Vision and Language: A Video Perspective

University of Bonn
MIT-IBM Watson AI Lab

BMVA Symposium on Vision and Language
Wednesday 17 January 2024

All opinions are my own

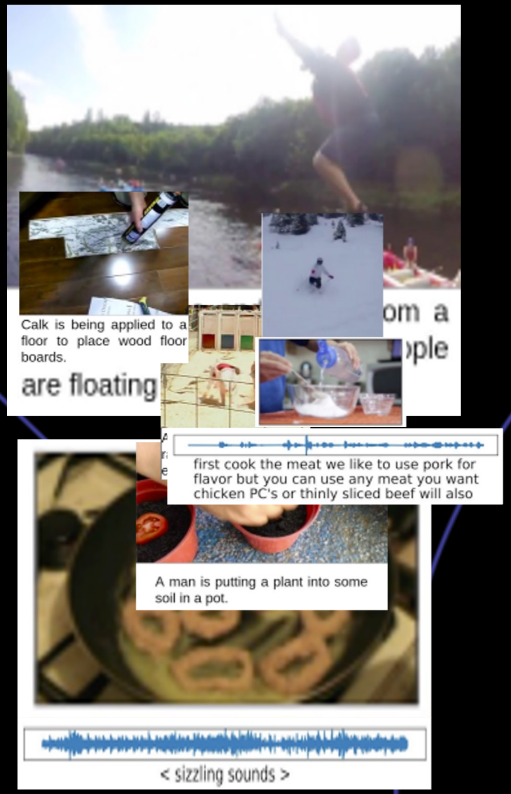
UNIVERSITÄT BONN

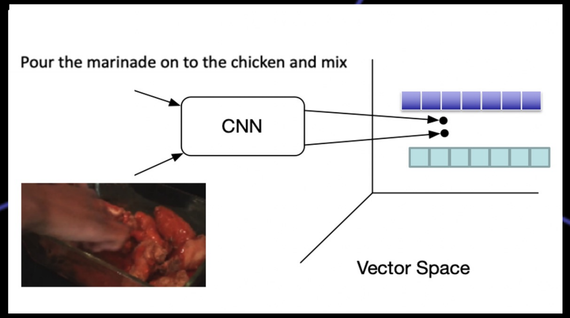
MIT-IBM Watson AI Lab

Vision-Language Learning

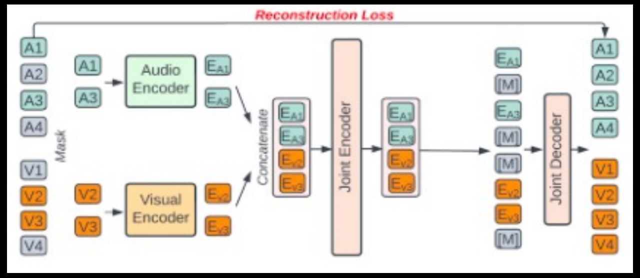
Vision-Language learning is the foundation of recent AI breakthroughs: CLIP, Stable diffusion, BLIP, and many more.

Allows for image generation, captioning, text-to image/video retrieval etc ...

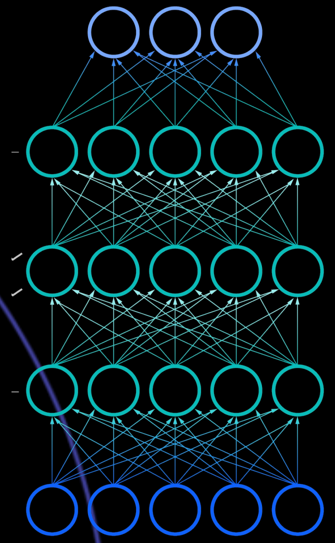




Via Contrastive Loss



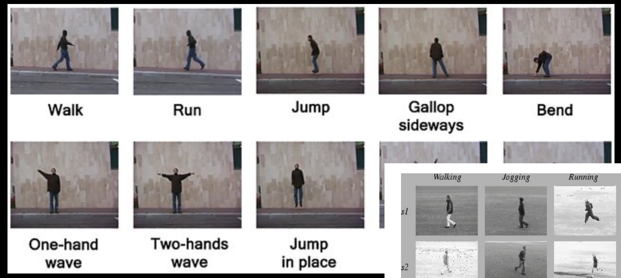
Via Reconstruction Loss



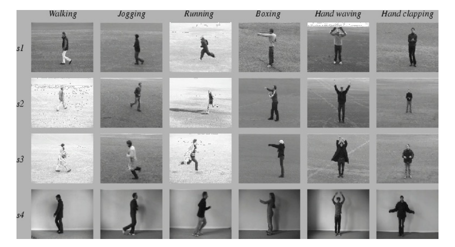
Source https://openai.com/research/dall-e

Video Understanding – A short history in datasets

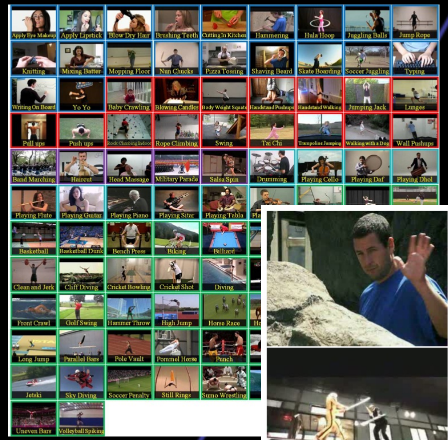
Classification



Weizman



KTH



UCF101

HMDB51

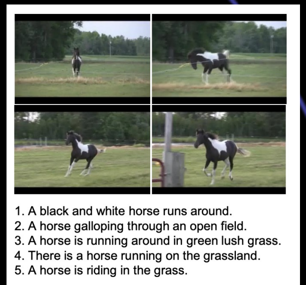
Kinetics 400 / 600 / 700

2000

2010

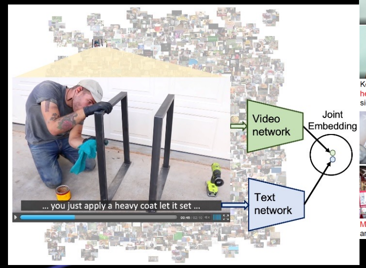
2020

Year

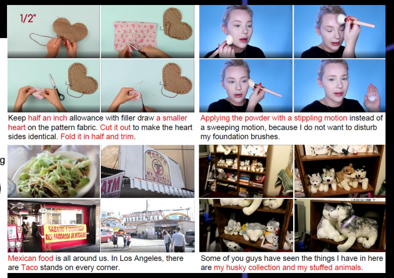


MSR-VTT

Youcook



HT100M



HD Vila

InternVid

VL-Retreival

Video understanding – What and where?

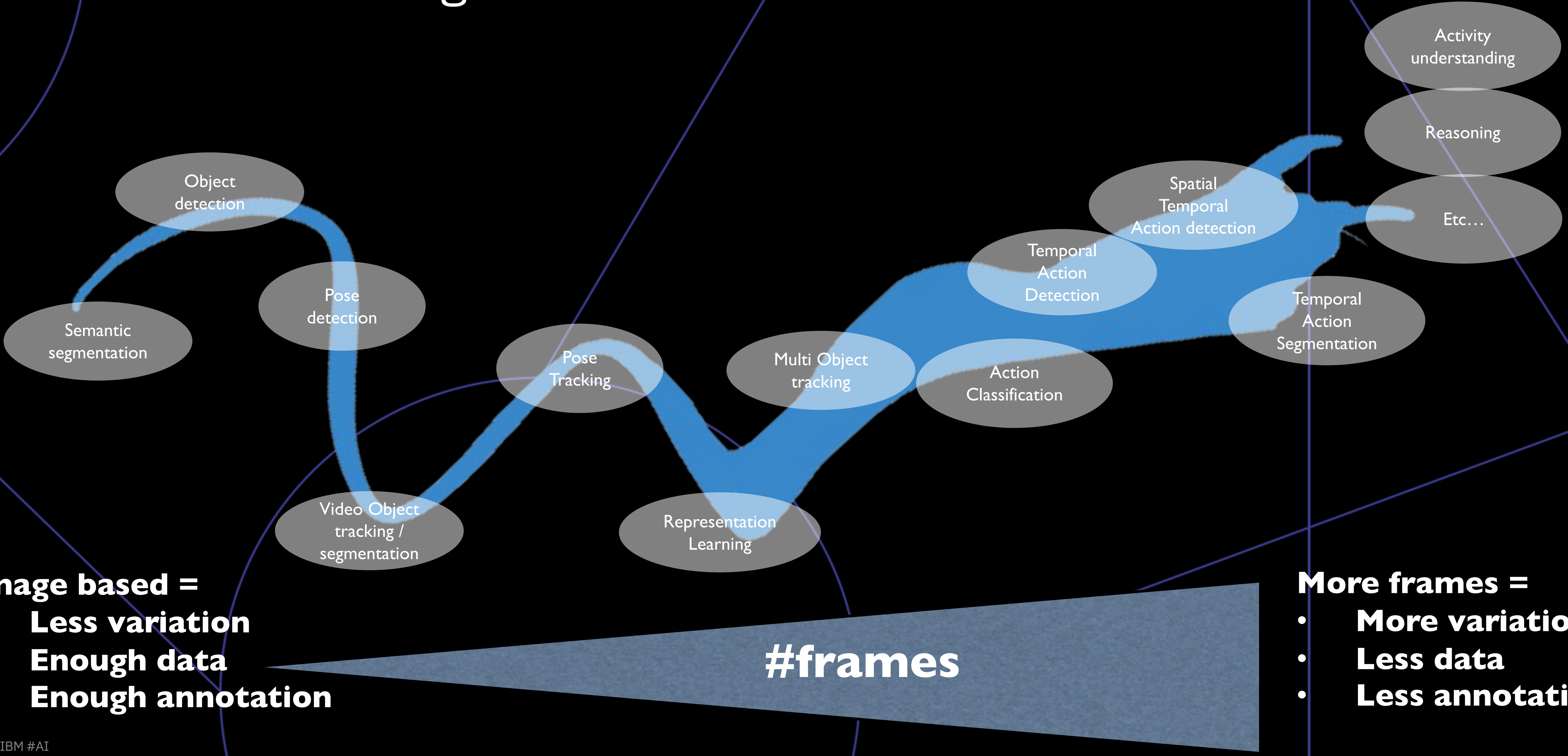


Image based =
- **Less variation**
- **Enough data**
- **Enough annotation**

More frames =
- **More variation**
- **Less data**
- **Less annotation**

UNIVERSITÄT BONN

Video understanding – How to scale?

The bigger the datasets, the harder to annotate ...

For a top-down dataset, you need to define action classes.

For a larger dataset, you ...

1) Need to define and find **more** action classes

2) Need to define **more distinct** action classes

3) Need to find/record **more videos** with **distinct** action classes

Video understanding – How to scale?

The bigger the datasets, the harder to annotate ...

Outlook

Weakly supervised

Idea: many videos (espe
Since ~2006 YouTube o
Since ~ 2009 YouTube p

Make something out of t

E.g. … How many video

Weak YouTube dataset

universitätbonn

Finding action classes in natural language:

Use NLP tagging to define verb-object-pairs
Filter for cooking relevant classes:

~ 800 pairs → cluster to ~180 action classes

Old and new problems:

- Not all verbal comments refer to actions (Hit accuracy: ~60%)
- Homonym (multiple expression have the same meaning)
- Polysemy (one expression has multiple meanings)
- Inconsistent granularity
- Semantic clustering not always consistent
- Highly imbalanced distribution

Example ‚roll it'

Example ‚put it'

10.04.2017 Lab Talk - INRIA

Lessons learned (2017):

- Classification does not work bottom up
→ might not scale
- Gap between natural language and class labels

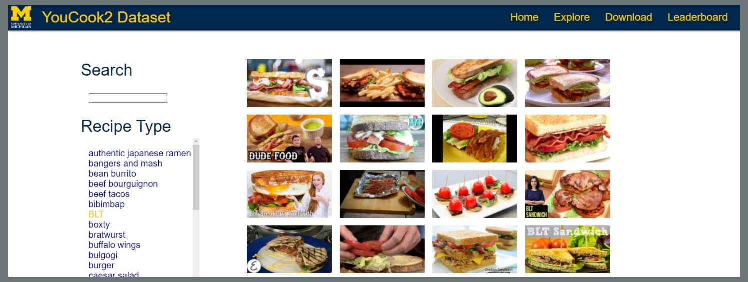
Vision-Language Learning for Video

- Vision-language models classify data without being trained on the test classes/datasets.

Setup:

Large-scale training data (e.g. HowTo100M)
on audio (A), video (V), text (T) ASR or caption

Zero shot testing:



YouCook2 (T→V / T→VA)

MSRVTT (T→V / T→VA)

MiningYouTube (localization)

CrossTask (localization)

Vision-Language embedding space

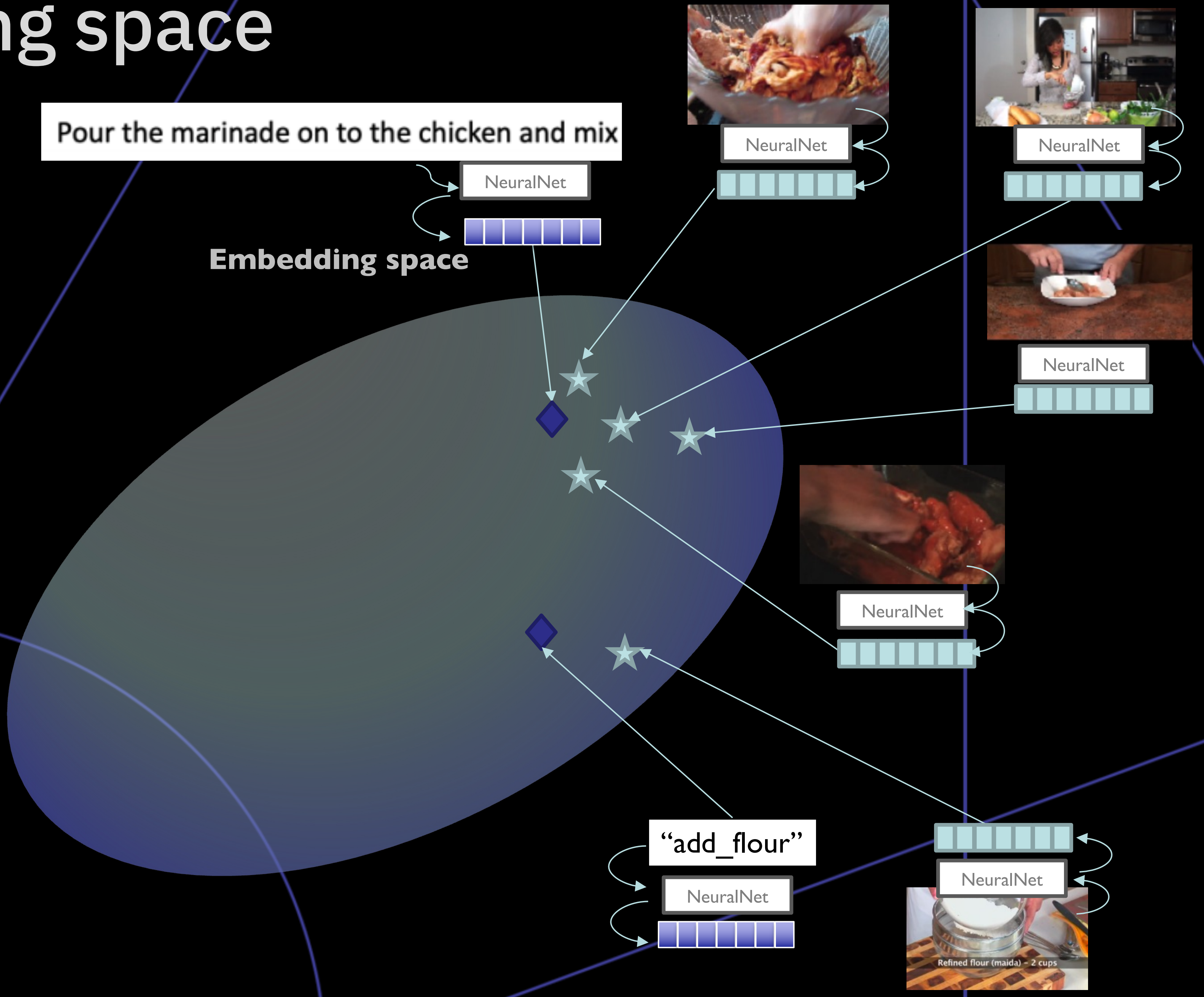
Tasks:

- Retrieval (cross-modal)

→ Based on distance between reference and test samples

Retrieval related:

- Zero-shot Classification
- Zero-shot Temporal detection/segmentation



Why Video Needs Language...

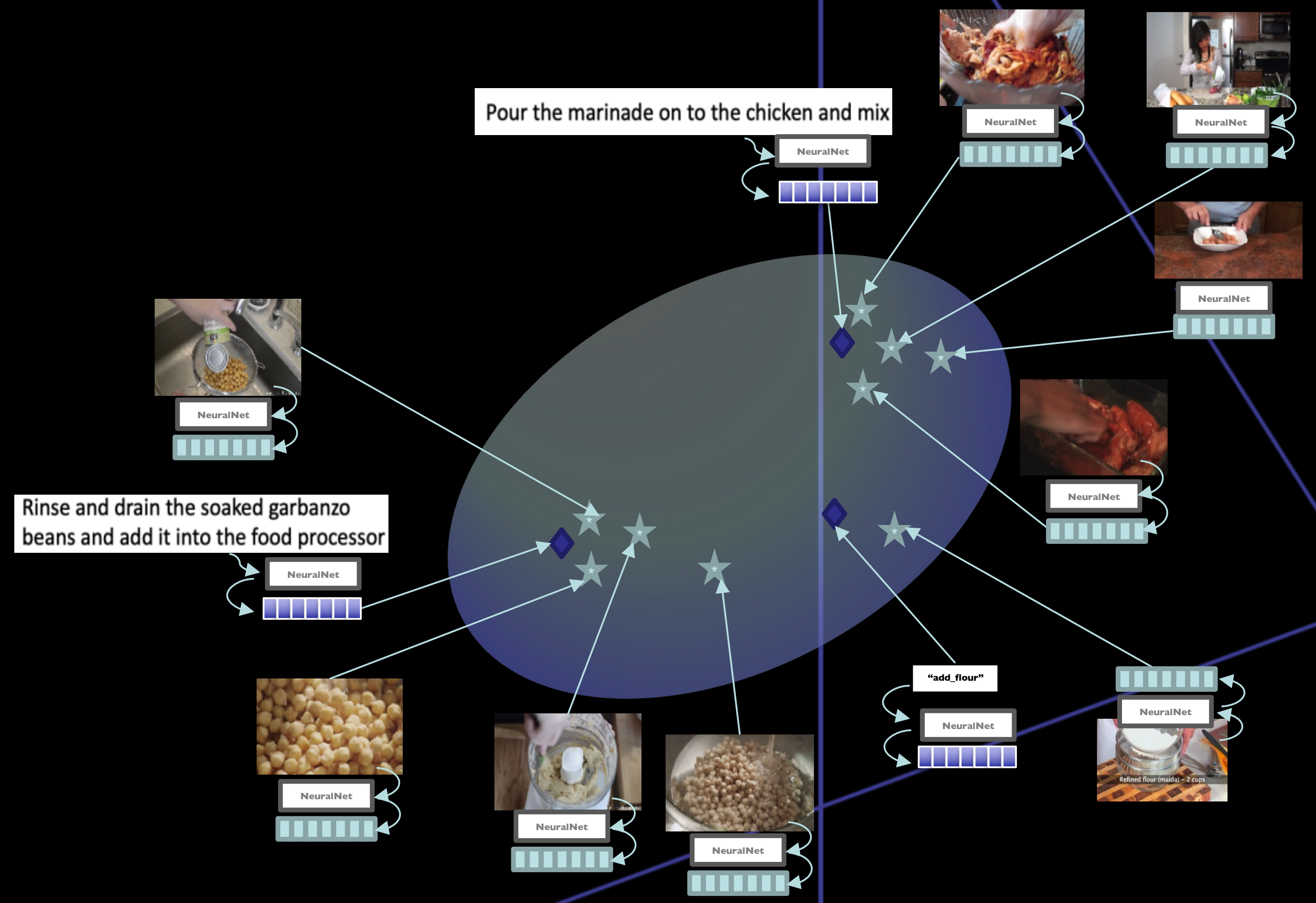
Real-world video understanding can be difficult ...

Actions are not well defined
- Perception/labels are subjective, depend on duration, expertise etc.

Actions are unconstrained
- They don't have a physical outline

There is no fix/complete "taxonomy" on actions
- Not possible to learn a vocabulary

Lack of annotated data
- We will never be able to label action data at a significant (real-world) scale

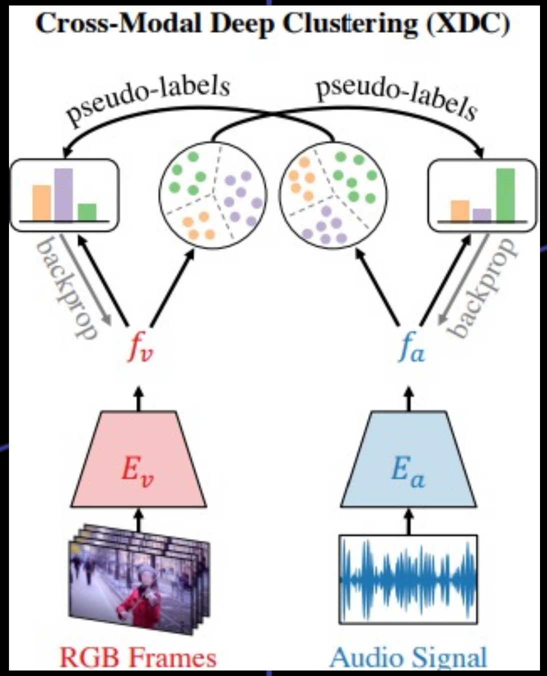


Recap of works in the field...

Incomplete list:

- HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips [Miech, ICCV 2019]
- End-to-end learning of visual representations from uncurated instructional videos. [Miech, CVPR 2020]
- Selfsupervised multimodal versatile networks. [Alayrac, NeurIPS 2020]
- Self-supervised learning by cross-modal audio-video clustering. [Alwassel, NeurIPS2020]
- Labelling unlabelled videos from scratch with multi-modal self-supervision. [Asano, NeurIPS 2020]
- Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval [ICCV 2021]
- MERLOT Reserve: Multimodal Neural Script Knowledge through Vision and Language and Sound [Zellers, CVPR2022]
- Learning Audio-Video Modalities from Image Captions [Nagrani,2022]
- Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset [Thapliyal, 2022]
- Many more ...

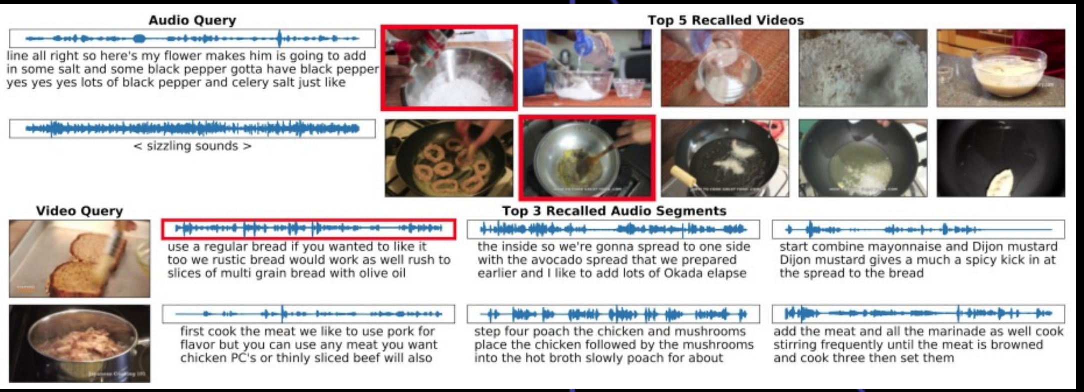
Miech, ICCV 2019



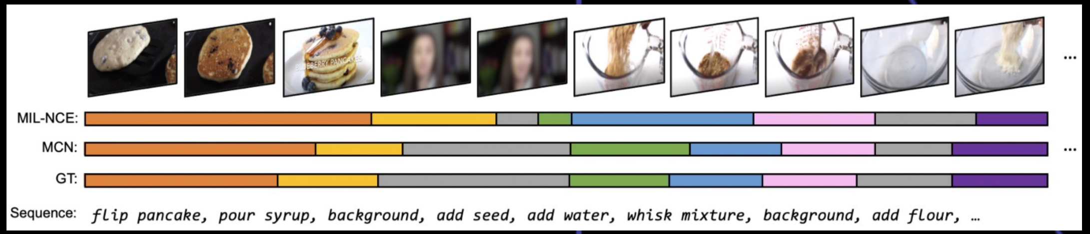
Alwassel, NeurIPS2020

Recap of our work in the field…

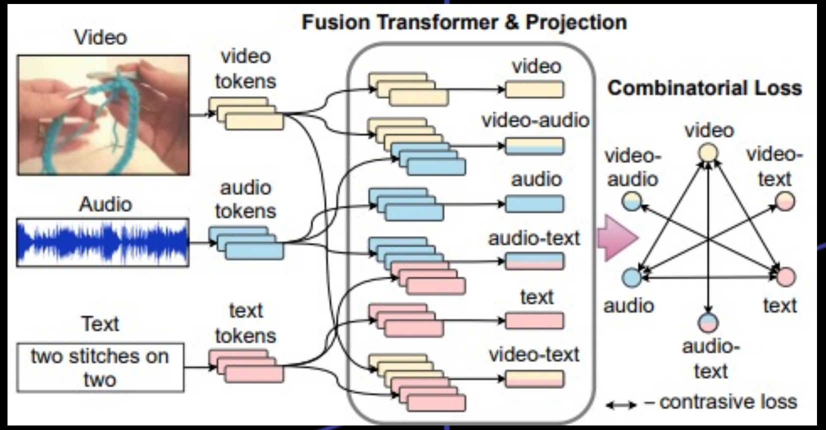
- **AVLnet: Learning Audio-Visual Language Representations from Instructional Videos** [Roudichenko et al., arxiv 2020, Interspeech *2021*]
- Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos [B. Chen et al., ICCV2021]
- Everything at Once – Multi-modal Fusion Transformer for Video Retrieval [N. Shvetsova et al., CVPR 2022]
- Preserving Modality Structure Improves Multi-Modal Learning [Sirnam et al., ICCV 2023]



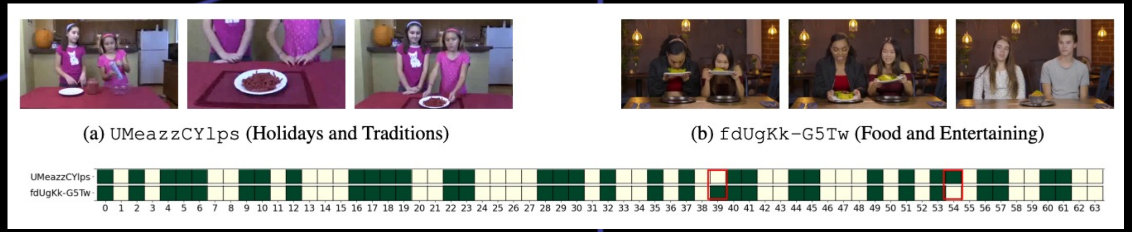
[Roudichenko et al., arxiv 2020, Interspeech *2021*]



[B. Chen et al., ICCV2021]



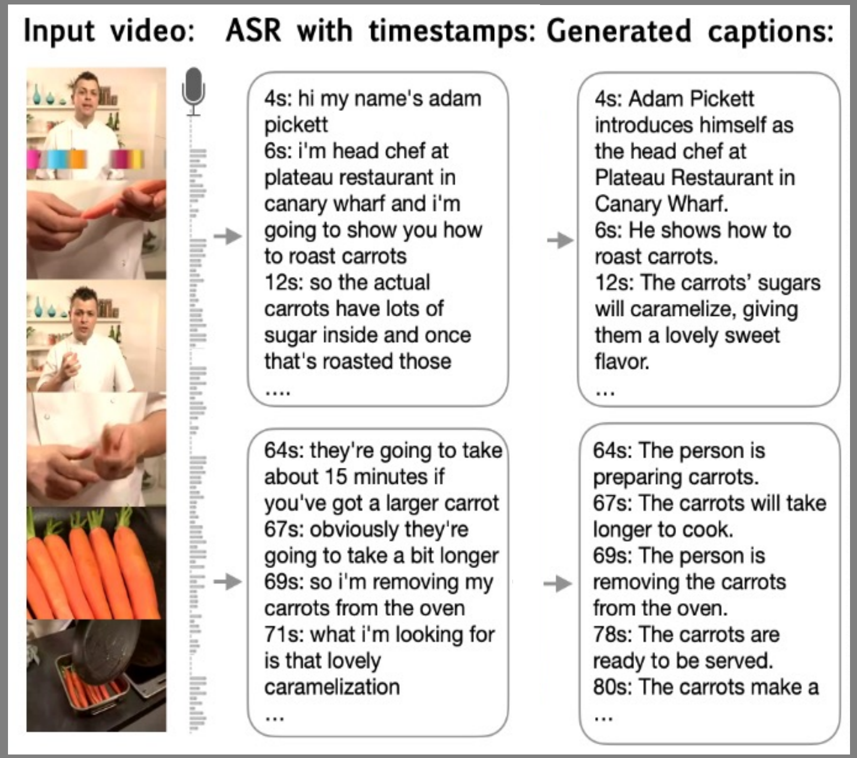
[N. Shvetsova et al., CVPR 2022]



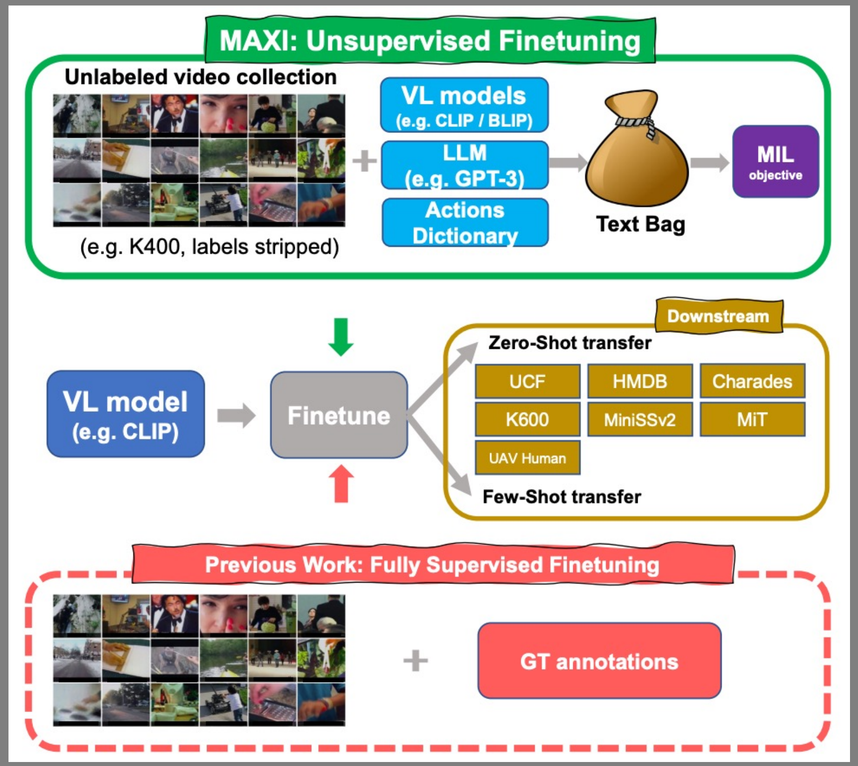
Sirnam et al., ICCV 2023

Vision-Language for (better) Video Understanding

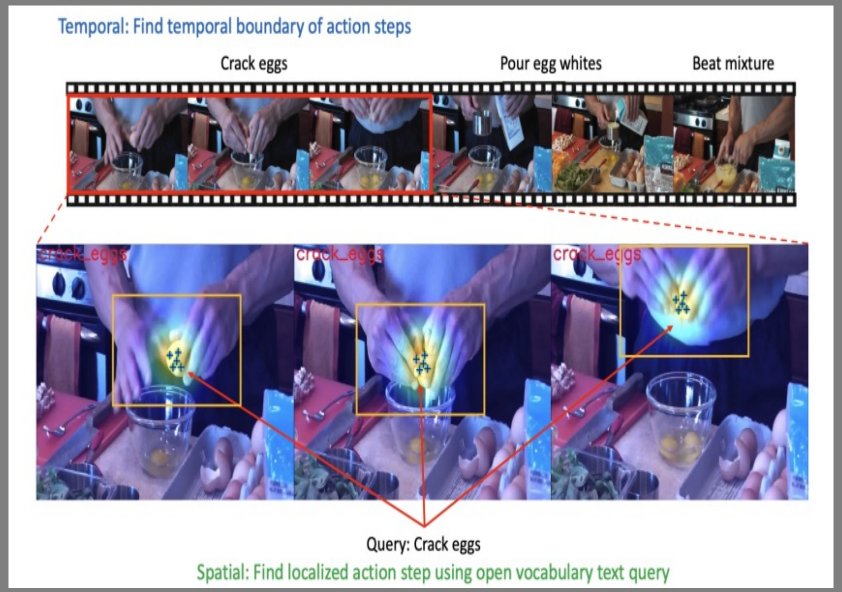
Fixing language for better multimodal learning



Action Classification in Times of Vision-Language Models



What, when, and where? Spatial-Temporal Grounding in Videos



Fixing language for better multimodal learning

What's wrong with language in video?

- Language (and topic) domain shift between downstream datasets
- Language domain shift between free training data (ASR subtitles) and downstream datasets (human annotated captions)

Dataset	Examples
MSR-VTT (∼43 symbols in a text)	1) The peoples are sharing their view on this car of different models 2) Someone is showing the ingredients for a dish they are going to make 3) A man is playing an instrument
YouCook2 (∼39 symbols in a text)	1) Combine macaroni sauce and cheese 2) Grate and cube potatoes 3) Stir in crushed tomatos
DiDeMo (∼147 symbols in a text)	1) A dog runs down a hill and stop behind a shrub. Dog sniffs and chews at patch of grass on rock. the dog approaches, then begins to sniff the cluster of plants first time hand is seen petting dog. 2) Only big screen is visible the camera first pans to the large screen. The view shifts from the basketball court to the fans in the seats across the stadium. Camera goes to the bigscreens the dancers are shown on the jumbotraun. 3) A bus stops. The bus stops at the end of the driveway. A kid is coming out of a school bus. School bus doors open.
MSVD (∼31 symbols in a text)	1) The cats are fighting 2) The lady sliced a vegetable 3) A man is eating a pizza
LSMDC (∼46 symbols in a text)	1) SOMEONE goes to the kitchen, wets a towel, comes back to the bed, kneels it, places the towel on SOMEONE's brow. 2) He slaps SOMEONE again. 3) SOMEONE moves off through the crowd.

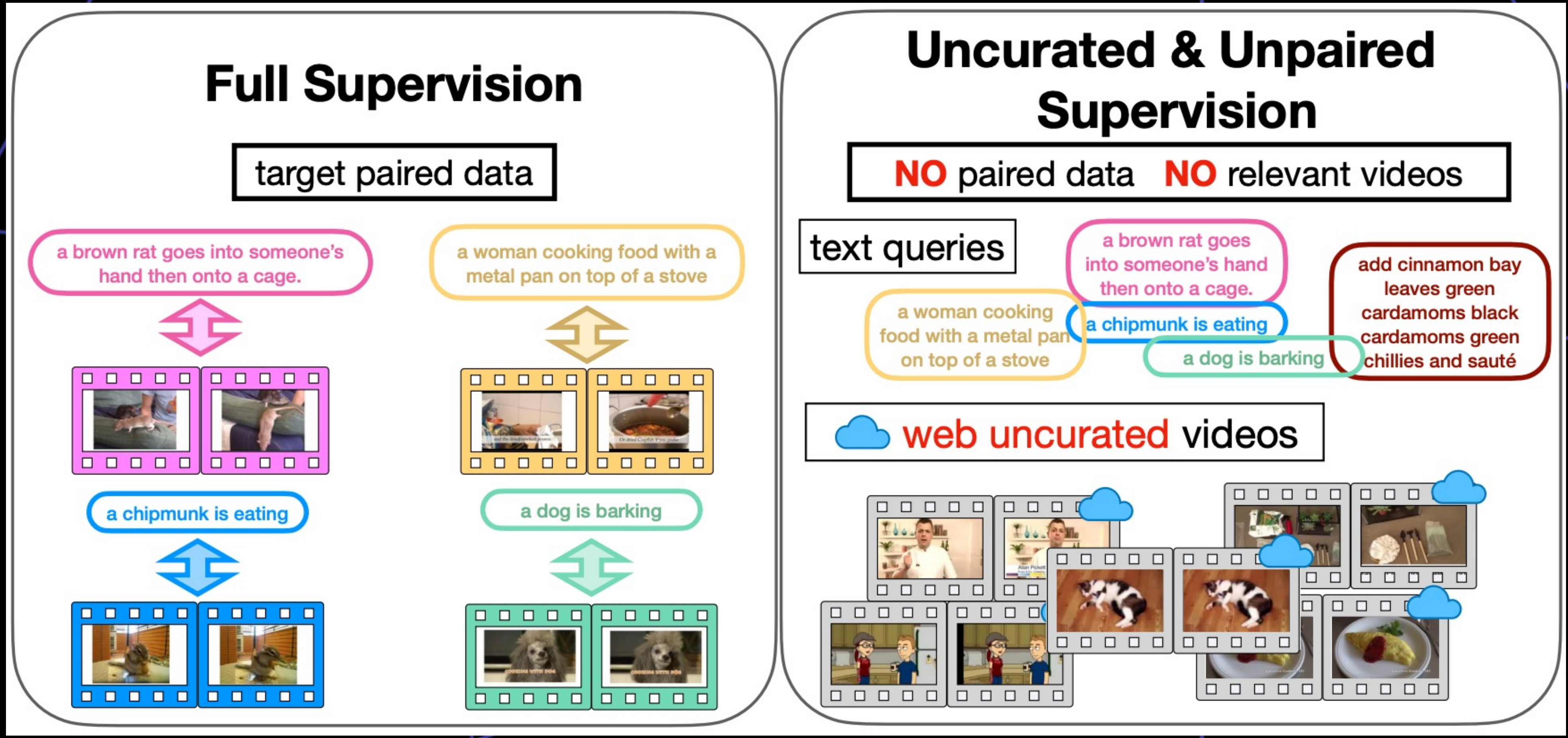
In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video Retrieval [N. Shvetsova & A. Kukleva et al., ICCV 2023]

Language domain shift in downstream datasets …

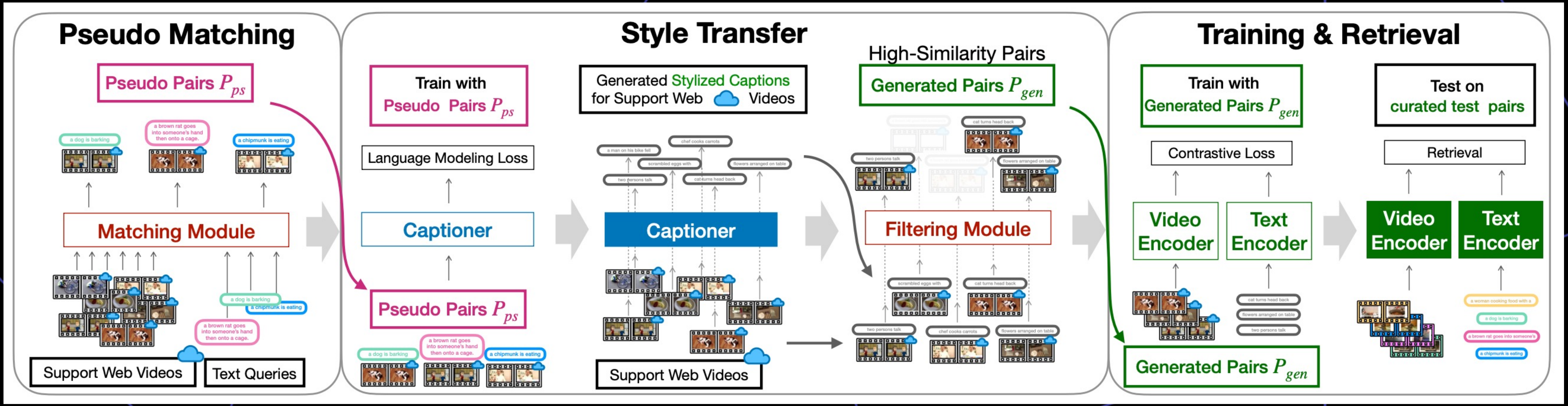
… is not a bug, it's life!

1) Deal with it!

2) Without training data!



In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video Retrieval [N. Shvetsova & A. Kukleva et al., ICCV 2023]

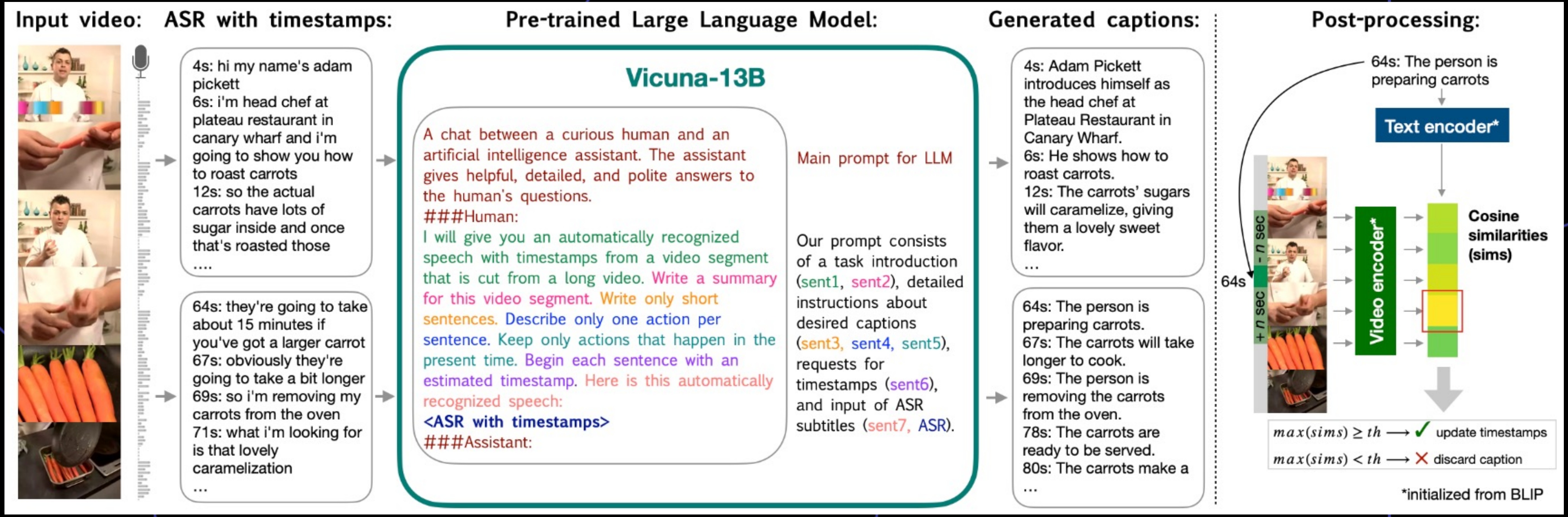


In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video Retrieval [N. Shvetsova & A. Kukleva et al., ICCV 2023]

Method	Image-Text Datasets	Video-Text Datasets	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC			
			R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
HowTo100M [43]	-	HowTo100M	7.5	21.2	29.6	38	6.1	17.3	24.8	46	-	-	-	-	-	-	-	-	-	-	-	-
SupportSet [48]	-	HowTo100M	8.7	23.0	31.1	31	-	-	-	-	-	-	-	-	8.9	26.0	37.9	18	-	-	-	-
VATT [1]		HowTo100M+AS	-	-	29.7	49	-	-	45.5	13	-	-	-	-	-	-	-	-	-	-	-	-
EAO§ [55]	-	HowTo100M	9.9	24.0	32.6	28	**19.8**	**42.9**	**55.1**	**8**	6.6	19.0	26.8	42	18.0	40.4	52.3	9	3.6	8.5	13.0	177
Nagrani et al. [45]	-	VideoCC3M	19.4	39.5	50.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Frozen in Time [3]	CC+COCO	WebVid-2M	24.7	46.9	57.2	7	-	-	-	-	21.1	46.0	56.2	7	-	-	-	-	-	-	-	-
CLIP-straight [49]	WIT	-	31.2	53.7	64.2	4	-	-	-	-	-	-	-	-	37.0	64.1	73.8	**2**	11.3	22.7	29.2	56.5
CLIP4CLIP [38]	WIT	HowTo100M	32.0	57.0	66.9	4	-	-	-	-	-	-	-	-	38.5	66.9	76.8	**2**	15.1	28.5	36.4	28
Nagrani et al. [45]	WIT	VideoCC3M	33.7	57.9	67.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BLIP$^{\|\|}$ [31]	CC+COCO+3more*	-	33.3	57.3	67.5	3.5	5.8	15.0	21.9	76	24.6	50.4	59.7	5.3	37.0	63.3	72.6	3	15.2	28.2	35.9	35
In-Style (ours) (CLIP)	WIT	HowTo100M†+VATEX‡	35.0	59.6	70.4	**3**	5.1	14.0	20.3	103	26.6	50.5	62.6	5	38.6	66.3	77.9	3	16.0	**31.6**	38.5	**26.5**
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M†+VATEX‡	**36.0**	**61.9**	**71.5**	**3**	6.8	16.7	24.5	63	**29.4**	**59.2**	**68.6**	**3.5**	**44.9**	**72.7**	**81.1**	**2**	16.4	30.1	38.7	28
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M†+WikiHow	34.2	59.6	69.0	**3**	7.3	19.2	27.1	46	29.7	56.2	67.4	4	42.8	70.2	79.1	**2**	**17.0**	30.8	**39.6**	27
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M†+Food.com	32.8	54.9	65.8	4	7.2	19.8	27.9	47	25.7	52.8	63.1	5	39.5	64.9	74.9	**2**	14.5	28.9	37.2	30.5
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M†+Target‡	36.2	61.8	71.9	3	8.6	21.6	30.0	37	32.1	61.9	71.2	3	44.8	72.5	81.2	2	16.1	33.6	39.7	25
In-Style (ours) (EAO)	-	HowTo100M+Target‡	16.4	35.8	48.9	10	20.3	46.4	58.8	7	13.2	31.6	44.0	15	23.4	50.0	62.4	5	4.9	12.3	16.7	94

HowToCaption: Prompting LLMs to Transform Video Annotations at Scale (N. Shvetsova & A. Kukleva et al., arxiv)

Convert noisy ASR subtitles of instructional videos into video captions
→ high-quality video captions at scale without human supervision



HowToCaption: Prompting LLMs to Transform Video Annotations at Scale (N. Shvetsova & A. Kukleva et al., arxiv)

HowToCaption – LLM Prompting + Filtering + Alignment

Caption Post-processing	YouCook2 R10↑	YouCook2 MR↓	MSR-VTT R10↑	MSR-VTT MR↓	MSVD R10↑	MSVD MR↓	LSMDC R10↑	LSMDC MR↓	Average R10↑	Average MR↓
Lower bound: original ASR as supervision	39.3	20	61.7	5	77.1	2	31.5	56	52.4	20.8
No post-processing	40.2	18	65.9	4	79.8	2	34.4	40	55.1	16.0
Filtering (using BLIP)	42.5	16	71.2	3	81.7	2	37.4	30	58.2	12.8
Filtering&alignment (using BLIP)	42.4	17	71.7	3	82.2	2	38.5	29.5	58.7	12.9
Filtering&alignment (with ours)	44.1	15	73.3	3	82.1	2	38.6	29	59.5	12.3

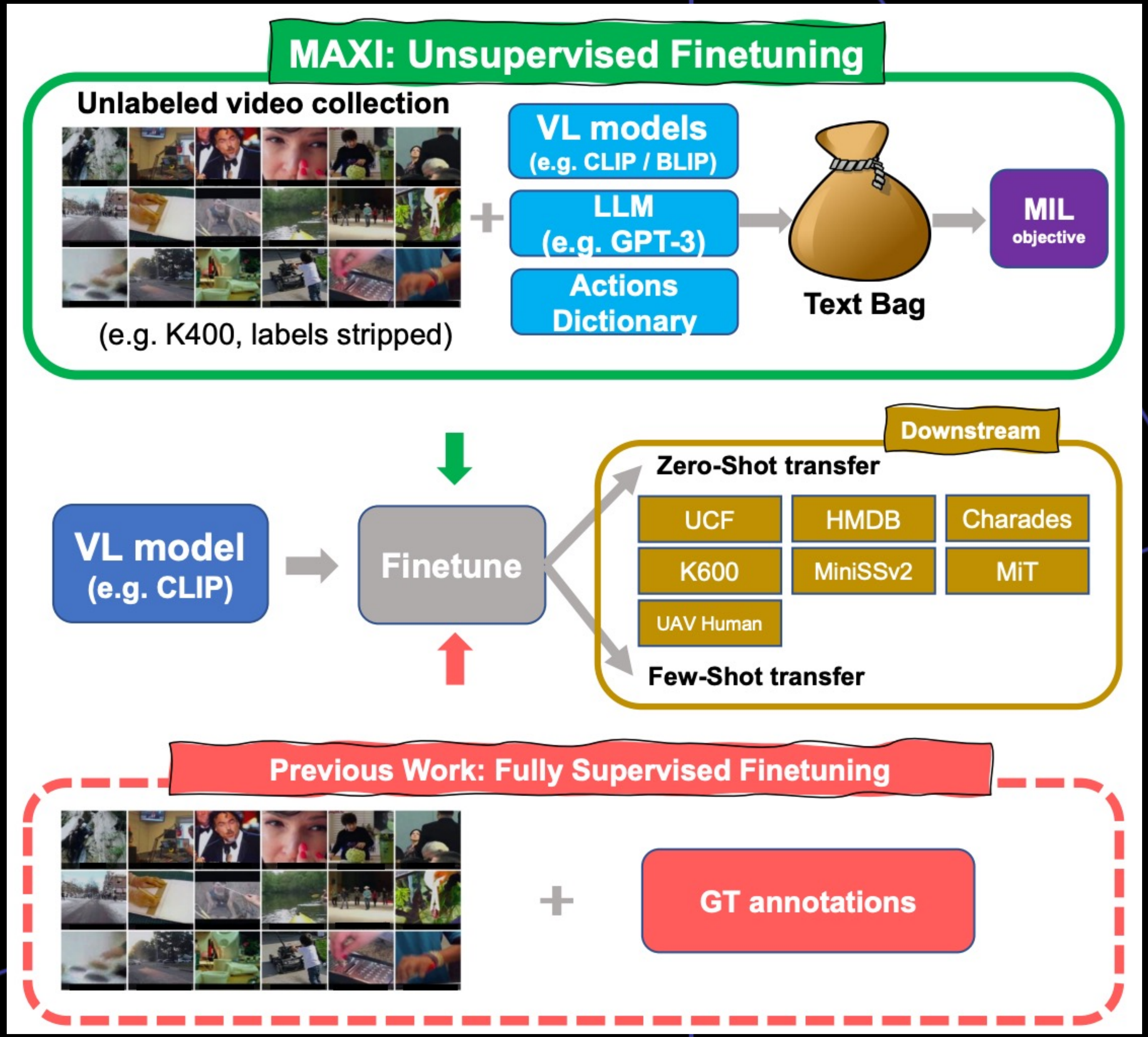
MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge [W.Lin et al. , ICCV 2023]

Problem: Why is CLIP bad on Kintetics?

→ Vocabulary gap between VL pretrained models and action classification

→ Usually fixed by fine-tuning with GT

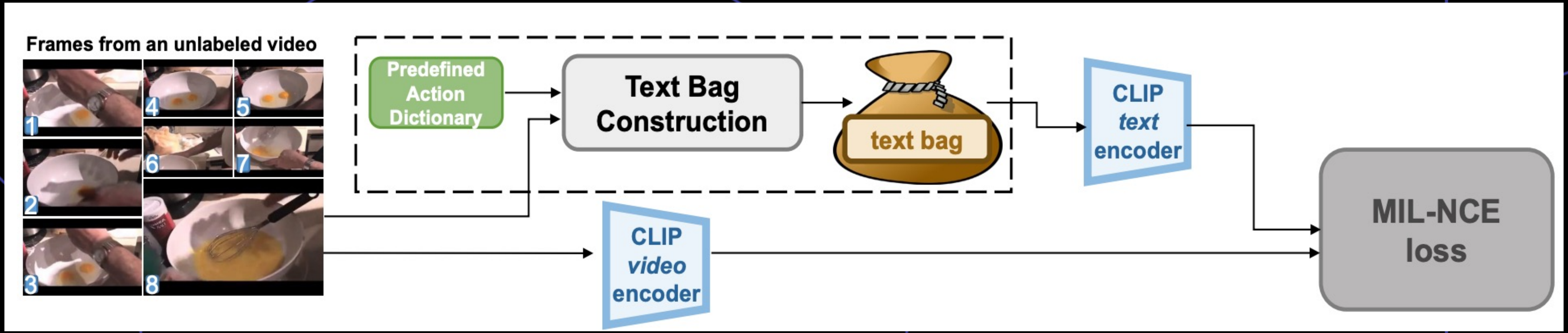
→ Idea: Can we fix it without annotations?



MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge [W.Lin et al. , ICCV 2023]

Input:
- Videos (without labels), Vocabulary (without videos), Pretrained VL / LLM model

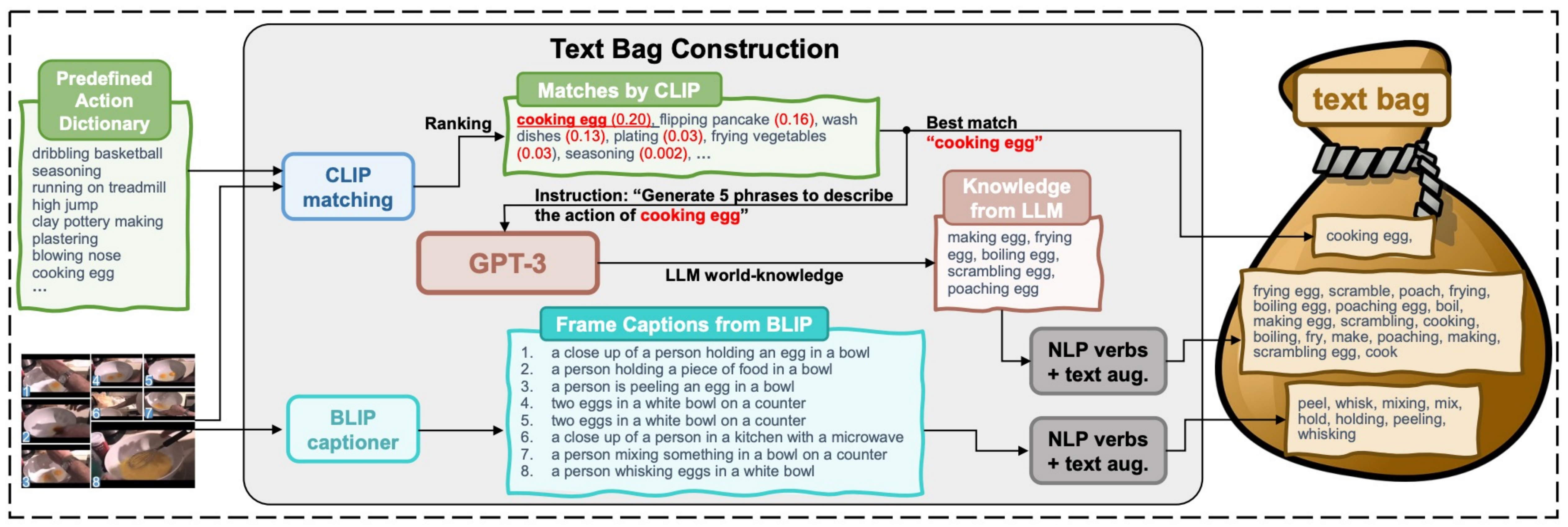
Idea:
- Construct a bag of text samples from vocabulary
- Match bag of text samples via Multiple Instance Loss → MIL-NCE



MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge [W.Lin et al. , ICCV 2023]

Text Bag Options:

1) Preselect best vocabulary matches via VL model
2) Use LLM to create synonyms, rephrasing etc.
3) Use captioner to generate more samples



MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge [W.Lin et al. , ICCV 2023]

Results:

Method	gt	language	vis.encoder	frames	UCF101	HMDB51	K600 Top1	K600 Top5
ER-ZSAR [7]	yes	Manual description	TSM	16	51.8 ± 2.9	35.3 ± 4.6	42.1 ± 1.4	73.1 ± 0.3
JigsawNet [34]	yes	Manual description	R(2+1)D	16	56.0 ± 3.1	38.7 ± 3.7	-	-
ActionCLIP [47]	yes	K400 dict.	ViT-B/16	32	58.3 ± 3.4	40.8 ± 5.4	66.7 ± 1.1	91.6 ± 0.3
XCLIP [33]	yes	K400 dict.	ViT-B/16	32	72.0 ± 2.3	44.6 ± 5.2	65.2 ± 0.4	86.1 ± 0.8
A5 [18]	yes	K400 dict.	ViT-B/16	32	69.3 ± 4.2	44.3 ± 2.2	55.8 ± 0.7	81.4 ± 0.3
ViFi-CLIP [38]*	yes	K400 dict.	ViT-B/16	16	74.9 ± 0.6	50.9 ± 0.7	67.7 ± 1.1	90.8 ± 0.3
ViFi-CLIP [38]	yes	K400 dict.	ViT-B/16	32	76.8 ± 0.7	51.3 ± 0.6	71.2 ± 1.0	92.2 ± 0.3
Text4Vis [50]	yes	K400 dict.	ViT-L/14	16	-	-	68.9 ± 1.0	-
CLIP [36]	no	-	ViT-B/16	16	69.9 ± 1.3	38.0 ± 1.7	63.5 ± 0.4	86.8 ± 0.4
MAXI	no	K400 dict.	ViT-B/16	16	76.6 ± 0.9	50.5 ± 0.9	70.4 ± 0.8	91.5 ± 0.3
MAXI	no	K400 dict, GPT3 verbs	ViT-B/16	16	$\underline{77.8} \pm 0.3$	51.6 ± 0.9	$\mathbf{71.6} \pm 1.0$	92.3 ± 0.3
MAXI	no	K400 dict, GPT3 verbs	ViT-B/16	16/32	$\underline{77.8} \pm 0.5$	51.9 ± 1.1	$\mathbf{71.6} \pm 1.0$	$\underline{92.4} \pm 0.3$
MAXI	no	K400 dict, GPT3 verbs, BLIP verbs	ViT-B/16	16	$\mathbf{78.2} \pm 0.8$	$\underline{52.2} \pm 0.6$	71.4 ± 0.9	$\mathbf{92.5} \pm 0.3$
MAXI	no	K400 dict, GPT3 verbs, BLIP verbs	ViT-B/16	16/32	$\mathbf{78.2} \pm 0.8$	$\mathbf{52.3} \pm 0.7$	$\underline{71.5} \pm 0.8$	$\mathbf{92.5} \pm 0.4$

Zero-shot action recognition on UCF101, HMDB51 and K600, CLIP fine-tuned with K400 vocabulary + videos

MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge [W.Lin et al. , ICCV 2023]

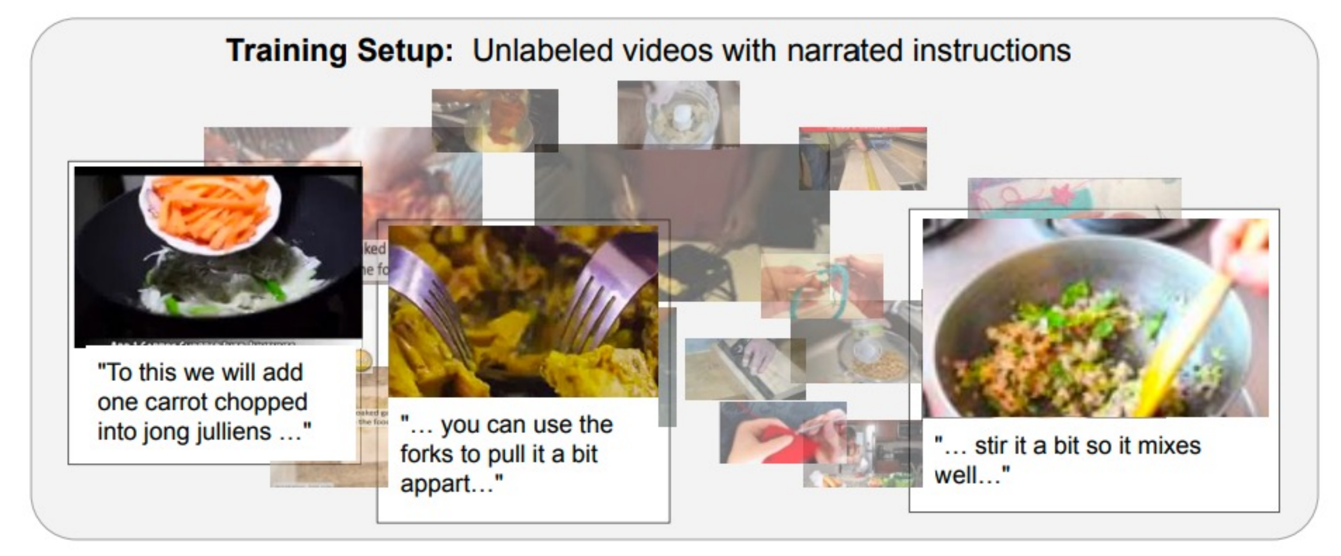
Results:

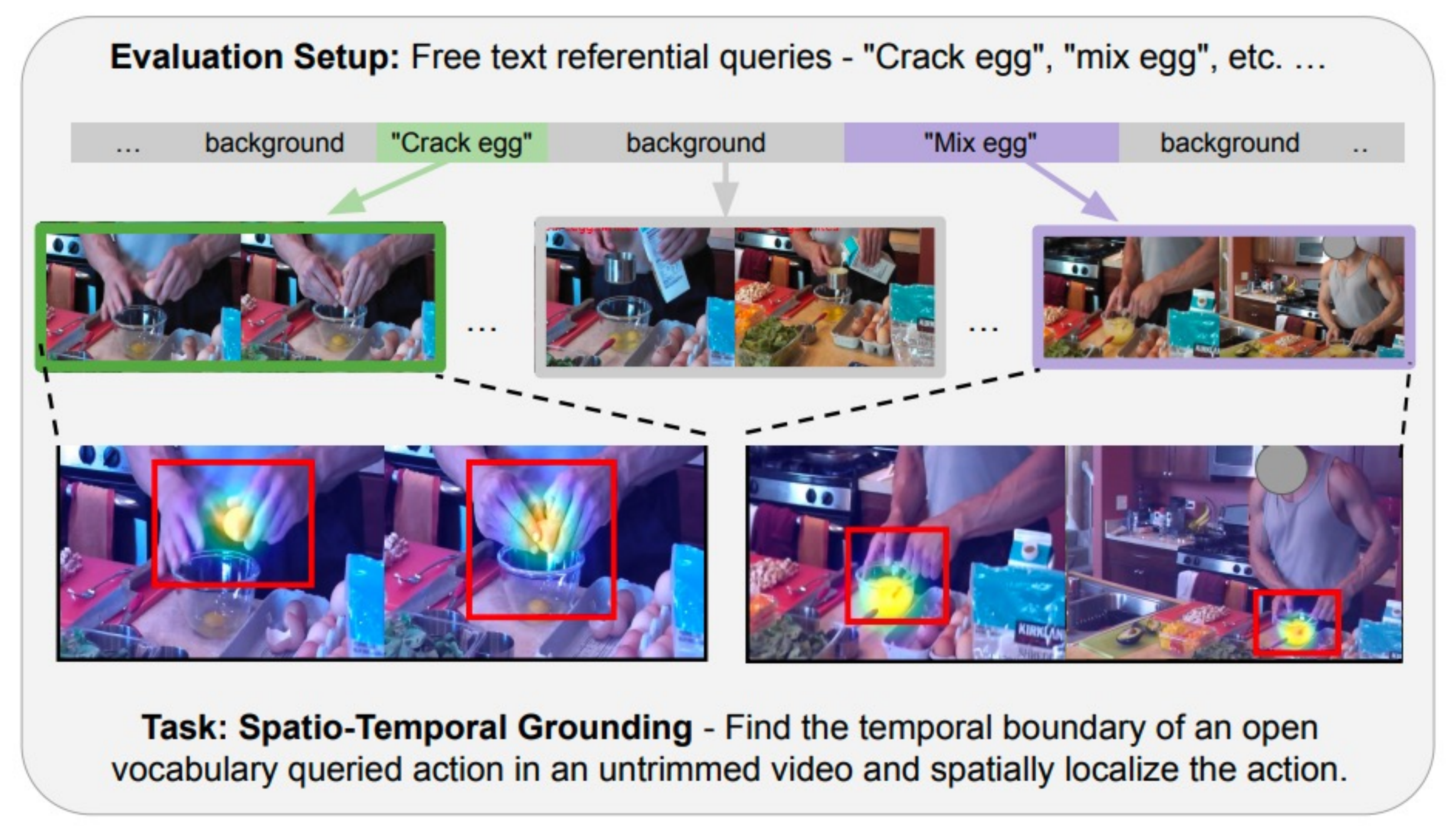
Action dictionary	dictionary size	UCF101	HMDB51	K600	MiniSSv2	Charades	UAV Human	Moments-in-time
CLIP [36] (w/o finetune) Zero-Shot		69.93 / 92.7	38.02 / 66.34	63.48 / 86.80	3.96 / 14.42	19.80	1.79 / 7.05	20.11 / 40.81
K400	400	**78.18 / 96.03**	**50.35 / 77.10**	**70.78 / 92.17**	5.74 / 17.70	**23.89**	**3.06 / 9.46**	22.41 / 45.83
MiniKinetics	200	75.10 / 95.82	48.34 / 76.95	69.23 / 90.92	**6.50 / 18.76**	22.70	2.40 / 8.04	**22.50 / 46.01**
K400+WebVid2.5M	800	75.99 / 96.00	45.97 / 73.94	69.14 / 91.13	4.81 / 15.79	22.67	2.11 / 8.00	20.92 / 43.99

Zero-shot action recognition with CLIP fine-tuned with K400 videos + other vocabulary (mAP on Charades and Top1/Top5 accuracy on other datasets).

How to understand what's going on?

What, when, and where? - Self-Supervised Spatio-Temporal Grounding in Untrimmed Multi-Action Videos from Narrated Instructions [B. Chen et al ., arxiv]





What, when, and where? - Self-Supervised Spatio-Temporal Grounding [B. Chen et al., arxiv]

Datasets:

YouCook2-Interactions Dataset

→frame-level bounding boxes for instructional cooking videos

→ annotations for on YouCook2 validation split

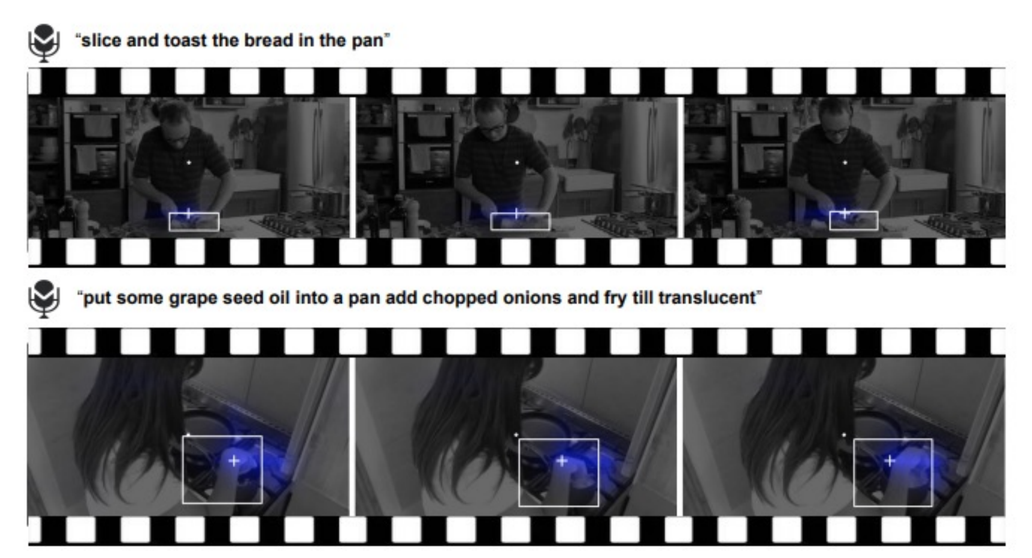
→ trimmed clips only

Grounding YouTube (coming soon)

→ frame-level point clouds and bounding boxes for cooking videos

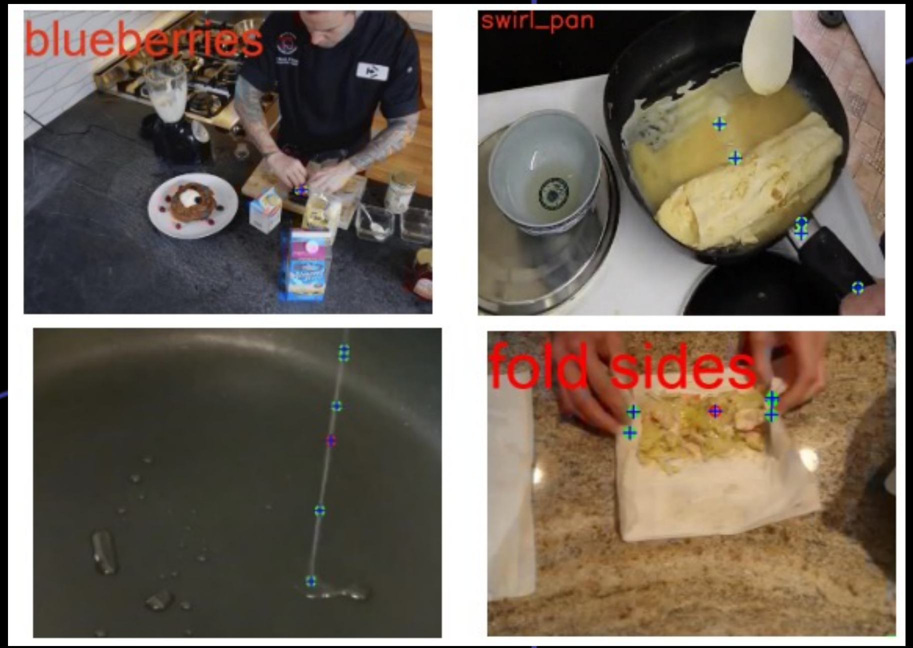
→ annotations for mining YouTube

→ ***Untrimmed spatial-temporal grounding***



Look at What I am Doing: Self-Supervised Spatial Grounding of Narrations in Instructional Videos; Reuben Tan, Bryan A. Plummer, Kate Saenko, Hailin Jin, Bryan Russell, NeurIPS2021

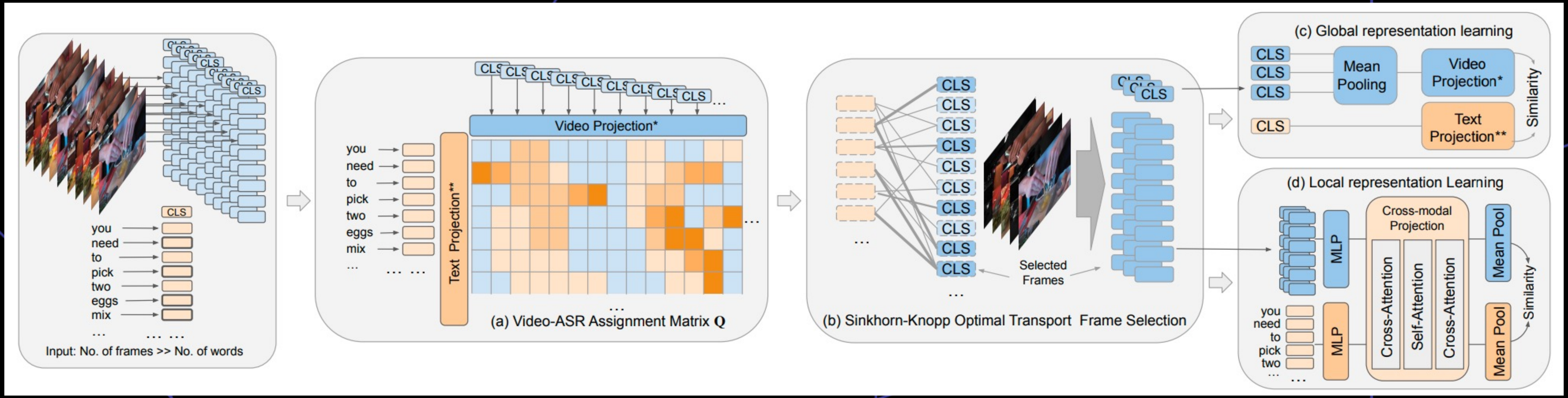
https://cs-people.bu.edu/rxtan/projects/grounding_narrations/



What, when, and where? - Self-Supervised Spatio-Temporal Grounding [B. Chen et al., arxiv]

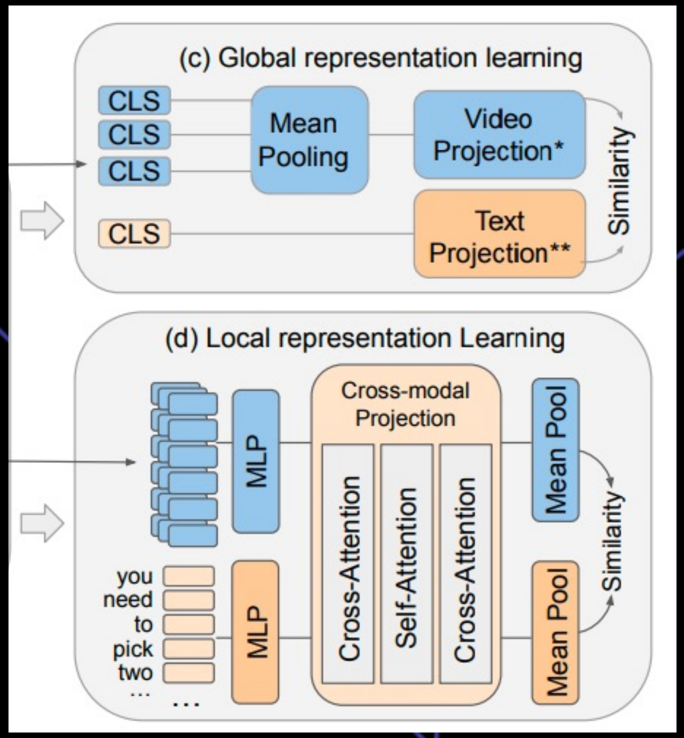
Idea:

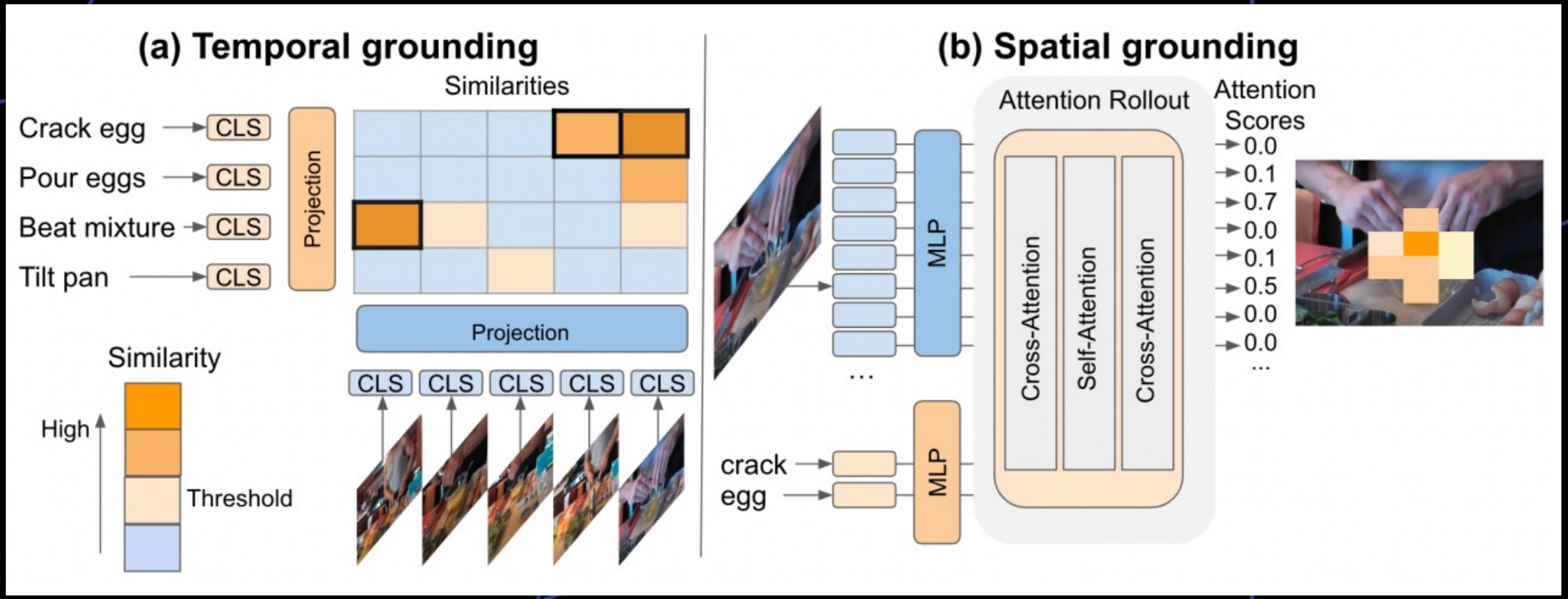
- Local information better at capturing spatial information
- Global information better at capturing temporal information
- Add frame selection for efficiency



What, when, and where? - Self-Supervised Spatio-Temporal Grounding [B. Chen et al ., arxiv]

Challenge: Capture (long) temporal and single-frame spatial boundaries
--> One branch for global representation learning → start-end frame
--> One branch for local, spatial representation learning → bounding box





What, when, and where? - Self-Supervised Spatio-Temporal Grounding [B. Chen et al., arxiv]

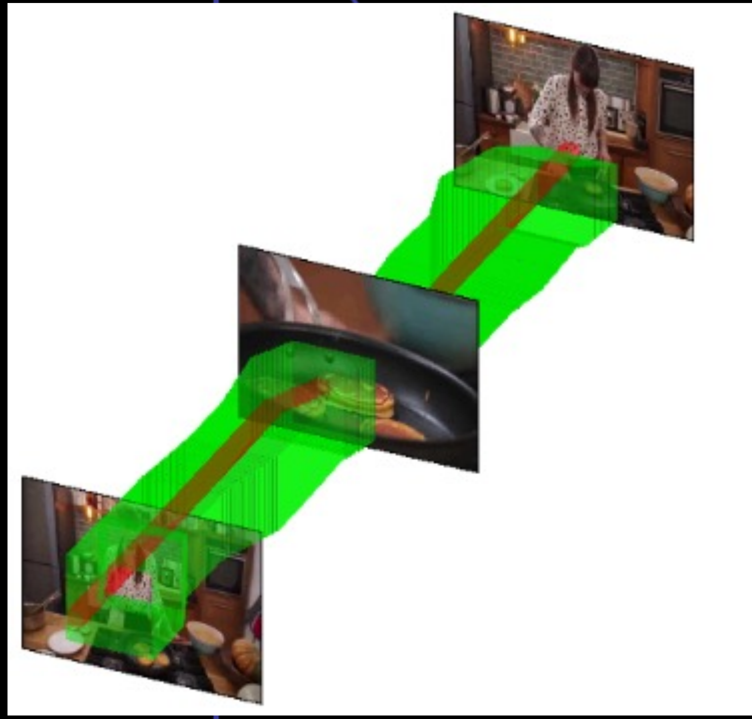
Results: Spatial-Temporal Grounding

Method	Backbone	DataSet	Supervision	Modality	GroundingYoutube IoU+Point	mAP 0.1	mAP 0.2	mAP 0.3	mAP 0.4	mAP 0.5	mAP 0.1:0.5
CoMMA† [45]	S3D	HT250K	Self	VT	1.02	2.18	1.72	1.11	0.93	0.37	1.26
MIL-NCE [35]	S3D*	HT100M	Self	VT	4.67	33.94	25.16	12.65	3.42	0.41	15.11
Ours	S3D	HT100M	Self	VT	**9.12**	**42.70**	**35.49**	**25.16**	**16.22**	**10.05**	**25.92**
GLIP [30]	Swin-L*	Cap24M	Weak	IT	1.24	2.83	2.10	1.52	0.96	0.37	1.56
CoMMA‡ [45]	CLIP	HT100M	Self	VT	1.68	3.51	2.32	1.88	0.99	0.40	1.82
CLIP [37]	CLIP	HT100M	Self	IT	3.59	29.54	22.15	9.16	2.48	0.39	12.74
RegionCLIP [61]	ResNet-101*	CC3M	Weak	IT	5.65	35.65	27.43	15.69	4.31	0.86	16.78
Ours	CLIP	HT100M	Self	VT	10.09	42.81	36.05	25.84	17.10	11.35	26.63
Ours	CLIP*	HT100M	Self	VT	**11.53**	**43.64**	**36.94**	**26.78**	**19.45**	**14.61**	**28.26**
MIL-NCE(temp.)+RegionCLIP(spa.)	-	-	-	VT	9.21	40.54	34.97	22.38	13.79	9.18	22.33

What, when, and where? - Self-Supervised Spatio-Temporal Grounding [B. Chen et al., arxiv]

Results: Spatial Grounding only

Method	Backbone	Data	Super.	Mod.	YC-Inter Acc	GroundingYT Acc	GroundingYT mAP	V-HICO Acc	V-HICO mAP	Daly Acc	Daly mAP
MIL-NCE [35]	S3D*	HT100M	Self	VT	23.67	27.45	8.21	12.65	11.23	13.84	24.23
CoMMA† [45]	S3D	HT250K	Self	VT	48.63	47.68	23.38	40.97	21.45	54.48	33.39
Ours	S3D	HT100M	Self	VT	**53.98**	**60.62**	**44.93**	**44.32**	**24.31**	**66.35**	**45.93**
CLIP [37]	CLIP	HT100M	Self	IT	14.10	12.50	3.49	29.23	12.51	18.02	27.28
CoMMA‡ [45]	CLIP	HT100M	Self	VT	52.65	47.56	36.42	55.20	34.54	61.06	44.37
RegionCLIP [61]	RN50x4*	CC3M	Weak	IT	51.56	52.84	23.42	57.92	37.82	67.12	48.62
GLIP [30]	Swin-L*	Cap24M	Weak	IT	52.84	53.62	24.73	**66.05**	41.17	-	-
Ours	CLIP	HT100M	Self	VT	57.10	55.49	43.12	60.71	39.28	70.08	50.56
Ours	CLIP*	HT100M	Self	VT	**58.35**	**56.98**	**45.32**	62.34	**41.56**	**71.35**	**52.78**
TubeDETR [53]	MDETR	Vid-STG	Full	VT	51.63	53.24	41.76	63.23	40.87	84.21	62.98
STCAT [23]	ResNet-101	Vid-STG	Full	VT	54.47	55.90	44.21	65.34	41.10	85.42	63.94



Vision-Language in Video – What’s next?

Pro:

No more labels! No more annotation!

→ Natural language requests for video systems (retrieval, detection, etc.)

→ Natural language representations of video

… will lead to new applications in video understanding

Vision-Language in Video – What's next?

Con:

- No more simple metrics! (retrieval might already be ceiling)
- No more simple comparability!

Before: Classification accuracy on 2-3 standard datasets

Now: Various mixtures of pretraining and downstream testing

→ How do we know what works better?

People at a glance ...


Nina Shvetsova
Goethe University Frankfurt


Anna Kukleva
MPII Saarbruecken


Wei Lin
TU Graz


Brian Chen
Meta/Columbia University


Bernt Schiele
MPII Saarbrucken

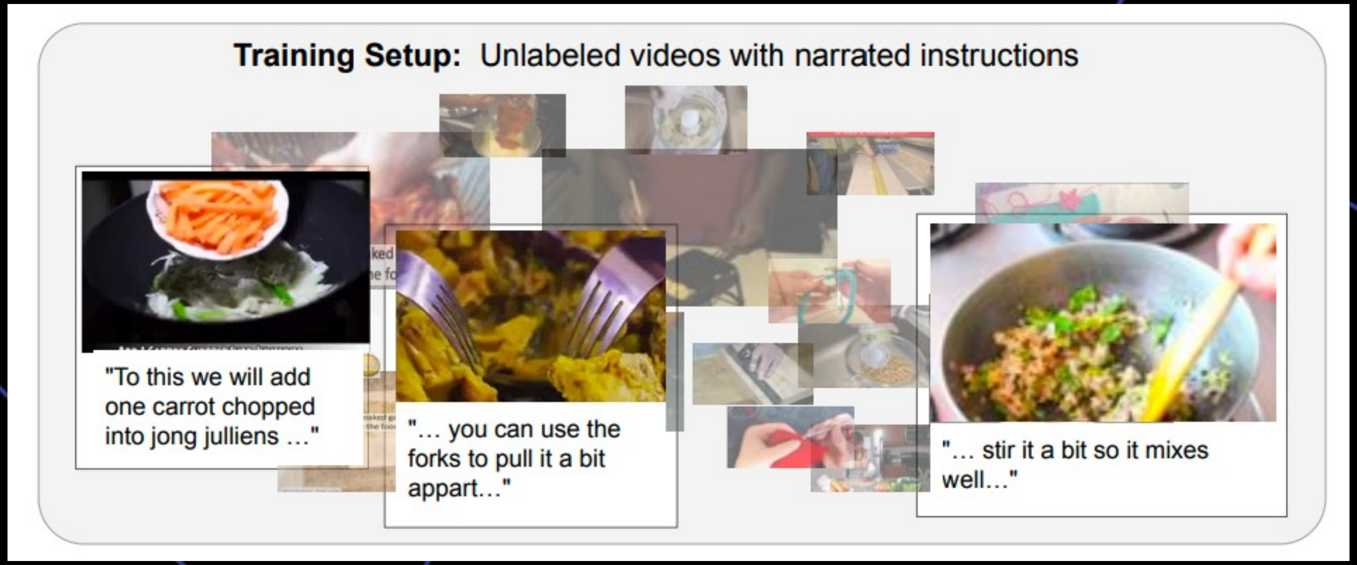

Christian Rupprecht
Oxford

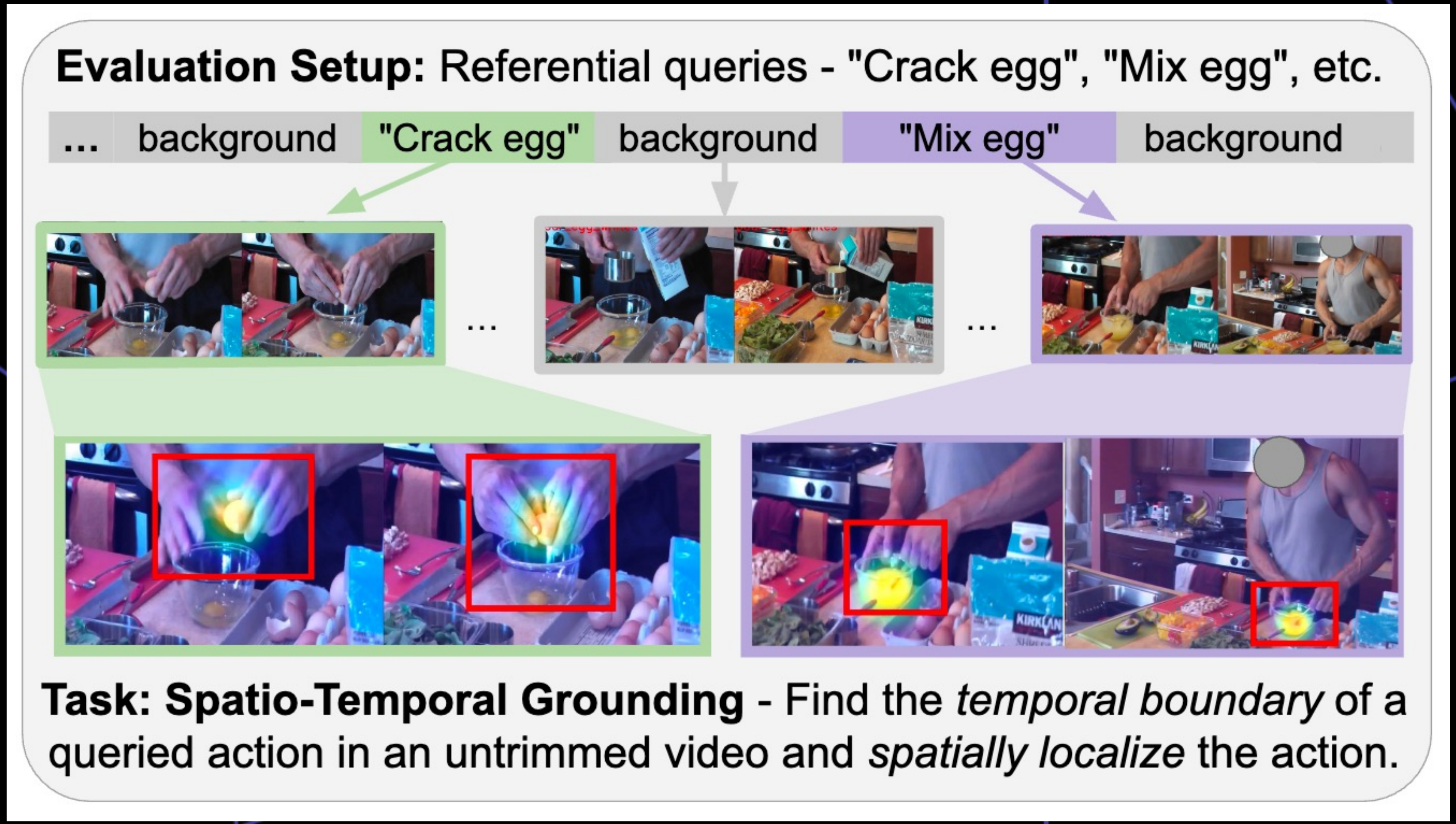

Horst Bishof
TU Graz

Thanks for listening!

How to understand what's going on?

What, when, and where? - Self-Supervised Spatio-Temporal Grounding in Untrimmed Multi-Action Videos from Narrated Instructions [B. Chen et al., arxiv 2023]





What, when, and where? - Self-Supervised Spatio-Temporal Grounding [B. Chen et al., arxiv 2023]

Datasets:

YouCook2-Interactions Dataset

→ frame-level bounding boxes for instructional cooking videos

→ annotations for on YouCook2 validation split

→ trimmed clips only

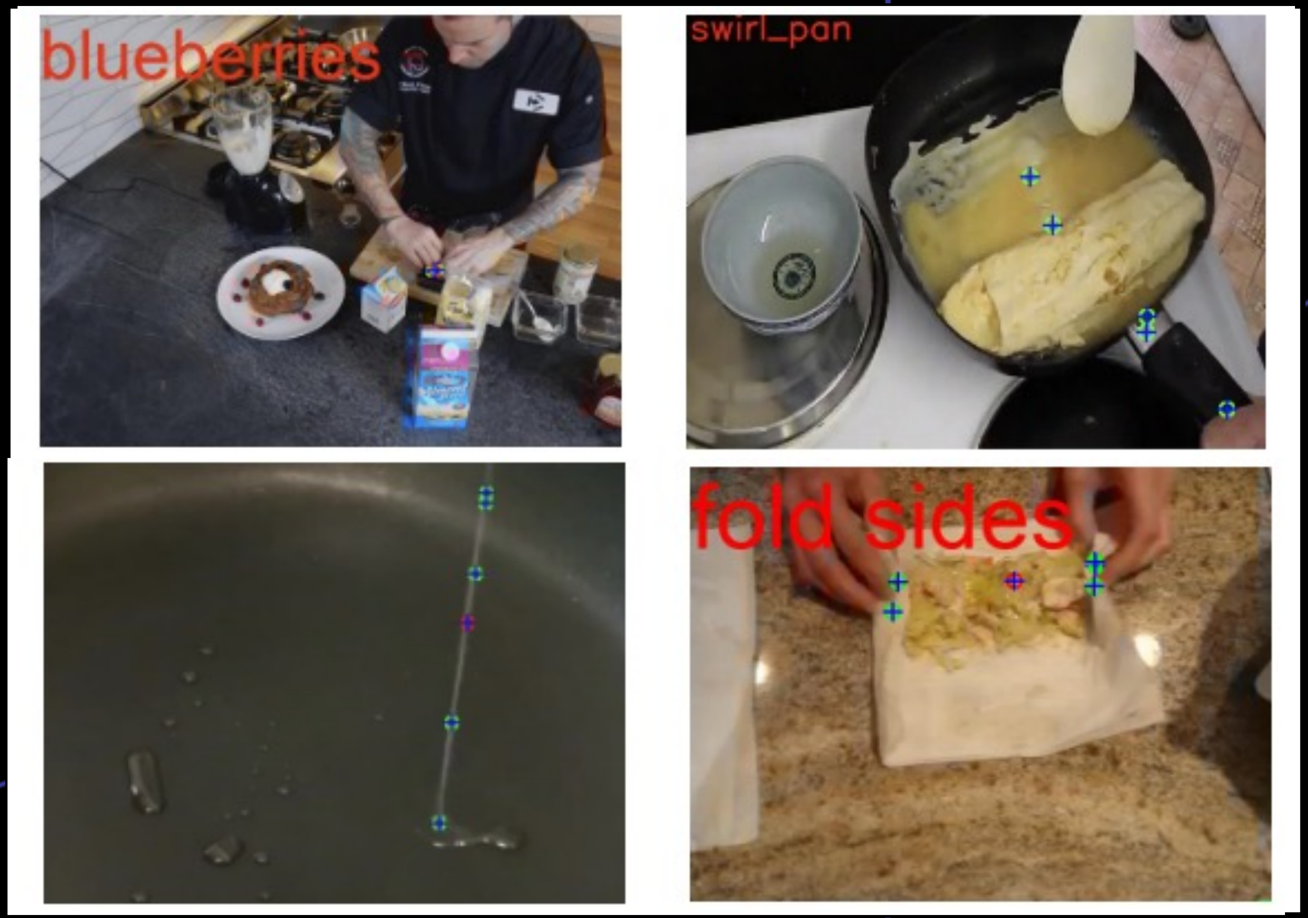
Grounding YouTube

→ frame-level point clouds and bounding boxes for cooking videos

→ annotations for mining YouTube

→ ***Untrimmed spatial-temporal grounding***

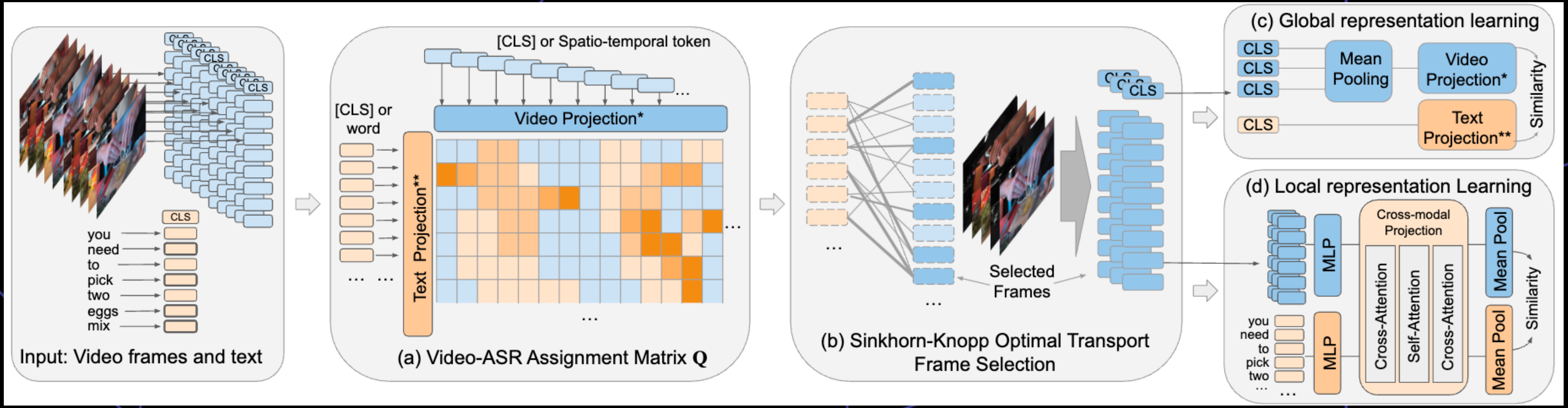
Look at What I am Doing: Self-Supervised Spatial Grounding of Narrations in Instructional Videos; Tan et al., NeurIPS2021



What, when, and where? - Self-Supervised Spatio-Temporal Grounding [B. Chen et al., arxiv]

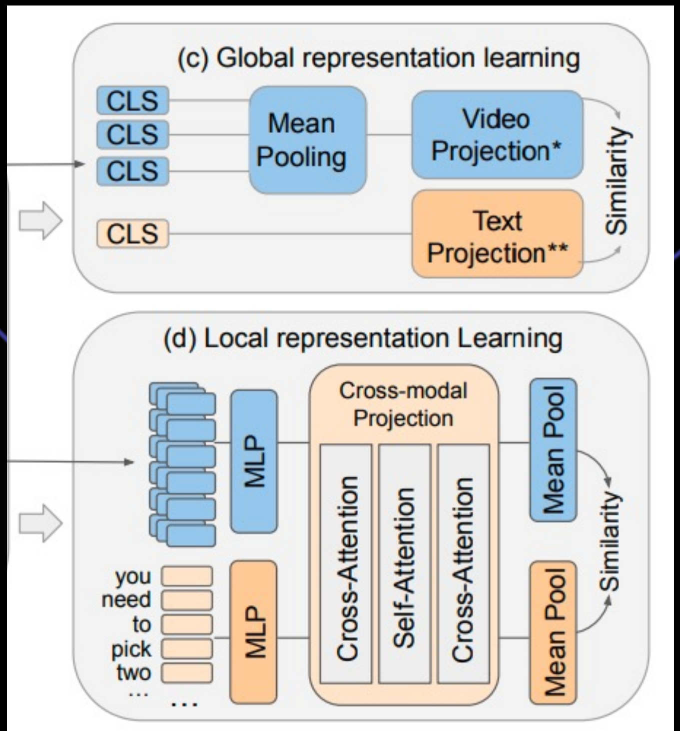
Idea:

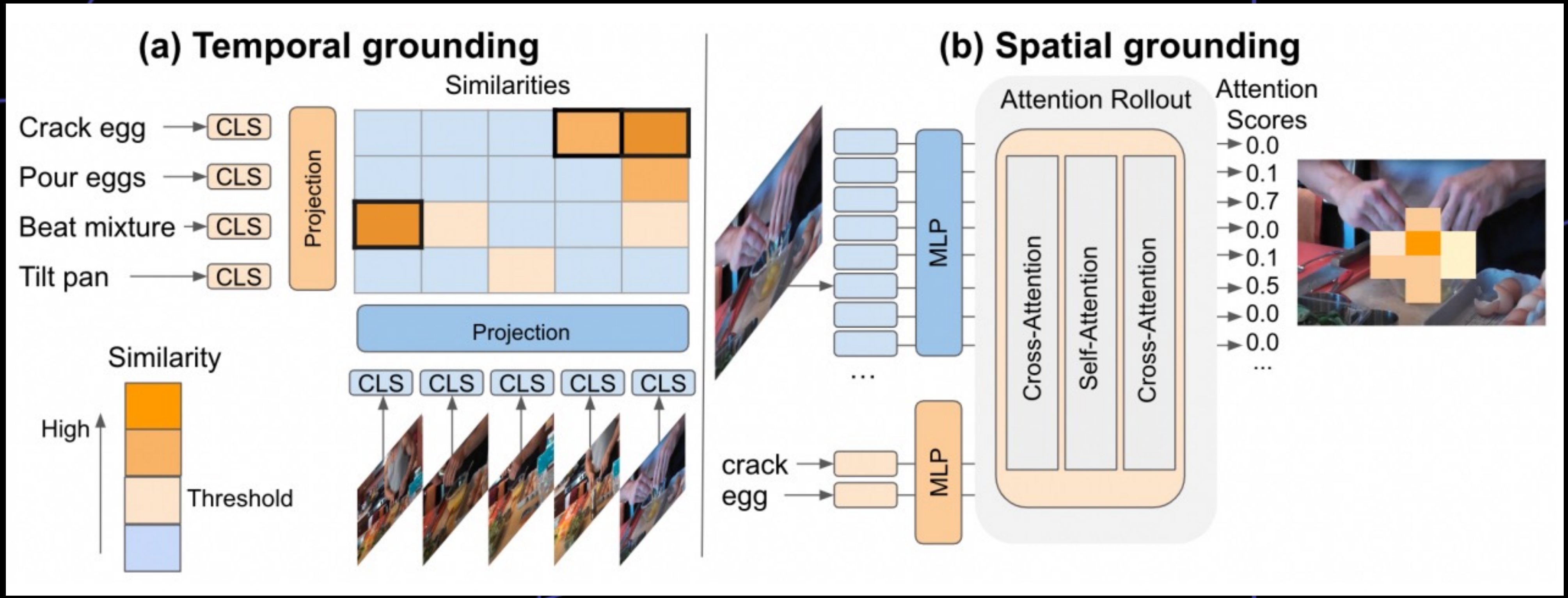
- Local information better at capturing spatial information
- Global information better at capturing temporal information
- Add frame selection for efficiency



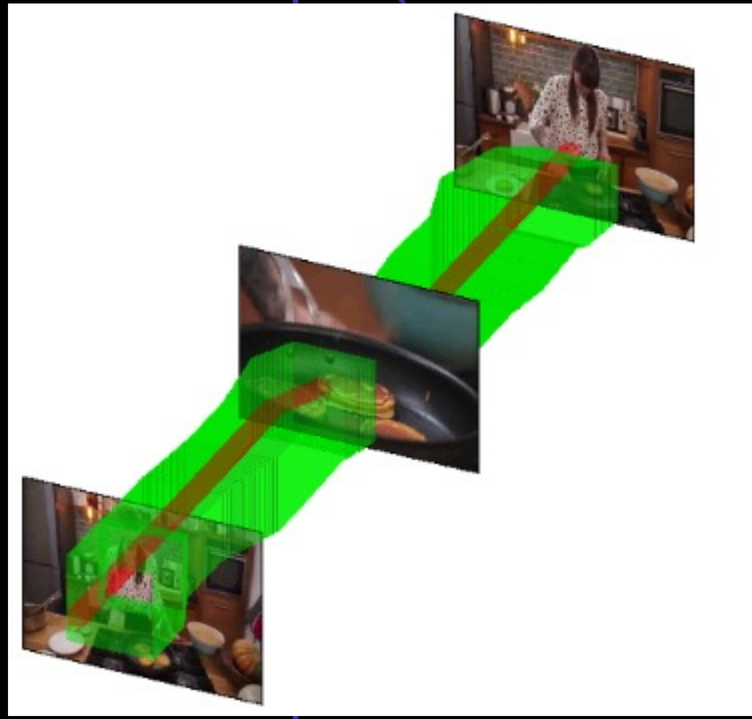
What, when, and where? - Self-Supervised Spatio-Temporal Grounding [B. Chen et al., arxiv]

Challenge: Capture (long) temporal and single-frame spatial boundaries
--> One branch for global representation learning → start-end frame
--> One branch for local, spatial representation learning → bounding box





What, when, and where? - Self-Supervised Spatio-Temporal Grounding [B. Chen et al., arxiv]



Results:

- Sota results for spatio-temporal localization in untrimmed videos
- Global + local information also single tasks
- Smart frame selection helps

Method	Backbone	DataSet	Supervision	Modality	GroundingYoutube IoU+Point	mAP 0.1	0.2	0.3	0.4	0.5	0.1:0.5
CoMMA† (Tan et al., 2021)	S3D	HT250K	Self	VT	1.02	2.18	1.72	1.11	0.93	0.37	1.26
MIL-NCE (Miech et al., 2020)	S3D*	HT100M	Self	VT	4.67	33.94	25.16	12.65	3.42	0.41	15.11
Ours	S3D	HT100M	Self	VT	**9.12**	**42.70**	**35.49**	**25.16**	**16.22**	**10.05**	**25.92**
GLIP (Li et al., 2022a)	Swin-L*	Cap24M	Weak	IT	1.24	2.83	2.10	1.52	0.96	0.37	1.56
CoMMA‡ (Tan et al., 2021)	CLIP	HT100M	Self	VT	1.68	3.51	2.32	1.88	0.99	0.40	1.82
CLIP (Radford et al., 2021)	CLIP	HT100M	Self	IT	3.59	29.54	22.15	9.16	2.48	0.39	12.74
RegionCLIP (Zhong et al., 2022)	ResNet-101*	CC3M	Weak	IT	5.65	35.65	27.43	15.69	4.31	0.86	16.78
Ours	CLIP	HT100M	Self	VT	10.09	42.81	36.05	25.84	17.10	11.35	26.63
Ours	CLIP*	HT100M	Self	VT	**11.53**	**43.64**	**36.94**	**26.78**	**19.45**	**14.61**	**28.26**
MIL-NCE(temp.)+RegionCLIP(spa.)	-	-	-	VT	9.21	40.54	34.97	22.38	13.79	9.18	22.33

	GroundingYT Spatio-temporal	MiningYT Temporal	YouCook-Inter. Spatial
None	15.1	17.8	55.5
Global selection	15.7	18.5	54.3
Local selection	15.6	18.1	56.3
Sinkhorn	17.1	19.9	57.1
only Local loss	5.7	4.5	54.3
only Global loss	7.6	18.8	32.5
w/ Both loss	17.1	19.9	57.1