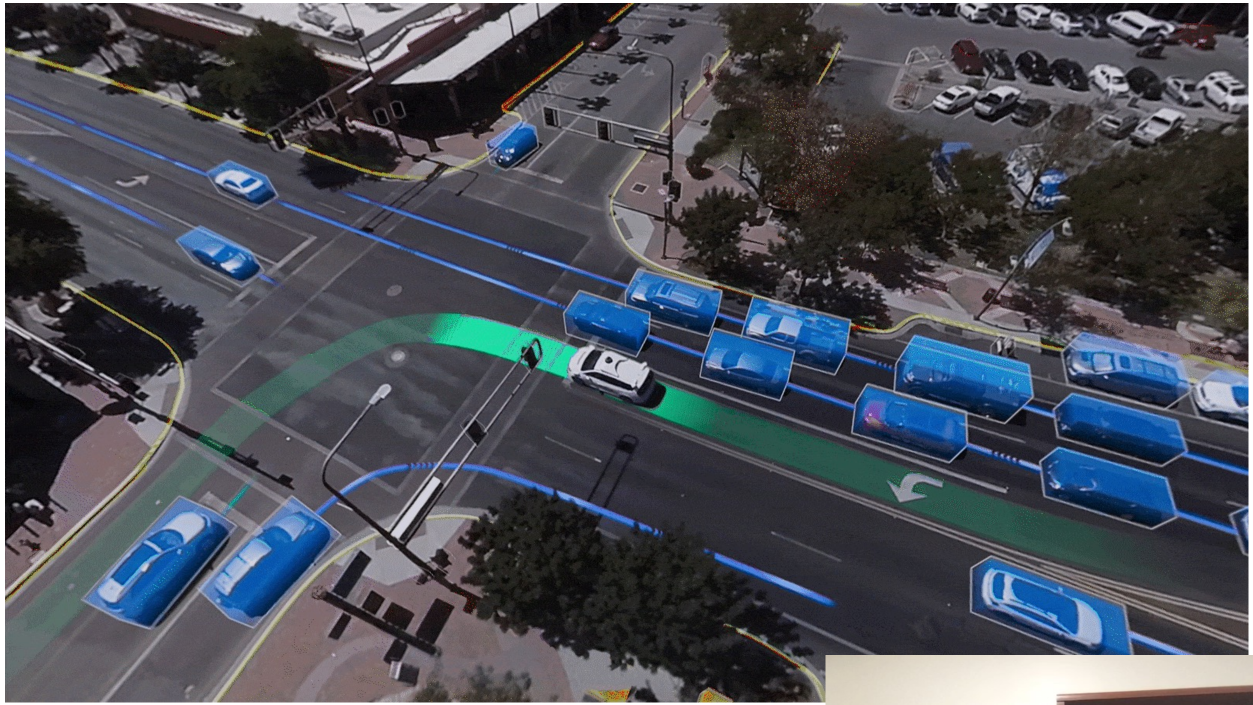
What **else** can we learn from videos?

How to Turn  into:



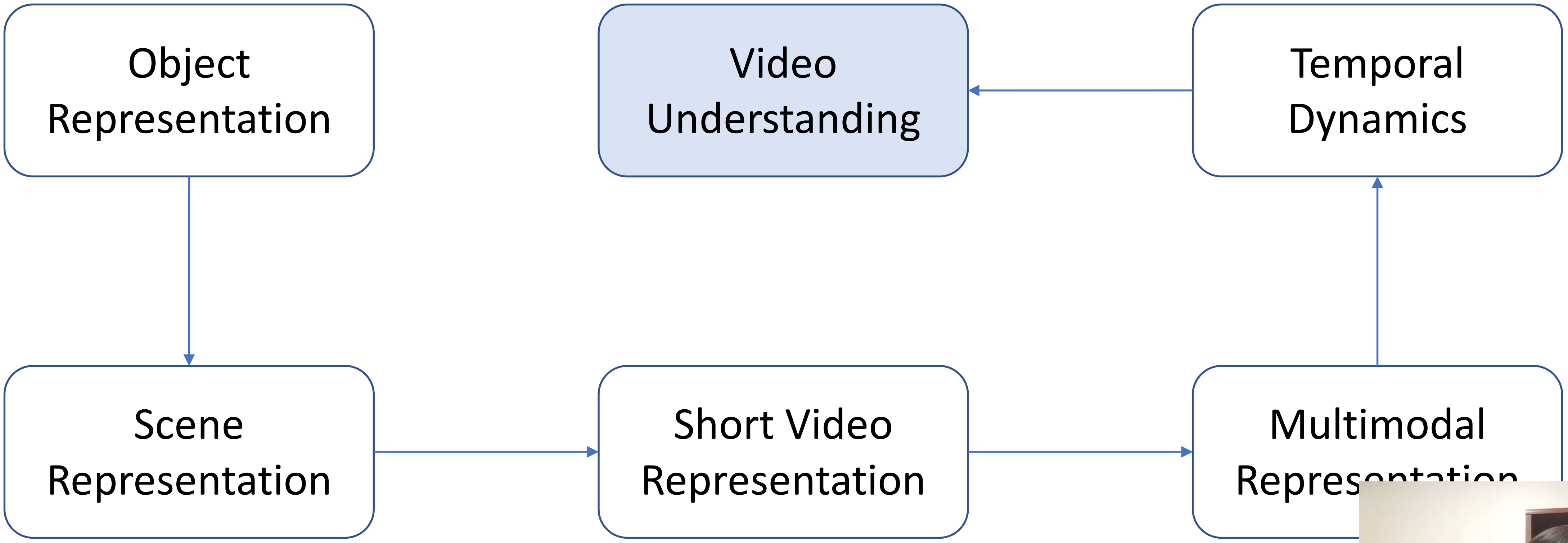
What **else** can we learn from videos?





Transfer what has been learned from passive observations

A RoadMap Towards Video Understanding



Hard to label all of them

Need to learn from unlabeled videos!!

Scene-level Contrastive Learning

View 2: Augmented image

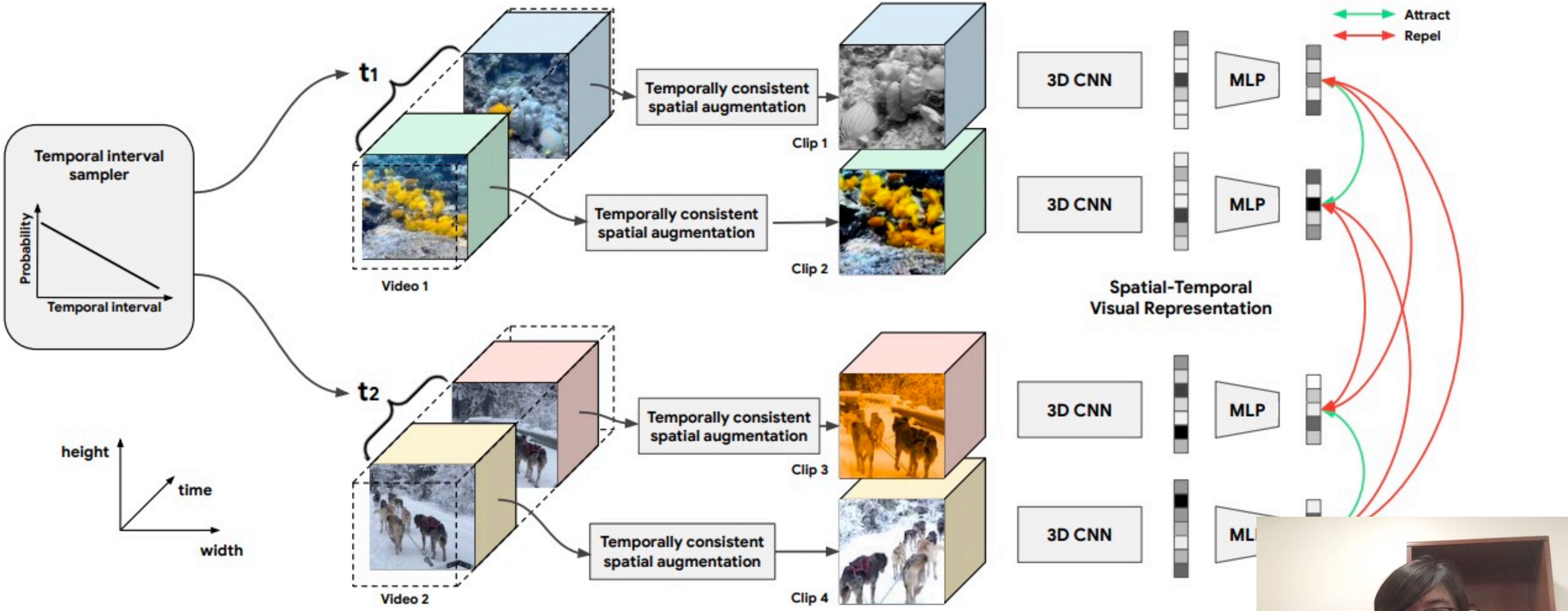
View 1: Augmented image

Similar

Different

DistInst
CPC
CMC
SimCLR
MoCo
...

Contrastive Learning for Videos

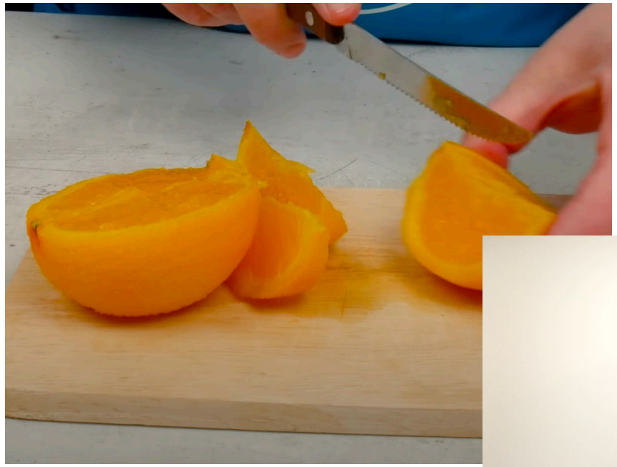


Qian and Meng et al., Spatiotemporal Contrastive Video Representation Learning, CV

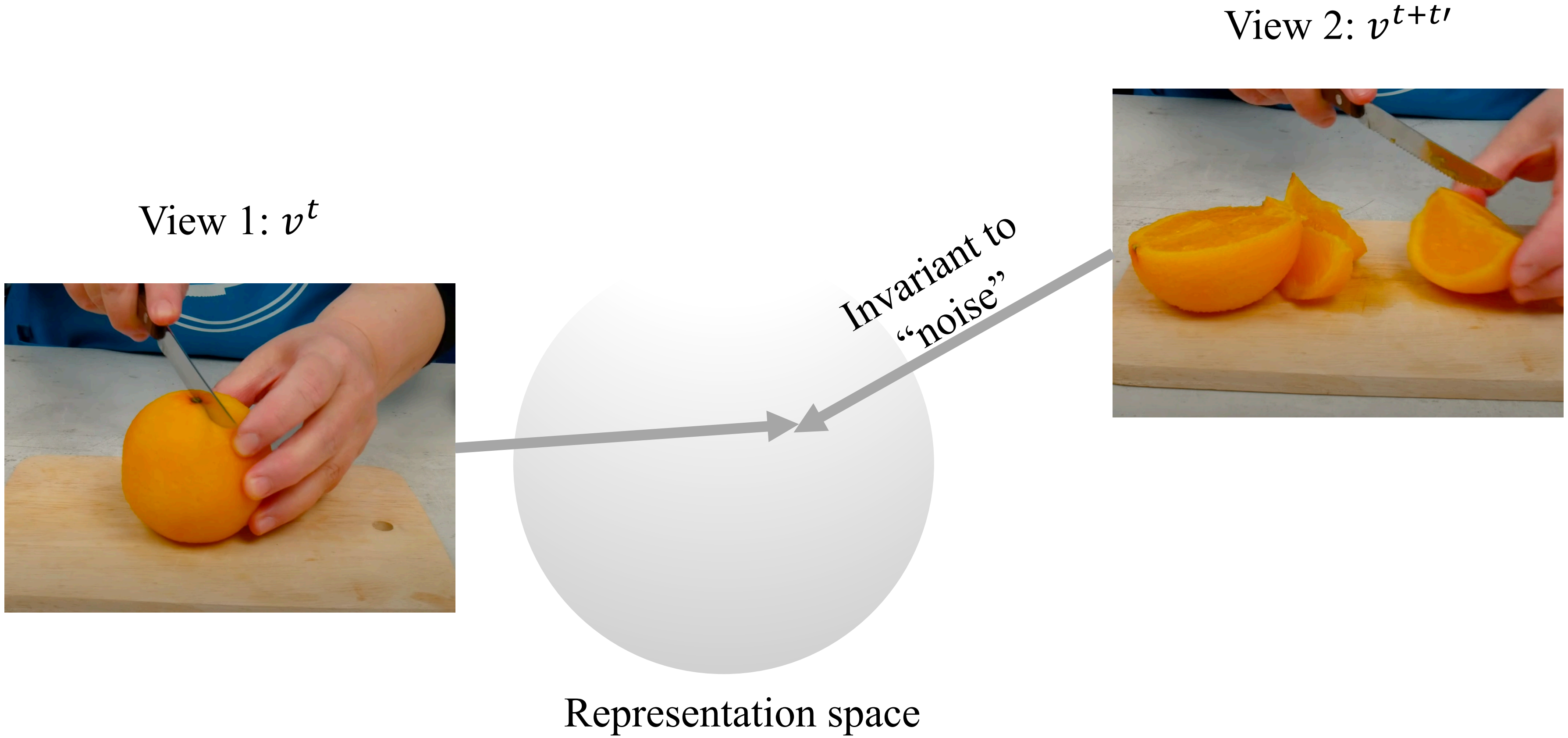
What should consist positive pairs?

For images:
Preserve objects

For videos:
?

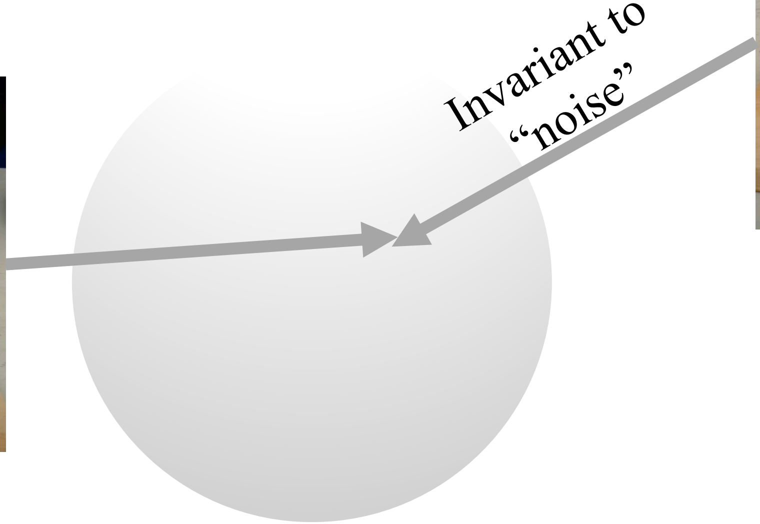


Natural views introduce undesired invariances



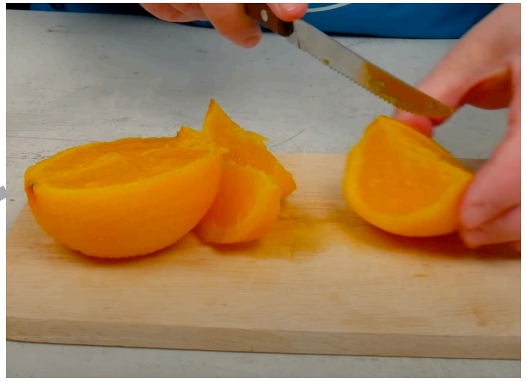
Natural views introduce undesired invariances

View 1: v^t



Representation space

View 2: $v^{t+t'}$

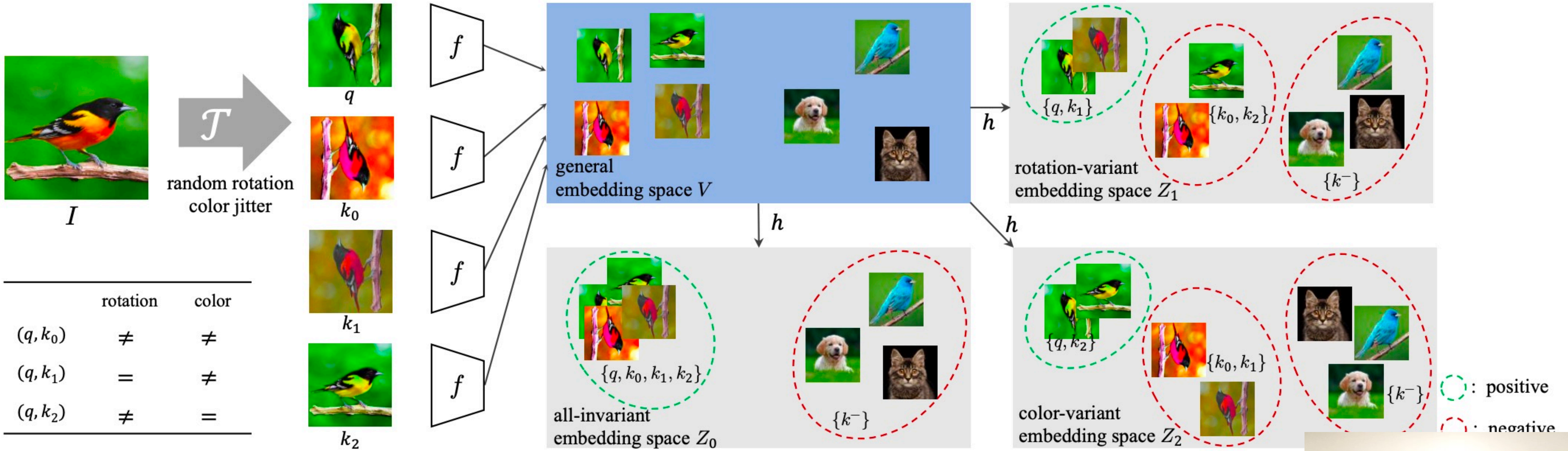


Signal:
Color, local flow

“Noise”:
Shape deformation

Loses temporal info

Solution 1: Construct many pairs of views

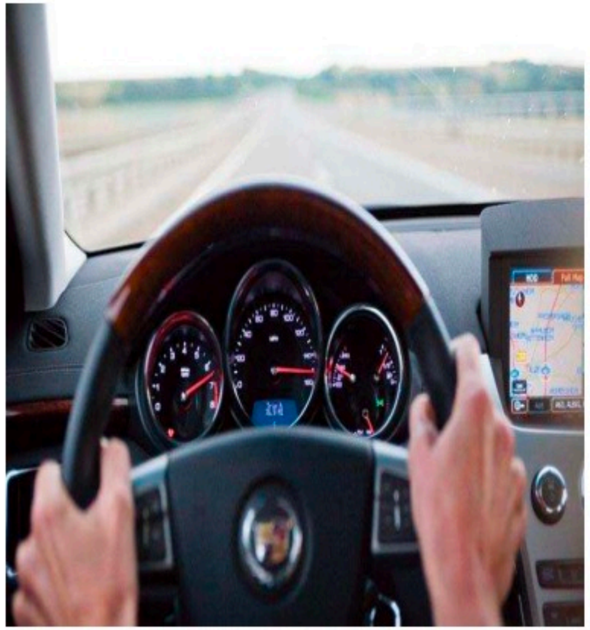


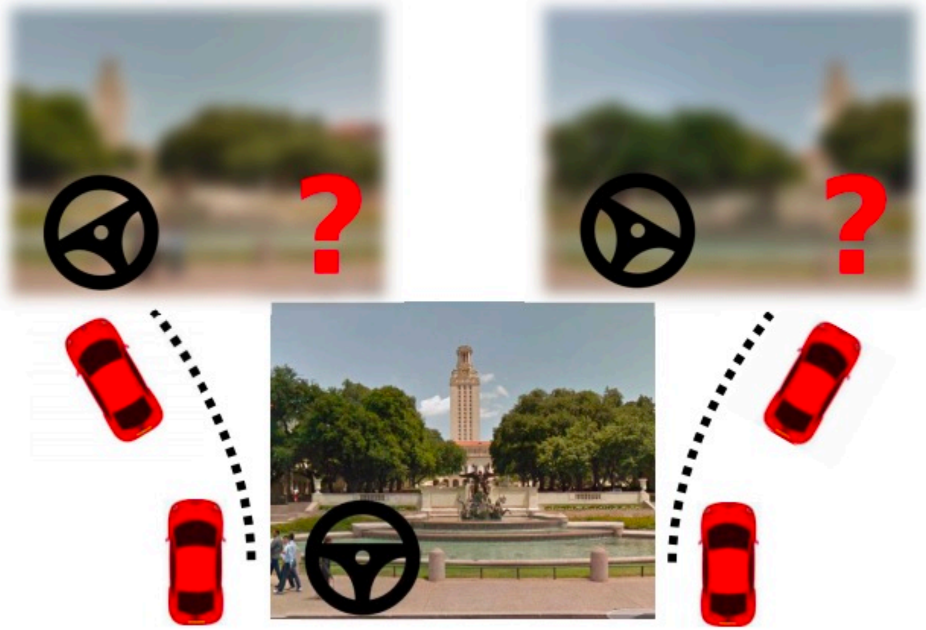
	rotation	color
(q, k_0)	$\neq$	$\neq$
(q, k_1)	$=$	$\neq$
(q, k_2)	$\neq$	$=$

May not scale well

Xiao et al., What Should Not Be Contrastive in Contrastive Learning, ICLR 202

Solution 2: Equivariant representations



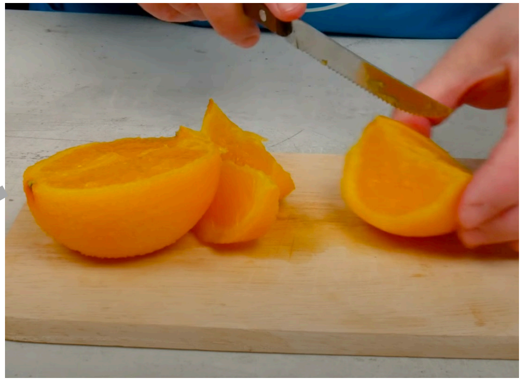


Not necessary for many tasks

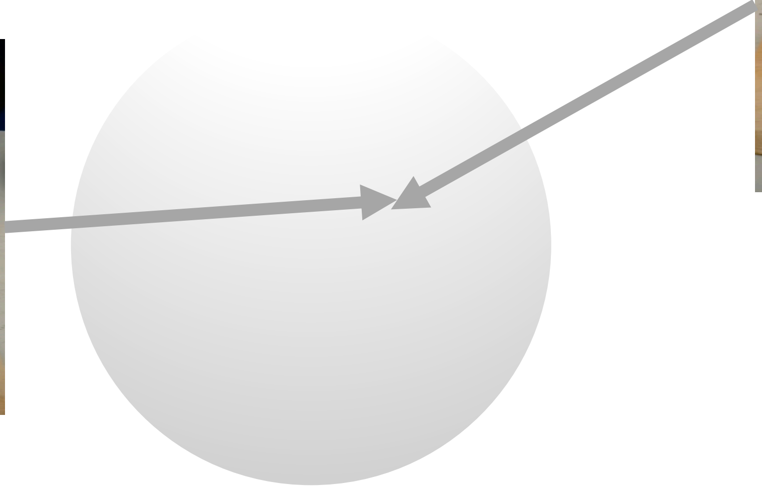
Jayaraman and Grauman, Learning image representations tied to ego-motion, ICC

Our solution: Simply encode the augmentations

View 2: $v^{t+t'}$

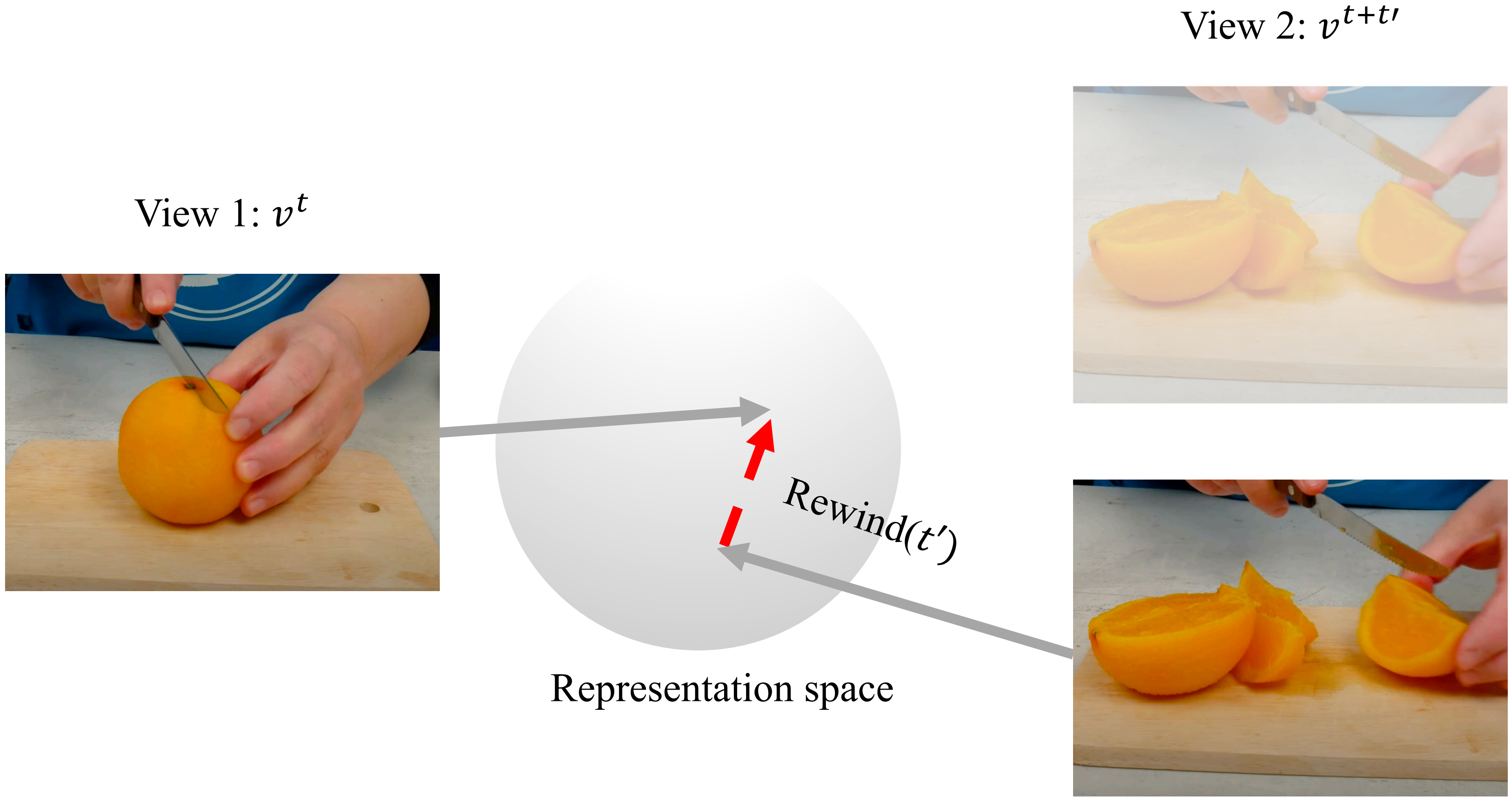


View 1: v^{t}



Representation space

Our solution: Simply encode the augmentations



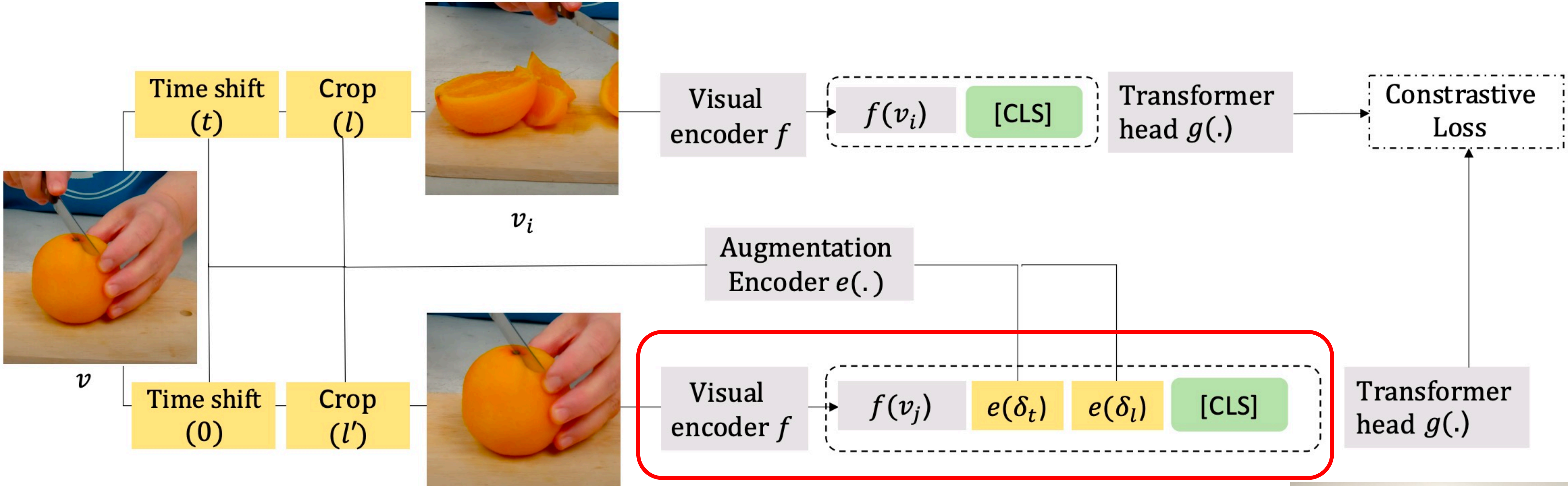
Learn an implicit "prediction" model of t'

Shared and **predictable** information can be preserved:
color, shape, etc.

Unpredictable is still
"noise
car

Special cases: view-invariant coding, view-predictive coding

Composable AugmenTation Encoding (CATE)



Projection head is now a Transforr
encodes a sequence of augmentat

Sun, Nagrani, Tian, and Schmid, Composable augmentation encoding for video representation l

The Something-Something Dataset

Classes	
Putting something on a surface	4,081
Moving something up	3,750
Covering something with something	3,530
Pushing something from left to right	3,442
Moving something down	3,242
Pushing something from right to left	3,195
Uncovering something	3,004
Taking one of many similar things on the table	2,969

Fine-grained actions that rely on
the arrow of ti

Augmentation encoding is helpful

Encoded	τ	Dropout	Top-1 Acc.	Top-5 Acc.
No	-	-	26.5	55.9
Crop	$\delta_{x,y}$	✗	27.2	56.7
Crop	$\delta_{x,y}$	✓	28.1	58.0
Time	$\text{sgn}(\delta_t)$	✗	28.1	57.9
Time	δ_t	✗	31.3	62.4
Time	δ_t	✓	31.2	61.4

Encode Time	τ	Time Offset Acc.
✗	-	5.7
✓	$\text{sgn}(\delta_t)$	65.7
✓	δ_t	**99.9**

Table 5: **Time Shift Classification on SSv1**. Encoding time significantly helps on this proxy task, validating the intuition that our model retains useful time information.

Augmentation encoding is composable

Enc. Crop	Enc. Time	Top-1 Acc.	Top-5 Acc.
✗	✗	26.5	55.9
✓	✗	28.1	58.0
✗	✓	31.2	61.4
✓	✓	**32.2**	**62.4**

Table 2: **Composing spatial (crop) and temporal encodings** for Something-Something v1. Each individual encoding outperforms the no encoding baseline (SimCLR++). Composing them together yields the best performance.

Per-class comparison (temporal aug.)

Arrow of time barely matters:

Label	ΔAP
Lifting something up completely, then letting it drop down	21.0
Pulling two ends of something so that it gets stretched	19.8
Moving something and something closer to each other	18.5
Taking one of many similar things on the table	17.2
Pushing something so that it almost falls off but doesn't	16.7
Poking something so lightly that it doesn't move	-4.6
Pretending to pour something out of something	-5.4
Poking a stack of something without the stack collapsing	-5.5
Pretending to spread air onto something	-7.8

Table 4: Classes that benefit the most and the least with **time encoding** on SSv1. We sort the classes by their differences on Average Precision.

t-SNE

CATE

No encoding

Comparison on other benchmarks

Method	top 1	top 5	top 10	top 20	top 50
OPN [32]	19.9	28.7	34.0	40.6	51.6
SpeedNet [5]	13.0	28.1	37.5	49.5	65.0
VCP [34]	19.9	33.7	42.0	50.5	64.4
Temporal SSL [25]	26.1	48.5	59.1	69.6	82.8
MemDPC† [18]	40.2	63.2	71.9	78.6	-
CATE	**54.9**	**68.3**	**75.1**	**82.3**	**89.9**

Table A6: Nearest neighbor retrieval evaluation on UCF-101 split 1. †: with Flow

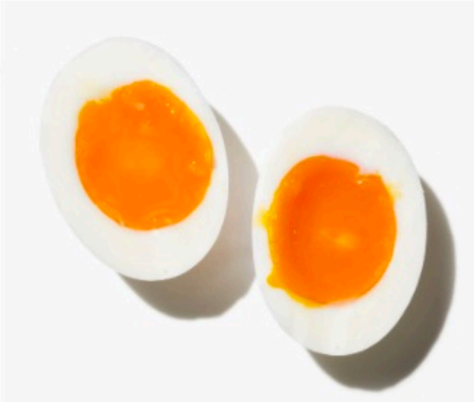
Method	top 1	top 5	top 10	top 20	top 50
VCP [34]	6.7	21.3	32.7	49.2	73.3
MemDPC† [18]	15.6	37.6	52.0	65.3	-
CATE	**33.0**	**56.8**	**69.4**	**82.1**	**92.8**

Table A7: Nearest neighbor retrieval evaluation on HMDB-51 split 1. †: with Flow

Checkpoints are released!

https://github.com/google-research/google-research/tree/master/cate

The egg problem

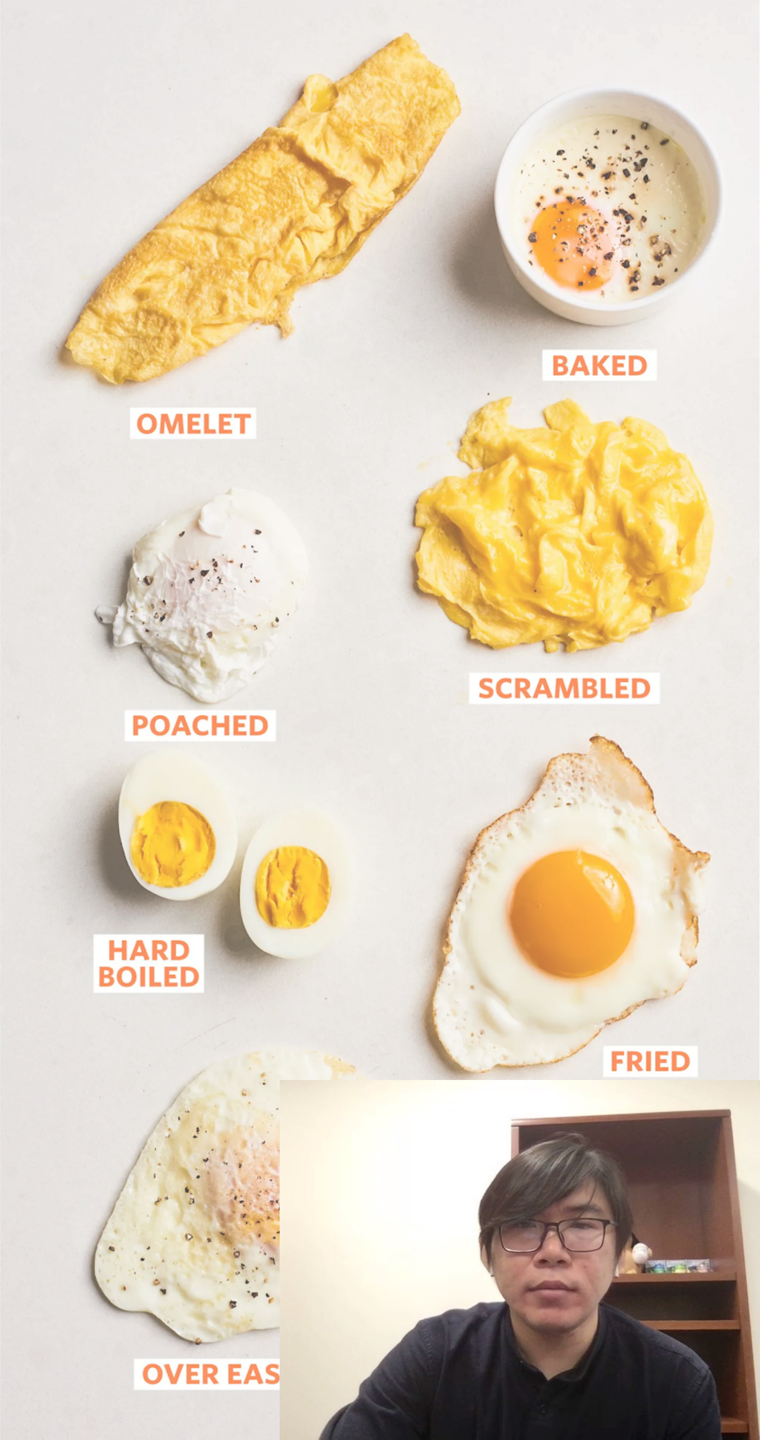
$f($  $, \text{boil}) =$ 

A more compact representation for videos:

Actions as object state transitions

(Action recognition, object tracking, ...,

Visual Commonsense)

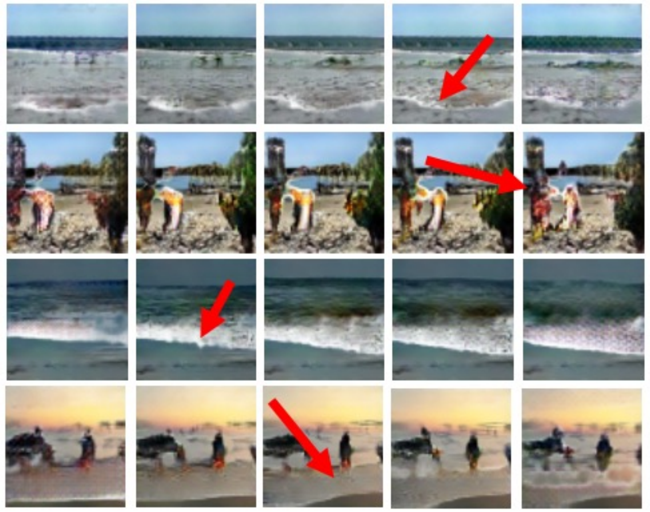


But why?

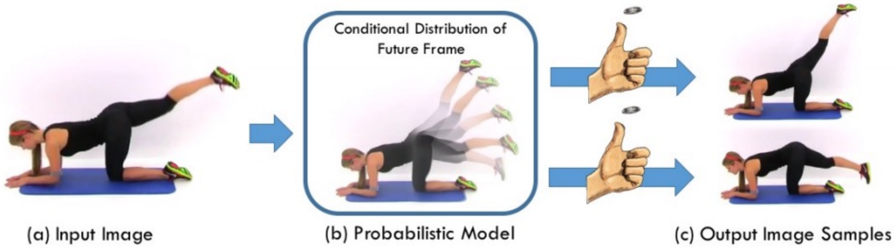
- Towards Long Video Understanding
  - Only use “key moments”
  - Video summarization
- Structured Representation
  - Objects
  - Their state transitions over time (visual dynamics)
- Modeling temporal dynamics is itself important

How to predict the future?

Generate images…



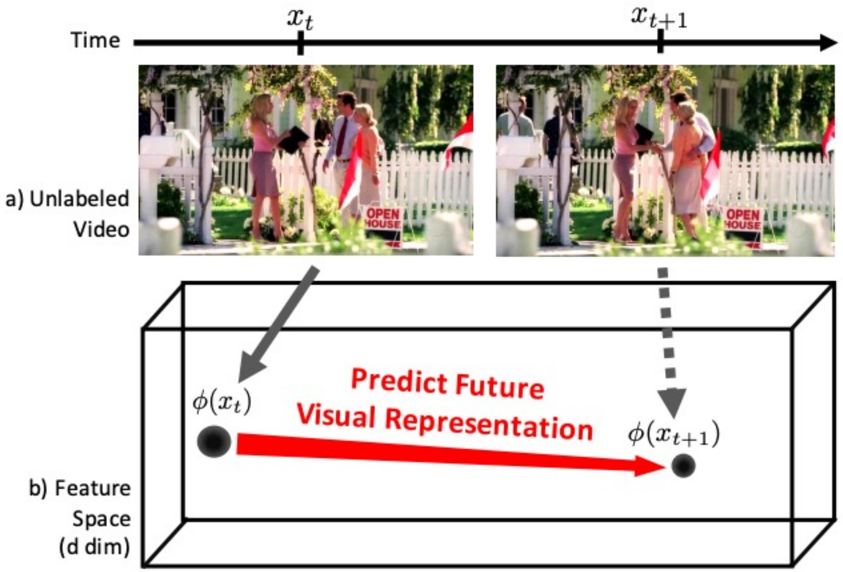
Vondrick et al., 2016



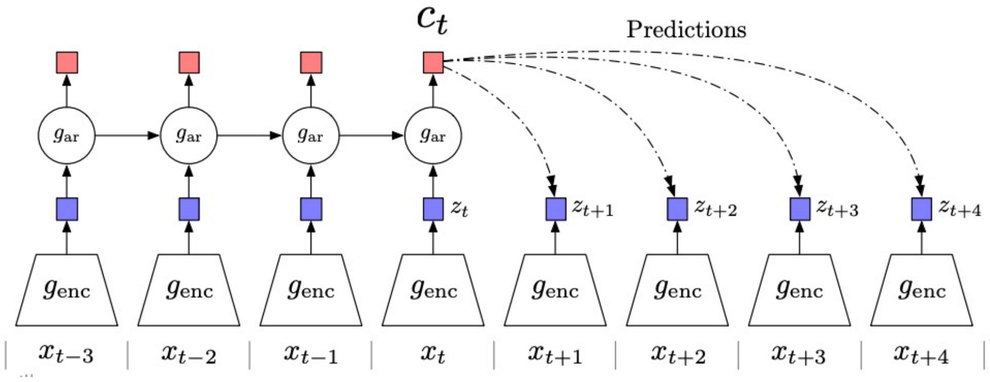
Xue et al., 2016

How to predict the future?

Generate representations…



Vondrick et al., 2015



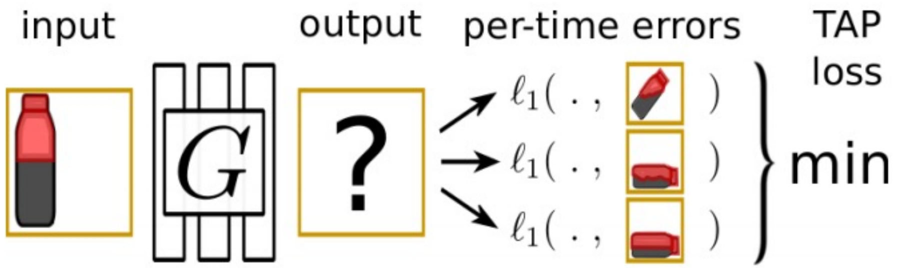
van den Oord et al., 2018

Problem solved?

Not quite…

Predict at fixed offset into future = deal with high uncertainty!

Could let network output most predictable moment in near future



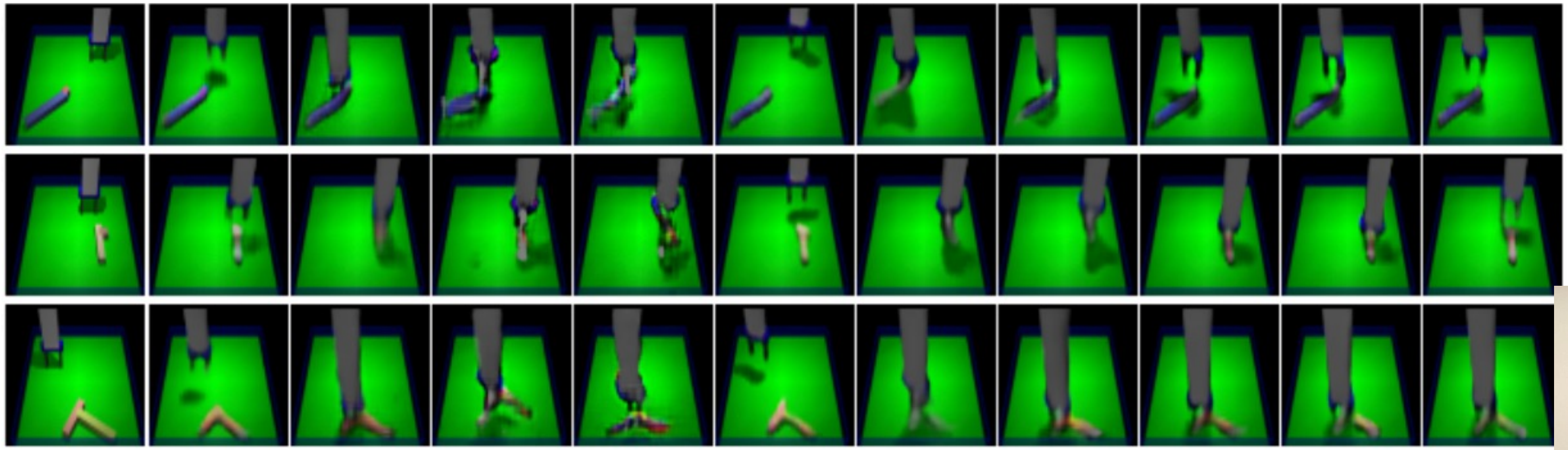
Jayaraman et al., 2018

Okay, problem solved now?

Not quite…

Very short-term prediction – a few seconds into future at most

Limited to simple, low-level visual data



Jayaraman et al., 2018

The ideal future prediction

Dynamic, rather than at a fixed offset into the future

High-level, e.g., mixing eggs and flour → rolling out dough

Unsupervised, to take advantage of large unlabeled datasets

(a) Time = **t**

“go ahead and pour the cream in”

Better future predictions

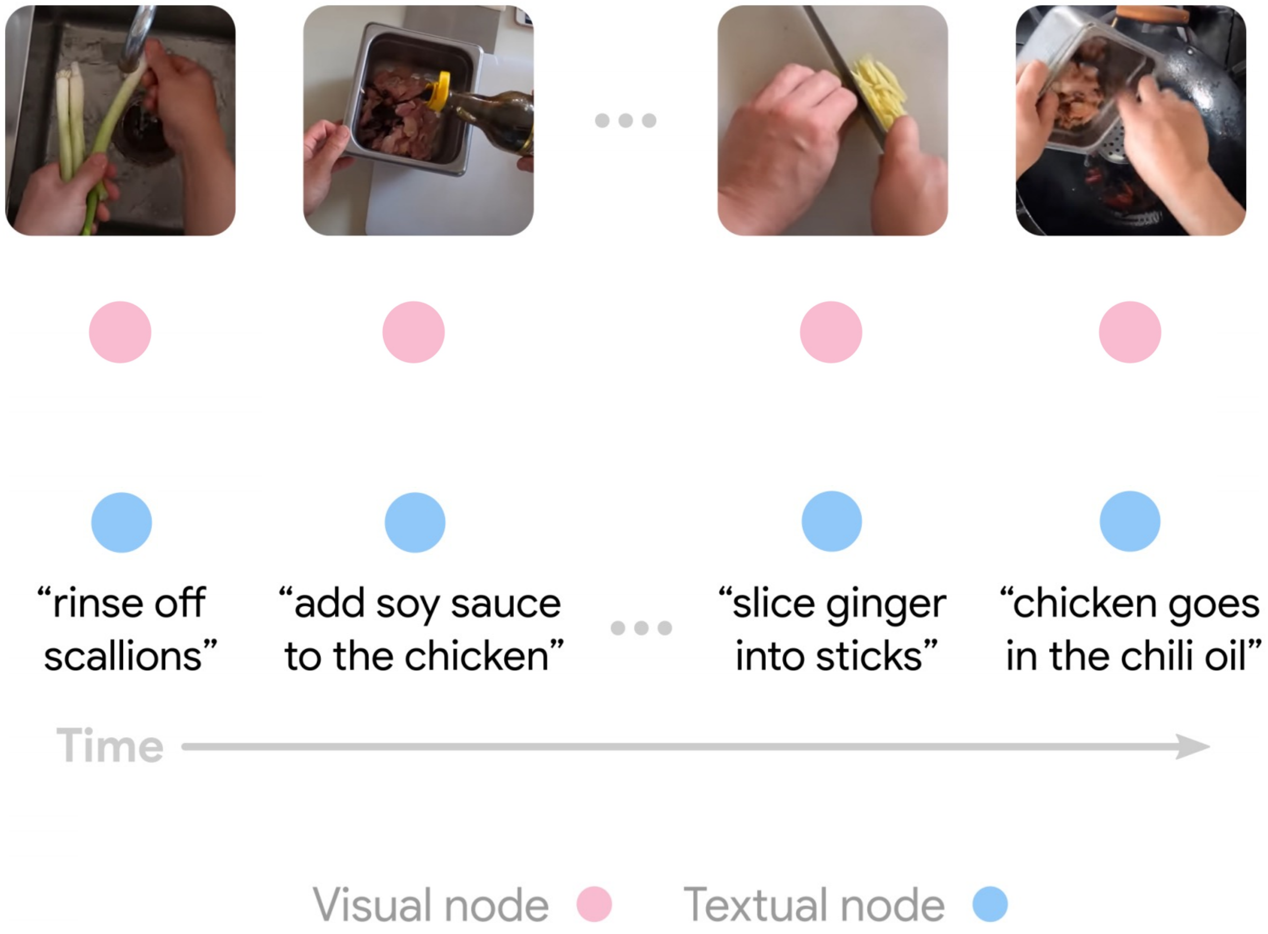



"rinse off scallions" "add soy sauce to the chicken" ••• "slice ginger into sticks" "chicken goes in the chili oil"

Time

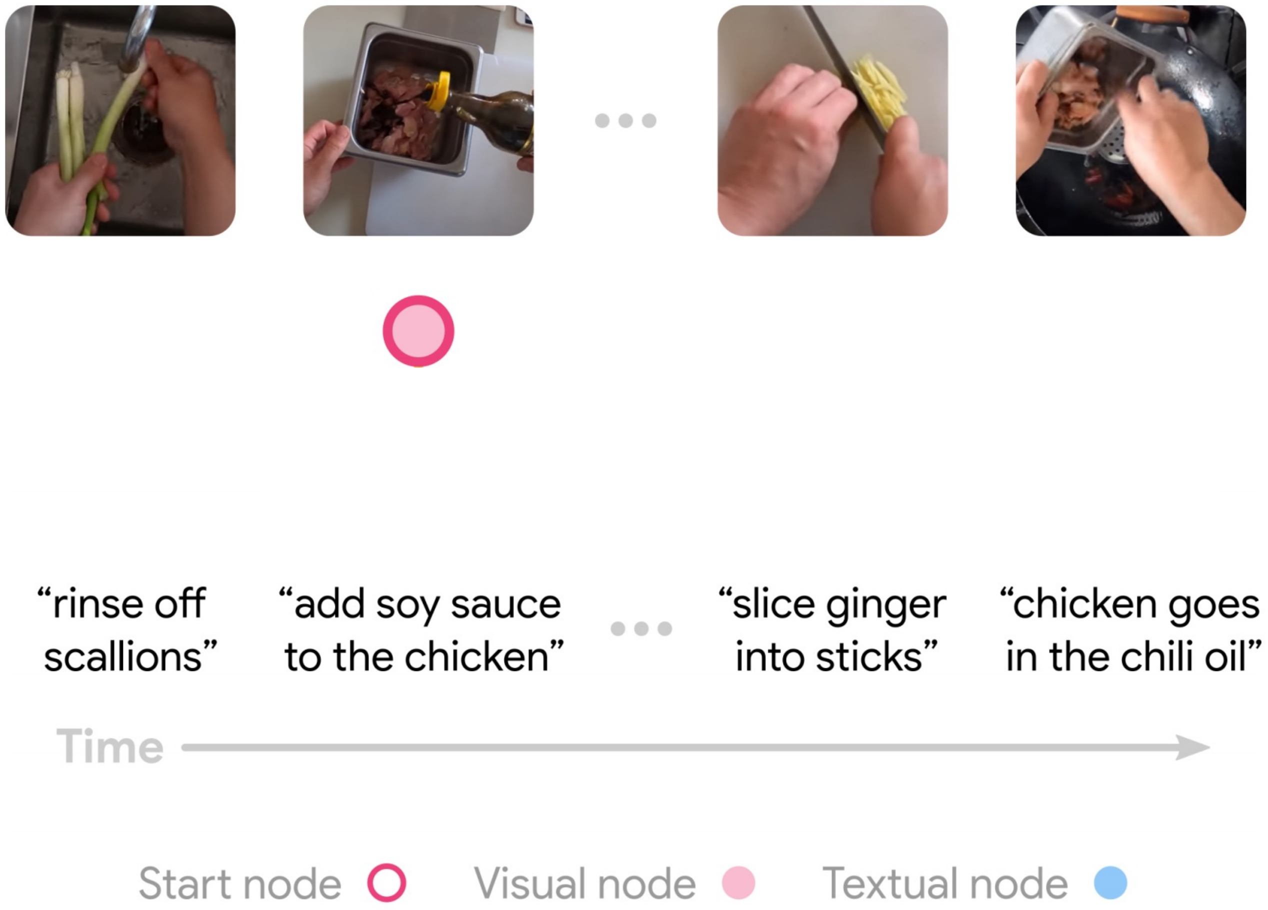
Epstein, Wu, Schmid, and Sun, Learning Temporal Dynamics from Cycles in Narrated Vide

Better future predictions



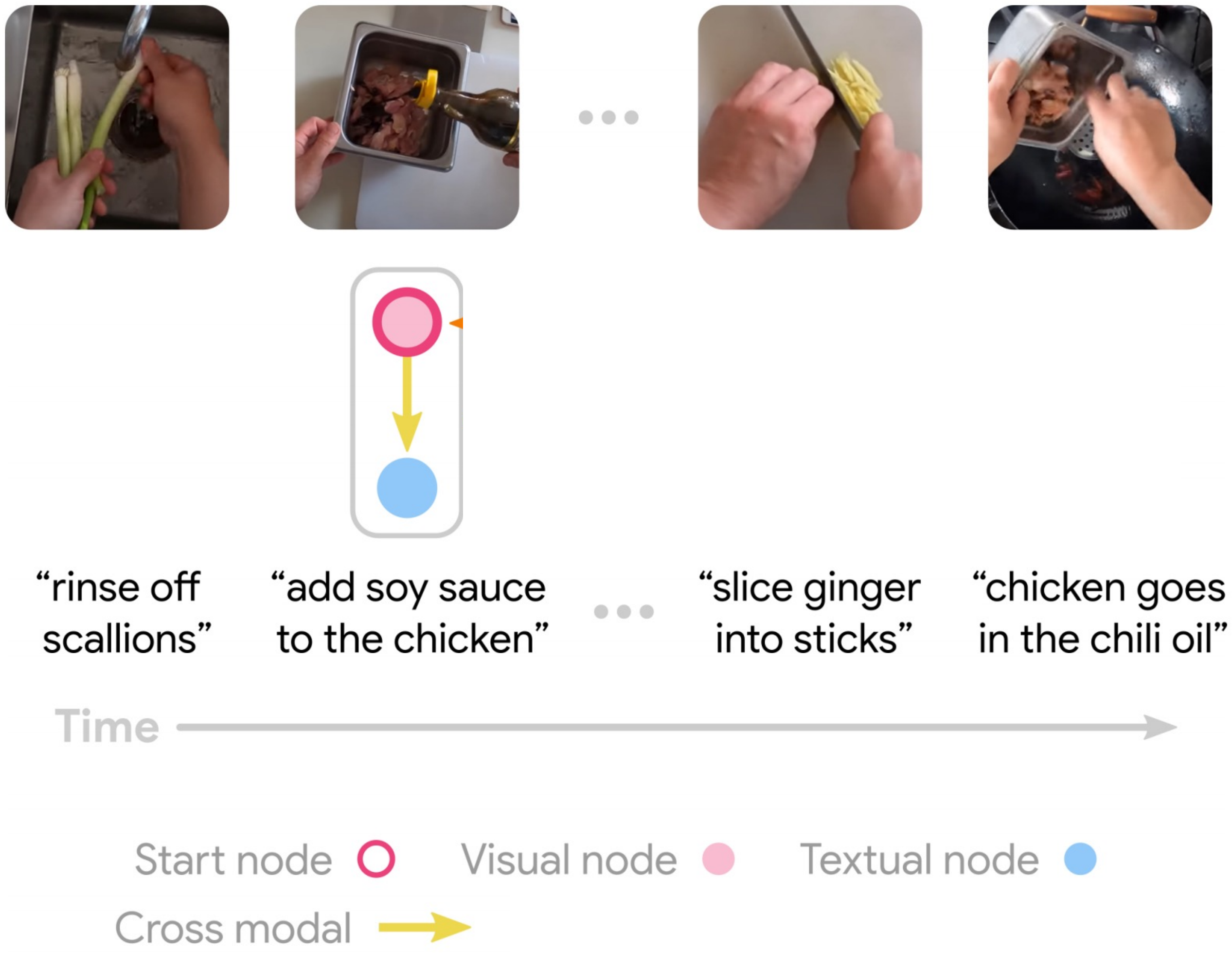
Epstein, Wu, Schmid, and Sun, Learning Temporal Dynamics from Cycles in Narrated Vide

Cycling through video

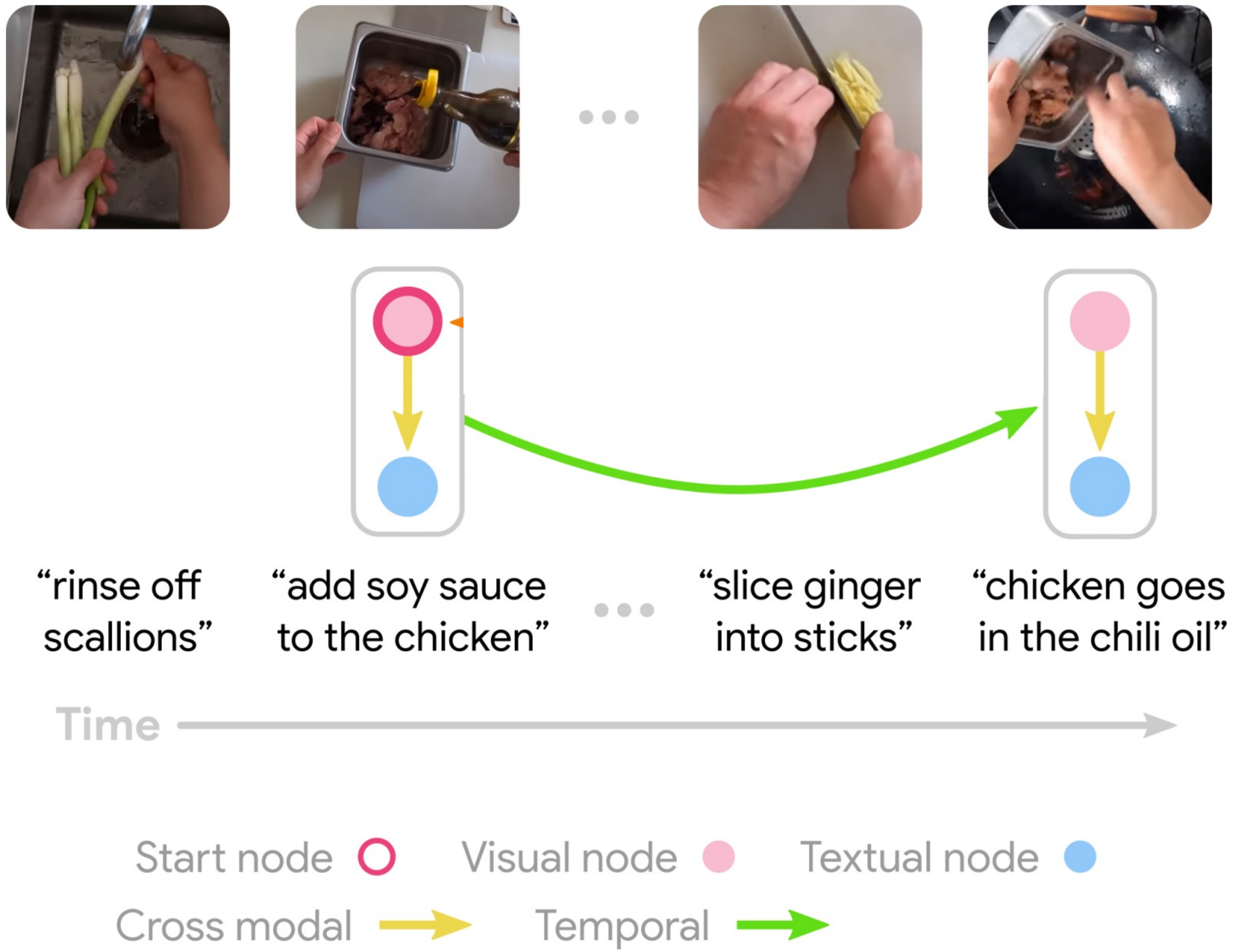


Epstein, Wu, Schmid, and Sun, Learning Temporal Dynamics from Cycles in Narrated Vide

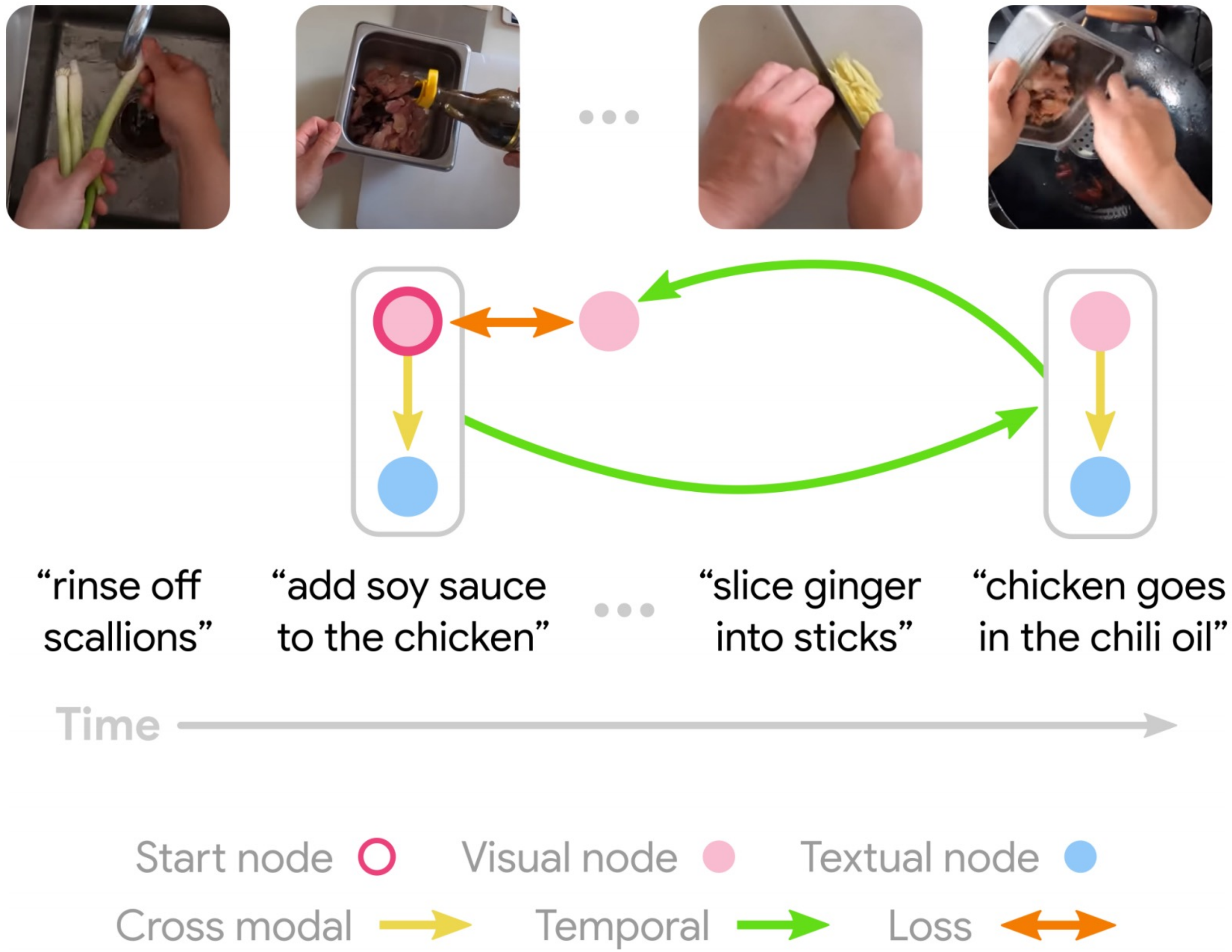
Cycling through video



Cycling through video



Cycling through video



Cycling through video - intuition

(a) Time = **t** | (b) Time = **t+1** | (c) Time = **t+22** | (d) Time = **t+35**

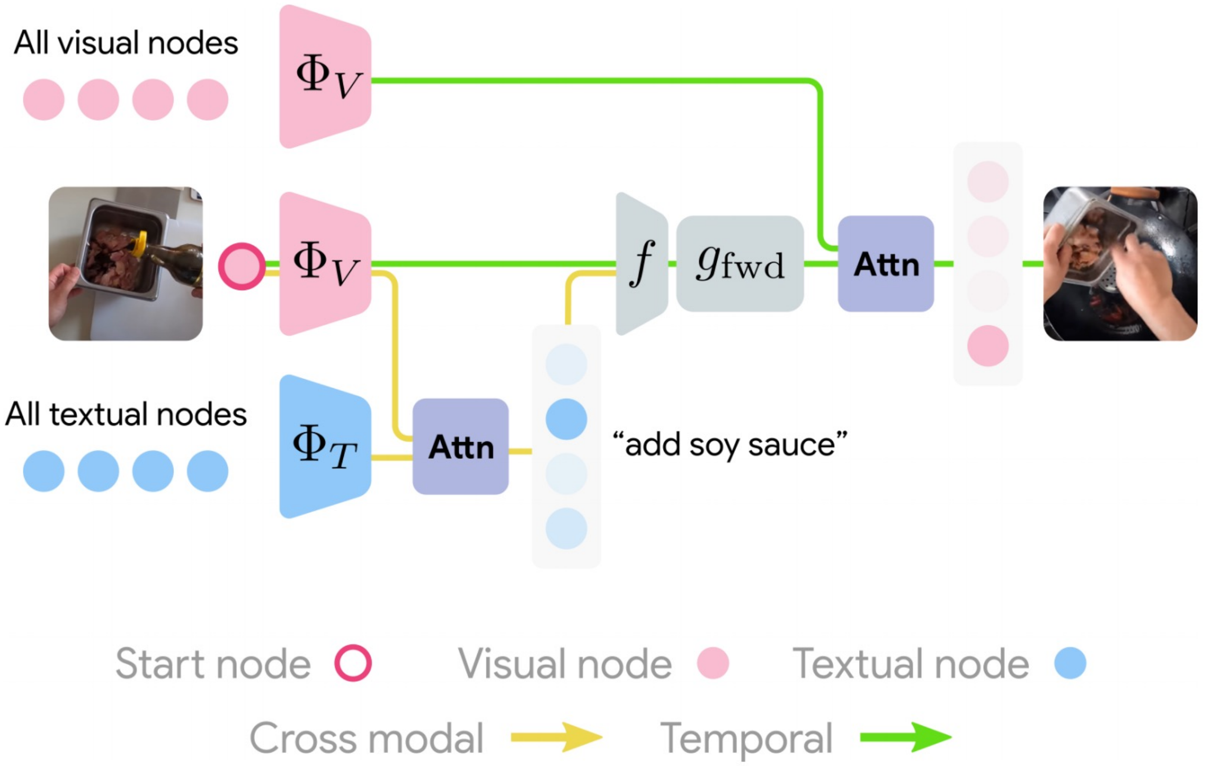
(a) "go ahead and pour the cream in"

(b) "go ahead and pour the cream in"

(c) "we'll be back in 30 minutes"

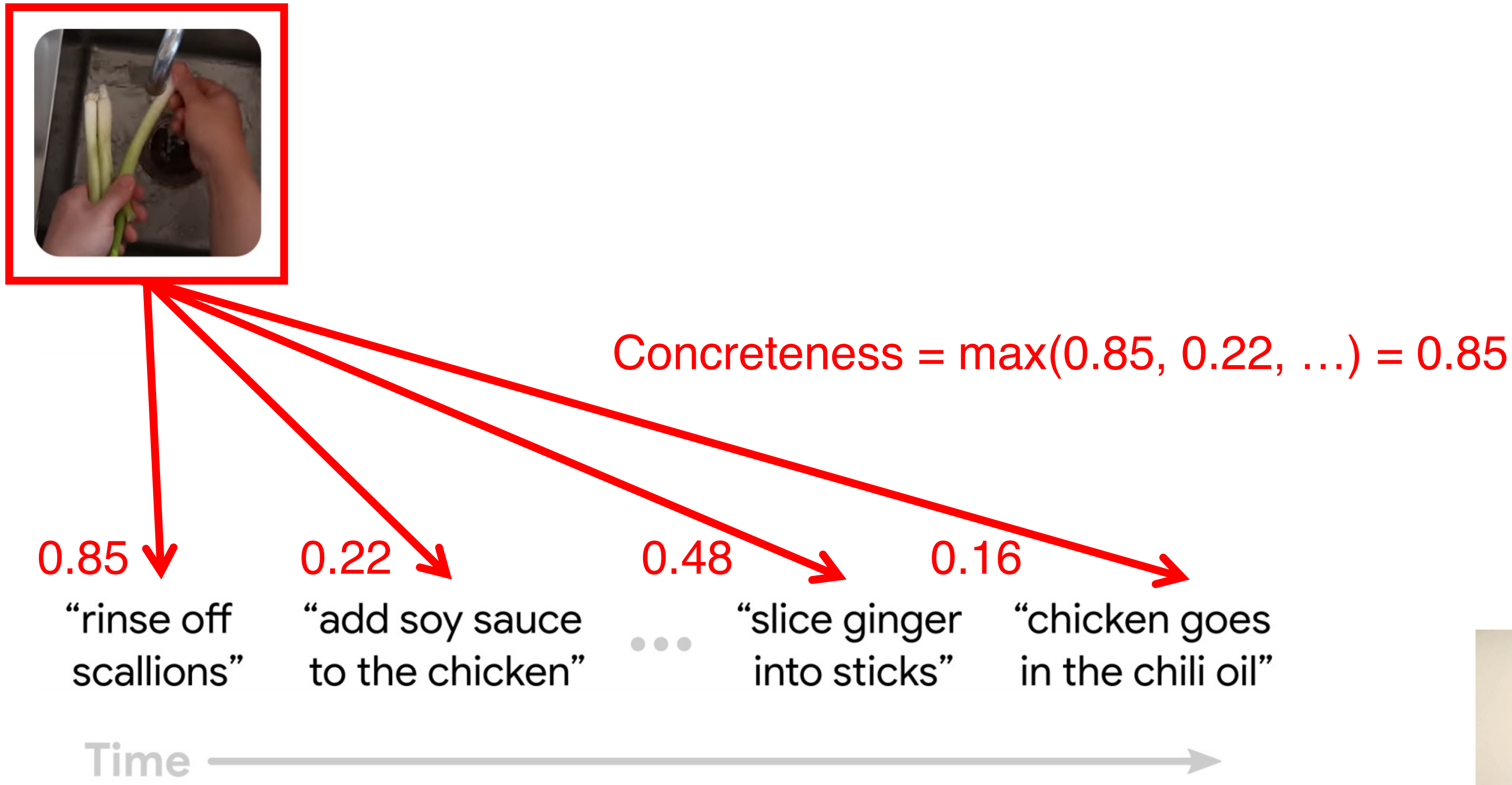
(d) "we have soft-serve ice cream"

Cycling through video - implementation

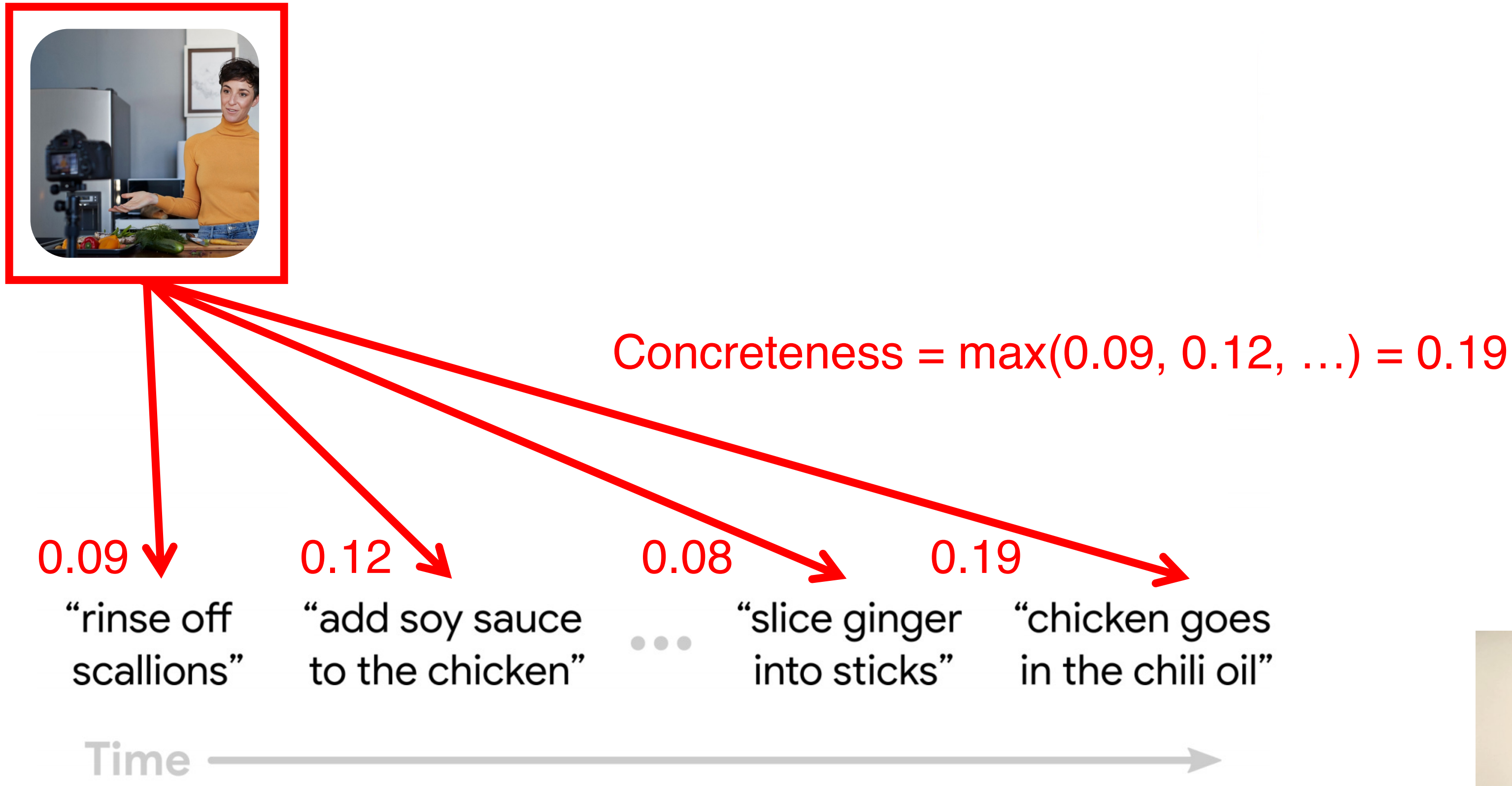


Epstein, Wu, Schmid, and Sun, Learning Temporal Dynamics from Cycles in Narrated Vide

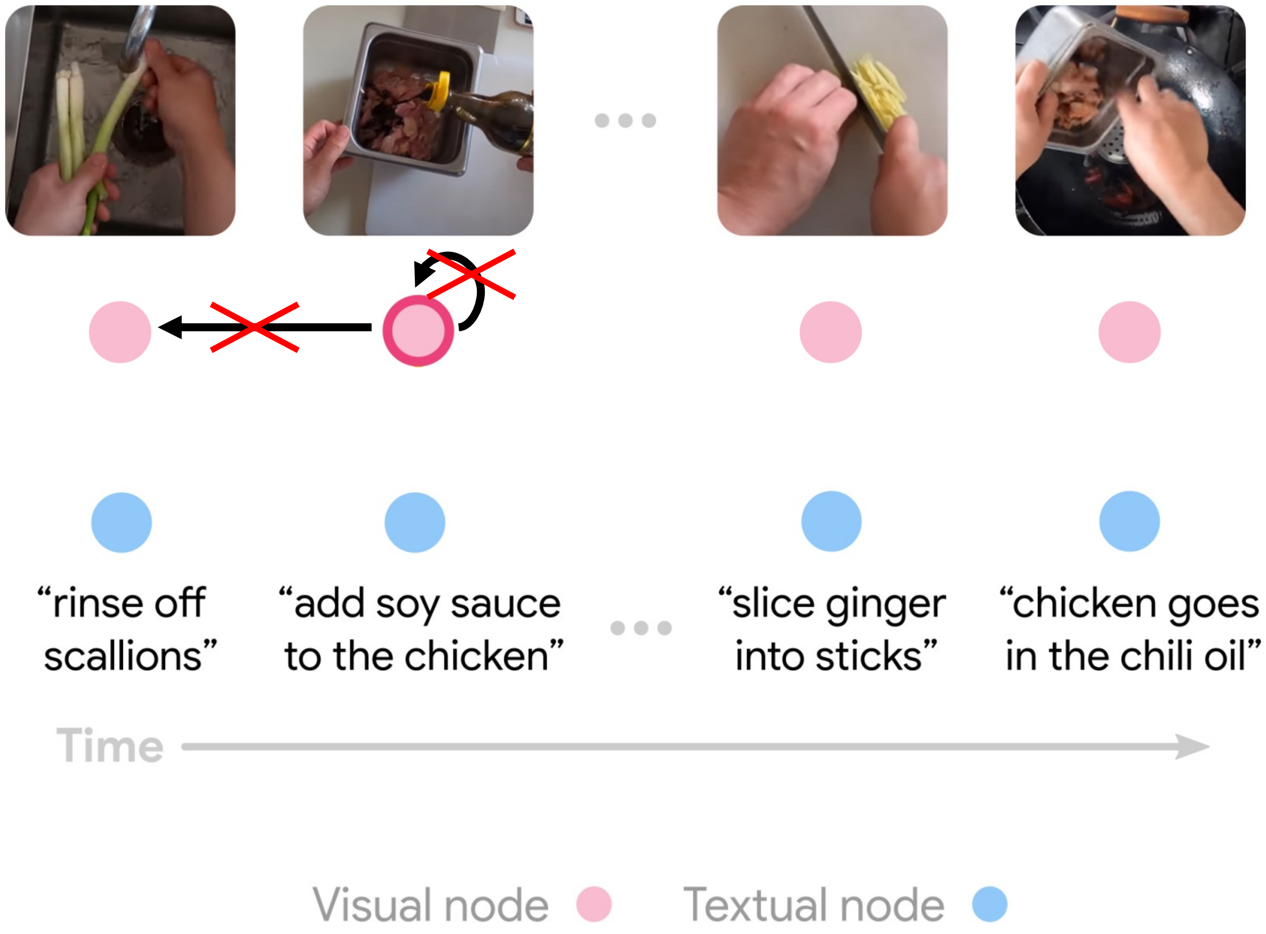
Selecting start nodes



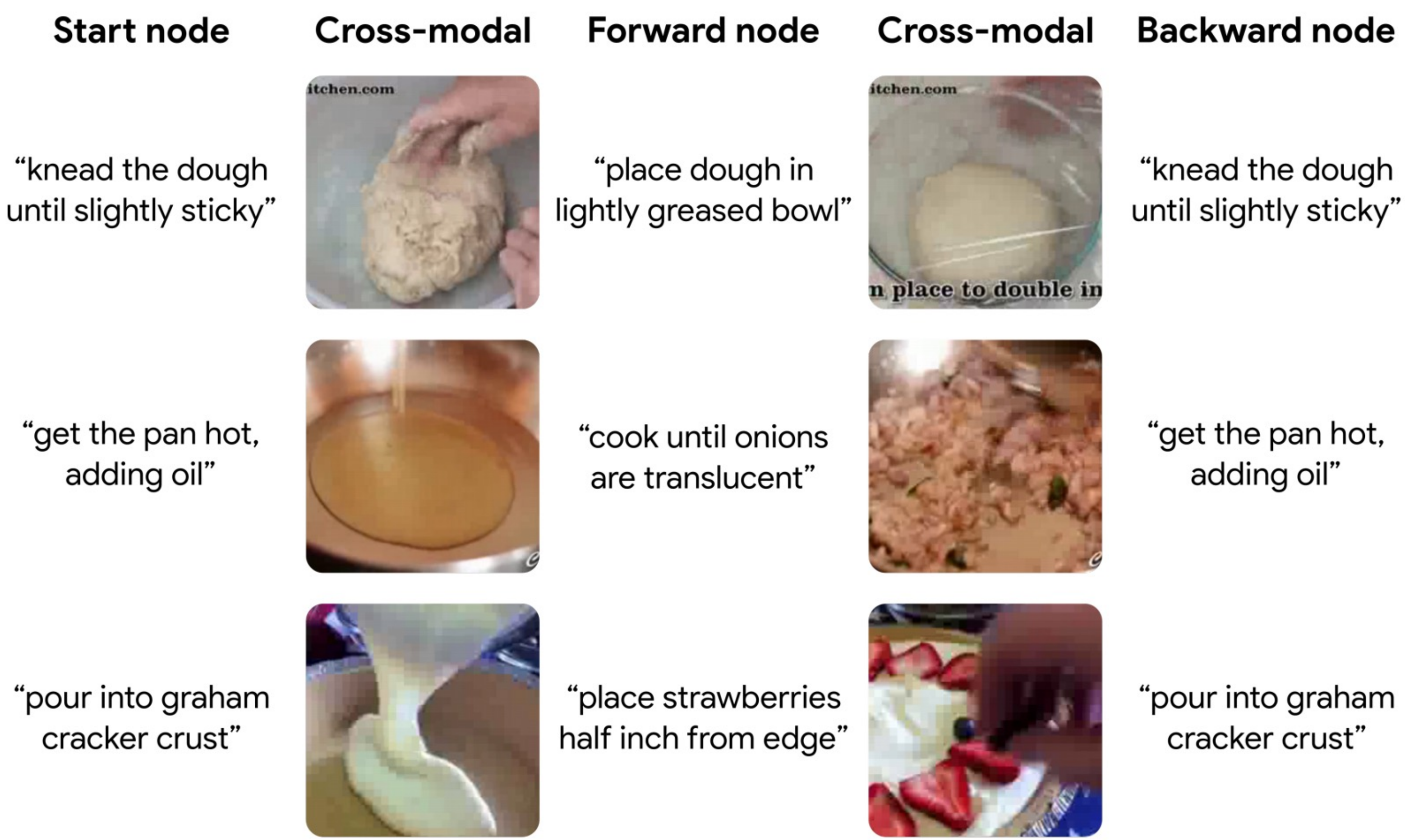
Selecting start nodes



Constraining temporal attention



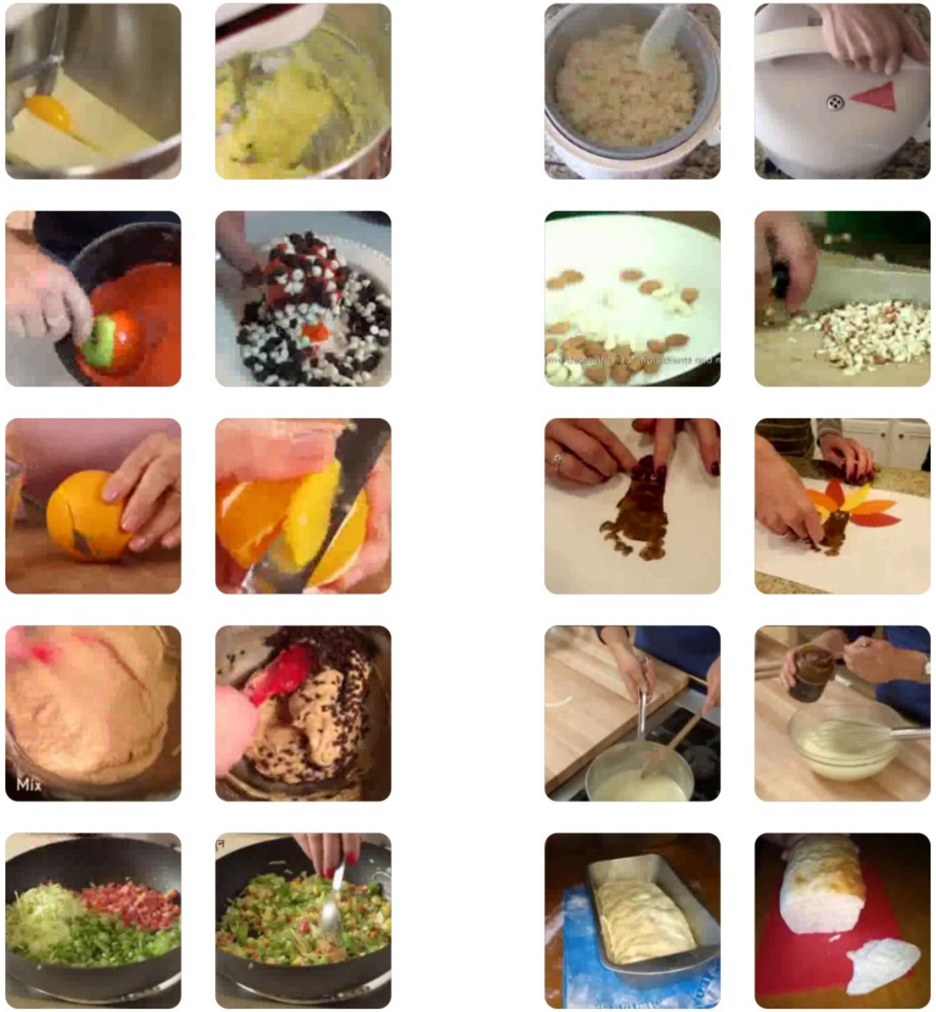
Discovering cycles in video



Finding cycles

Discovering transitions in video

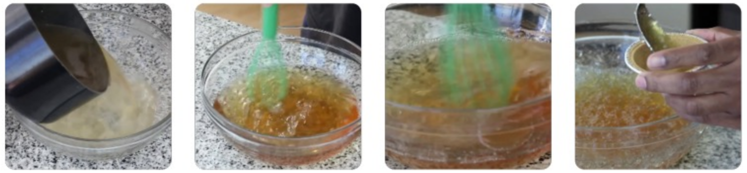
From To From To



Temporally ordering image collections

1 2 3 4 5

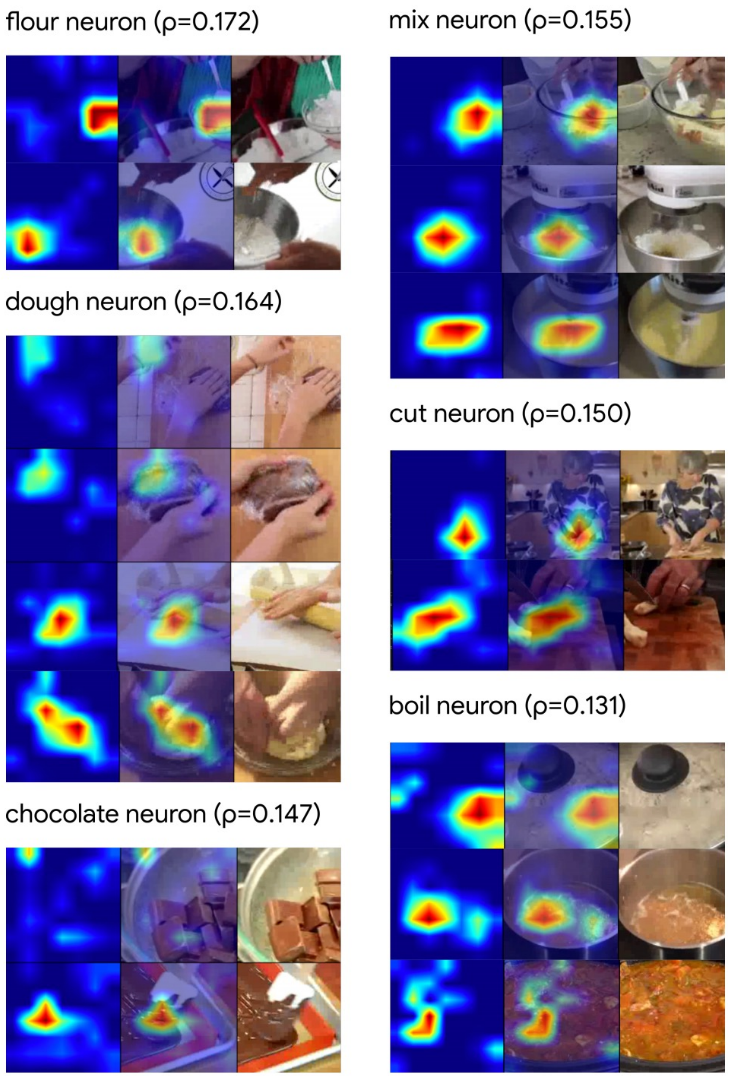
1 3 5 2 6



4 7 8 9

1 2 3 5 4

Action and object neurons emerge



Vision-Language Navigation

ALFRED

ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks; Shridhar et al., 2019

VLN as a Benchmark

- Natural testbed for multimodal representations
  - Joint model visual observations, language instructions, etc.
  - From passive observation to active exploration
- The Transfer Learning Game
  - What to teach an agent before entering an environment?
  - Language and object grounding
  - Not always ideal to learn “end-to-end” and “from scra

Focus One: language representations

$x_{1:L}$

move to the large black end table against the wall
pick up the phone sitting on top of the end table with the blue case
carry the phone to the foot of the bed
place the phone on the bed to the right of the cushion

$y_{1:M}$

goto table pickup cellphone
goto bed put cellphone bed

Often easier to collect

Can be “pre-trained” without a specific environment.

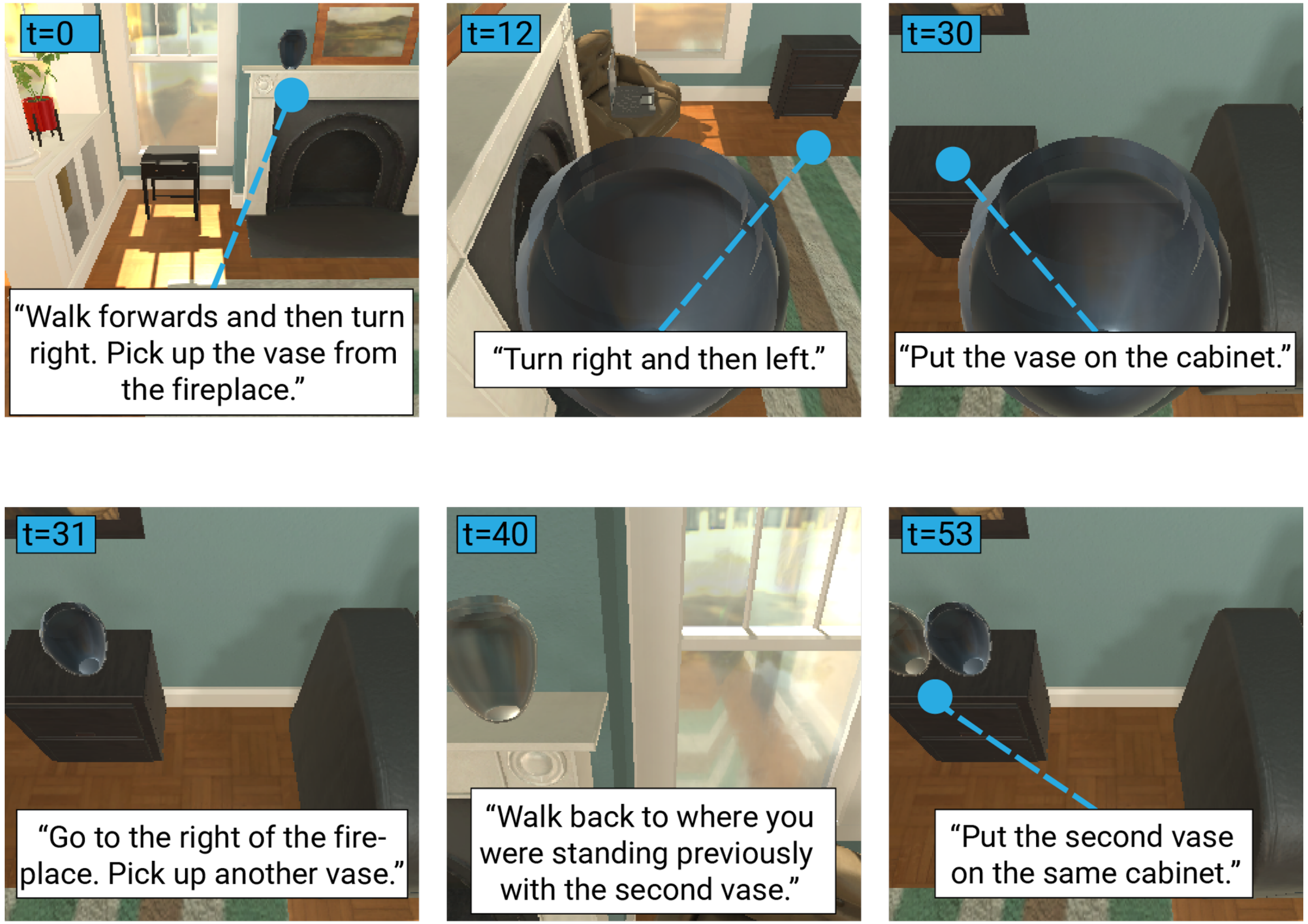
Pashevich, Schmid, and Sun, Episodic Transformer for Vision-and Language Navigation,

Focus Two: Long-term dependencies

Goal: “put two vases on a cabinet”

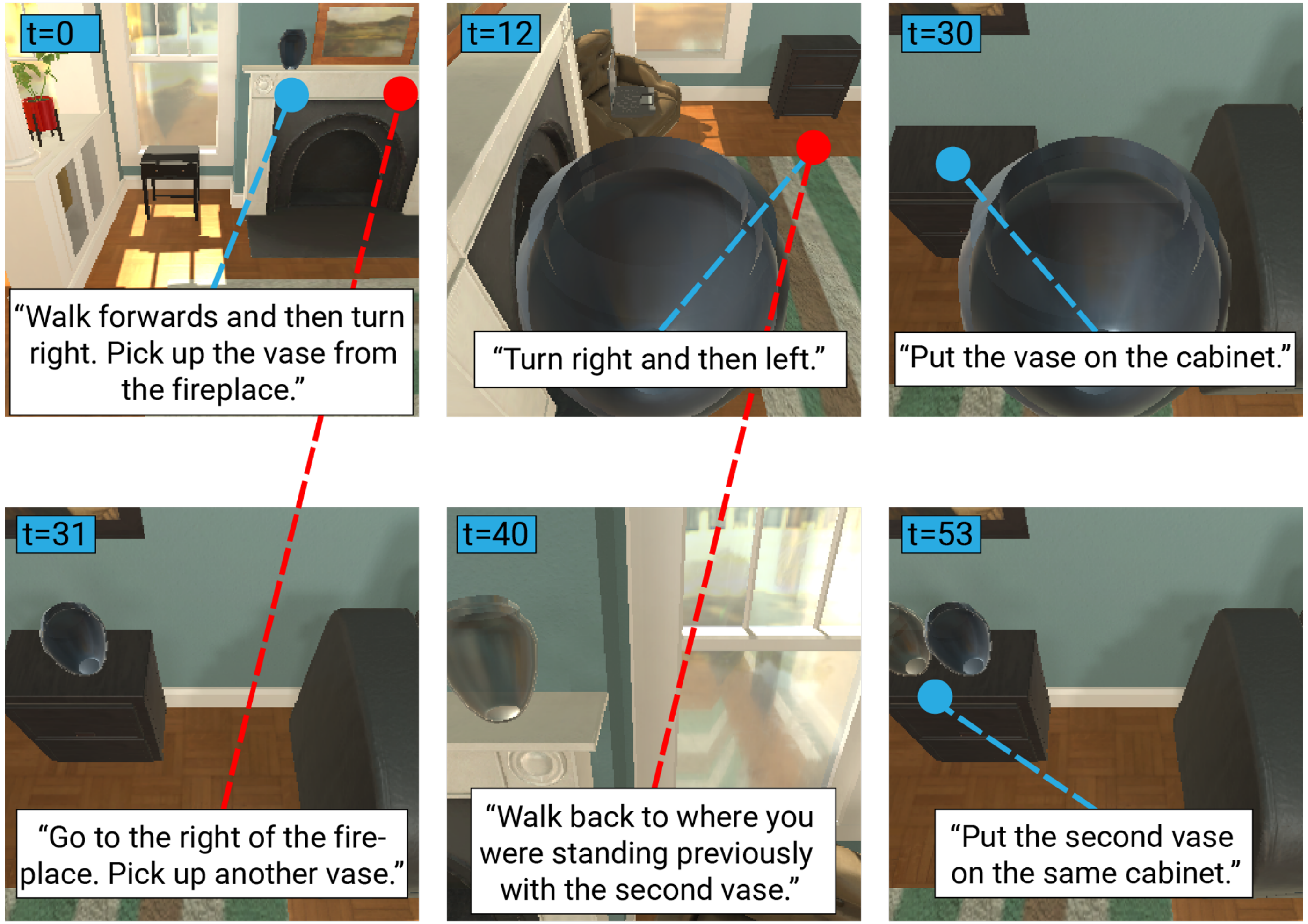
Focus Two: Long-term dependencies

Goal: "put two vases on a cabinet"



Focus Two: Long-term dependencies

Goal: “put two vases on a cabinet”



Results: comparison with state-of-the-art

Model	Validation		Test	
	Seen	Unseen	Seen	Unseen
Shridhar *et al.* [50]	3.70	0.00	3.98	0.39
Nguyen *et al.* [58]	N/A	N/A	12.39	4.45
Singh *et al.* [52]	19.15	3.78	22.05	5.30
E.T. (ours)	33.78	3.17	28.77	5.04
E.T. (ours) + synth. data	**46.59**	**7.32**	**38.42**	**8.57**
Human	-	-	-	91.00

Comparison with state-of-the-art models.

Self-attention to capture long-term dependency

Goal: Grab an apple, cook it and put it in the sink. **Instructions:** Turn to your left twice so that are facing the fridge. Open the fridge, grab an apple from the shelf and close the fridge door. *to the left of the fridge to face the microwave.* Put the apple in the microwave and cook it for a seconds before taking it back out and closing the microwave. Turn to face your left. Put the ap in the sink.

Code and checkpoints are released!

https://github.com/alexpashevich/E.T.

Summary

- Many interesting tasks for detailed video understanding
  - Video is encyclopedia of multimedia contents!
- From manual annotation to “automatic” supervision
  - Self-supervised: Contrastive Learning
  - Cross-modal supervised: Cross-modal cycle consistency
- Many interesting applications of detailed video understanding
  - Structured multimodal representations for navigation
  - Better interpretable, more generalizable models

Collaborators