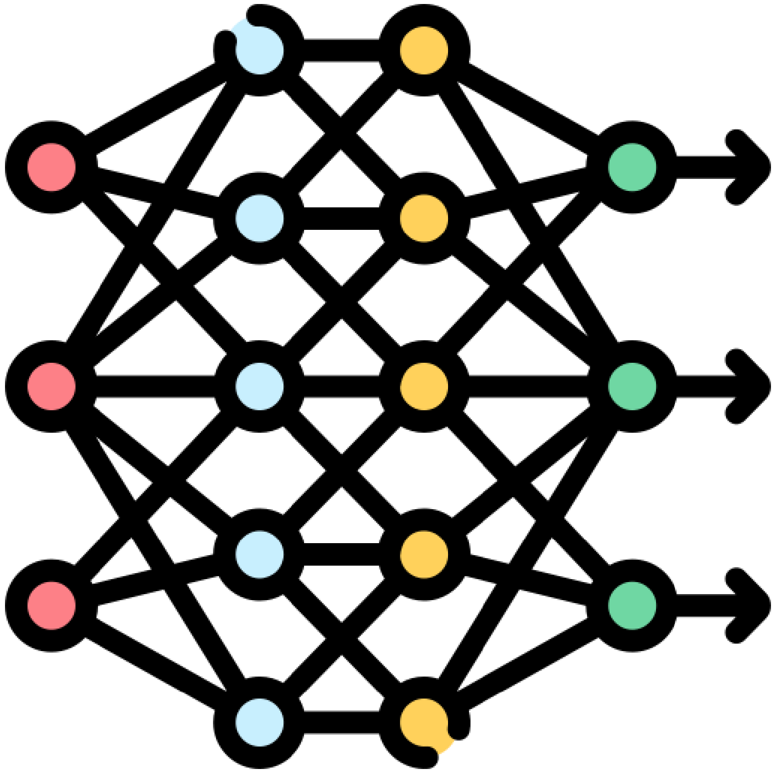
Data and Evaluation in Video Understanding

Hazel Doughty

Universiteit Leiden
The Netherlands

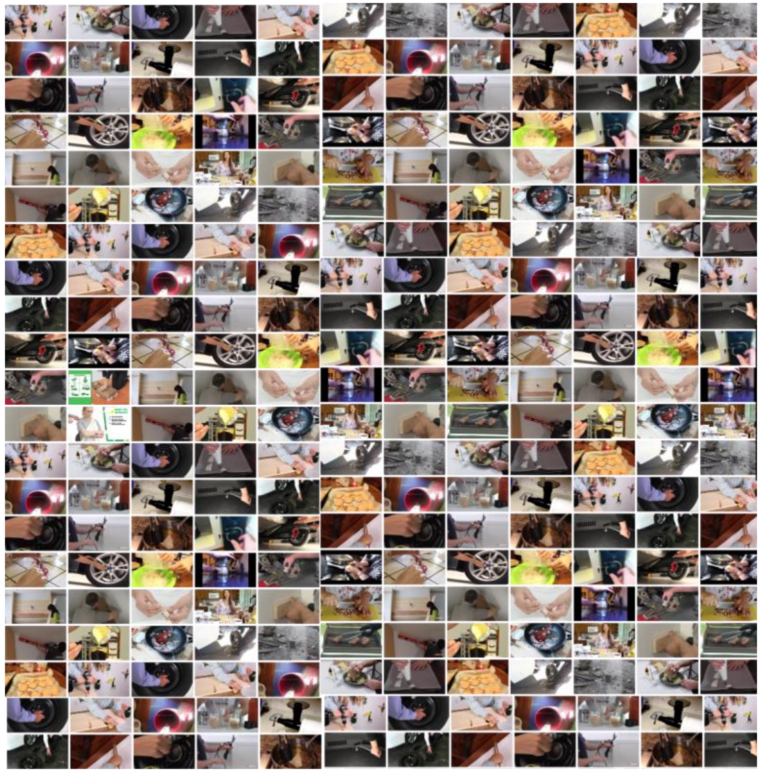
Computer Vision by Learning

Model

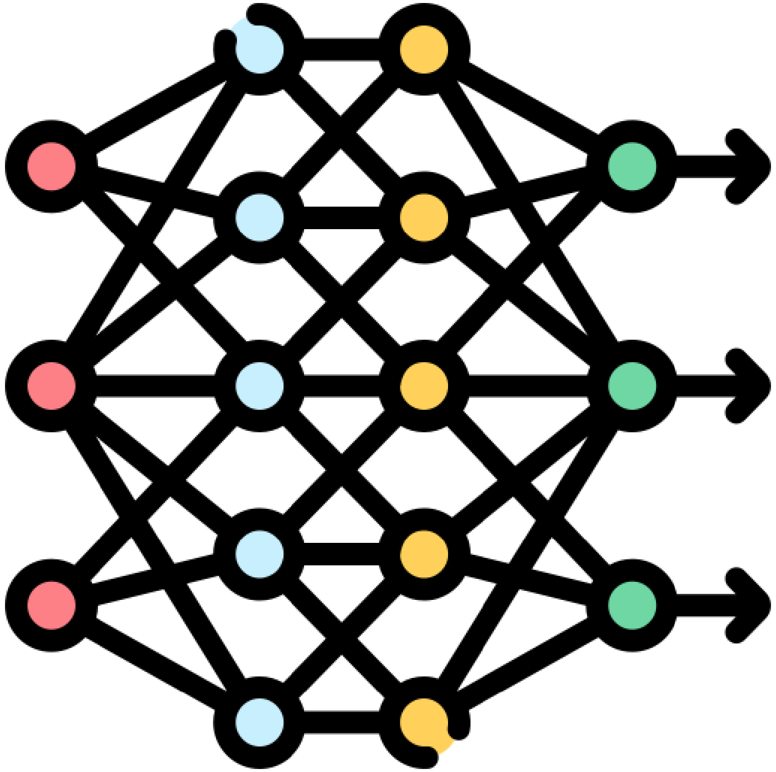


Computer Vision by Learning

Data



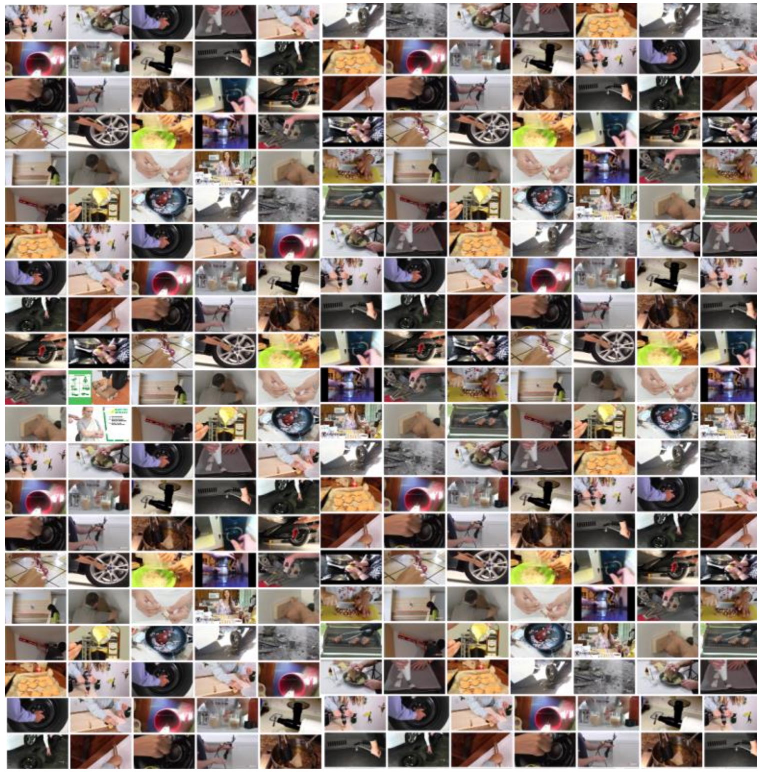
Model



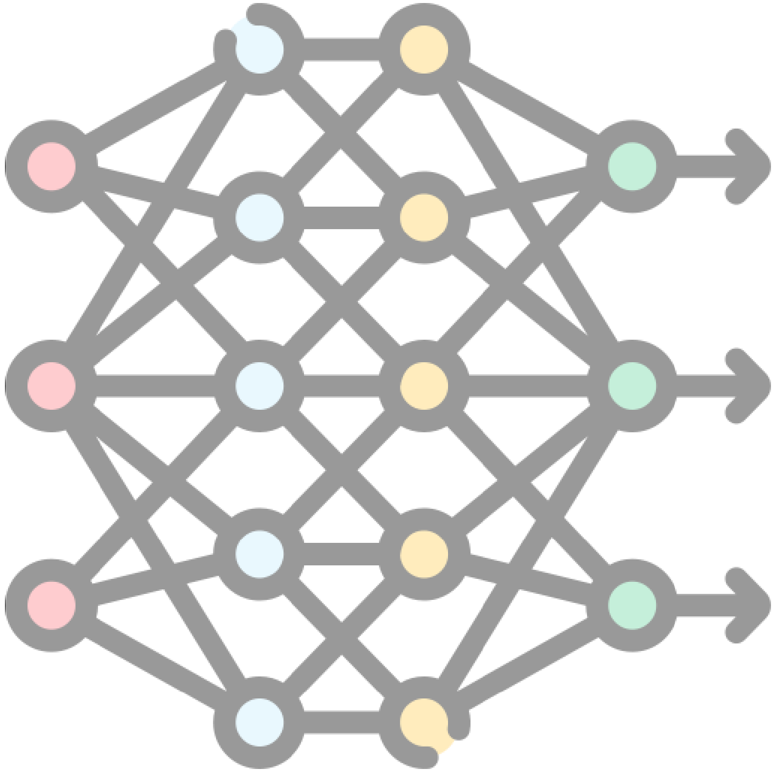
Evaluation

Data & Evaluation are Important

Data



Model



Evaluation

On What Task?

Performing CPR

What?	Press chest	Tilt head	Blow mouth
How?	Press down on chest firmly and regularly	Tilt head backwards carefully	Blow into mouth while pinching nose

Video-to-Text Retrieval

Query:

Retrieved Text:

1. Press chest
2. Tilt head
3. Blow mouth

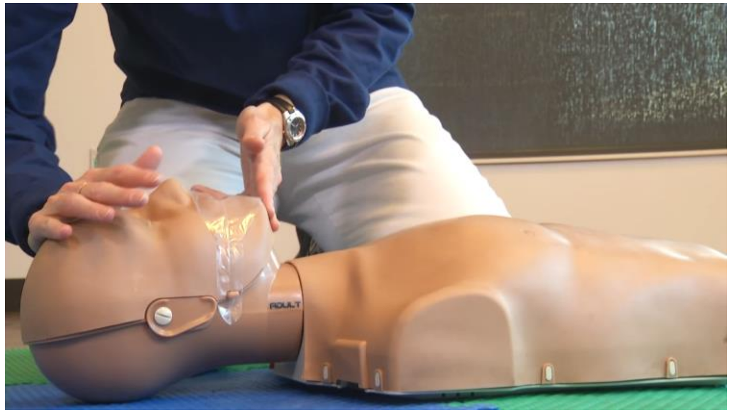
Text-to-Video Retrieval

Query:

Tilt head

Retrieved Videos:

1



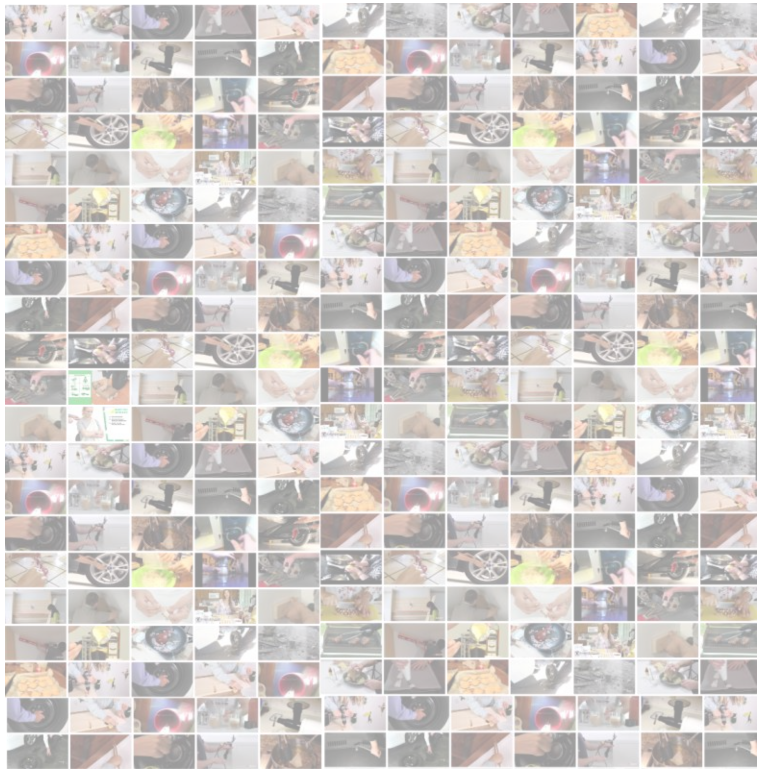
2

…

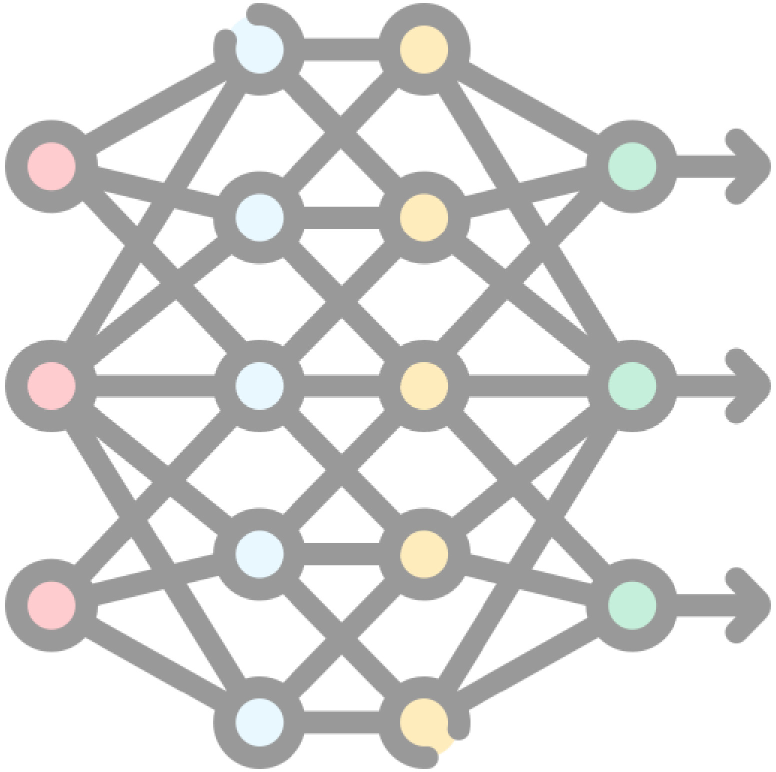
100

Evaluation

Data



Model



Evaluation

Two Main Problems in Video Retrieval

Peel and cut the potatoes

Peel the potatoes and cut them

Problem 1: Captions are very coarse-grained

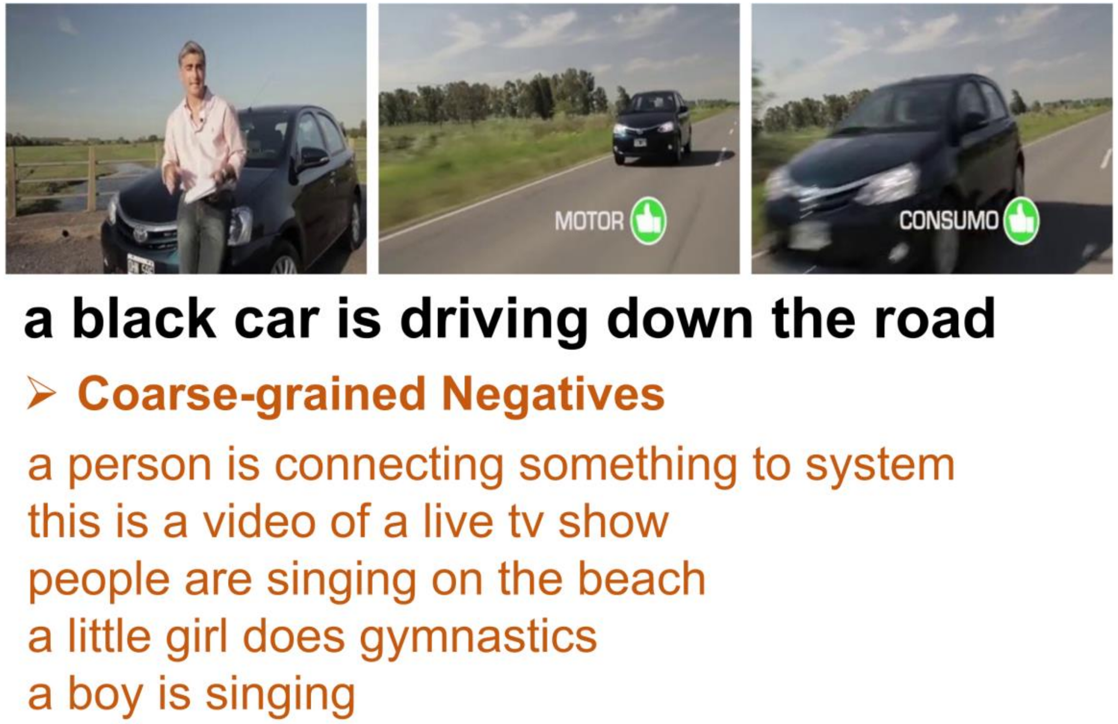
Two Main Problems in Video Retrieval

Peel and cut the potatoes

Peel the potatoes and cut them

Problem 2: Videos only have one ground-truth caption

Video-Text Retrieval Datasets are Coarse-Grained



a black car is driving down the road

➢ **Coarse-grained Negatives**

a person is connecting something to system
this is a video of a live tv show
people are singing on the beach
a little girl does gymnastics
a boy is singing

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

We Go Beyond Coarse-Grained Negatives

a black car is driving down the road

- **Coarse-grained Negatives**

a person is connecting something to system
this is a video of a live tv show
people are singing on the beach
a little girl does gymnastics
a boy is singing

- **Fine-grained Negatives**

a **white** car is driving down the road
a black **motorcycle** is driving down the road
a black car is **parked** down the road
a black car is driving **across** the road
a black car is driving down the **mountain**

PoSRank: Fine-Grained Evaluation

A man wearing a blue t-shirt and black pants, is leaning forward and slowly trimming the hair of another sheep with a trimming machine.

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

PoSRank: Fine-Grained Evaluation

A man wearing a blue t-shirt and black pants, is leaning forward and slowly trimming the hair of another sheep with a trimming machine.

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

PoSRank: Fine-Grained Evaluation

A **man** wearing a blue **t-shirt** and black pants, is leaning forward and slowly trimming the hair of another sheep with a trimming machine.

- ***noun***

A **woman** wearing a blue **t-shirt** and black pants…

A **man** wearing a blue **suit** and black pants….

A **girl** wearing a blue **t-shirt** and black pants….

PoSRank: Fine-Grained Evaluation

A **man** wearing a blue **t-shirt** and black pants, is leaning forward and slowly trimming the hair of another sheep with a trimming machine.

- ***noun***

A **woman** wearing a blue **t-shirt** and black pants…

A **man** wearing a blue **suit** and black pants….

A **girl** wearing a blue **t-shirt** and black pants….

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

PoSRank: Fine-Grained Evaluation

Caption t_i :

A **man** **wearing** a **blue** t-shirt and black pants, is leaning **forward** and **slowly** trimming the hair of another sheep with a trimming machine.

● *noun*
A **woman** wearing a blue **t-shirt** and black pants...
A **man** wearing a blue **suit** and black pants...
A **girl** wearing a blue **t-shirt** and black pants...

● *verb*
A man **eating** a blue t-shirt...
A man **folding** a blue t-shirt...
A man **busting** a blue t-shirt...

● *adjective*
...wearing a **green** t-shirt...
...wearing a **black** t-shirt...
...wearing a **fake** t-shirt...

● *preposition*
...is leaning **away**...
...is leaning **backward**...
...is leaning **back**...

● *adverb*
...and **rapidly** trimming the hair...
...and **speedily** trimming the hair...
...and **quickly** trimming the hair...

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

PoSRank: Fine-Grained Evaluation

Caption t_i :

A **man** **wearing** a **blue** t-shirt and black pants, is leaning **forward** and **slowly** trimming the hair of another sheep with a trimming machine.

● *noun*

A **woman** wearing a blue **t-shirt** and black pants...
A **man** wearing a blue **suit** and black pants...
A **girl** wearing a blue **t-shirt** and black pants...

● *verb*

A man **eating** a blue t-shirt...
A man **folding** a blue t-shirt...
A man **busting** a blue t-shirt...

● *adjective*

...wearing a **green** t-shirt...
...wearing a **black** t-shirt...
...wearing a **fake** t-shirt...

● *preposition*

...is leaning **away**...
...is leaning **backward**...
...is leaning **back**...

● *adverb*

...and **rapidly** trimming the hair...
...and **speedily** trimming the hair...
...and **quickly** trimming the hair...

Caption t_i :

A **person** is **putting** liquid in a cup with a **white** mug, moves his right hand **backward**, and then puts something in a cup with a **white** bowl.

● *verb*

A person is **divesting** liquid..., **moves** his right hand backward...
A person is **putting** liquid..., **keeps** his right hand backward...
A person is **putting** liquid..., and then **removes** something...

● *noun*

A **dog** is putting liquid...
A **machine** is putting...
A **shape** is putting...

● *adjective*

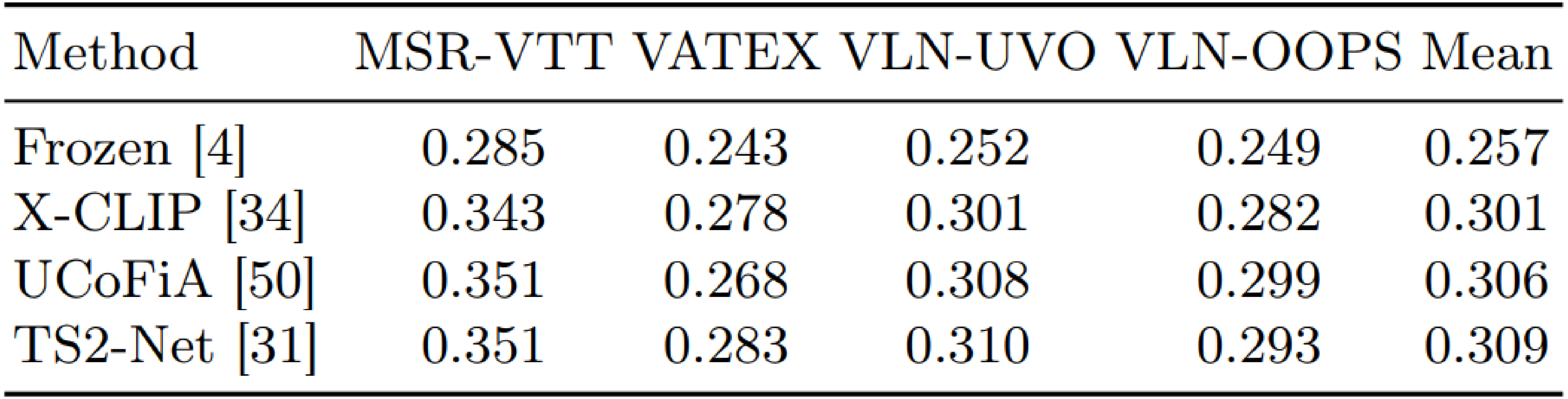
...with a **black** mug...
...a cup with a **black** bowl.
...with a **colorful** mug...

● *preposition*

...moves his right hand **forward**...
...moves his right hand **ahead**...
...moves his right hand **away**...

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

How Well Do Current Models Understand Fine-Grained Differences?

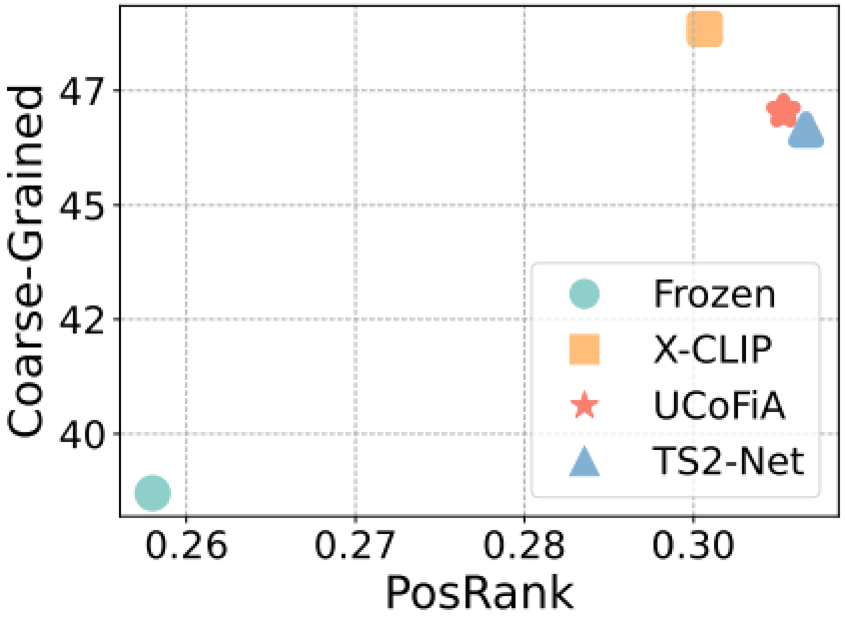


Method	MSR-VTT	VATEX	VLN-UVO	VLN-OOPS	Mean
Frozen [4]	0.285	0.243	0.252	0.249	0.257
X-CLIP [34]	0.343	0.278	0.301	0.282	0.301
UCoFiA [50]	0.351	0.268	0.308	0.299	0.306
TS2-Net [31]	0.351	0.283	0.310	0.293	0.309

Max score is 1.0

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

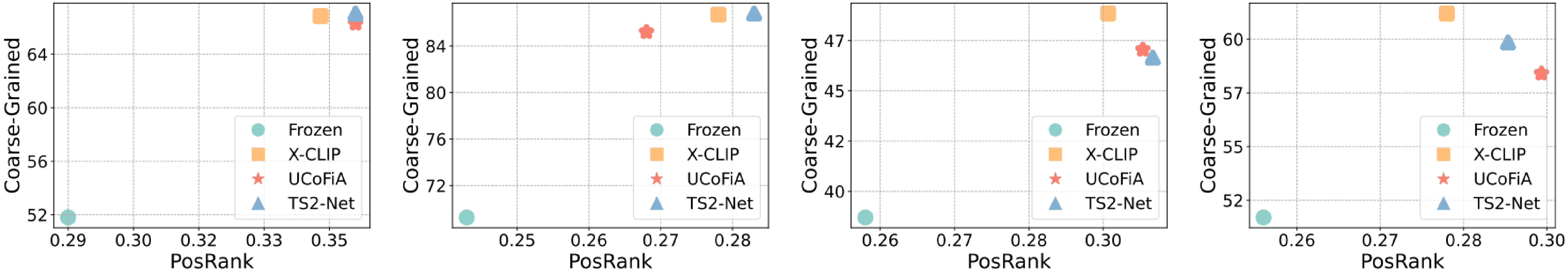
Does Performance in Fine-Grained Retrieval Correlate with Existing Coarse-Grained Metrics?



(c) VLN-UVO

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

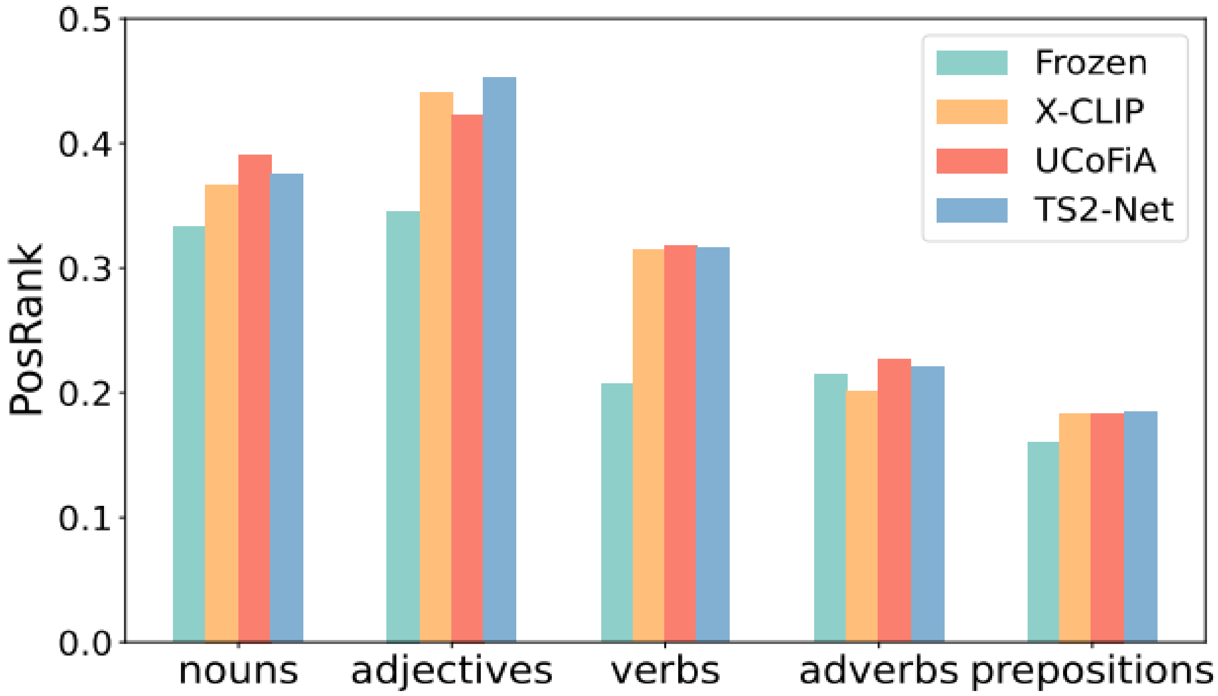
Does Performance in Fine-Grained Retrieval Correlate with Existing Coarse-Grained Metrics?



(a) MSR-VTT (b) VATEX (c) VLN-UVO (d) VLN-OOPS

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

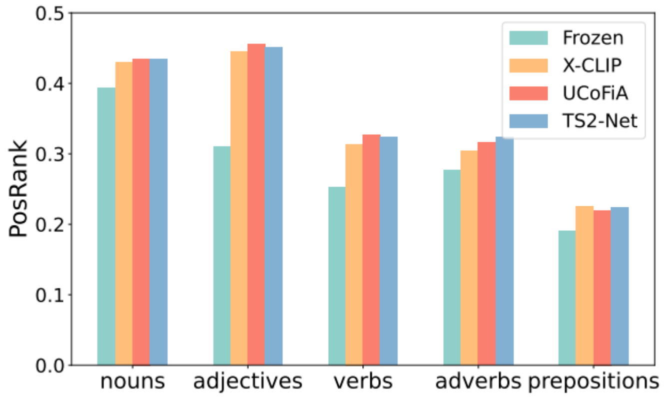
Are Certain Parts-of-Speech More Challenging Than Others?



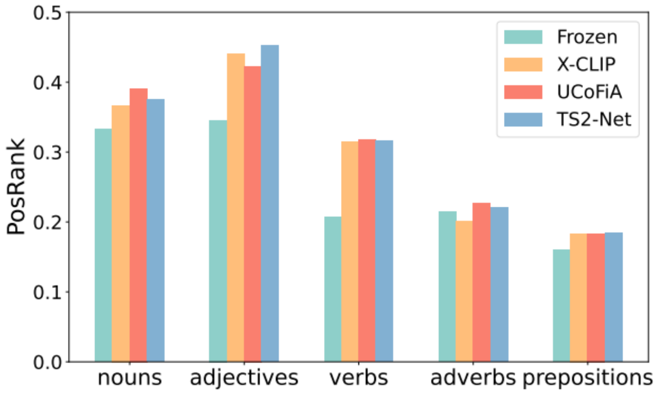
VLN-UVO

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

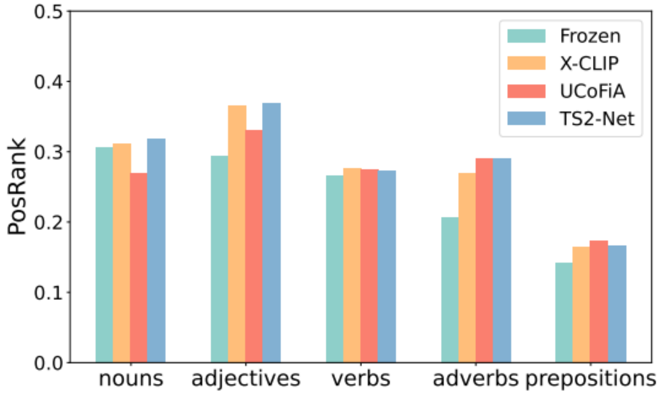
Are Certain Parts-of-Speech More Challenging Than Others?



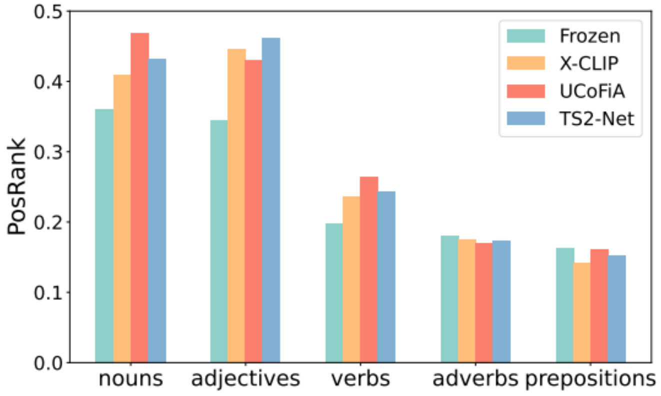
MSR-VTT



VLN-UVO

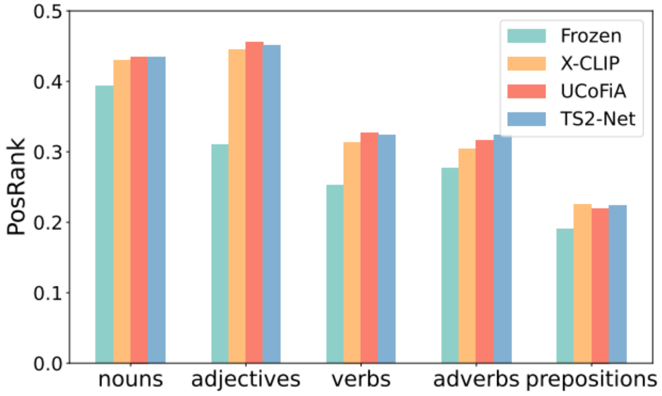


VATEX

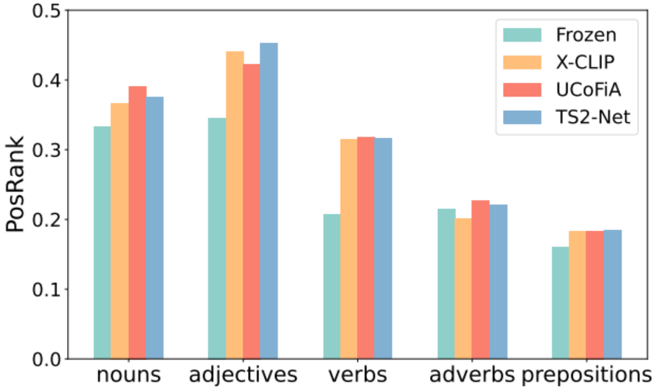


VLN-OOPS

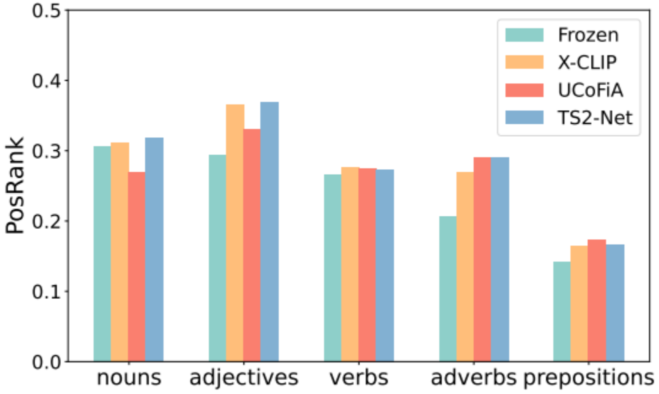
Are Certain Datasets More Suitable than Others for Fine-Grained Evaluation?



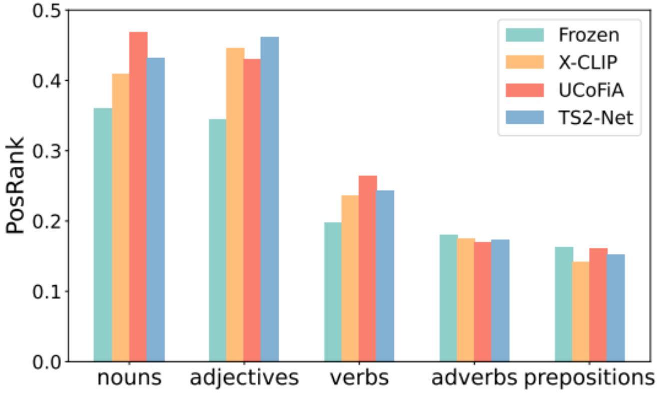
MSR-VTT



VLN-UVO



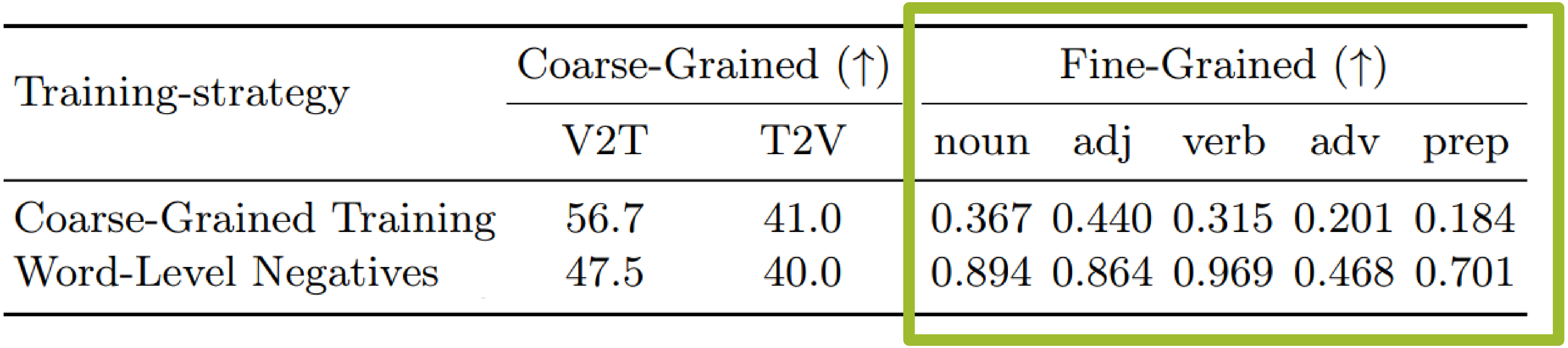
VATEX



VLN-OOPS

How Do We Improve This?

- Add fine-grained word-level negatives in training?

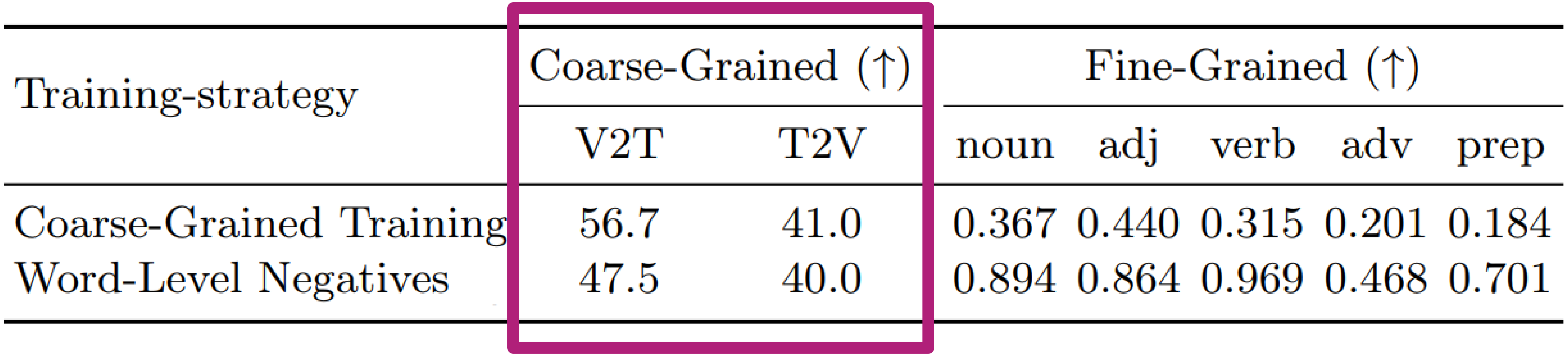


Training-strategy	Coarse-Grained (↑)		Fine-Grained (↑)				
	V2T	T2V	noun	adj	verb	adv	prep
Coarse-Grained Training	56.7	41.0	0.367	0.440	0.315	0.201	0.184
Word-Level Negatives	47.5	40.0	0.894	0.864	0.969	0.468	0.701

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

How Do We Improve This?

- Add fine-grained word-level negatives in training?



Training-strategy	Coarse-Grained (↑)		Fine-Grained (↑)				
	V2T	T2V	noun	adj	verb	adv	prep
Coarse-Grained Training	56.7	41.0	0.367	0.440	0.315	0.201	0.184
Word-Level Negatives	47.5	40.0	0.894	0.864	0.969	0.468	0.701

☹

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

How Do We Improve This?

- Phrase-level Negatives

Original Caption
A boy wearing a black t-shirt is tossing a basketball then throws the basketball towards the other boy

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

How Do We Improve This?

- Phrase-level Negatives

Original Caption
A boy wearing a black t-shirt is tossing a basketball then throws the basketball towards the other boy

Phrase-Level Negatives
A boy wearing a black t-shirt is **holding** a **football** then throws the basketball towards the other boy

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

How Do We Improve This?

- Phrase-level Negatives

Original Caption
A boy wearing a black t-shirt is tossing a basketball then throws the basketball towards the other boy

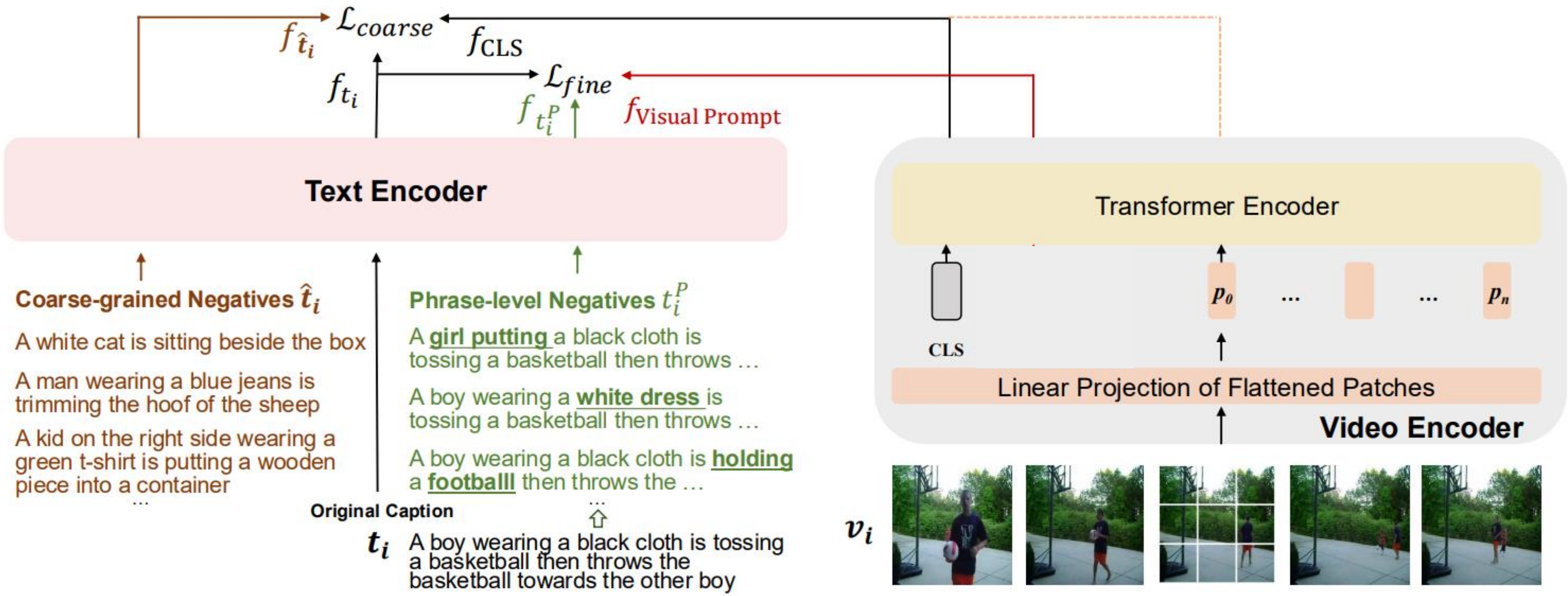
Phrase-Level Negatives
A boy wearing a black t-shirt is **holding** a **football** then throws the basketball towards the other boy

or

A boy wearing a black t-shirt is Holding a football then throws the basketball **away from** the other **girl**

How Do We Improve This?

- Fine-Grained Prompting



$\mathcal{L}_{coarse}$

$f_{\hat{t}_i}$

f_{CLS}

$\mathcal{L}_{fine}$

f_{t_i}

$f_{t_i^P}$

$f_{\text{Visual Prompt}}$

Text Encoder

Transformer Encoder

CLS

p_0 … … p_n

Linear Projection of Flattened Patches

Video Encoder

Coarse-grained Negatives $\hat{t}_i$

A white cat is sitting beside the box

A man wearing a blue jeans is trimming the hoof of the sheep

A kid on the right side wearing a green t-shirt is putting a wooden piece into a container

…

Phrase-level Negatives t_i^P

A **girl putting** a black cloth is tossing a basketball then throws …

A boy wearing a **white dress** is tossing a basketball then throws …

A boy wearing a black cloth is **holding** a **footballl** then throws the …

…

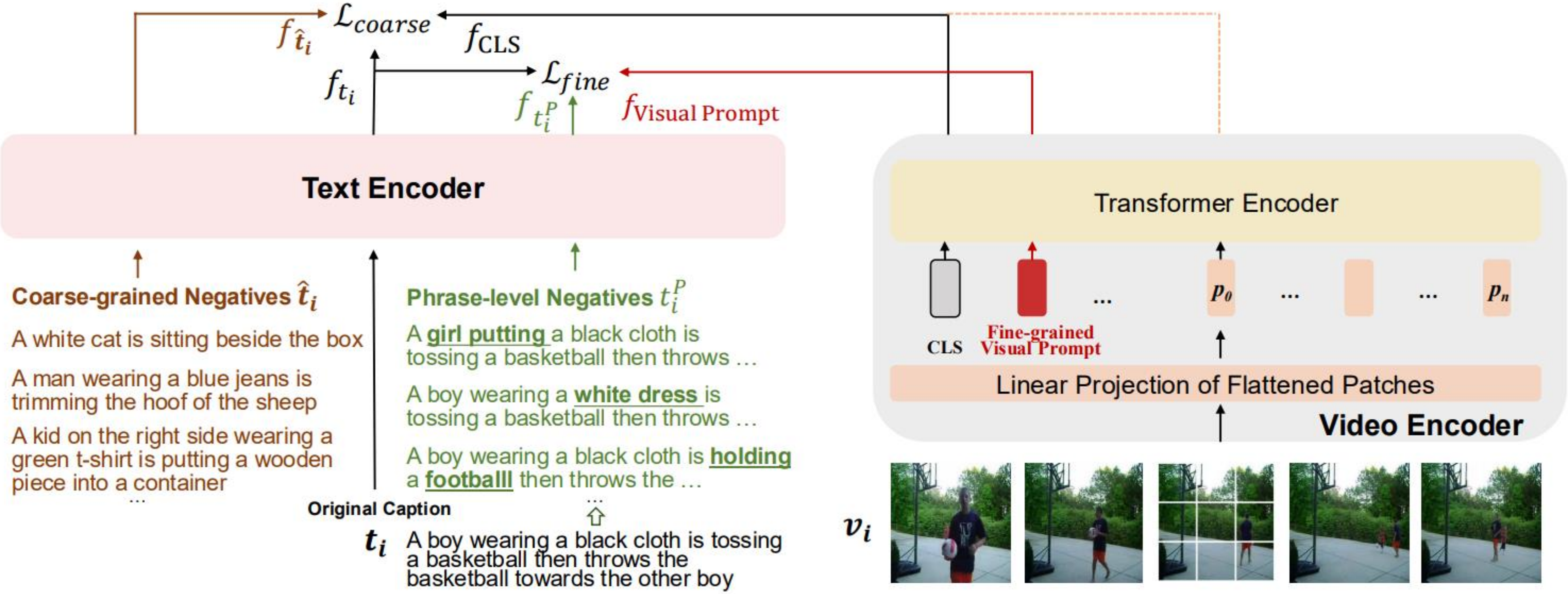
Original Caption

t_i A boy wearing a black cloth is tossing a basketball then throws the basketball towards the other boy

v_i

How Do We Improve This?

- Fine-Grained Prompting



Coarse-grained Negatives $\hat{t}_i$

A white cat is sitting beside the box

A man wearing a blue jeans is trimming the hoof of the sheep

A kid on the right side wearing a green t-shirt is putting a wooden piece into a container

...

Phrase-level Negatives t_i^P

A **girl putting** a black cloth is tossing a basketball then throws ...

A boy wearing a **white dress** is tossing a basketball then throws ...

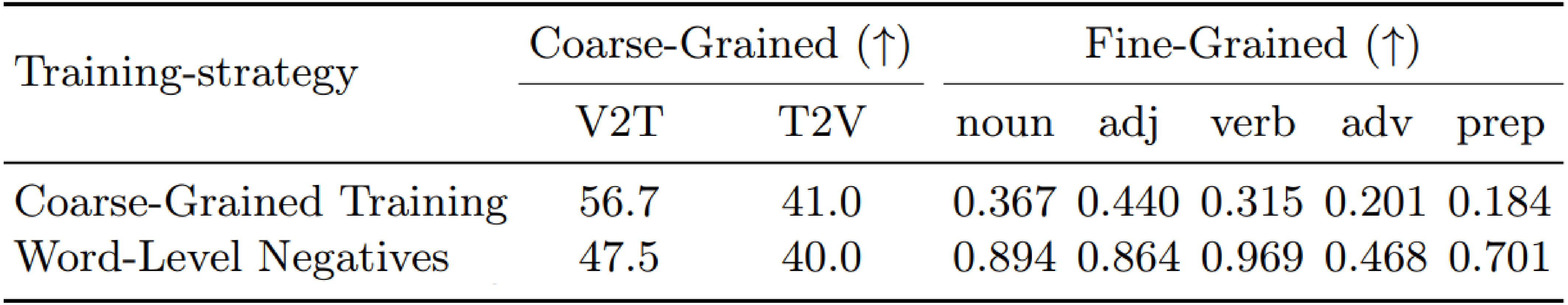
A boy wearing a black cloth is **holding** a **footballl** then throws the ...

...

Original Caption

t_i A boy wearing a black cloth is tossing a basketball then throws the basketball towards the other boy

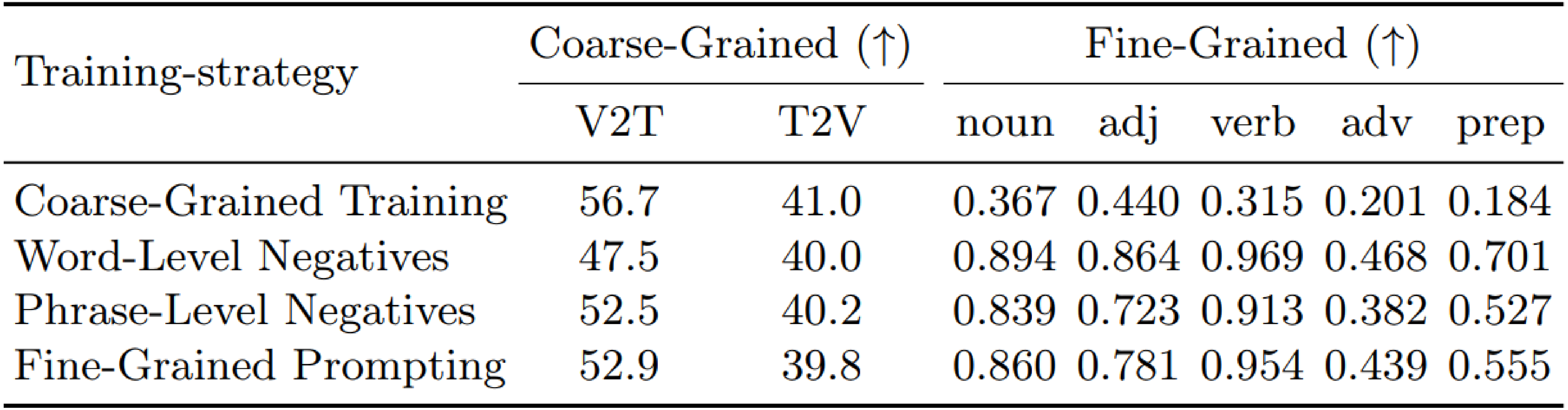
Results



Training-strategy	Coarse-Grained (↑)		Fine-Grained (↑)				
	V2T	T2V	noun	adj	verb	adv	prep
Coarse-Grained Training	56.7	41.0	0.367	0.440	0.315	0.201	0.184
Word-Level Negatives	47.5	40.0	0.894	0.864	0.969	0.468	0.701

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

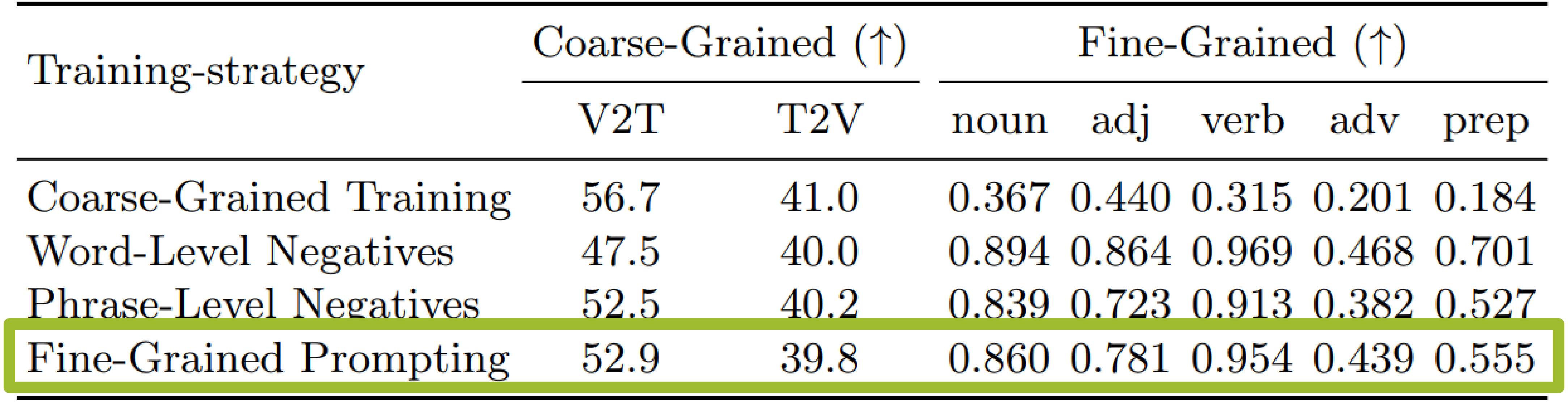
Results



Training-strategy	Coarse-Grained (↑)		Fine-Grained (↑)				
	V2T	T2V	noun	adj	verb	adv	prep
Coarse-Grained Training	56.7	41.0	0.367	0.440	0.315	0.201	0.184
Word-Level Negatives	47.5	40.0	0.894	0.864	0.969	0.468	0.701
Phrase-Level Negatives	52.5	40.2	0.839	0.723	0.913	0.382	0.527
Fine-Grained Prompting	52.9	39.8	0.860	0.781	0.954	0.439	0.555

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

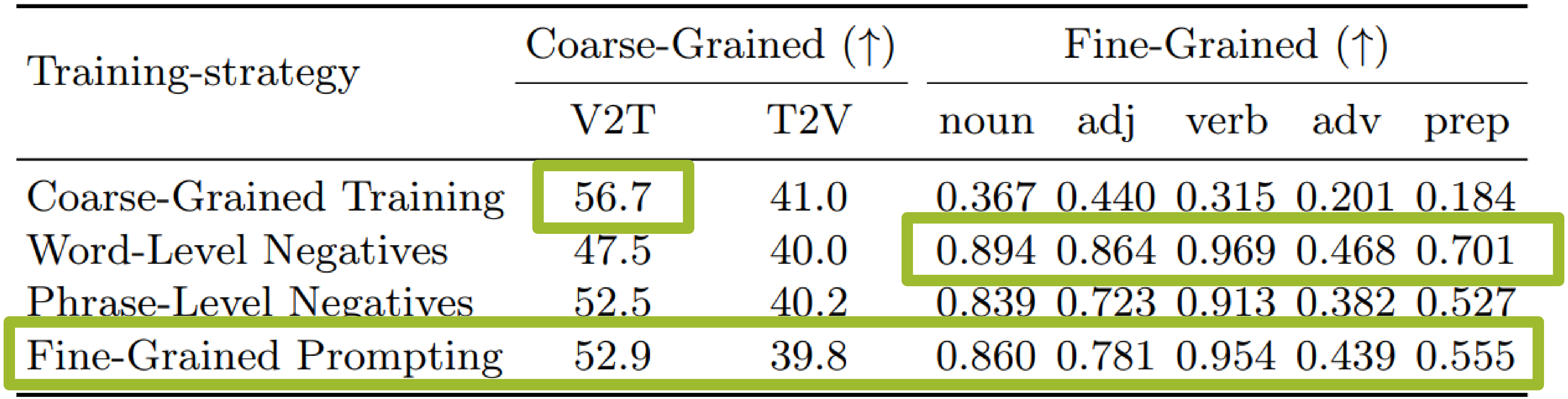
Results



Training-strategy	Coarse-Grained (↑)		Fine-Grained (↑)				
	V2T	T2V	noun	adj	verb	adv	prep
Coarse-Grained Training	56.7	41.0	0.367	0.440	0.315	0.201	0.184
Word-Level Negatives	47.5	40.0	0.894	0.864	0.969	0.468	0.701
Phrase-Level Negatives	52.5	40.2	0.839	0.723	0.913	0.382	0.527
Fine-Grained Prompting	52.9	39.8	0.860	0.781	0.954	0.439	0.555

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

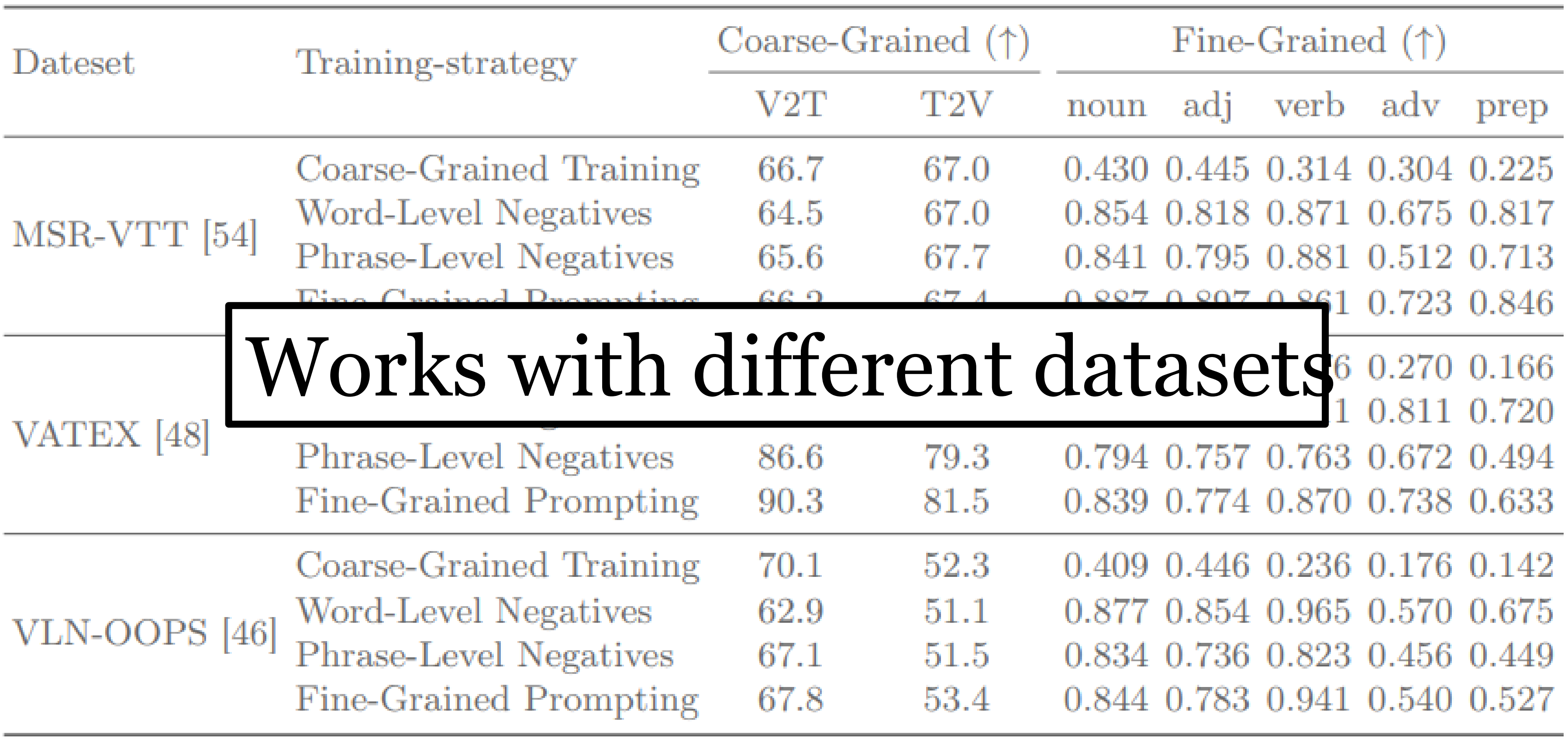
Results



Training-strategy	Coarse-Grained (↑)		Fine-Grained (↑)				
	V2T	T2V	noun	adj	verb	adv	prep
Coarse-Grained Training	56.7	41.0	0.367	0.440	0.315	0.201	0.184
Word-Level Negatives	47.5	40.0	0.894	0.864	0.969	0.468	0.701
Phrase-Level Negatives	52.5	40.2	0.839	0.723	0.913	0.382	0.527
Fine-Grained Prompting	52.9	39.8	0.860	0.781	0.954	0.439	0.555

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

Results

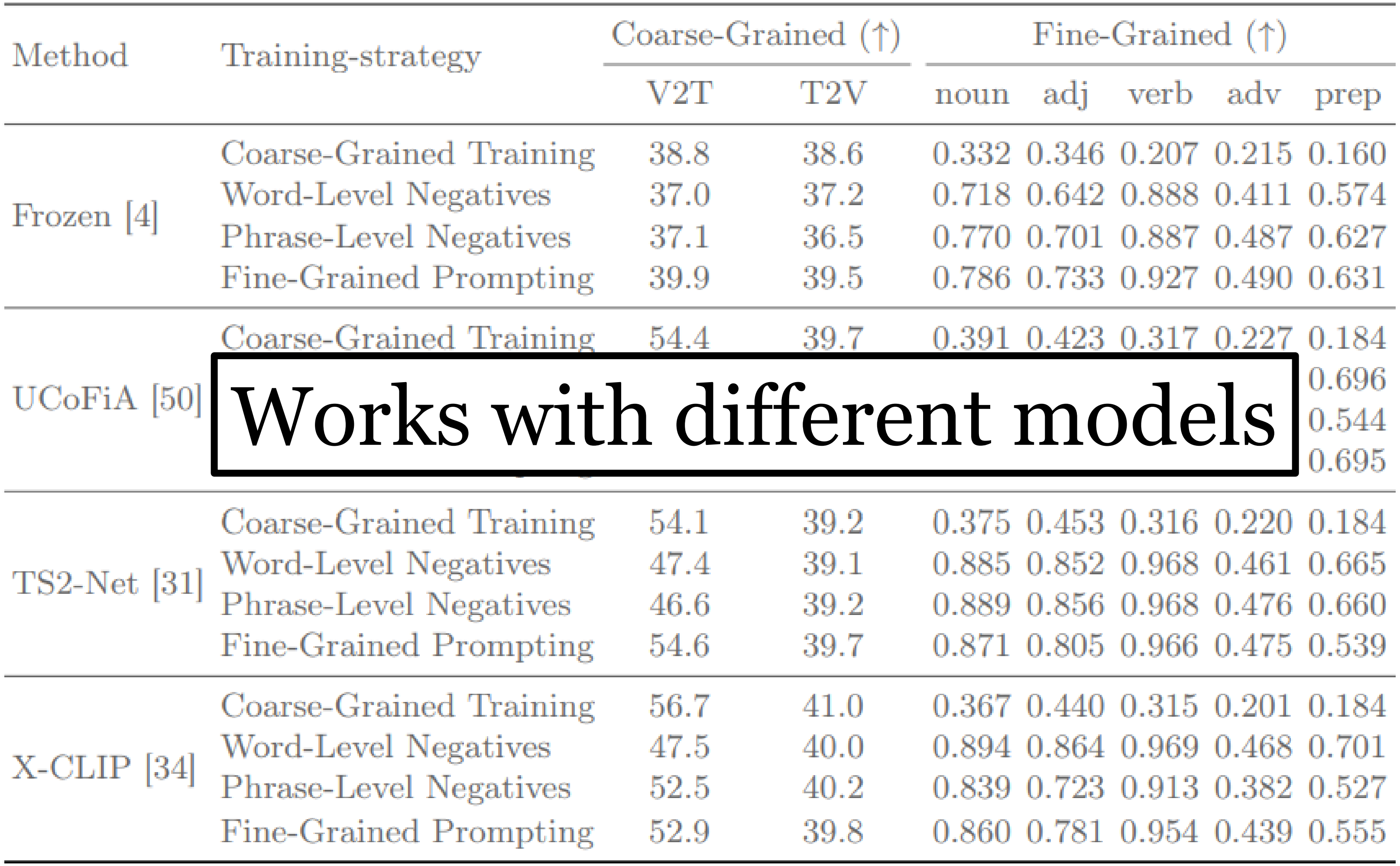


Dateset	Training-strategy	Coarse-Grained (↑)		Fine-Grained (↑)				
		V2T	T2V	noun	adj	verb	adv	prep
MSR-VTT [54]	Coarse-Grained Training	66.7	67.0	0.430	0.445	0.314	0.304	0.225
	Word-Level Negatives	64.5	67.0	0.854	0.818	0.871	0.675	0.817
	Phrase-Level Negatives	65.6	67.7	0.841	0.795	0.881	0.512	0.713
	Fine-Grained Prompting	66.2	67.4	0.887	0.897	0.861	0.723	0.846
VATEX [48]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.270	0.166
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.811	0.720
	Phrase-Level Negatives	86.6	79.3	0.794	0.757	0.763	0.672	0.494
	Fine-Grained Prompting	90.3	81.5	0.839	0.774	0.870	0.738	0.633
VLN-OOPS [46]	Coarse-Grained Training	70.1	52.3	0.409	0.446	0.236	0.176	0.142
	Word-Level Negatives	62.9	51.1	0.877	0.854	0.965	0.570	0.675
	Phrase-Level Negatives	67.1	51.5	0.834	0.736	0.823	0.456	0.449
	Fine-Grained Prompting	67.8	53.4	0.844	0.783	0.941	0.540	0.527

Works with different datasets

Aozhu Chen, Hazel Doughty, Xirong Li, Cees G. M. Snoek, Beyond Coarse-Grained Matching in Video-Text Retrieval. ACCV 2024.

Results



Method	Training-strategy	Coarse-Grained (↑)		Fine-Grained (↑)				
		V2T	T2V	noun	adj	verb	adv	prep
Frozen [4]	Coarse-Grained Training	38.8	38.6	0.332	0.346	0.207	0.215	0.160
	Word-Level Negatives	37.0	37.2	0.718	0.642	0.888	0.411	0.574
	Phrase-Level Negatives	37.1	36.5	0.770	0.701	0.887	0.487	0.627
	Fine-Grained Prompting	39.9	39.5	0.786	0.733	0.927	0.490	0.631
UCoFiA [50]	Coarse-Grained Training	54.4	39.7	0.391	0.423	0.317	0.227	0.184
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.696
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.544
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.695
TS2-Net [31]	Coarse-Grained Training	54.1	39.2	0.375	0.453	0.316	0.220	0.184
	Word-Level Negatives	47.4	39.1	0.885	0.852	0.968	0.461	0.665
	Phrase-Level Negatives	46.6	39.2	0.889	0.856	0.968	0.476	0.660
	Fine-Grained Prompting	54.6	39.7	0.871	0.805	0.966	0.475	0.539
X-CLIP [34]	Coarse-Grained Training	56.7	41.0	0.367	0.440	0.315	0.201	0.184
	Word-Level Negatives	47.5	40.0	0.894	0.864	0.969	0.468	0.701
	Phrase-Level Negatives	52.5	40.2	0.839	0.723	0.913	0.382	0.527
	Fine-Grained Prompting	52.9	39.8	0.860	0.781	0.954	0.439	0.555

Works with different models

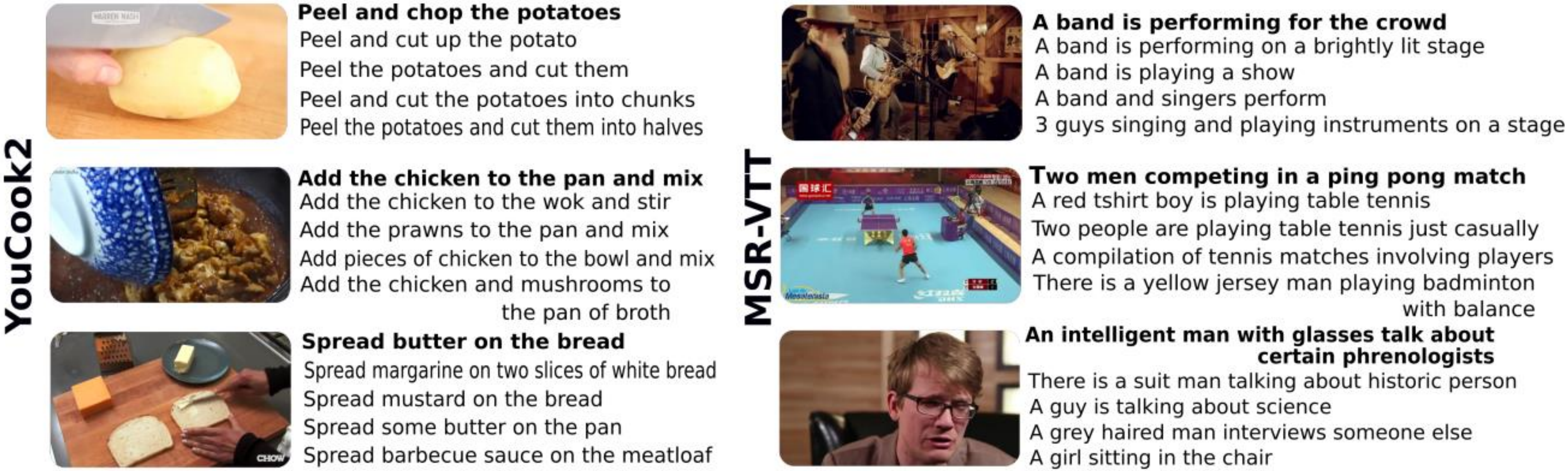
Two Main Problems in Video Retrieval

Peel and cut the potatoes

Peel the potatoes and cut them

Problem 2: Videos only have one ground-truth caption

Multiple Relevant Videos



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

Add the chicken to the pan and mix
Add the chicken to the wok and stir
Add the prawns to the pan and mix
Add pieces of chicken to the bowl and mix
Add the chicken and mushrooms to the pan of broth

Spread butter on the bread
Spread margarine on two slices of white bread
Spread mustard on the bread
Spread some butter on the pan
Spread barbecue sauce on the meatloaf

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

Two men competing in a ping pong match
A red tshirt boy is playing table tennis
Two people are playing table tennis just casually
A compilation of tennis matches involving players
There is a yellow jersey man playing badminton with balance

An intelligent man with glasses talk about certain phrenologists
There is a suit man talking about historic person
A guy is talking about science
A grey haired man interviews someone else
A girl sitting in the chair

Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

Multiple Relevant Videos

Problem: Current metrics don’t account for multiple relevant videos

Solution?

Let’s find a metric that does

nDCG = normalized discounted cumulative gain

Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

Multiple Relevant Videos

Problem: Current annotations don’t account for multiple relevant videos

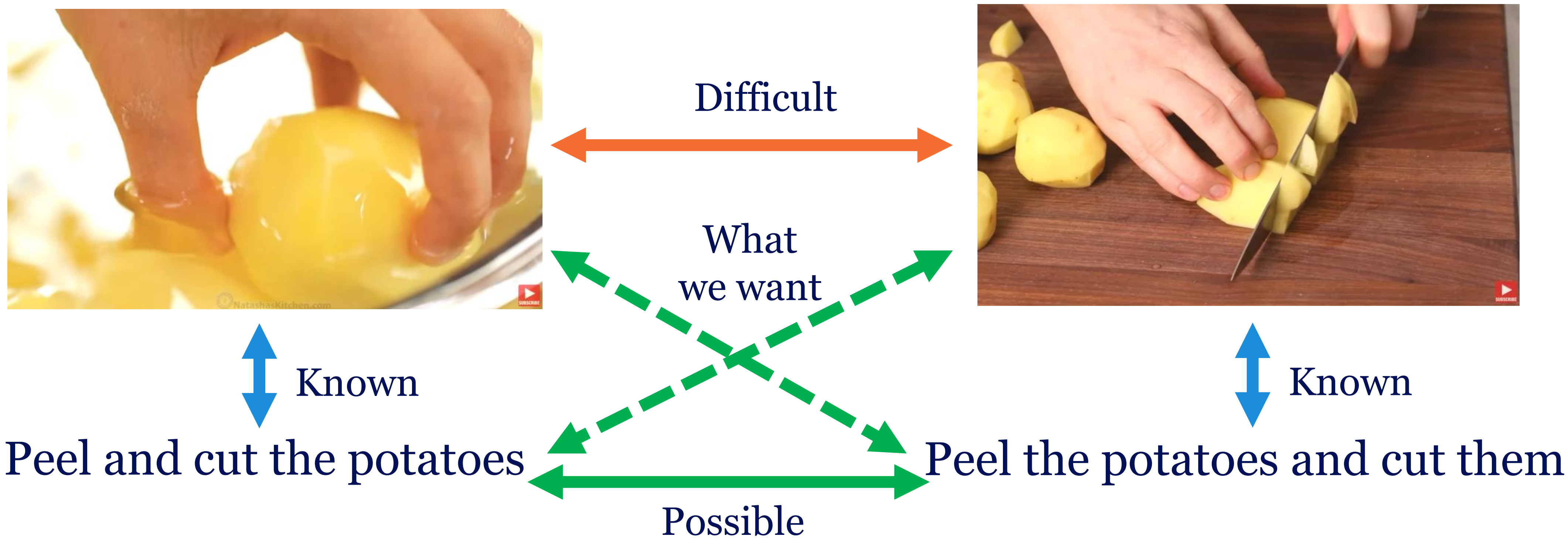
Solution?

~~Naïve solution: Annotate the relevance of each caption to all videos~~

Better solution: Approximate the relevance of each caption to all videos with information we already have

Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

Semantic Similarity



Difficult

What
we want

Known

Known

Peel and cut the potatoes

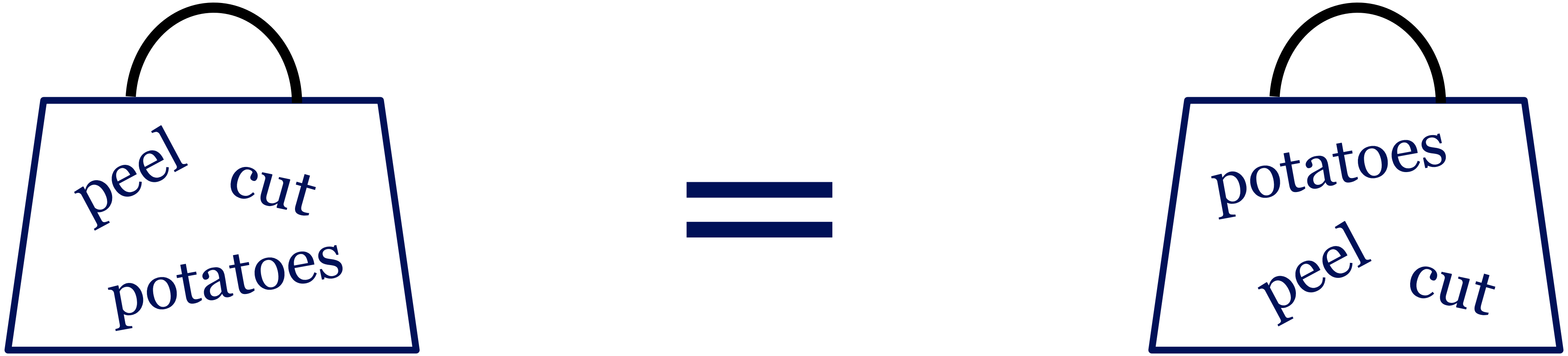
Peel the potatoes and cut them

Possible

Bag of Words Semantic Similarity

Peel a✗d cut t✗e potatoes

Peel t✗e potatoes a✗d cut th✗m

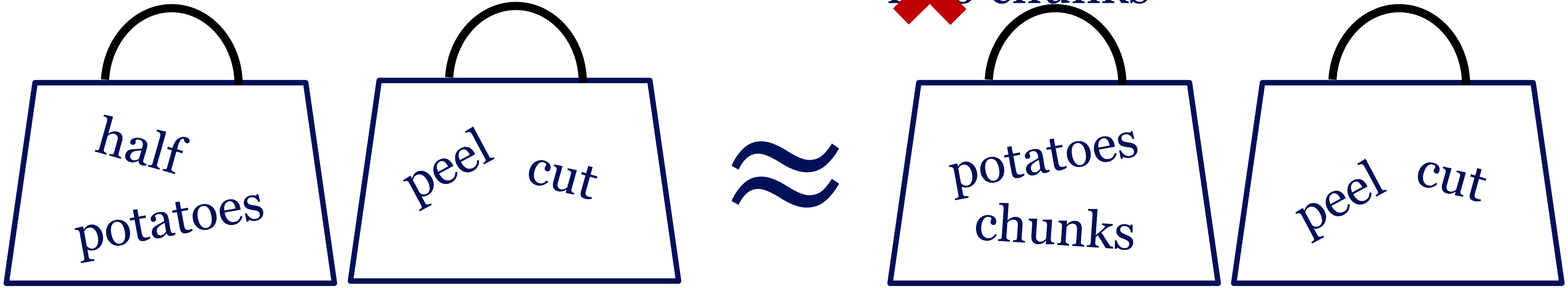


Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

Part-of-Speech Semantic Similarity

Peel ~~and~~ cut ~~the~~ potatoes ~~in~~ half

Peel ~~the~~ potatoes ~~and~~ cut ~~them~~ ~~into~~ chunks



half potatoes

peel cut

≈

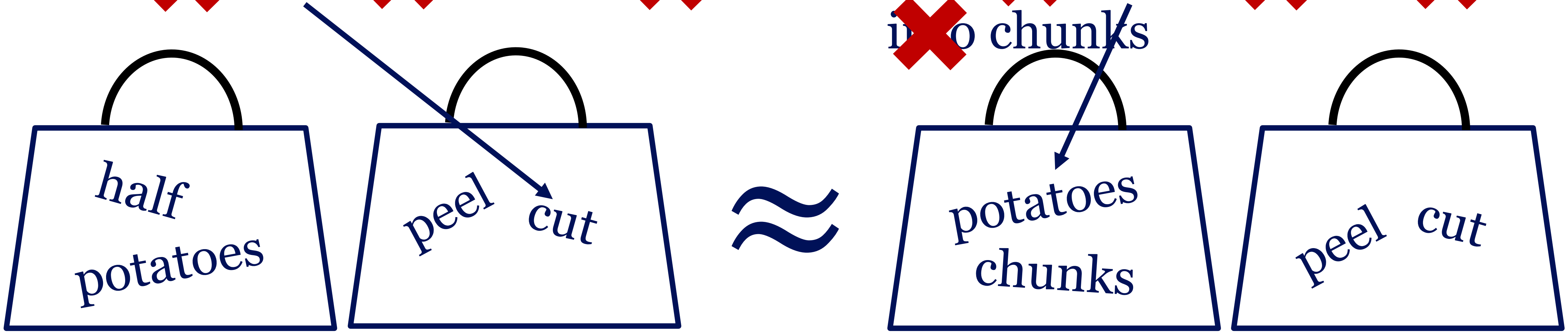
potatoes chunks

peel cut

Synset Semantic Similarity

Peel ~~and~~ chop ~~the~~ potatoes ~~in~~ half

Peel ~~the~~ spuds ~~and~~ cut ~~them~~ ~~into~~ chunks



half potatoes

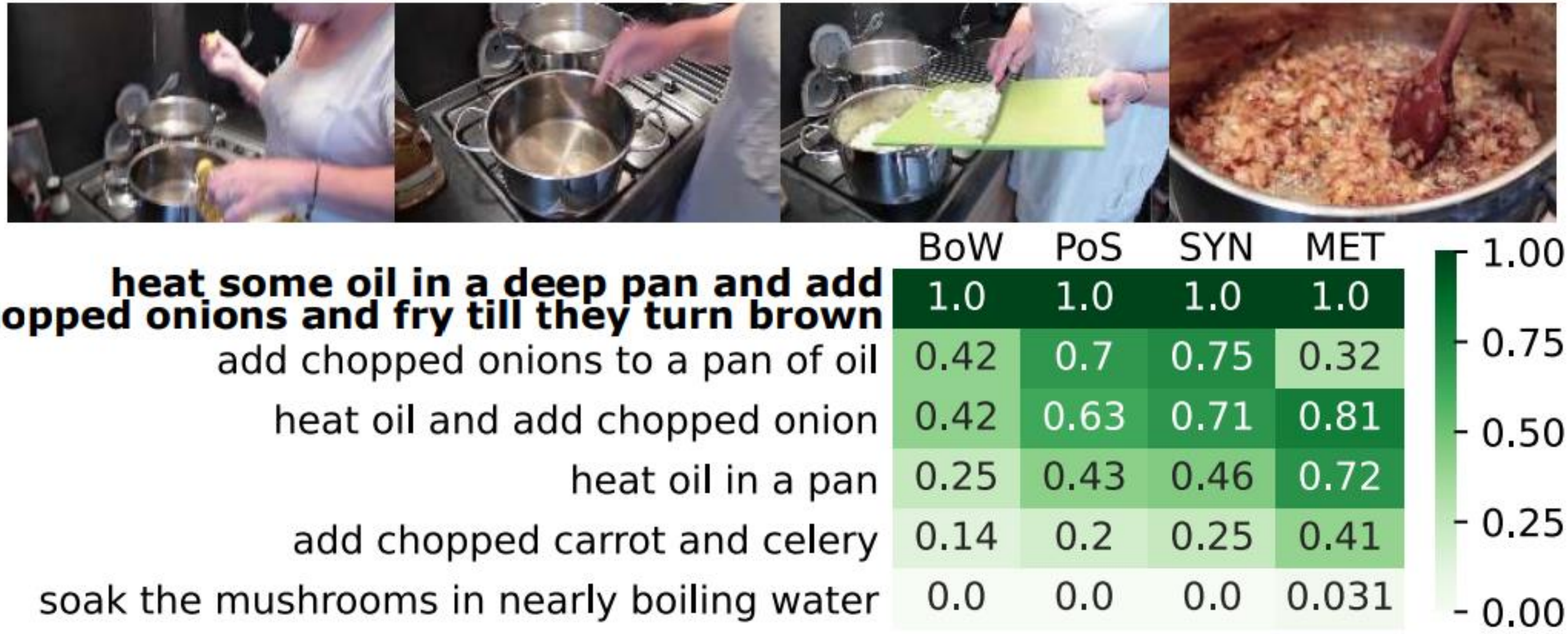
peel cut

≈

potatoes chunks

peel cut

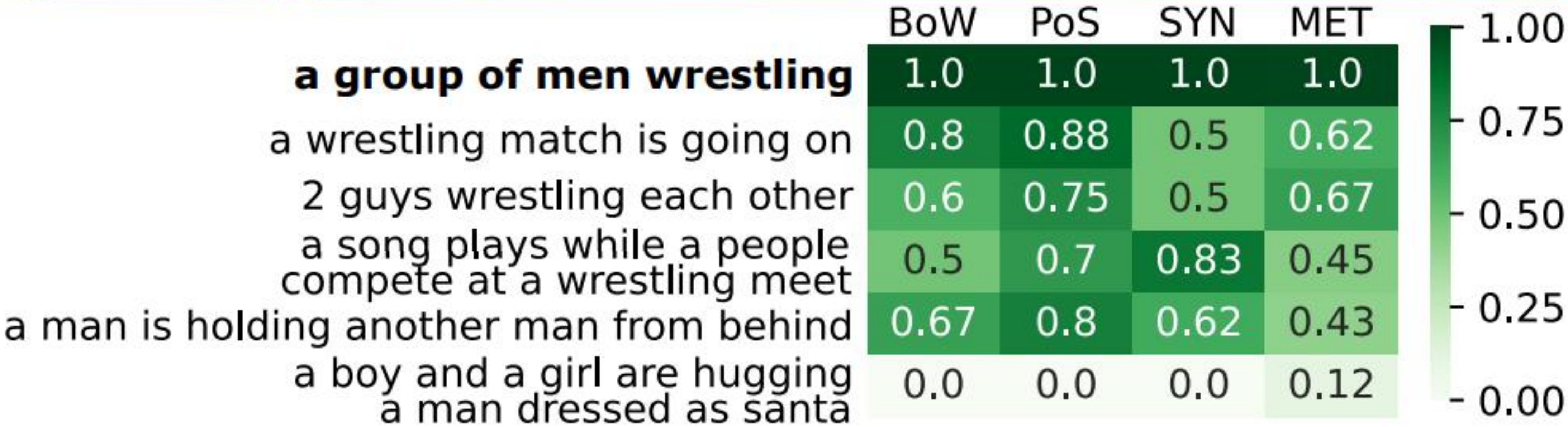
Semantic Similarity Examples



	BoW	PoS	SYN	MET
heat some oil in a deep pan and add chopped onions and fry till they turn brown	1.0	1.0	1.0	1.0
add chopped onions to a pan of oil	0.42	0.7	0.75	0.32
heat oil and add chopped onion	0.42	0.63	0.71	0.81
heat oil in a pan	0.25	0.43	0.46	0.72
add chopped carrot and celery	0.14	0.2	0.25	0.41
soak the mushrooms in nearly boiling water	0.0	0.0	0.0	0.031

Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

Semantic Similarity Examples

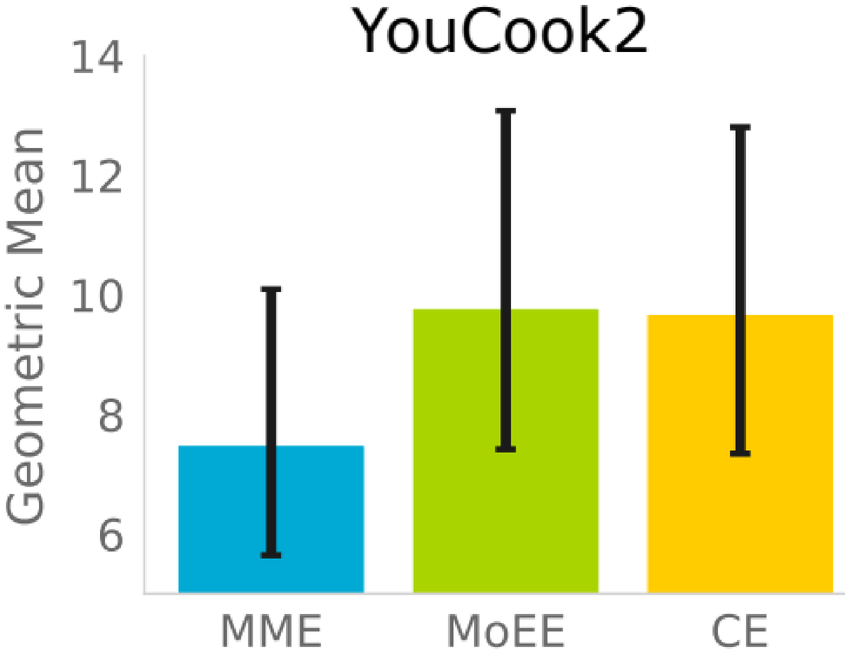


	BoW	PoS	SYN	MET
a group of men wrestling	1.0	1.0	1.0	1.0
a wrestling match is going on	0.8	0.88	0.5	0.62
2 guys wrestling each other	0.6	0.75	0.5	0.67
a song plays while a people compete at a wrestling meet	0.5	0.7	0.83	0.45
a man is holding another man from behind	0.67	0.8	0.62	0.43
a boy and a girl are hugging a man dressed as santa	0.0	0.0	0.0	0.12

Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

Impact on Performance

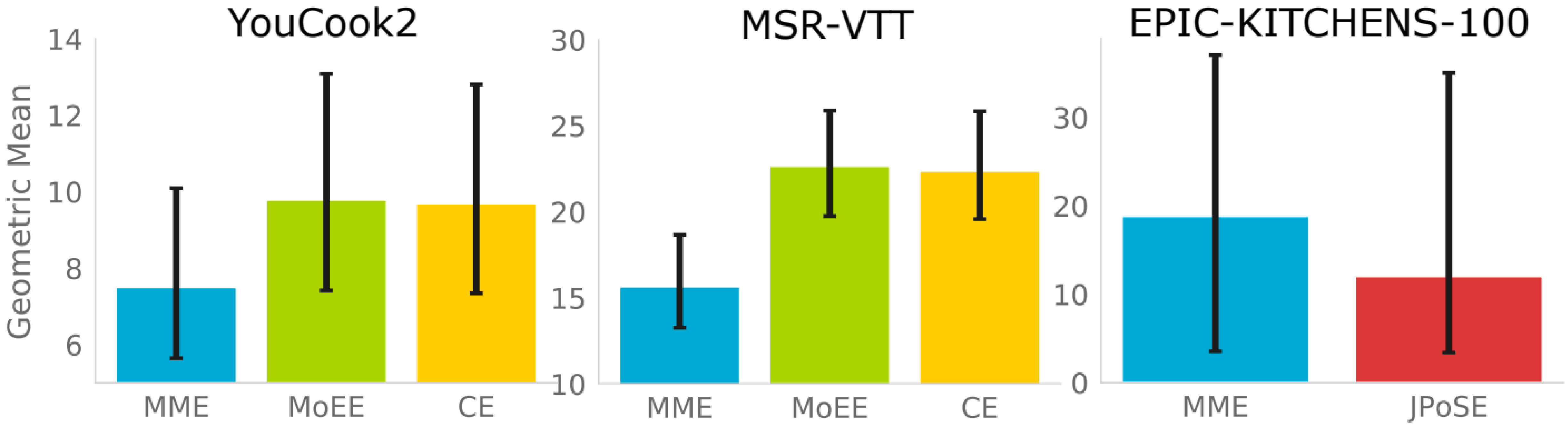
- Min and max performance on standard metrics when considering highly similar captions as ground-truth



Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

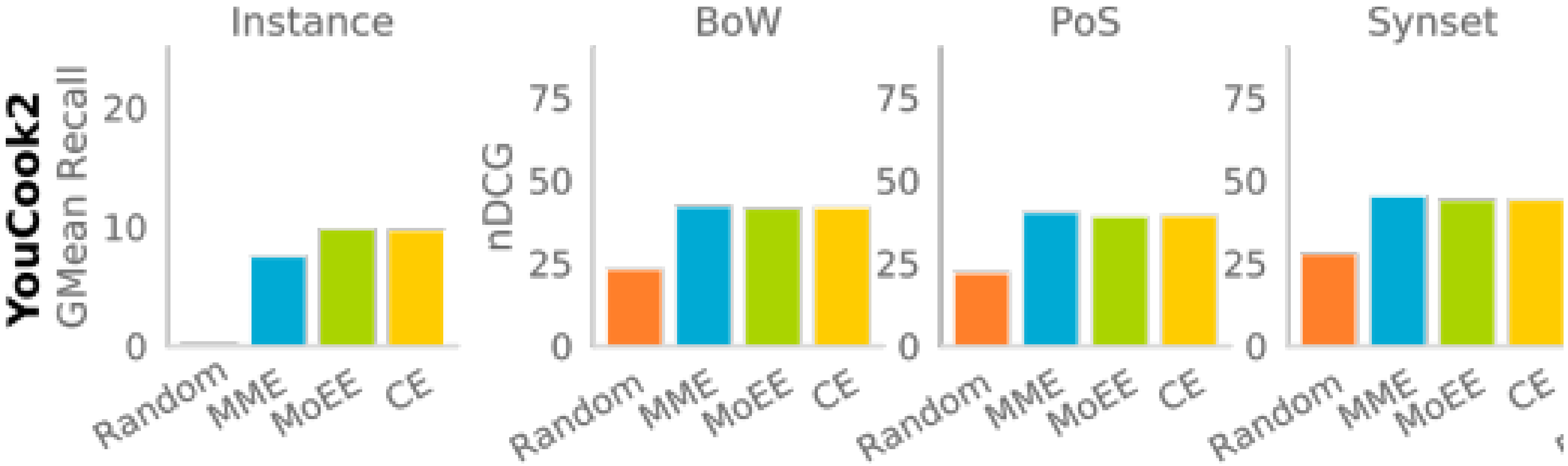
Impact on Performance

- Min and max performance on standard metrics when considering highly similar captions as ground-truth



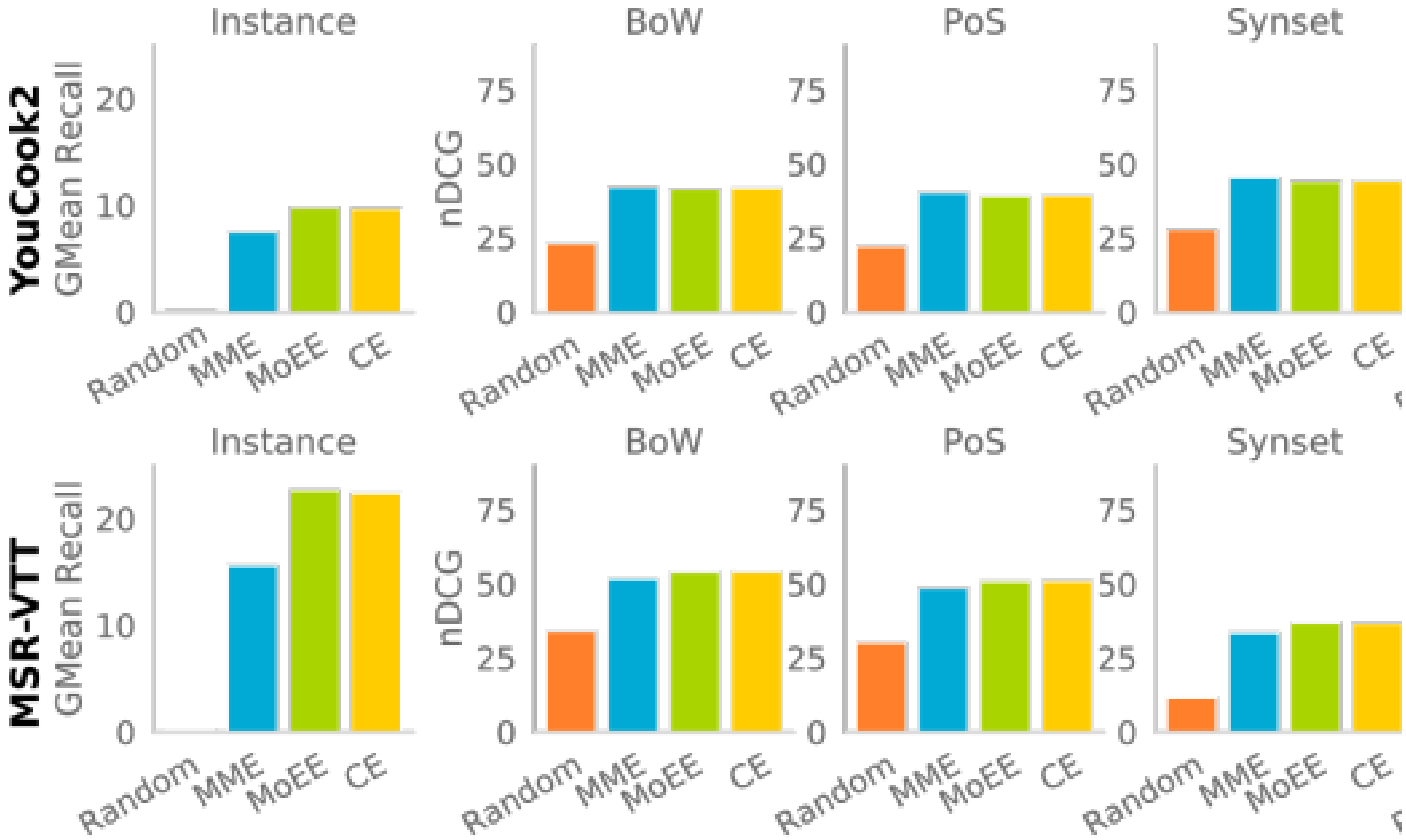
Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

Results with Semantic Similarity Metric



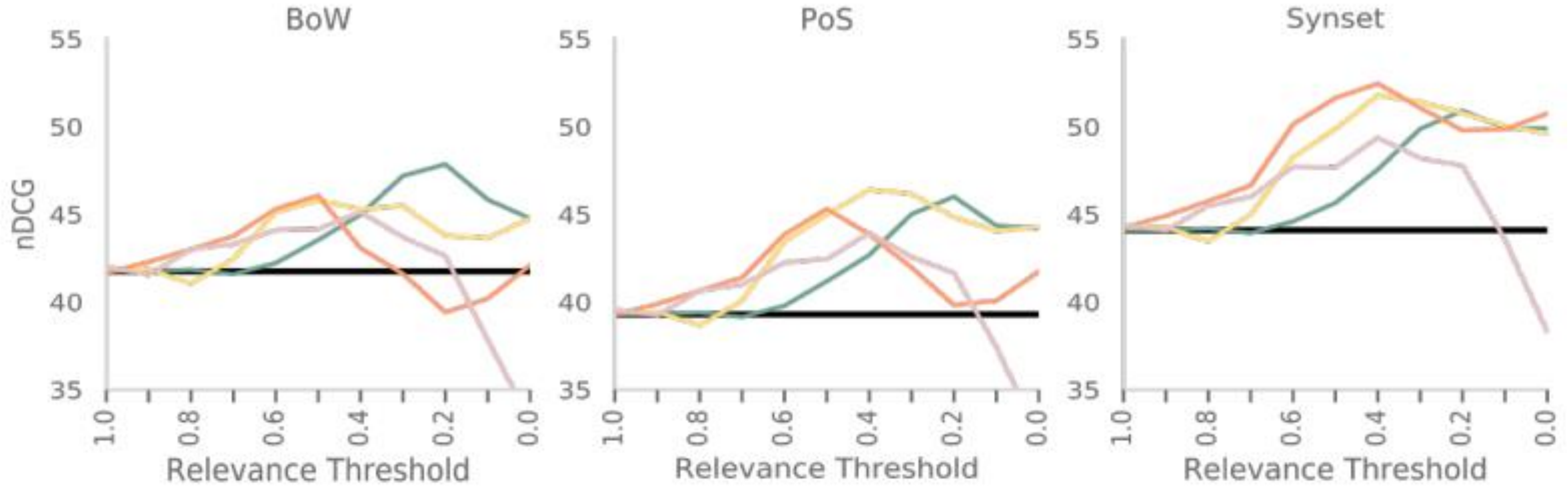
Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

Results with Semantic Similarity Metric



How to Improve?

Use similarity metric in training



Michael Wray, Hazel Doughty, Dima Damen. On Semantic Similarity in Video Retrieval. CVPR 2021

Evaluation Conclusion

- Two large issues even just in video retrieval

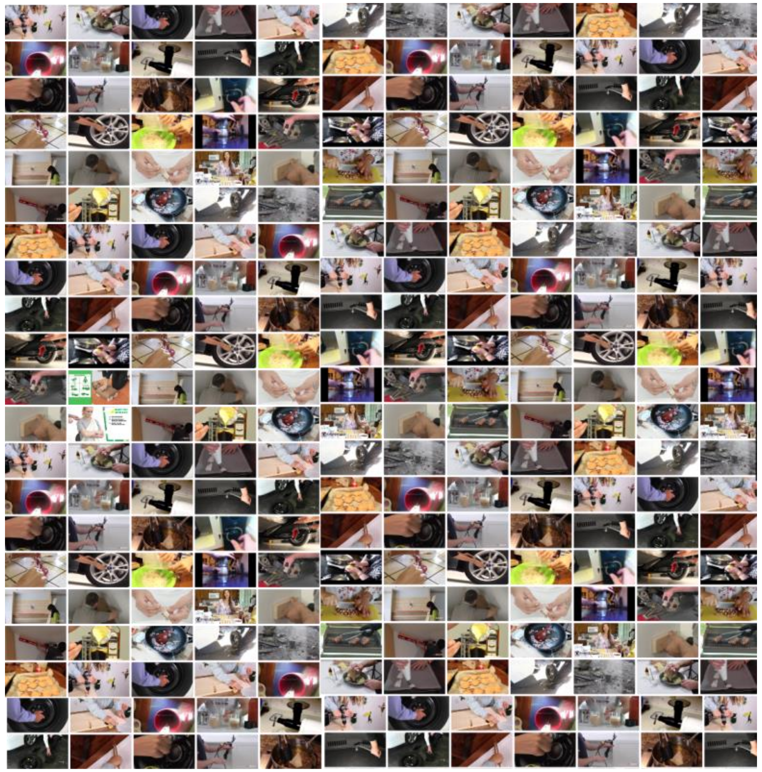
- Both issues relate to the labels, not the videos themselves

- Consider what your metric is really evaluating

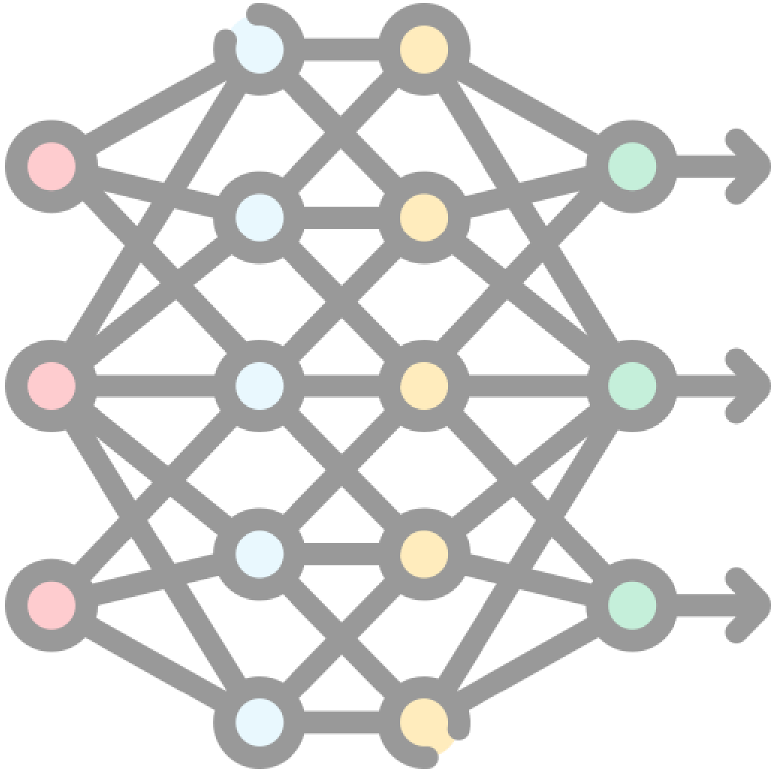
- Consider what your test set contains

Data

Data



Model



Evaluation

Motion in Video-Language Datasets

A group of people playing kites together on a beach

Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Motion in Video-Language Datasets

A group of people playing kites together on a beach

Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Motion in Video-Language Datasets

A group of people playing kites together on a beach

Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Motion in Video-Language Datasets

A group of people playing kites together on a beach

Cockatoos on the fence

Billiards, concentrated young woman playing in club

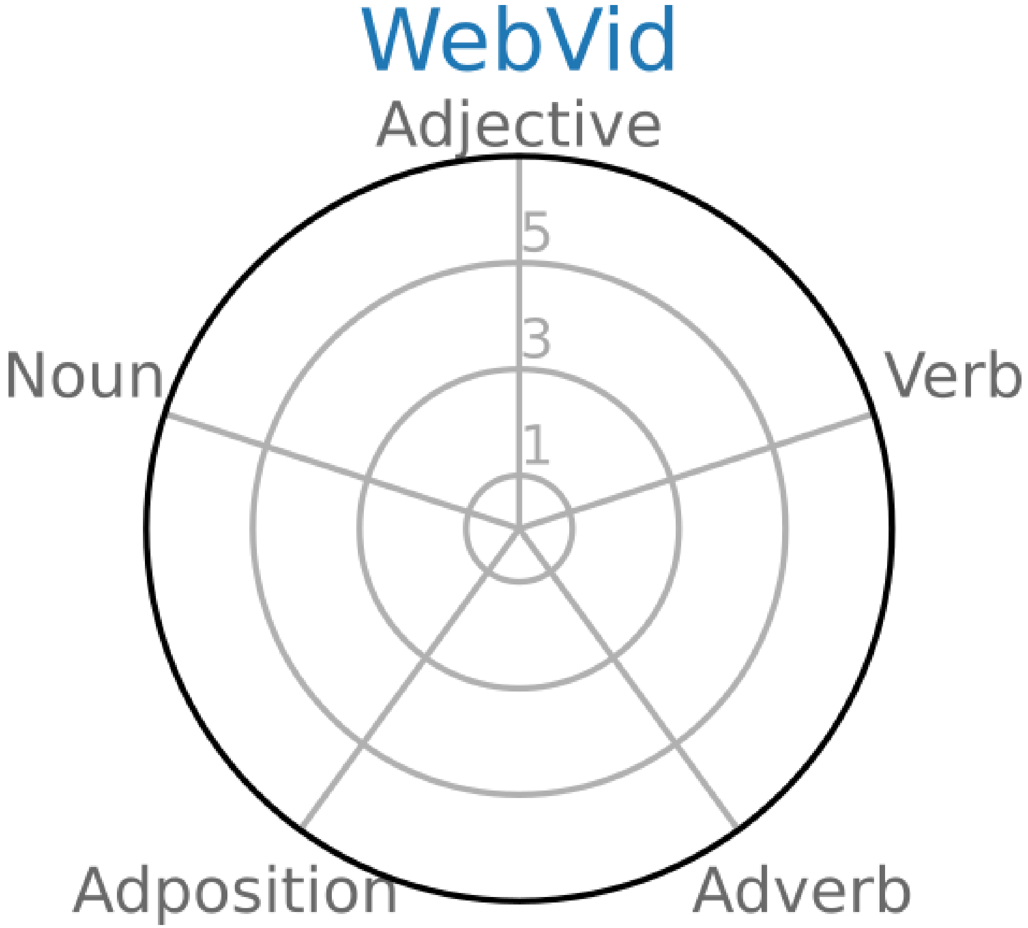
Female cop talking on walkietalkie, responding emergency call, crime prevention

A child with a suitcase. a happy little girl sits on a suitcase with a passport and money

Ontario, canada january 2014 heavy pretty snow on tree branches

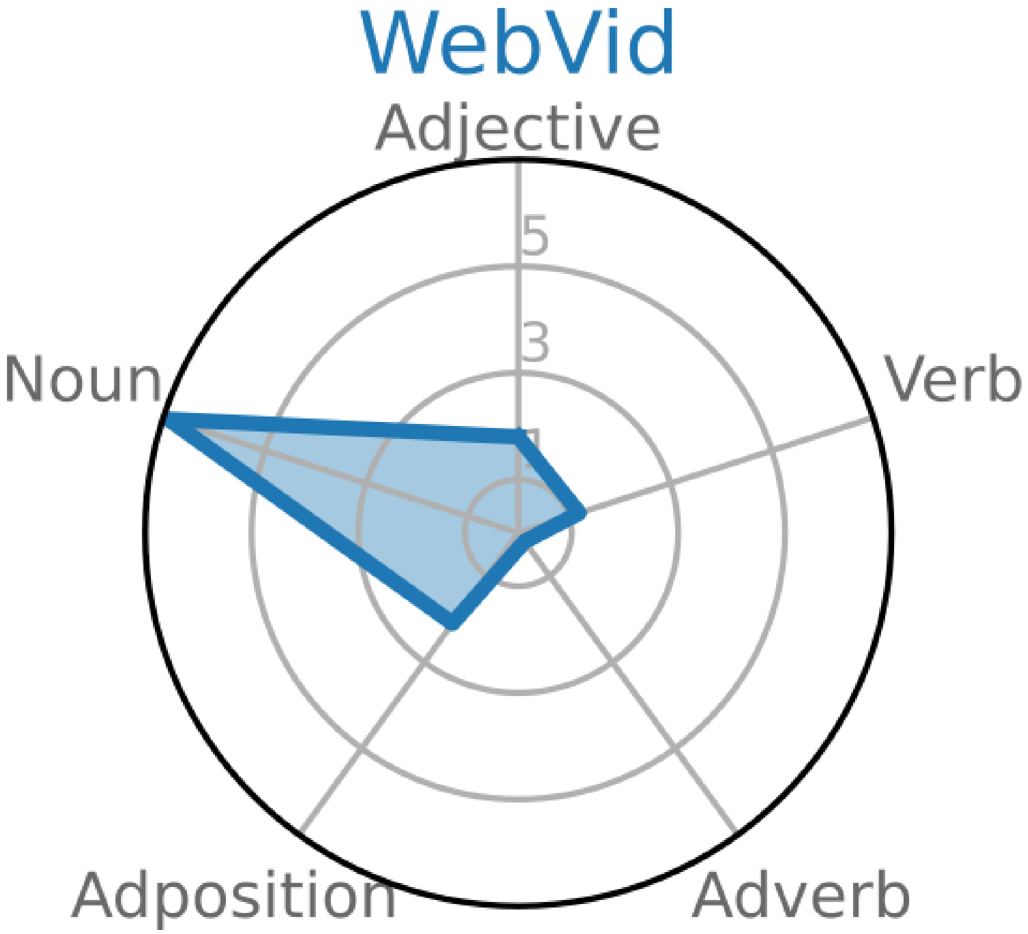
Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Motion in Video-Language Datasets



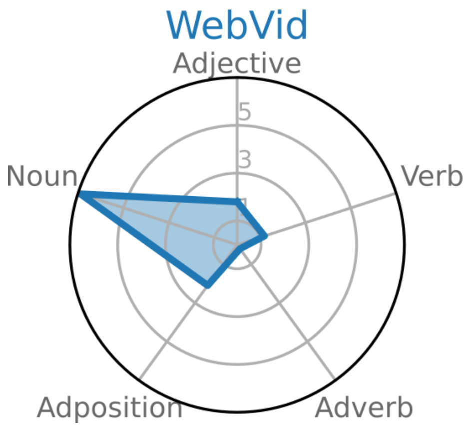
Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Motion in Video-Language Datasets

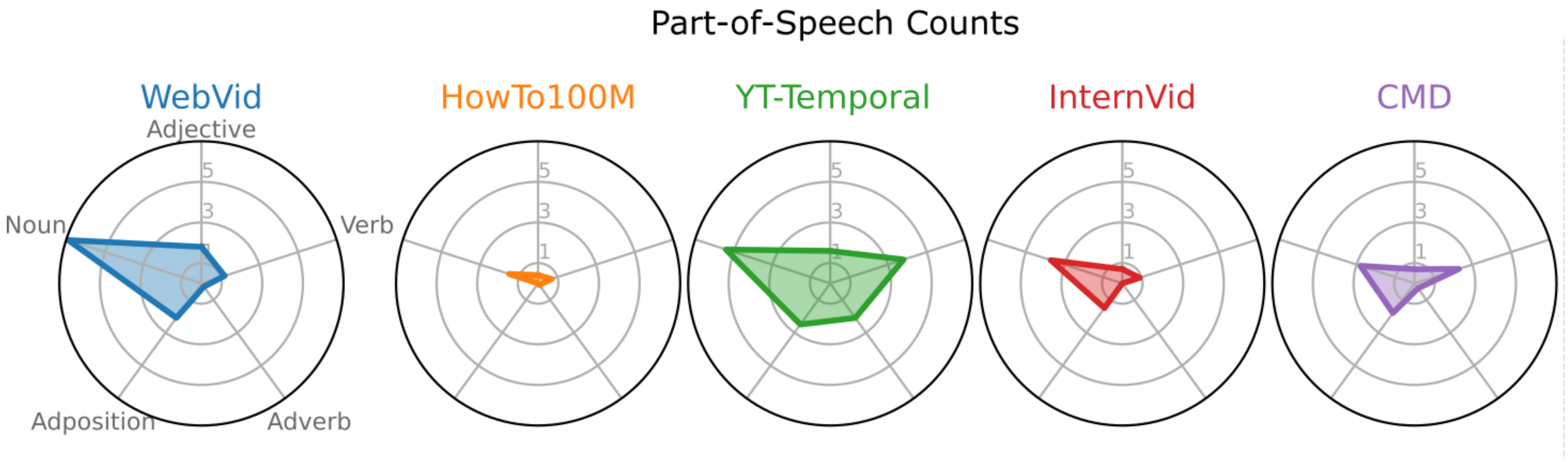


Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Motion in Video-Language Datasets

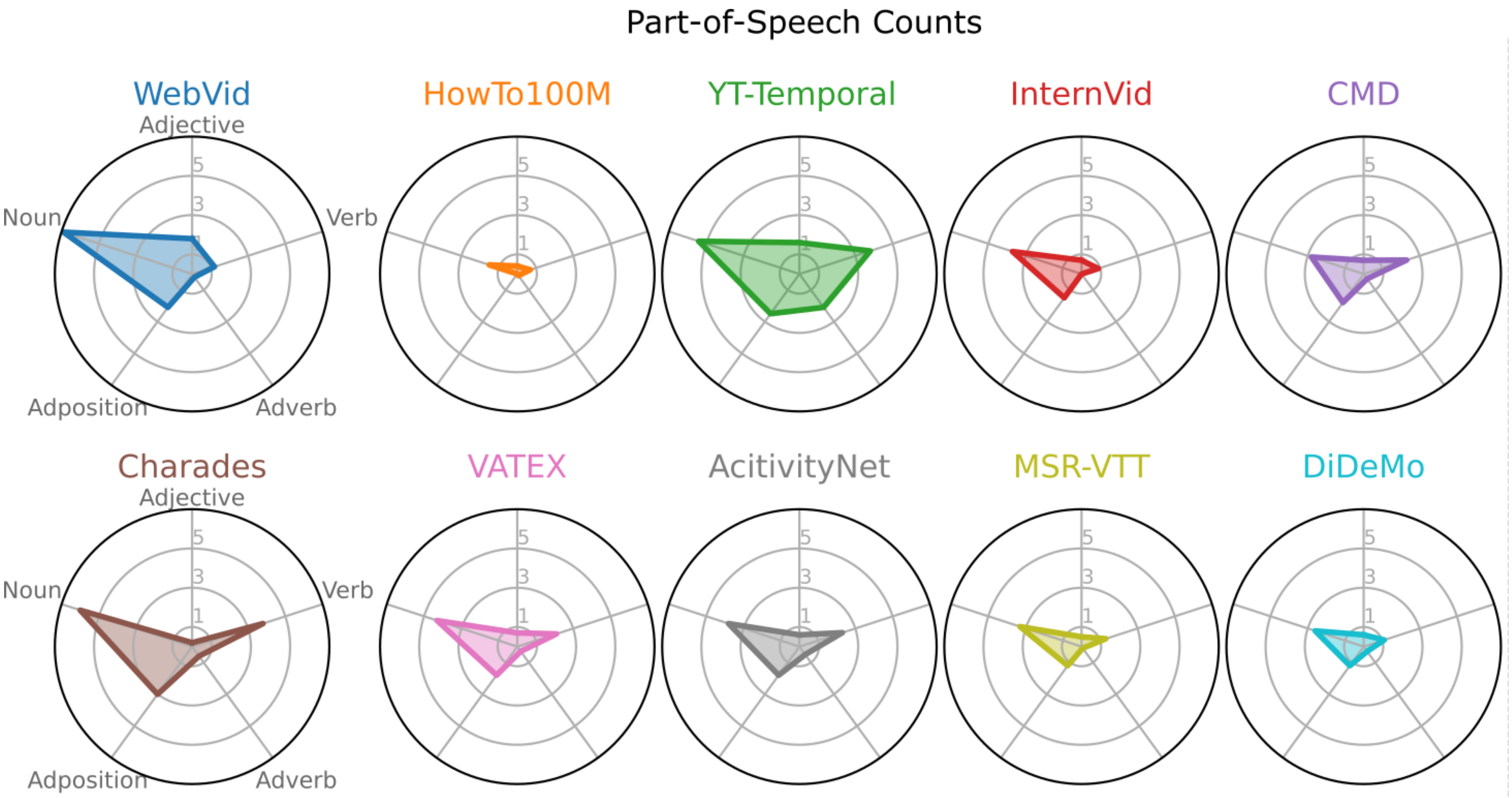


Motion in Video-Language Datasets



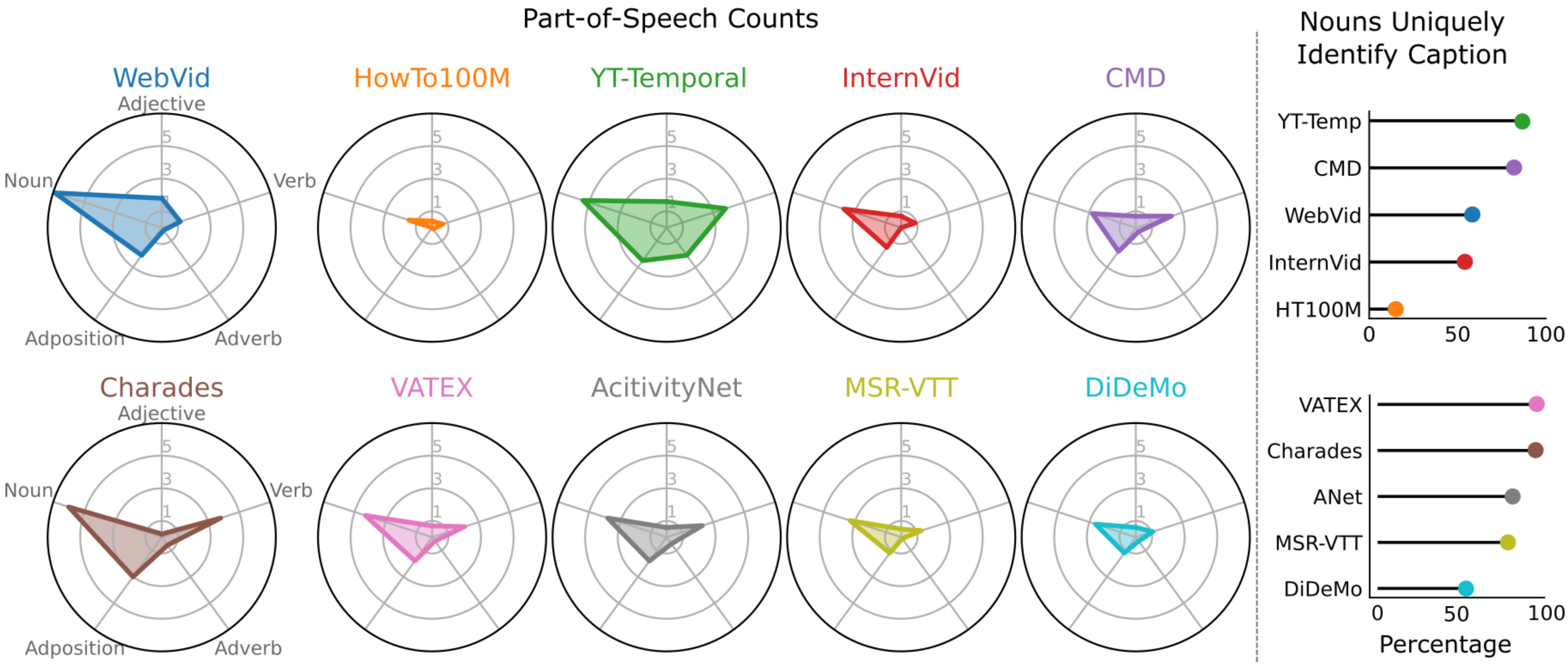
Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Motion in Video-Language Datasets



Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Motion in Video-Language Datasets



Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

VLMs Struggle to Capture Motion

The individual appears to be casually strolling in the park, occasionally looking towards the camera and then towards a fountain, which is the focal point of the video.

The video shows no significant motion. The orchid remains static throughout the entire clip, and the leaves in the background are similarly still.

We Need Motion-Focused Video Language Representations

Spatial-focused Video-Text Pair

~~A group of people playing kites together on the beach~~.

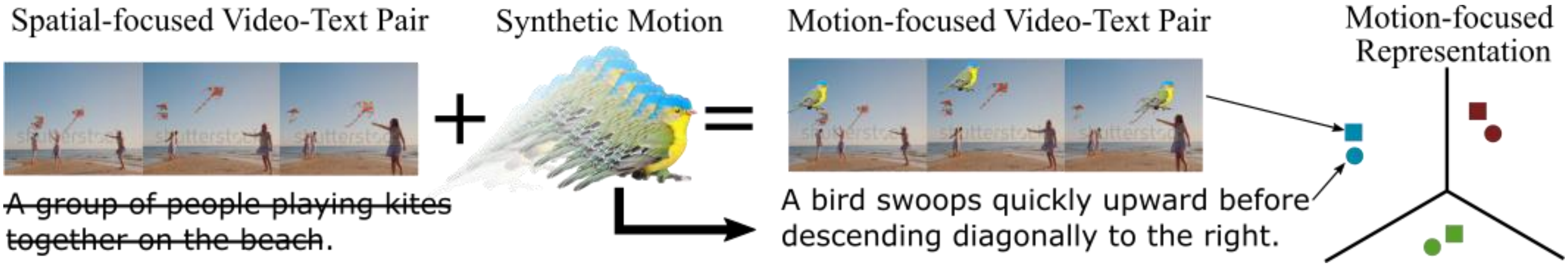
Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

We Need Motion-Focused Video Language Representations

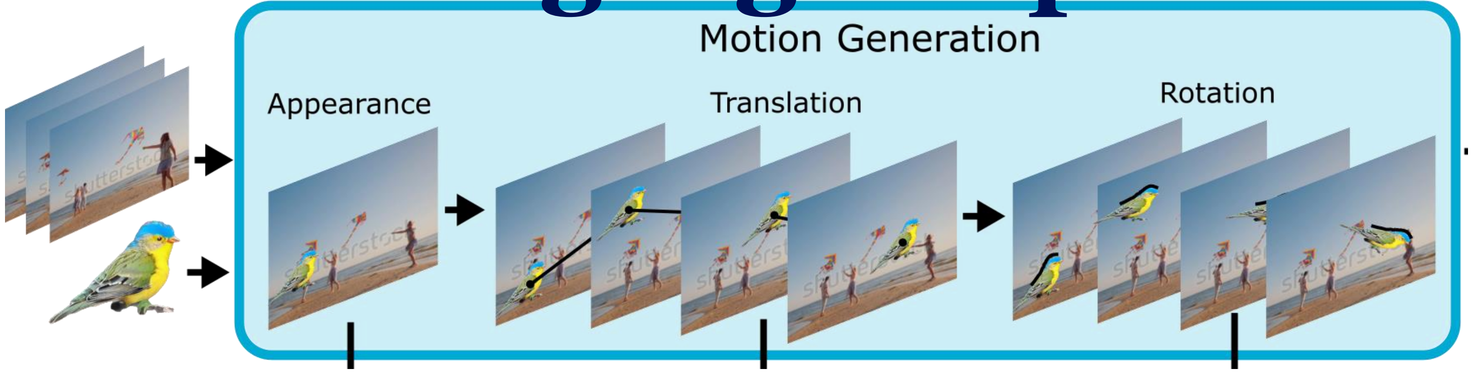
~~A group of people playing kites together on the beach~~.

Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

We Need Motion-Focused Video Language Representations

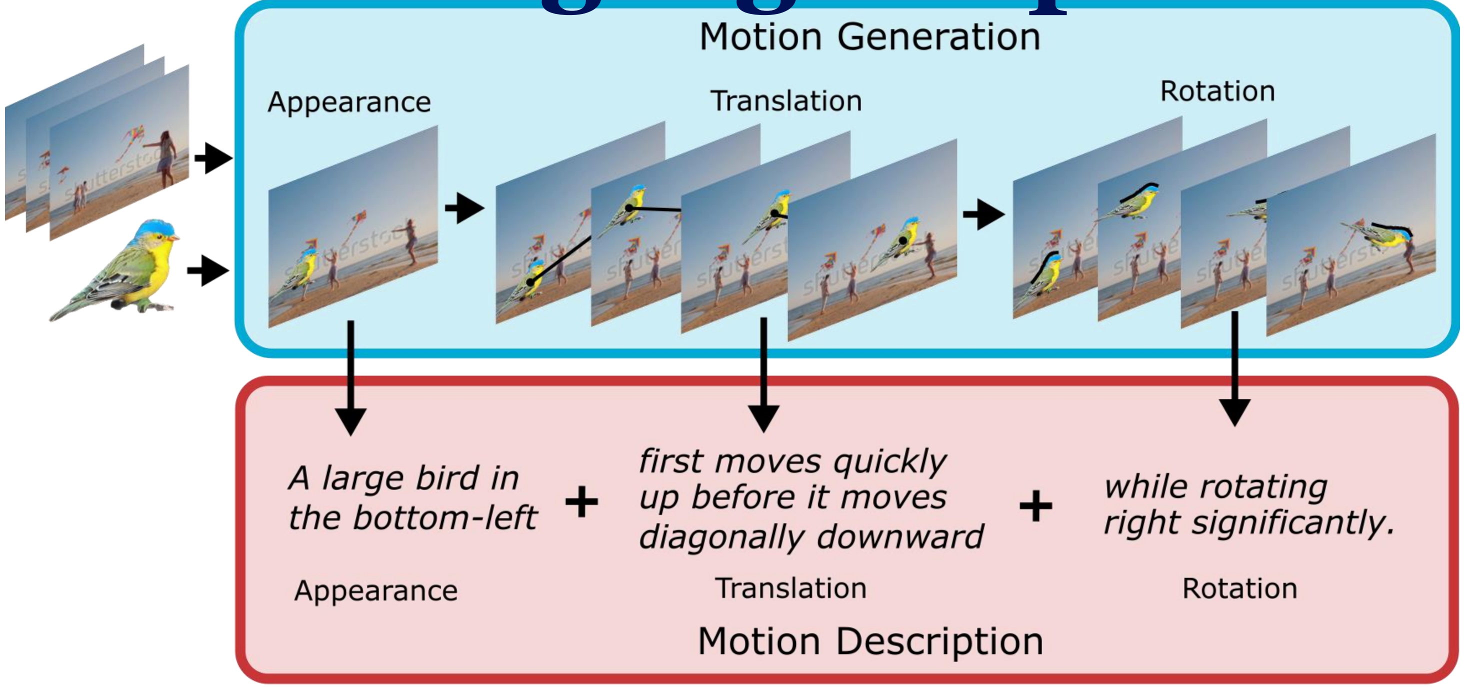


LocoMotion: Learning Motion-Focused Video Language Representations



Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

LocoMotion: Learning Motion-Focused Video Language Representations



Motion Generation

Appearance

Translation

Rotation

A large bird in the bottom-left + *first moves quickly up before it moves diagonally downward* + *while rotating right significantly.*

Appearance

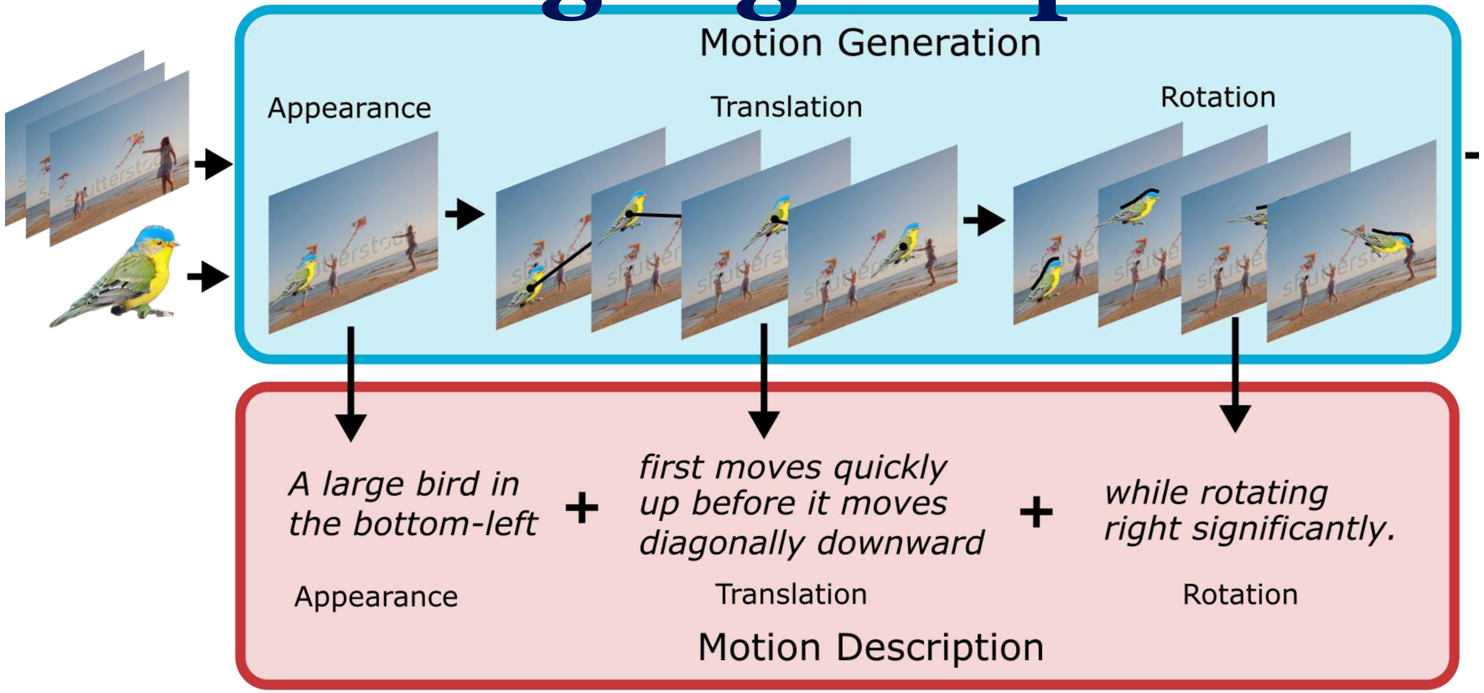
Translation

Rotation

Motion Description

Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

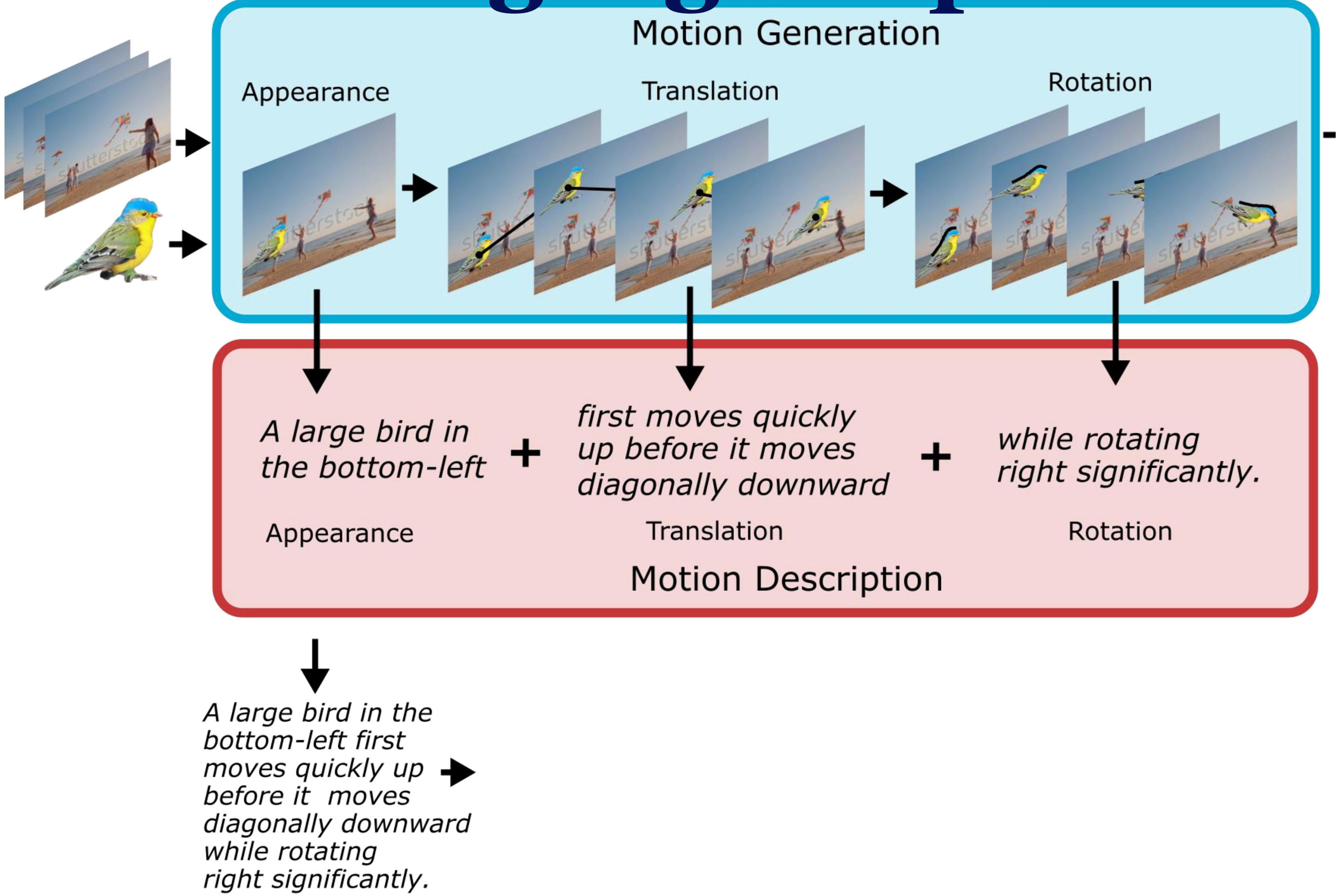
LocoMotion: Learning Motion-Focused Video Language Representations



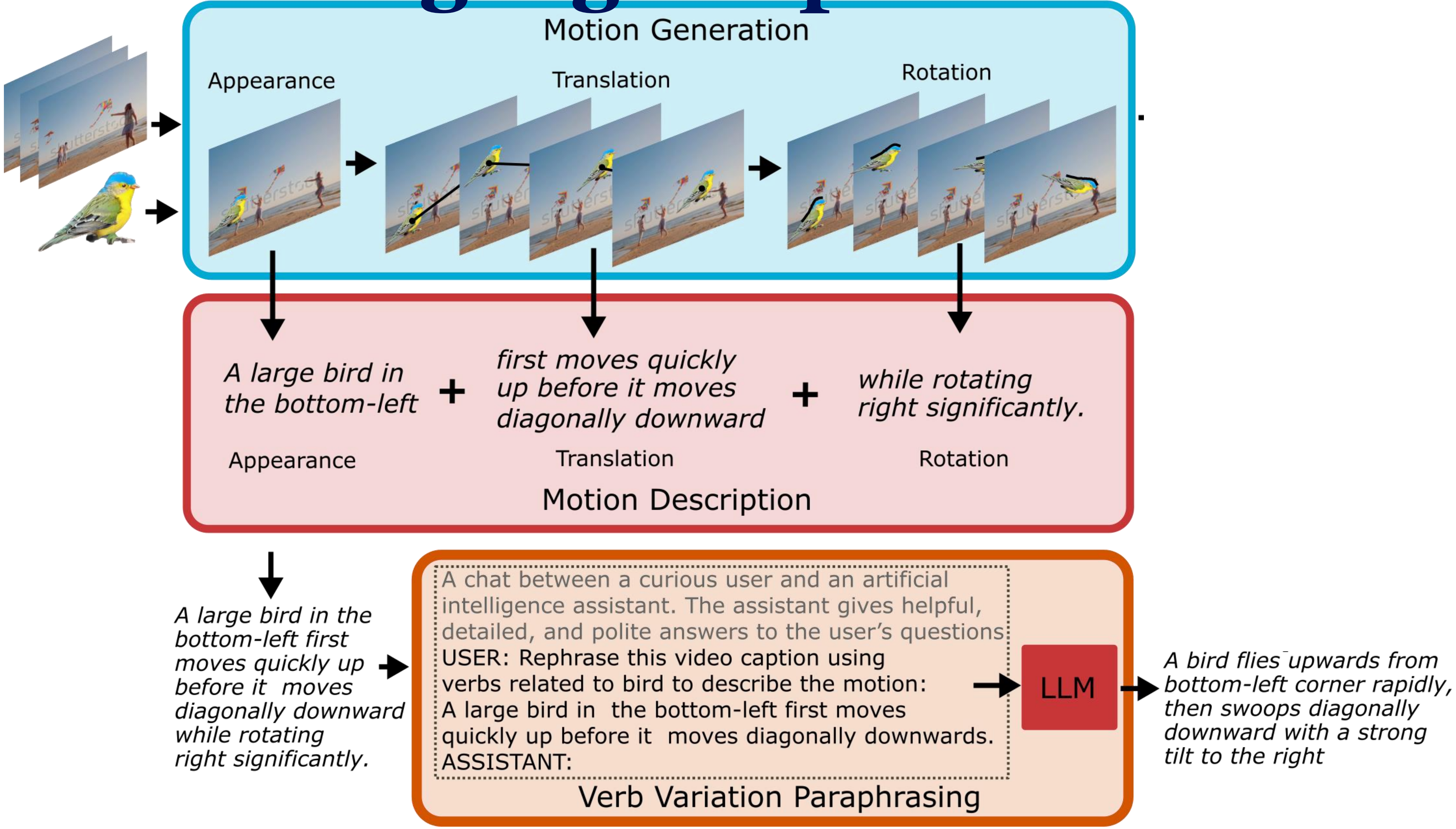
A large bird in the bottom-left + *first moves quickly up before it moves diagonally downward* + *while rotating right significantly.*

$$t_{translate} = \text{moves} + \underbrace{\left\{\begin{matrix}\text{quickly}\\ \\ \text{slowly}\end{matrix}\right.}_{\text{speed}} + \underbrace{\left\{\text{diagonally}\right. + \left\{\begin{matrix}\text{upwards}\\ \text{right}\\ \text{downwards}\\ \text{left}\end{matrix}\right.}_{\text{direction}} + \underbrace{\left\{\begin{matrix}\text{a lot}\\ \\ \text{a little}\end{matrix}\right.}_{\text{distance}}$$

LocoMotion: Learning Motion-Focused Video Language Representations



LocoMotion: Learning Motion-Focused Video Language Representations



Motion Generation

Appearance

Translation

Rotation

A large bird in the bottom-left + *first moves quickly up before it moves diagonally downward* + *while rotating right significantly.*

Appearance

Translation

Rotation

Motion Description

A large bird in the bottom-left first moves quickly up before it moves diagonally downward while rotating right significantly.

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions

USER: Rephrase this video caption using verbs related to bird to describe the motion:

A large bird in the bottom-left first moves quickly up before it moves diagonally downwards.

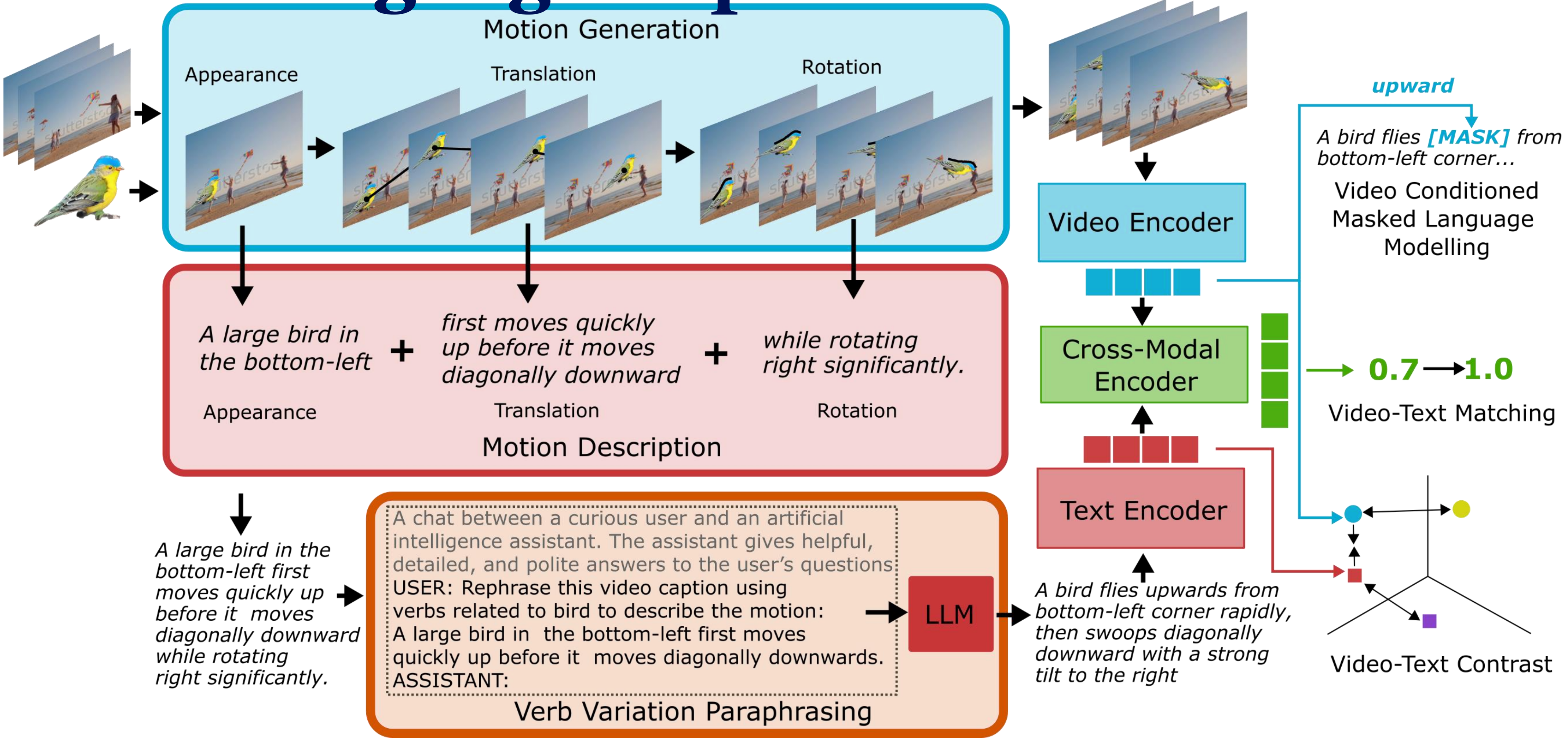
ASSISTANT:

LLM

A bird flies upwards from bottom-left corner rapidly, then swoops diagonally downward with a strong tilt to the right

Verb Variation Paraphrasing

LocoMotion: Learning Motion-Focused Video Language Representations



Motion Generation

Appearance

Translation

Rotation

A large bird in the bottom-left + *first moves quickly up before it moves diagonally downward* + *while rotating right significantly.*

Appearance

Translation

Rotation

Motion Description

A large bird in the bottom-left first moves quickly up before it moves diagonally downward while rotating right significantly.

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
USER: Rephrase this video caption using verbs related to bird to describe the motion:
A large bird in the bottom-left first moves quickly up before it moves diagonally downwards.
ASSISTANT:

LLM

Verb Variation Paraphrasing

Video Encoder

Cross-Modal Encoder

Text Encoder

A bird flies upwards from bottom-left corner rapidly, then swoops diagonally downward with a strong tilt to the right

upward

*A bird flies **[MASK]** from bottom-left corner...*

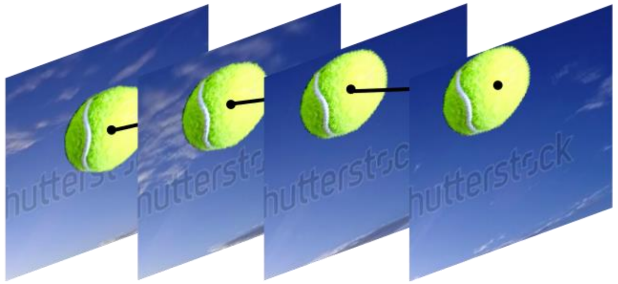
Video Conditioned Masked Language Modelling

0.7→1.0

Video-Text Matching

Video-Text Contrast

Example Motion-Focused Captions


A speeding tennis ball at the center diagonally soars upwards with a slight leftward twist.


A substantial vehicle at the top accelerates diagonally downward while simultaneously veering right.


A sled on the left initially ascends upward, followed by a gentle glide to the right while simultaneously pivoting right slightly.


A butterfly flutters in the center, darting diagonally right then left, while twirling right.


A scurrying rat in the bottom-left initially shifts right a tad before darting upwards rapidly with a slight leftward twist.


A sword in the center rises slightly before slicing diagonally to the right.

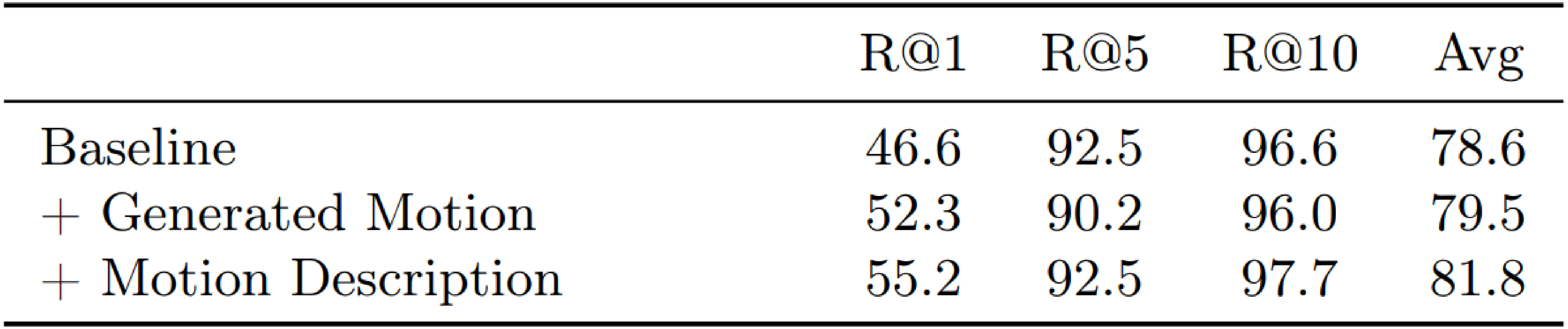
Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Model Ablation

	R@1	R@5	R@10	Avg
Baseline	46.6	92.5	96.6	78.6
+ Generated Motion	52.3	90.2	96.0	79.5

Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

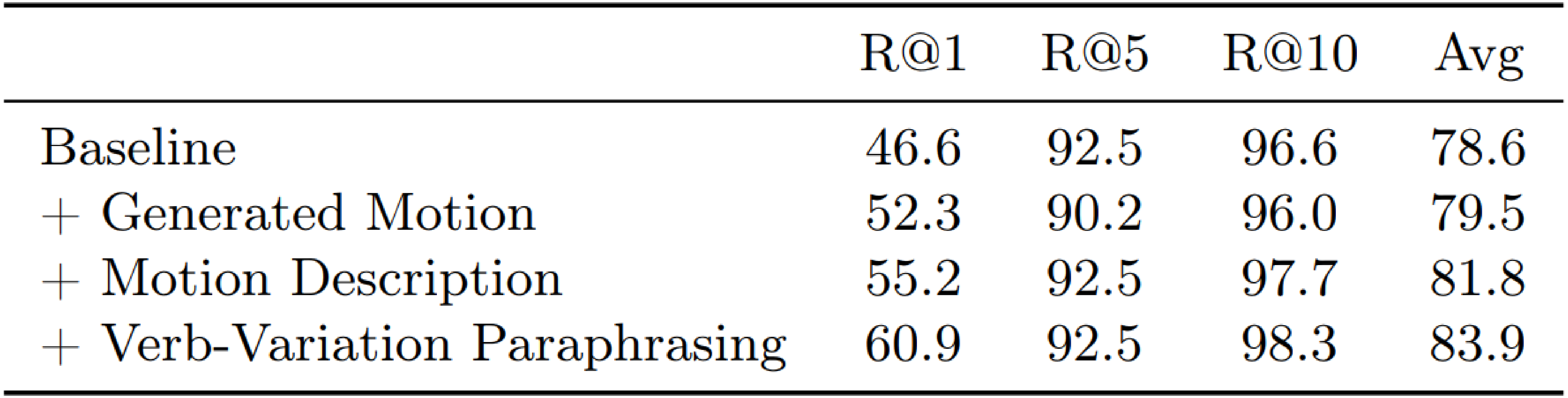
Model Ablation



	R@1	R@5	R@10	Avg
Baseline	46.6	92.5	96.6	78.6
+ Generated Motion	52.3	90.2	96.0	79.5
+ Motion Description	55.2	92.5	97.7	81.8

Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

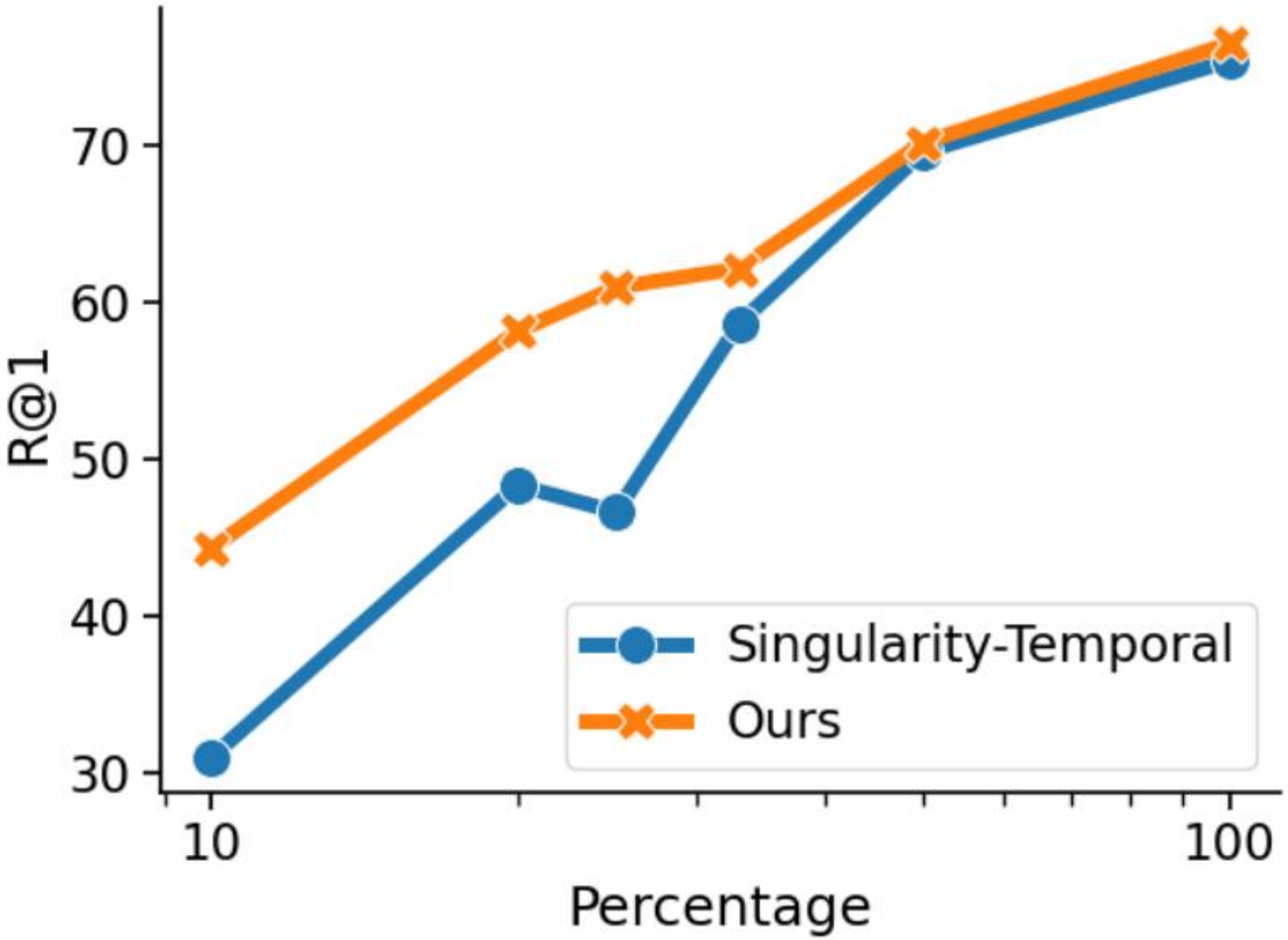
Model Ablation



	R@1	R@5	R@10	Avg
Baseline	46.6	92.5	96.6	78.6
+ Generated Motion	52.3	90.2	96.0	79.5
+ Motion Description	55.2	92.5	97.7	81.8
+ Verb-Variation Paraphrasing	60.9	92.5	98.3	83.9

Hazel Doughty, Fida Mohammad Thoker, Cees Snoek. LocoMotion: Learning Motion-Focused Video-Language Representations. ACCV 2024.

Data Efficient Fine-Tuning



We Don’t Have Enough Data

well-labelled

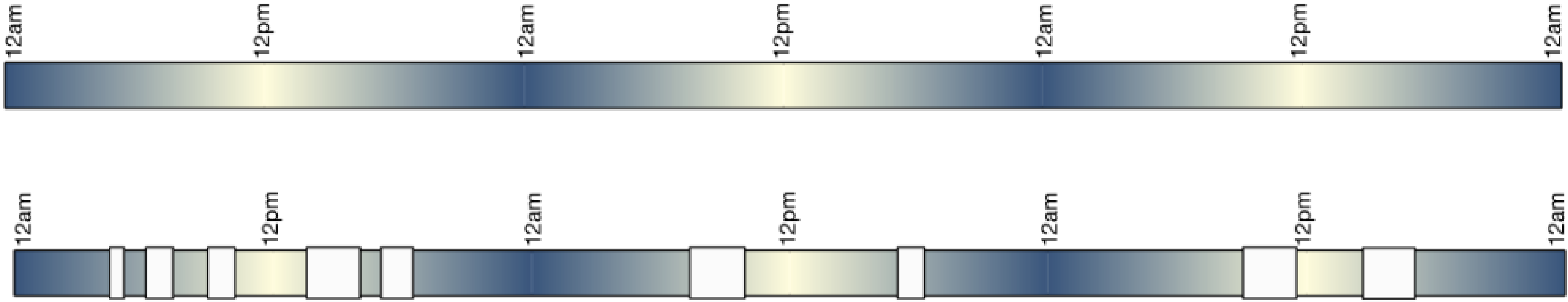
We Don’t^Have Enough Data

Sneak Preview

HD-EPIC

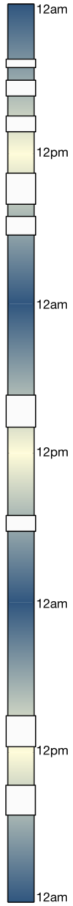
Perrett et al (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



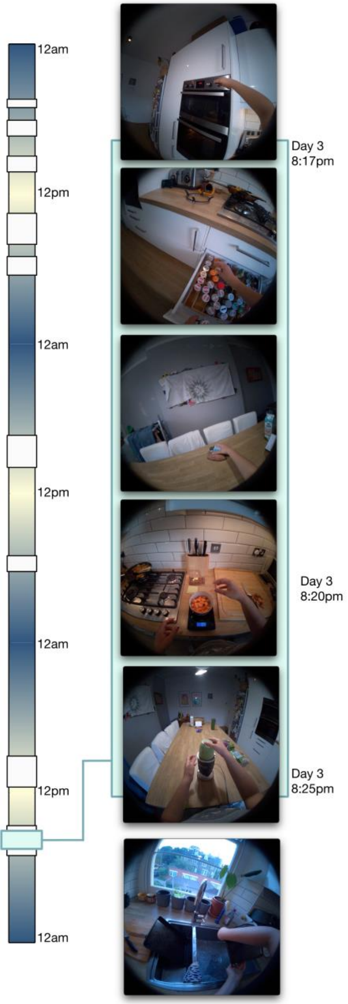
Perrett et al (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



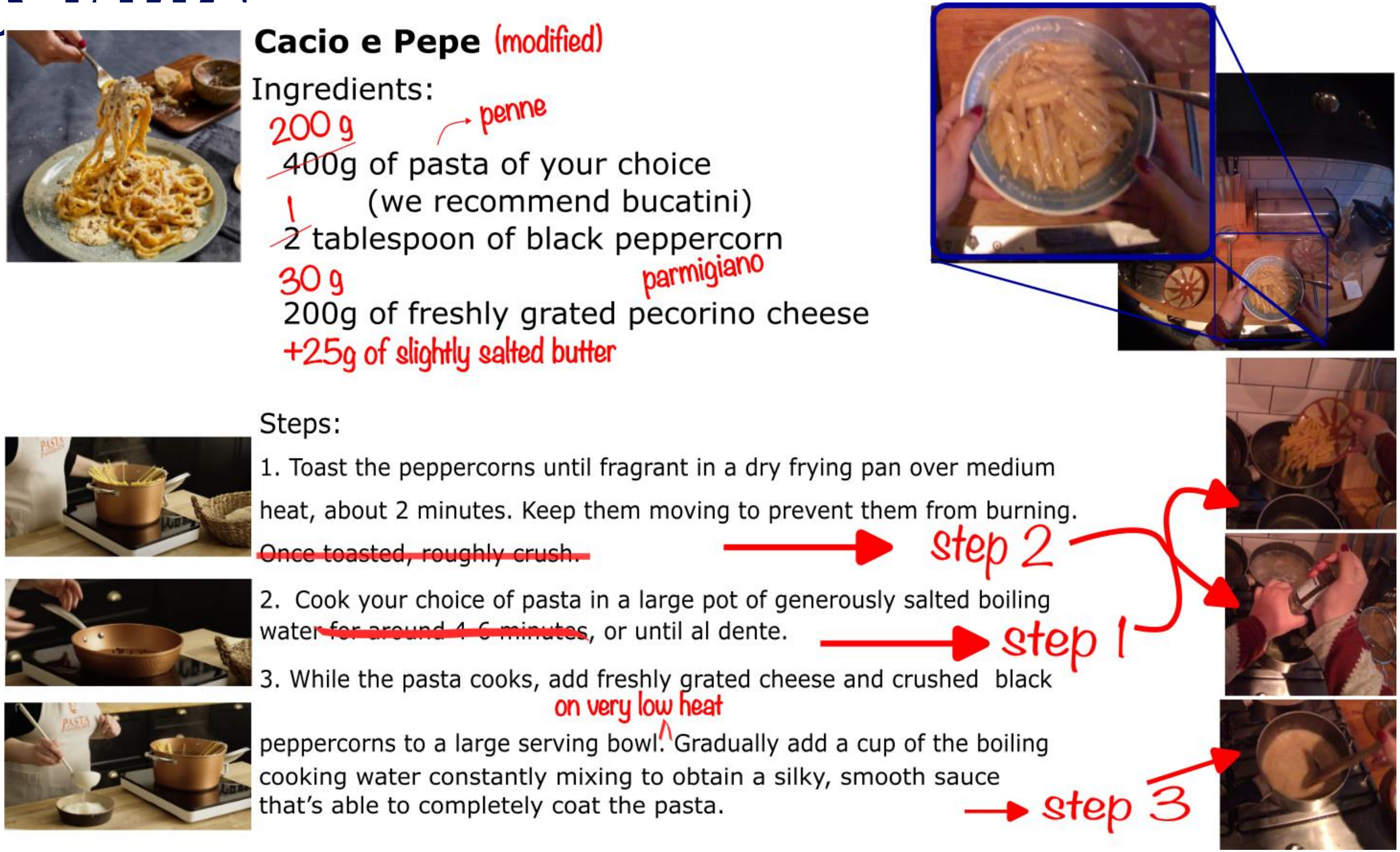
Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



Cacio e Pepe (modified)

Ingredients:

200 g → penne

~~400g~~ of pasta of your choice (we recommend bucatini)

1

~~2~~ tablespoon of black peppercorn

30 g parmigiano

200g of freshly grated pecorino cheese

+25g of slightly salted butter

Steps:

1. Toast the peppercorns until fragrant in a dry frying pan over medium heat, about 2 minutes. Keep them moving to prevent them from burning. ~~Once toasted, roughly crush.~~ → step 2

2. Cook your choice of pasta in a large pot of generously salted boiling water ~~for around 4-6 minutes~~, or until al dente. → step 1

3. While the pasta cooks, add freshly grated cheese and crushed black peppercorns to a large serving bowl. on very low heat Gradually add a cup of the boiling cooking water constantly mixing to obtain a silky, smooth sauce that's able to completely coat the pasta. → step 3

Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC

ay 3
:17pm

Recipe: Southwestern Salad

1: Preheat the oven to 400F

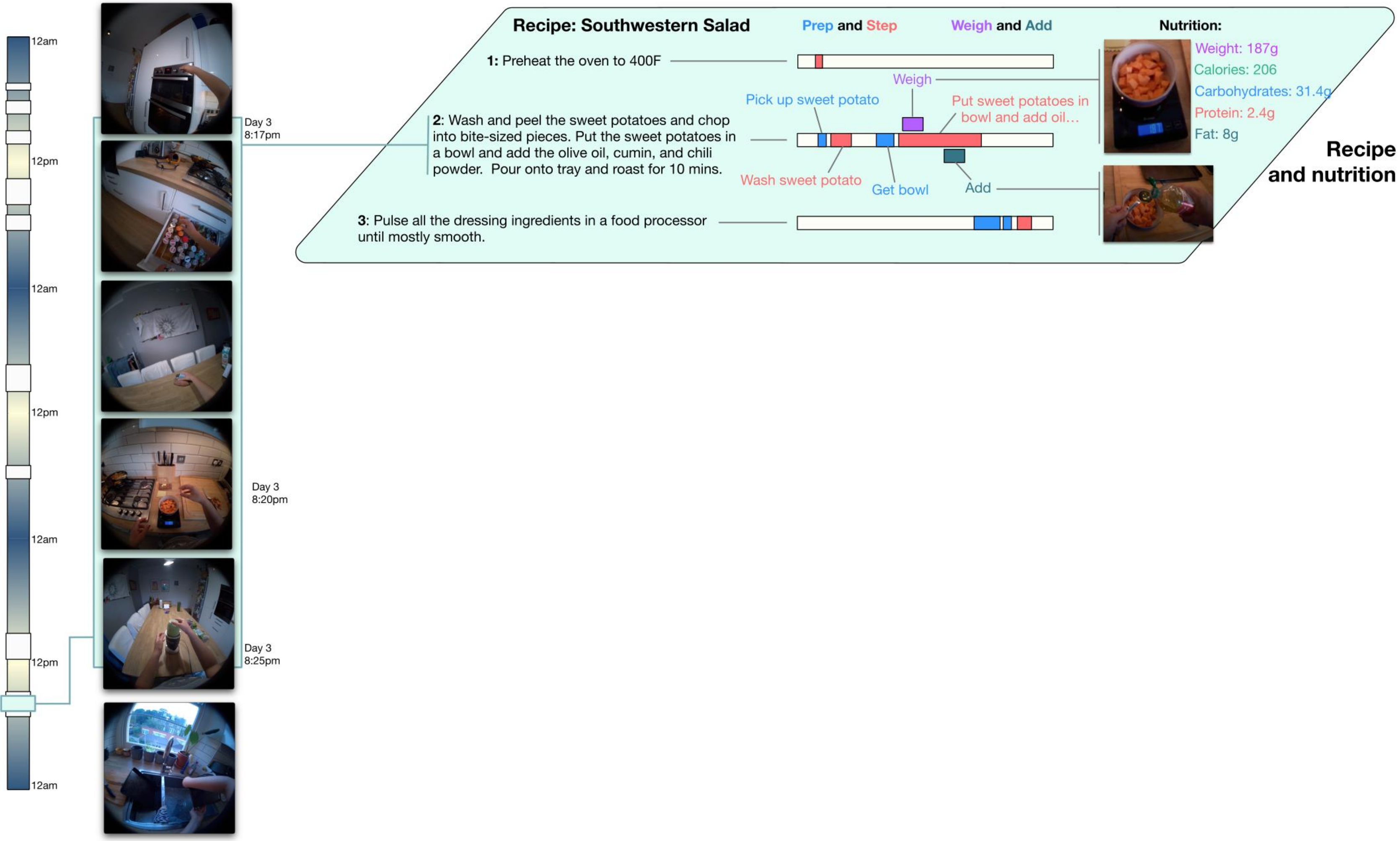
2: Wash and peel the sweet potatoes and chop into bite-sized pieces. Put the sweet potatoes in a bowl and add the olive oil, cumin, and chili powder. Pour onto tray and roast for 10 mins.

3: Pulse all the dressing ingredients in a food processor until mostly smooth.

Recipe and nutrition

Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



Recipe: Southwestern Salad

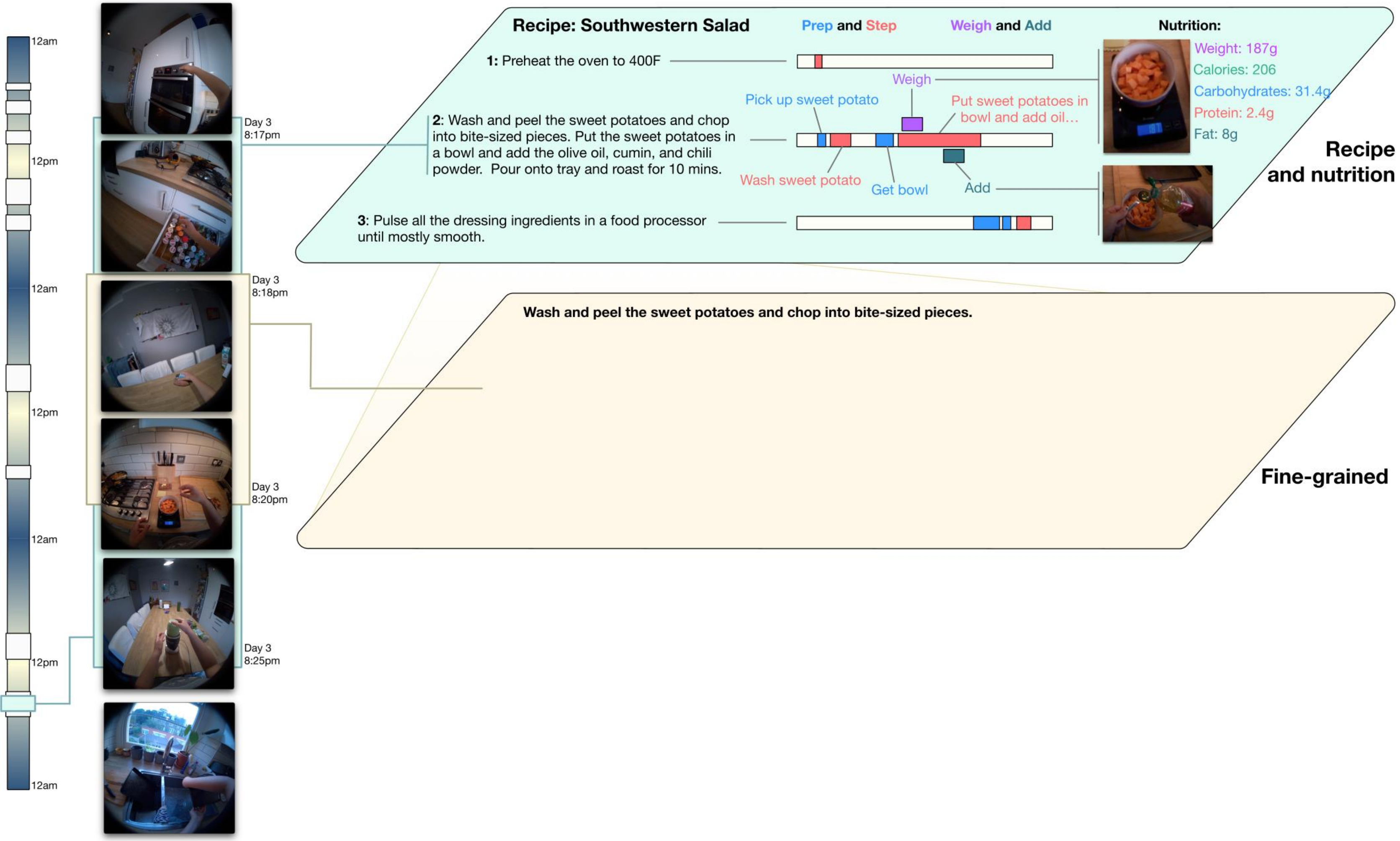
1: Preheat the oven to 400F

2: Wash and peel the sweet potatoes and chop into bite-sized pieces. Put the sweet potatoes in a bowl and add the olive oil, cumin, and chili powder. Pour onto tray and roast for 10 mins.

3: Pulse all the dressing ingredients in a food processor until mostly smooth.

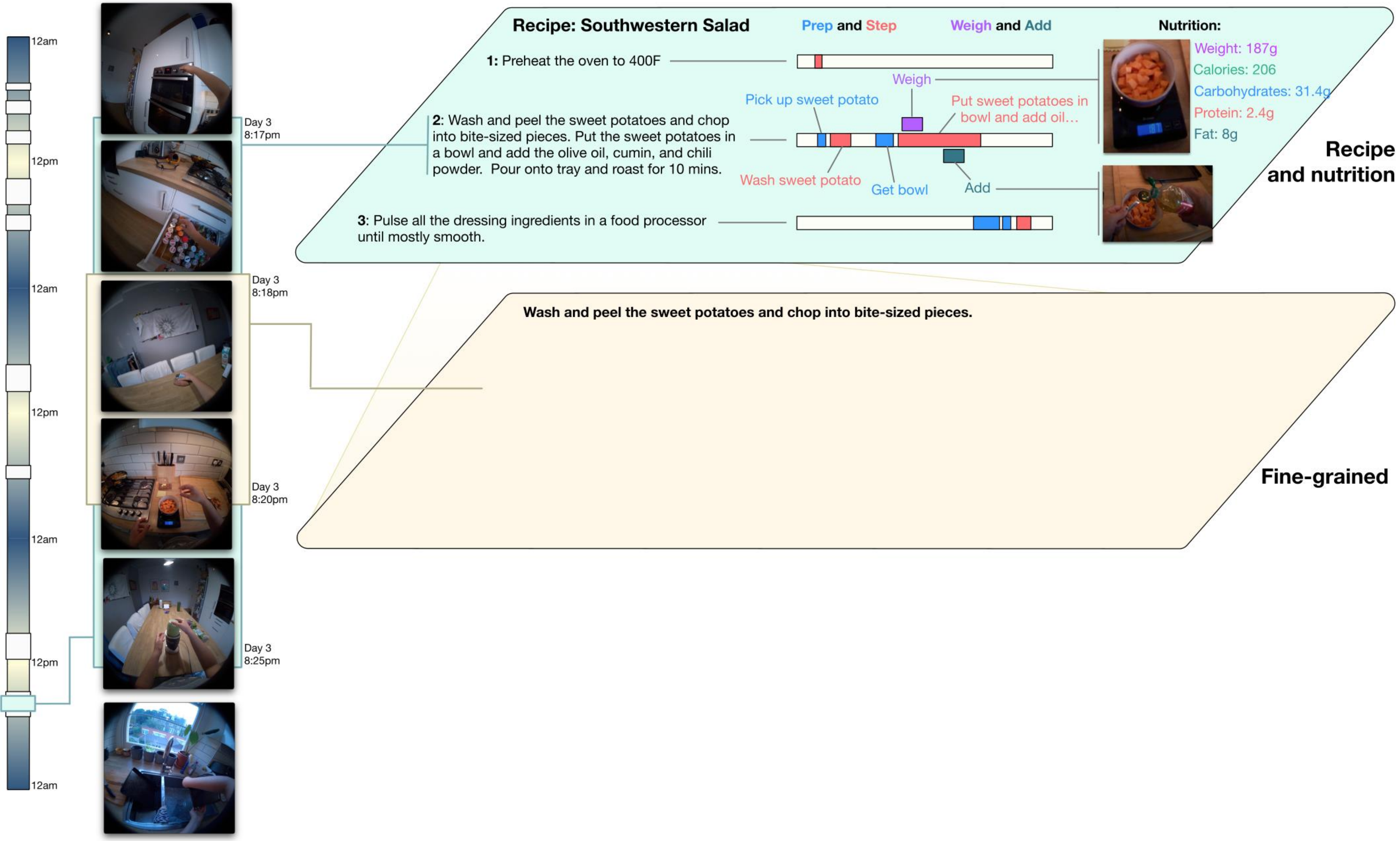
Perrett et al (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



Recipe: Southwestern Salad

Prep and Step

Weigh and Add

Nutrition:

1: Preheat the oven to 400F

2: Wash and peel the sweet potatoes and chop into bite-sized pieces. Put the sweet potatoes in a bowl and add the olive oil, cumin, and chili powder. Pour onto tray and roast for 10 mins.

3: Pulse all the dressing ingredients in a food processor until mostly smooth.

Weight: 187g
Calories: 206
Carbohydrates: 31.4g
Protein: 2.4g
Fat: 8g

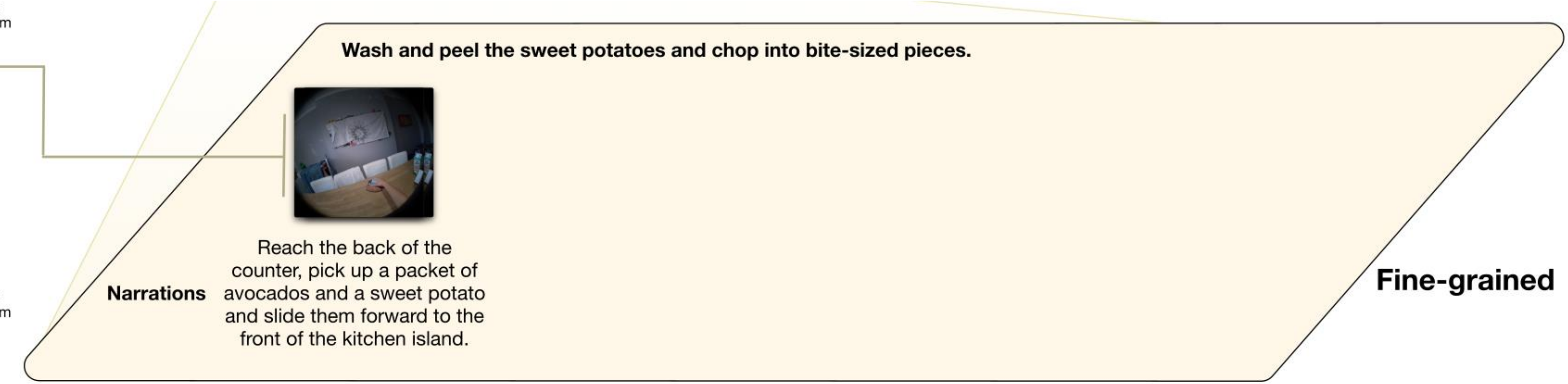
Recipe and nutrition

Wash and peel the sweet potatoes and chop into bite-sized pieces.

Fine-grained

Perrett et al (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



Wash and peel the sweet potatoes and chop into bite-sized pieces.

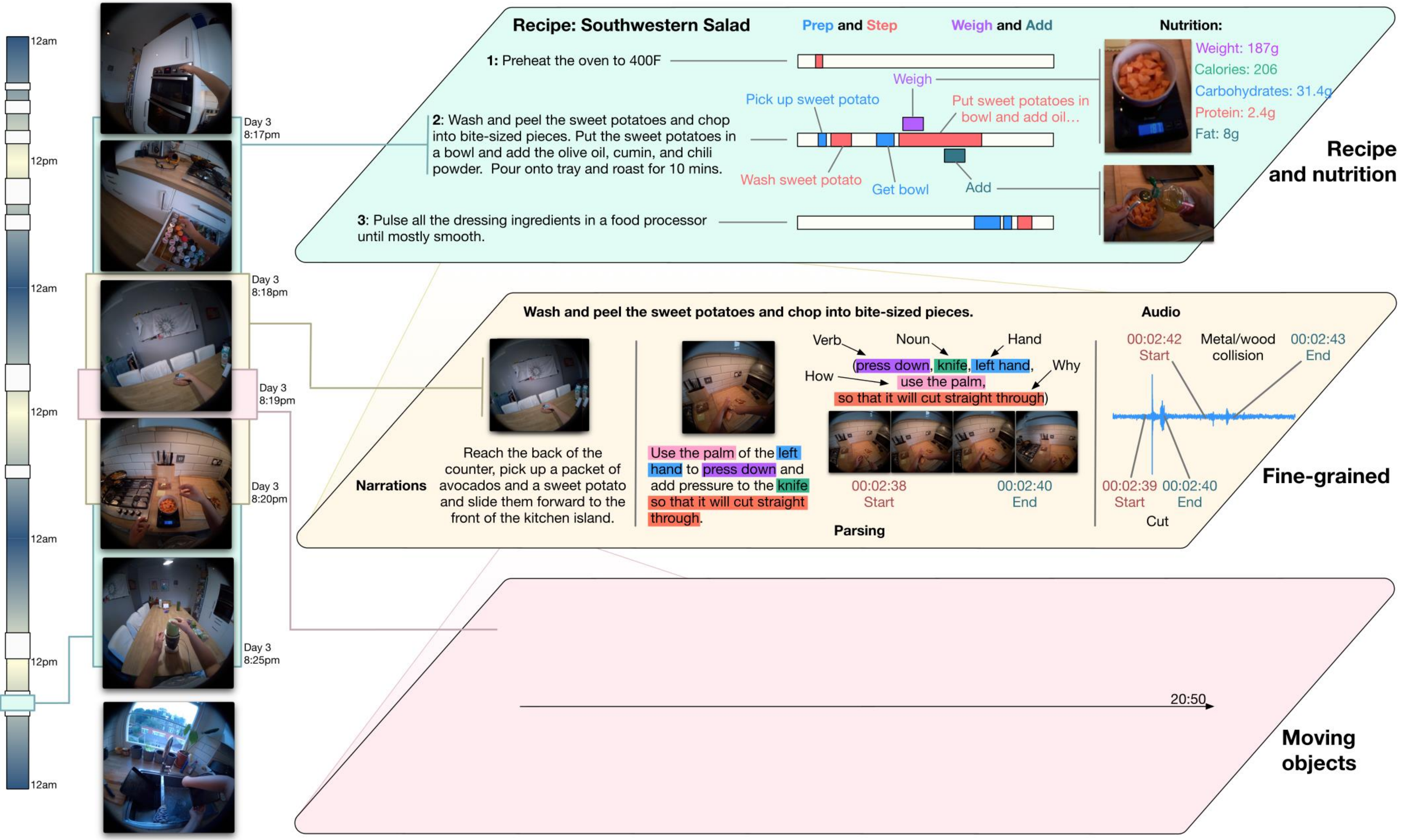
Narrations

Reach the back of the counter, pick up a packet of avocados and a sweet potato and slide them forward to the front of the kitchen island.

Fine-grained

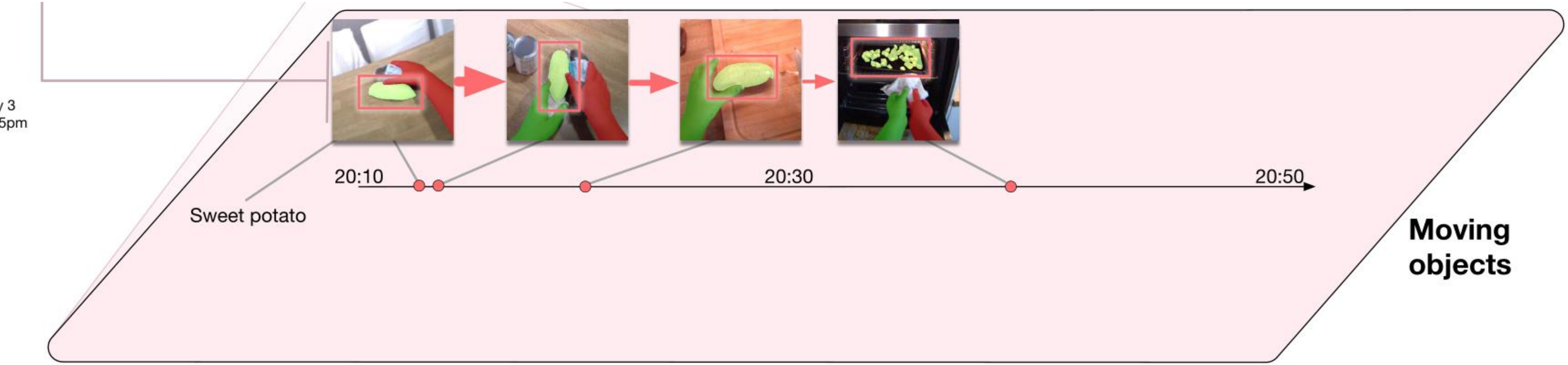
Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



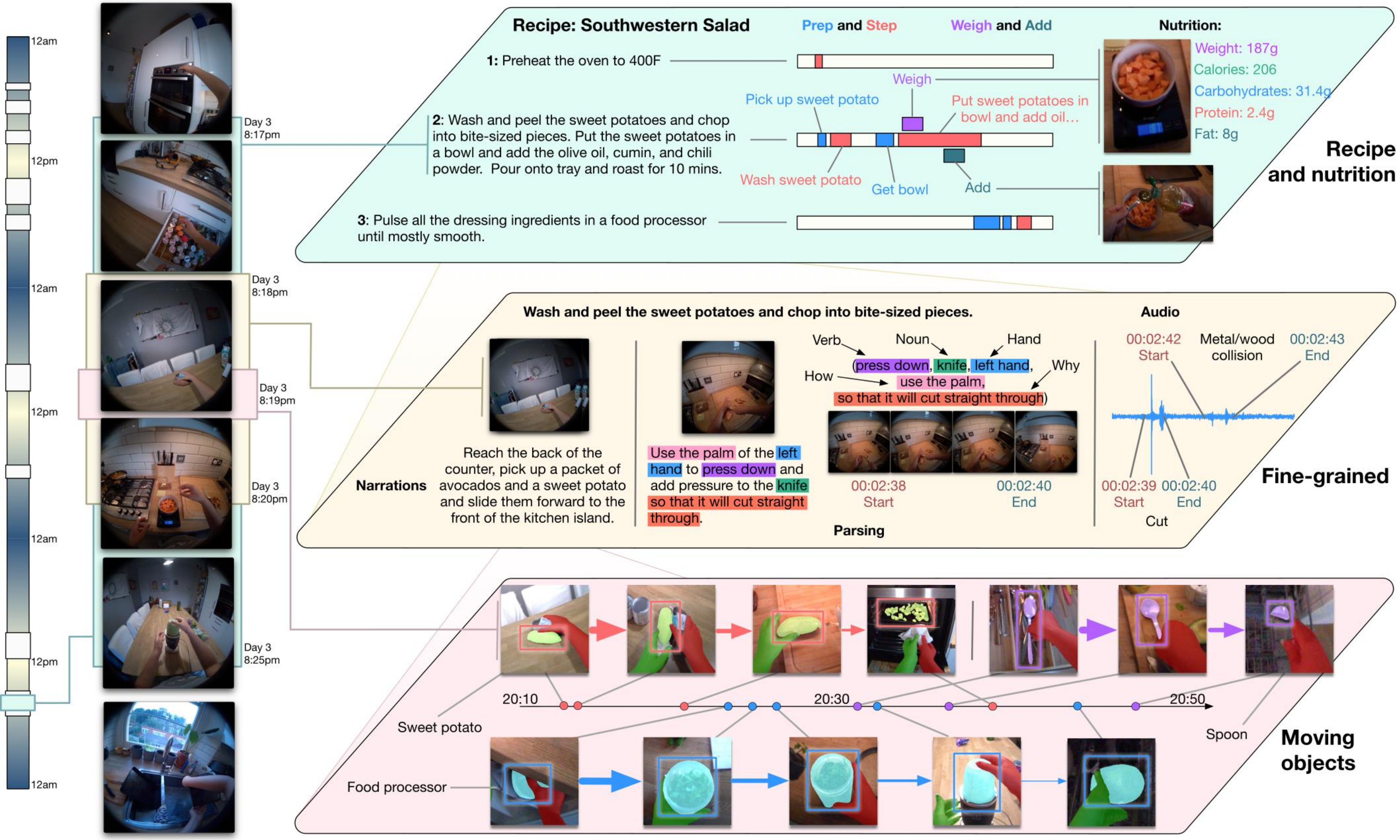
Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



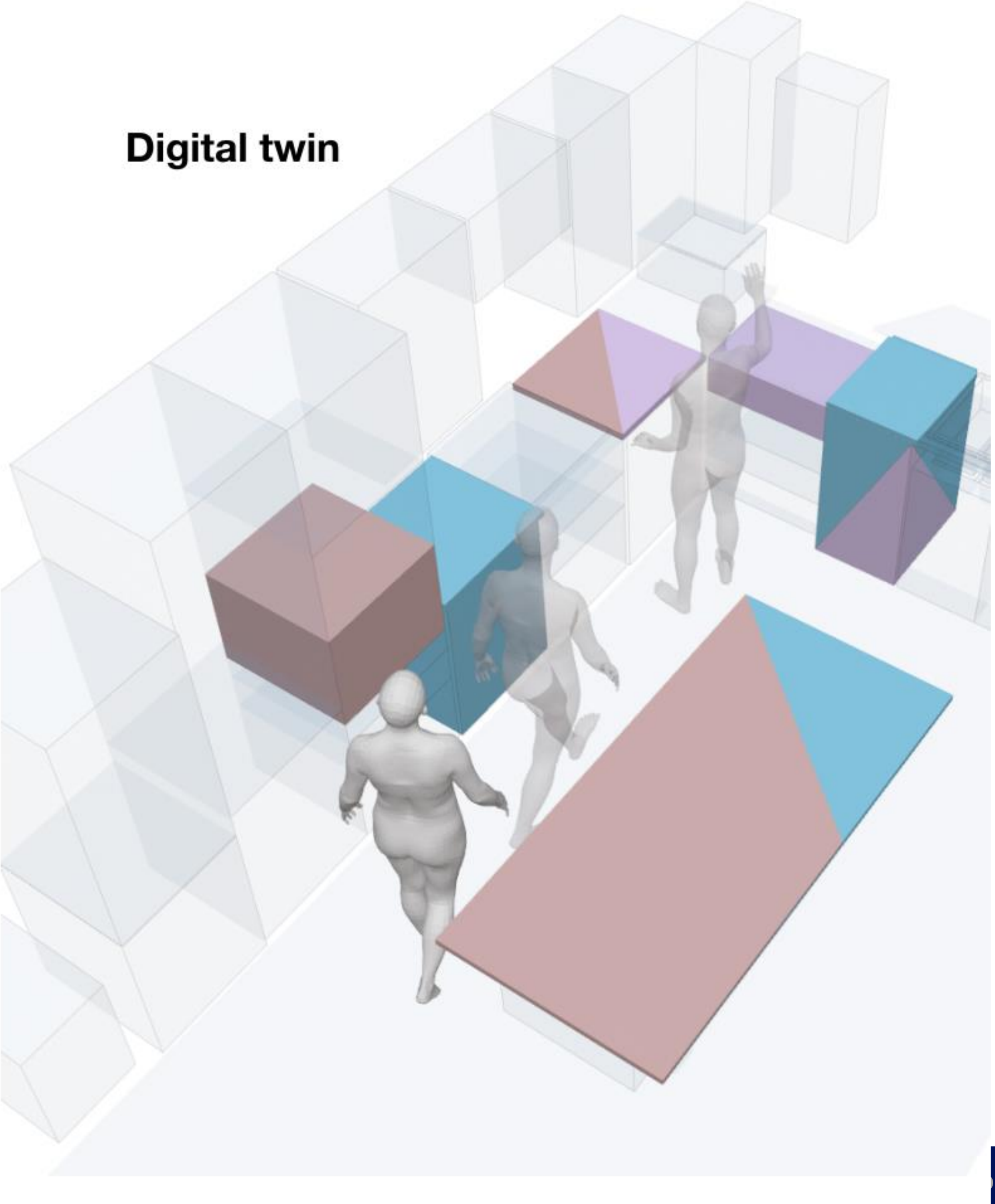
Perrett et al (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



Perrett et al (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

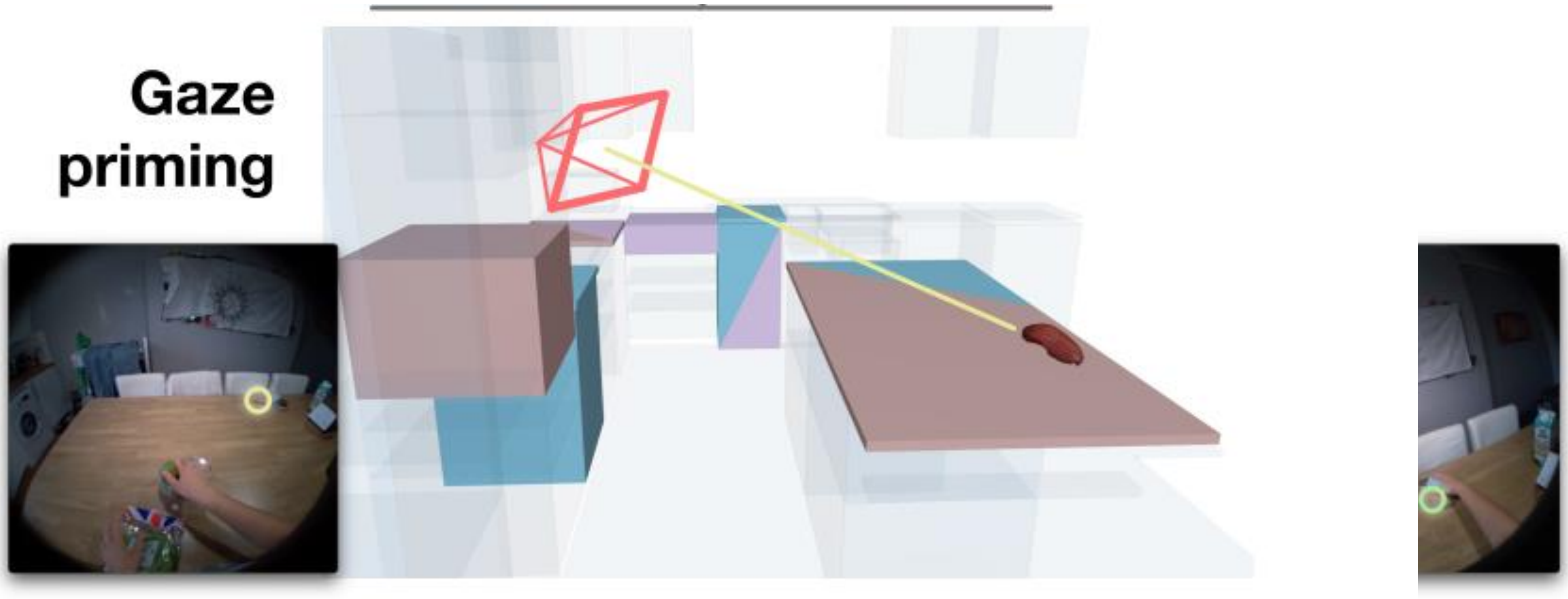
HD-EPIC



Perrett et

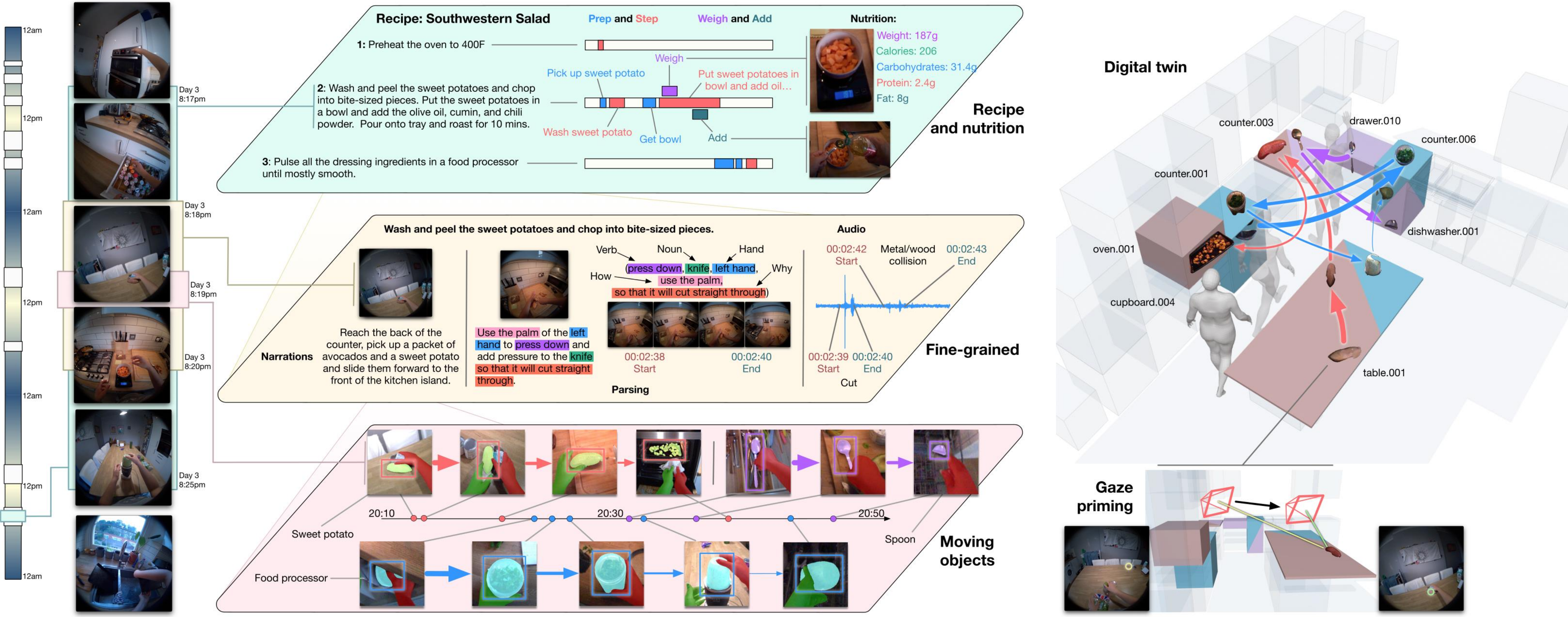
on]

HD-EPIC



Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC

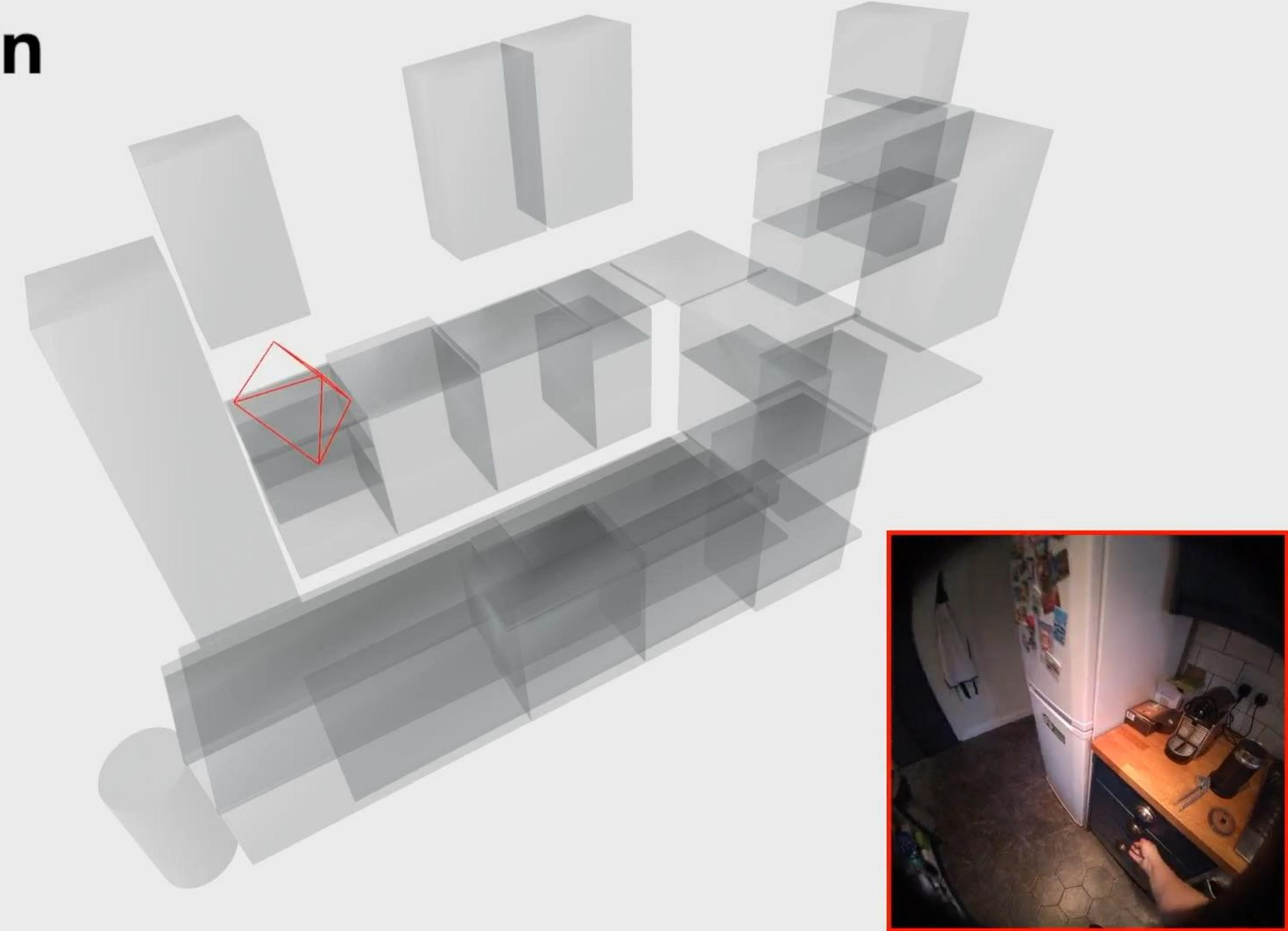


Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

Digital Twin

Fixtures

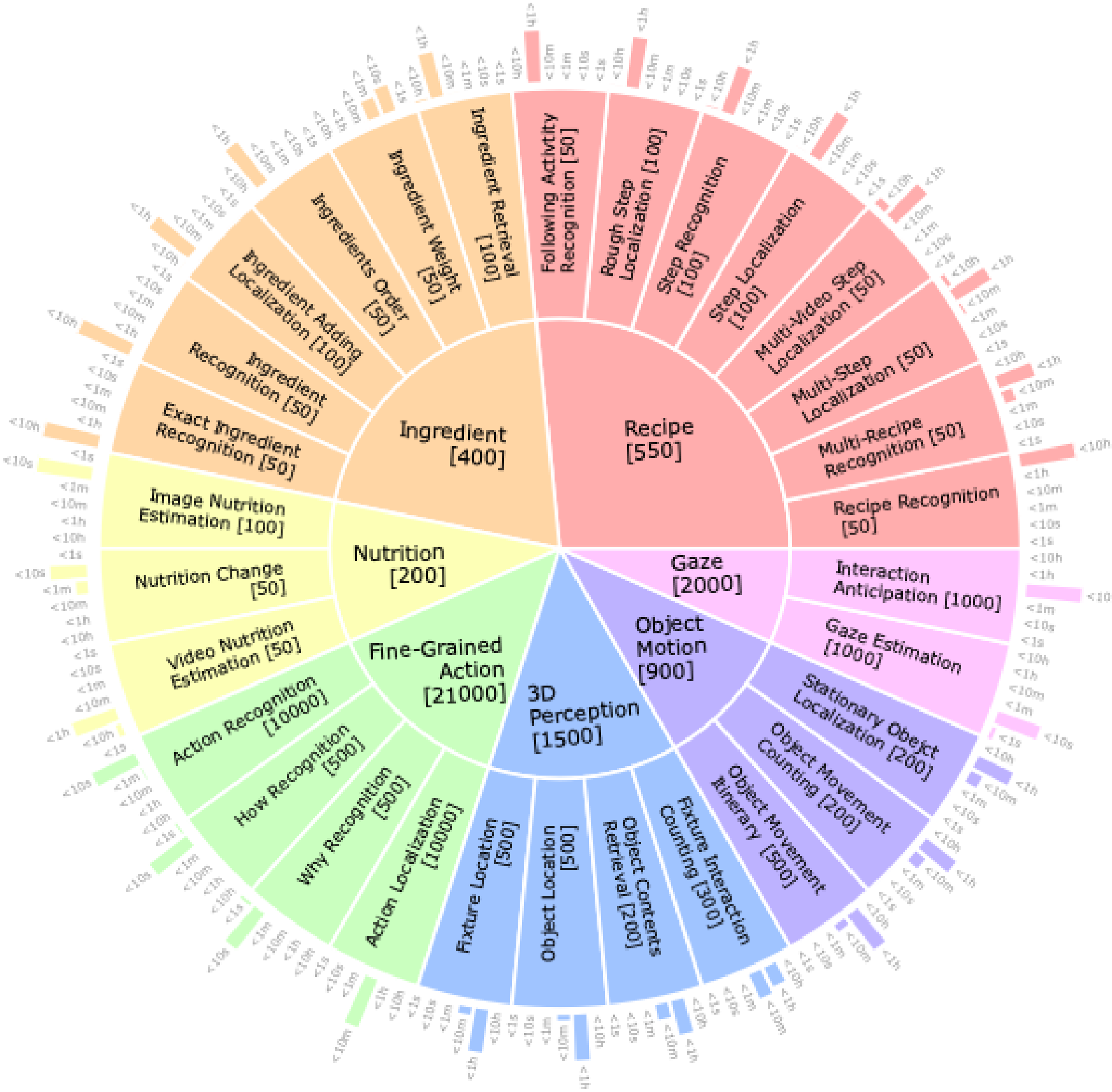
Open drawer



Highly-Detailed Narrations

HD-EPIC

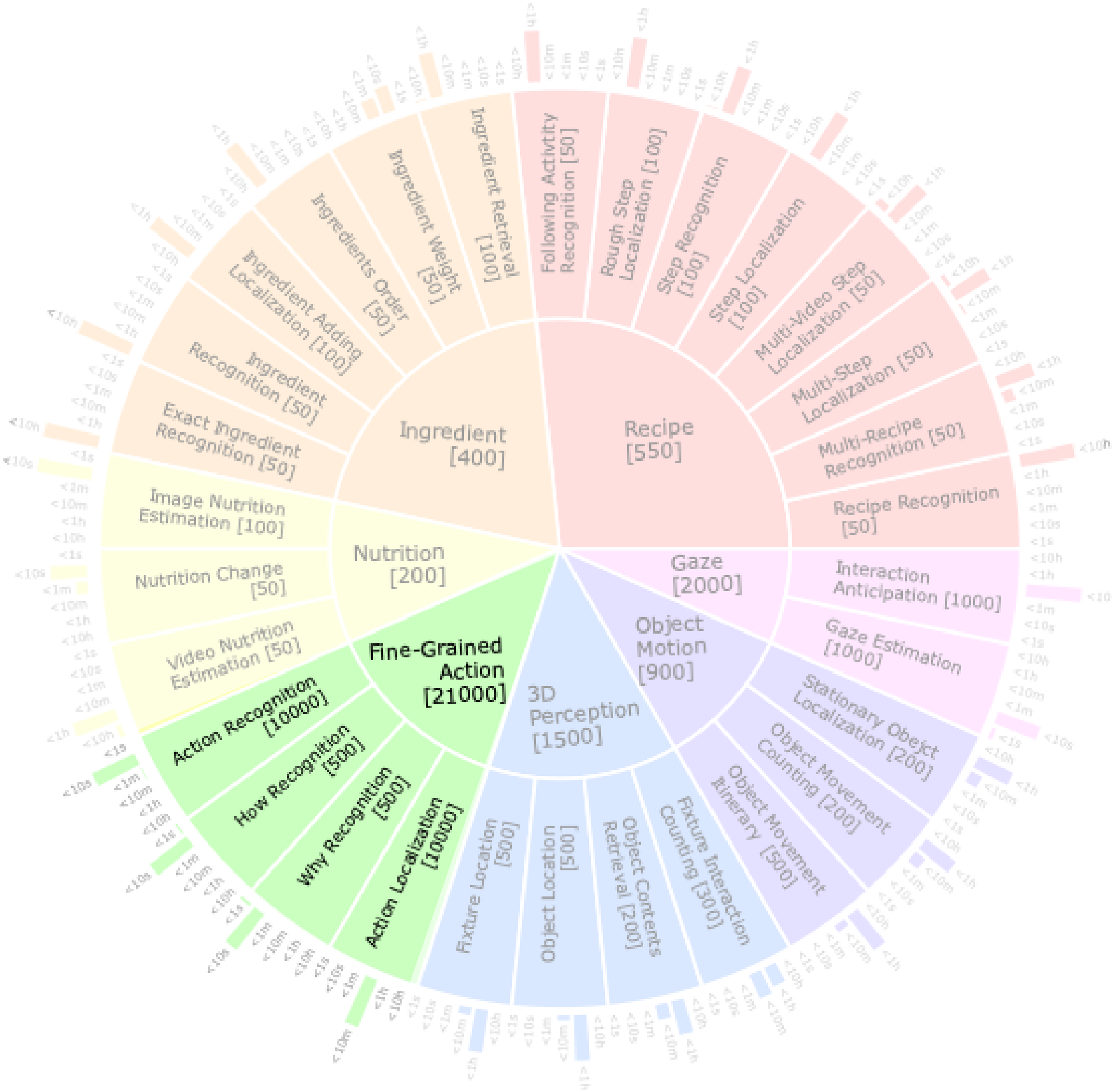
1. Recipe. Questions on temporally localising, retrieving, or recognising recipes and their steps.
2. Ingredient. Questions on the ingredients used, their weight, their adding time and order.
3. Nutrition. Questions on nutrition of ingredients and nutritional changes as ingredients are added to recipes.
4. Fine-grained action. What, how, and why of actions and their temporal localisation.
5. 3D perception. Questions that require the understanding of relative positions of objects in the 3D scene.
6. Object motion. Questions on where, when and how many times objects are moved across long videos.
7. Gaze. Questions on estimating the fixation on large landmarks and anticipating future object interactions.



Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC

1. Recipe. Questions on temporally localising, retrieving, or recognising recipes and their steps.
2. Ingredient. Questions on the ingredients used, their weight, their adding time and order.
3. Nutrition. Questions on nutrition of ingredients and nutritional changes as ingredients are added to recipes.
4. Fine-grained action. What, how, and why of actions and their temporal localisation.
5. 3D perception. Questions that require the understanding of relative positions of objects in the 3D scene.
6. Object motion. Questions on where, when and how many times objects are moved across long videos.
7. Gaze. Questions on estimating the fixation on large landmarks and anticipating future object interactions.



Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC

What is the best description for how the person carried out the action pick up bowl of coconut milk in this video segment? [00:18:44 - 00:18:46]

● **A.** Using both hands holding the bowl from bowl rim.

B. By holding both sides using the oven gloves.

●●● **C.** using the right hand and lift the large white bowl up.

D. using left hand and removing the fork used to stir it using right hand.

E. using both hands from the kitchen top above the dishwasher.

Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC

What is the best description for why the person performed the action turn tap in video ? [00:12:08 - 00:12:09]

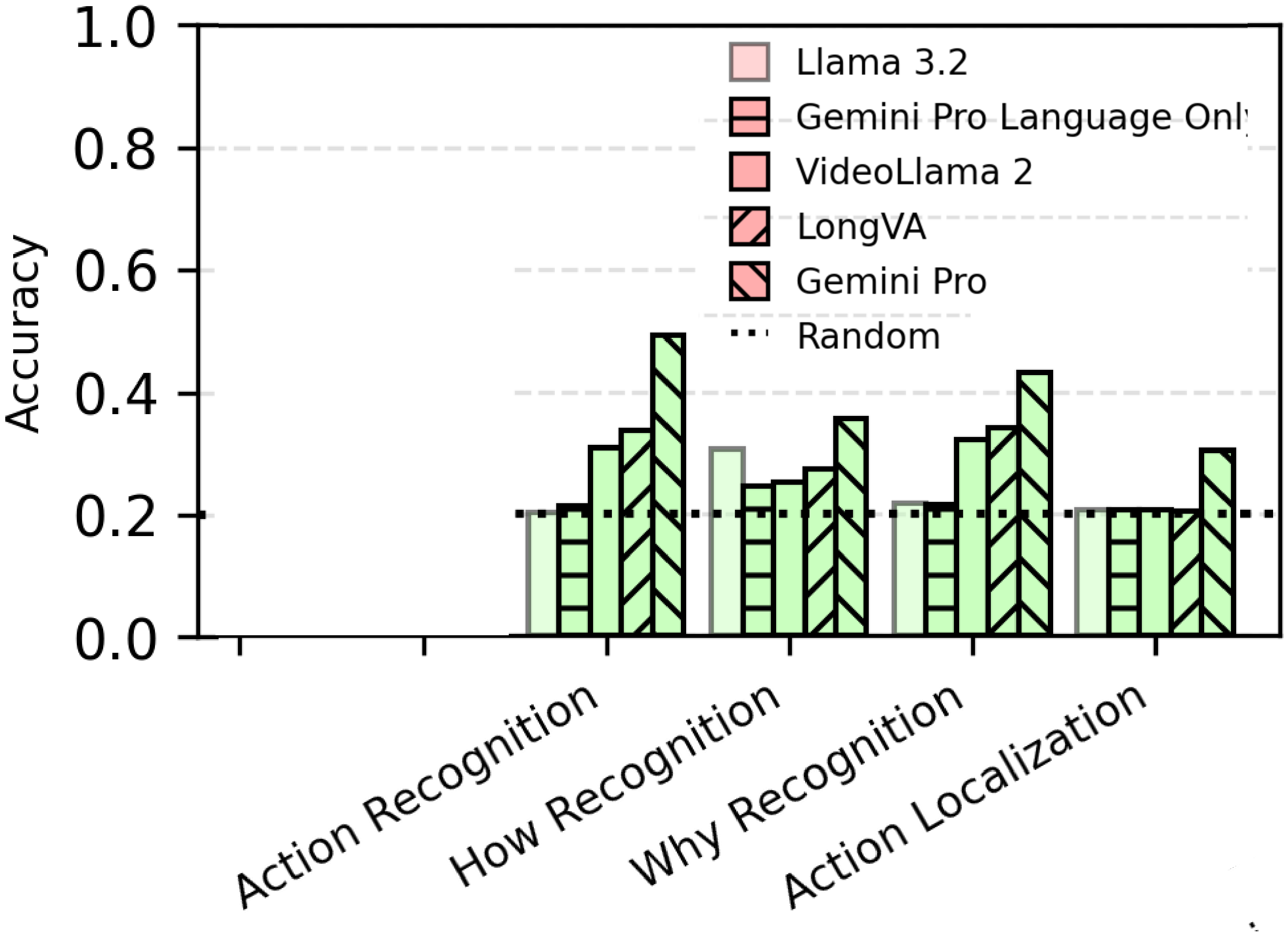
●● **A.** To increase the flow of water to speed up filling up the glass.

● **B.** So that the tap is above the sink strainer. **C.** To pour water ... the sink.

● **D.** Tap water falls on... inside the sink. **E.** To reduce water flow.

Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

HD-EPIC



Perrett et al. (2024). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. [Coming Soon]

Can’t be Done Alone

Aozhu Chen

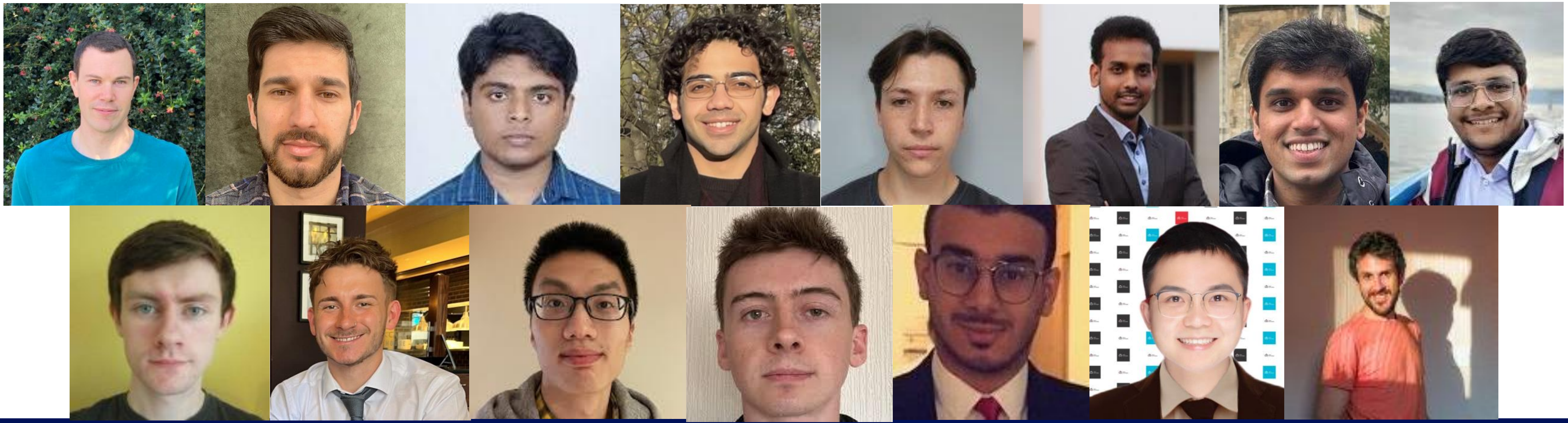
Fida Thoker

Cees Snoek

Michael Wray

Dima Damen

Kaiting Liu



Conclusion

- Data and evaluation are as important, if not more important than models

- Considering new data and new evaluation can be the key to needing new models

- Properly considering the task is also crucial

- Many assumptions made about common tasks, datasets and evaluation metrics that are worth questioning