Leveraging Temporal Context in Low Representational Power Regimes

WED-AM-235

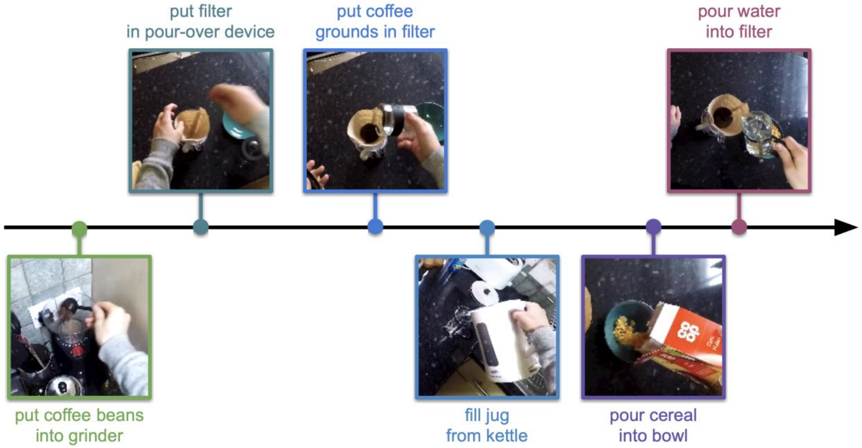
Camilo Fosco, SouYoung Jin, Emilie Josephs, Aude Oliva
Olivalab, MIT CSAIL

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

The problem

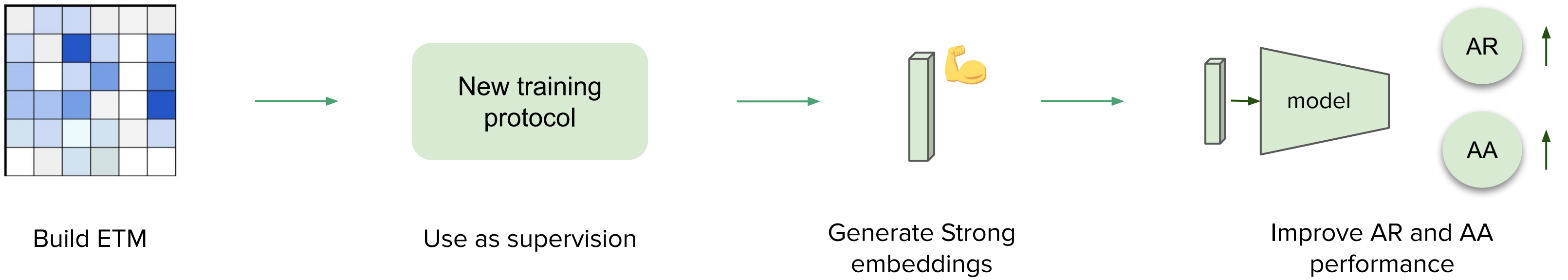
Using temporal information is crucial to understand videos. Yet, current models don't explicitly attempt to leverage temporal regularities in datasets with long videos.

- Can we leverage the **statistics in temporal sequences** of video datasets to **improve performance** in downstream tasks?
- Can we build richer embeddings with this information?
- Where does this type of information **help the most**?



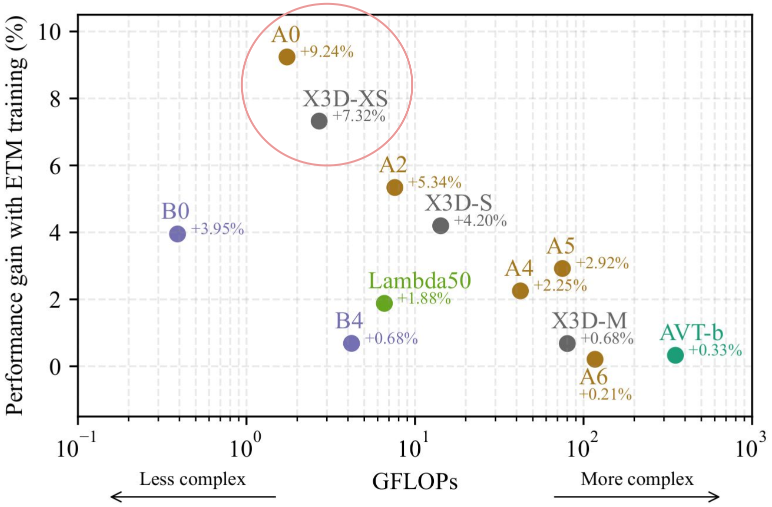
The idea

- We propose to build an **Event Transition Matrix**: a representation that captures typical transition probabilities between actions in long video sequences
- We use this matrix as supervision in a new training protocol to generate **strong embeddings for video snippets**
- We leverage these embeddings to **improve action recognition and action anticipation performance,** especially on low complexity models.



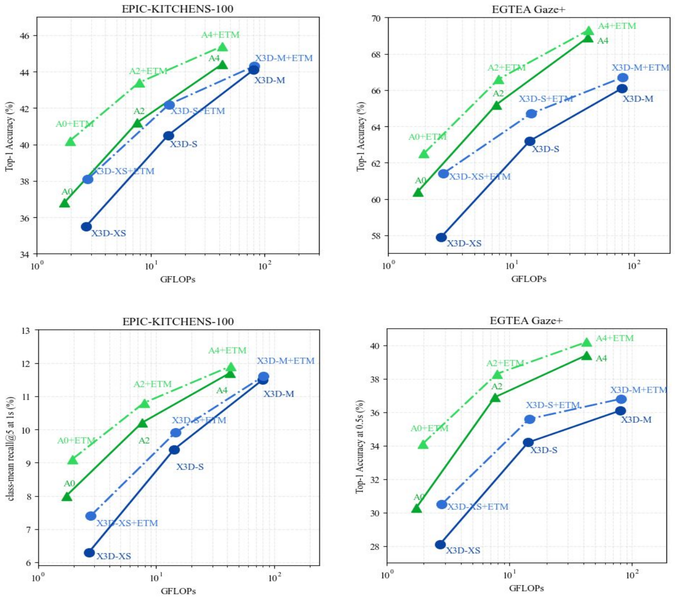
Key results

Our model-agnostic framework helps **low complexity models improve performance** on action recognition and action anticipation across 3 datasets.



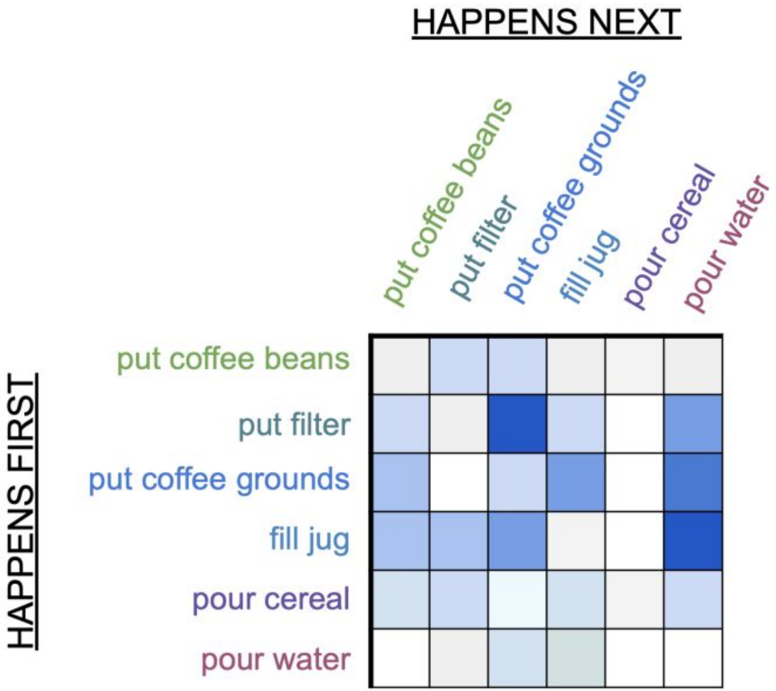
Action Recognition

Action Anticipation



The Event Transition Matrix

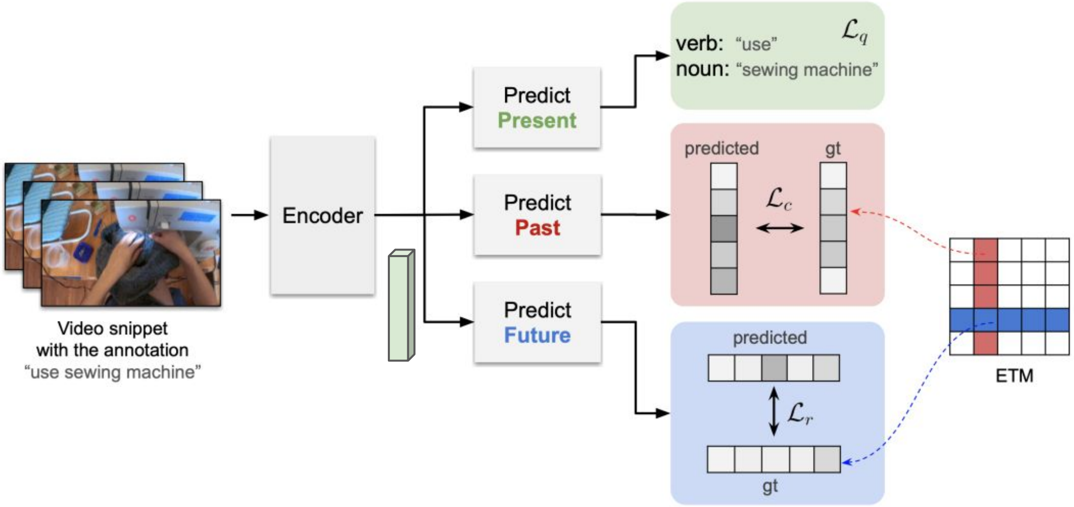
- Computed by looking at **all actions** happening after a given action, weighted by a decay function
- Square matrix, **not symmetric**
- Several postprocessing steps:
  - Dimensionality reduction
  - Decay definition
  - Normalization



How can we leverage this ETM?

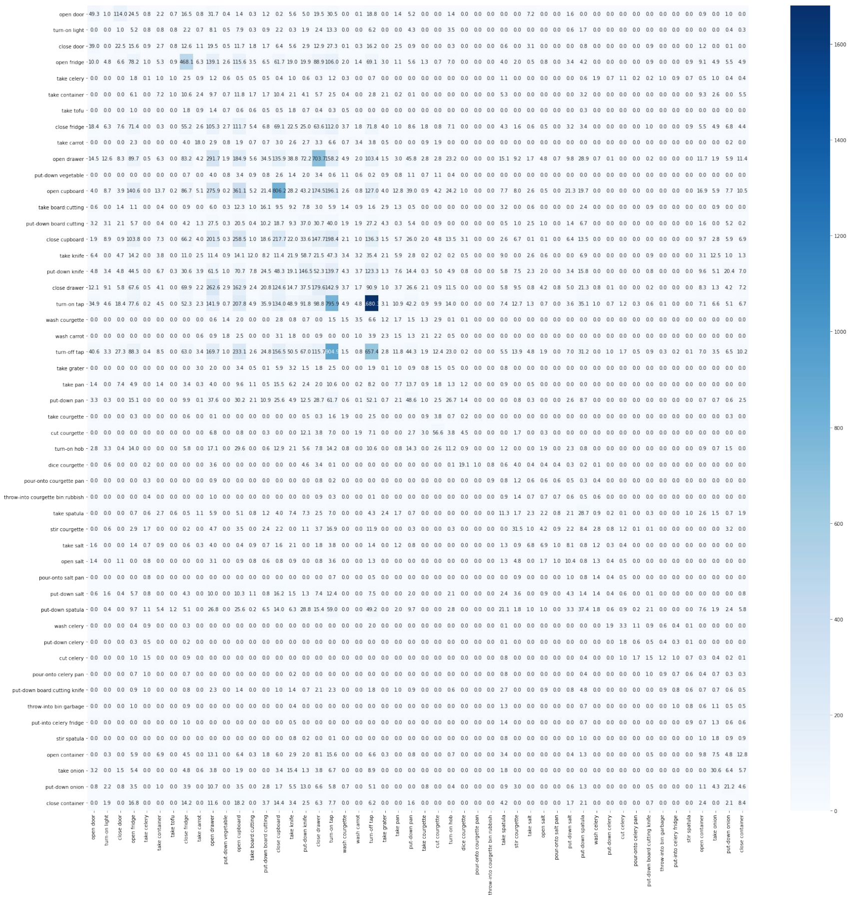
- We propose to use the rows and columns of the matrix as **targets in a regression problem**

- An encoder is tasked to **generate an embedding** that can predict the action + regress the past and future



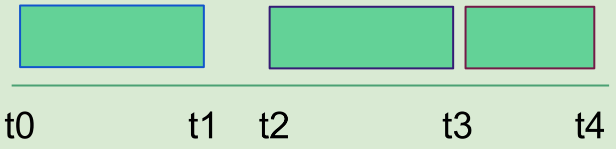
Building the ETM

- Large matrix with sparse entries
- **Dimensionality reduction:** do we use the full matrix, or reducing the number of actions considered help?
- **Decay:** how to we weight the contributions of actions happening later in the video?
- **Distance metric:** how do we measure the distance between two actions in a video?



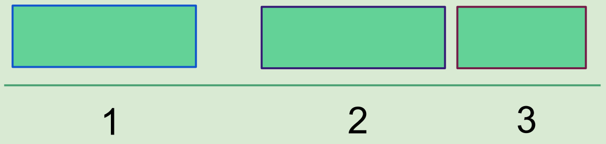
Types of distance metrics

Distance in frames/seconds between actions



- Takes into account temporal difference
- Can differentiate between end/start (but which one to choose?)
- But dependent on length of actions

Distance as index difference in ordinal sequence



- Only considers ordering
- Might be more adapted to causality concepts
- Doesn't depend on length of actions

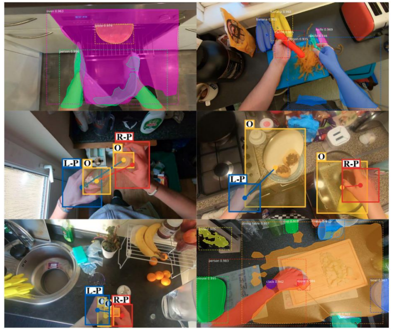
Testing ETM design choices

We test several design choices, including different decay functions, ETM sizes and distance metrics.

Size	Decay	Temp. Metric	**Present** (top-1 accuracy)		
			Verb	Noun	Action
13k	linear	time	0.551	0.462	0.288
13k	exponential	time	0.556	0.477	0.291
2.5k	exponential	time	0.586	0.488	0.313
2.5k	no decay	-	0.581	0.480	0.305
2.5k	linear	index	0.601	0.493	0.319
2.5k	exponential	index	**0.603**	**0.503**	**0.324**

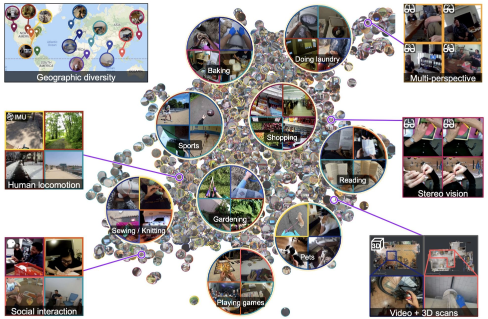
Datasets used

EPIC-KITCHENS-100



700 videos depicting cooking actions, totalling 100 hours.

EGO4D LTA



3670 hours of video from 71 different participants.

EGTEA Gaze+

10k segments annotated with 19 verbs, 51 nouns and 106 unique actions.

Tasks where we test our embeddings

Action Recognition

Receive a snippet, predict class label (here, verb and noun)

Roll Dough

Action Anticipation

Receive a collection of snippets, predict next action

Put Down Dough

Experimental results - Action recognition

Cross Dataset

Results with MoViNet A0

Dataset	Model	Present		
		Verb	Noun	Action
EK100 [14]	Baseline	64.8	47.4	36.8
	ETM(Ours)	**67.9**	**51.2**	**40.2**
EGO4D LTA [16]	Baseline	32.3	23.5	21.1
	ETM(Ours)	**32.9**	**24.2**	**22.0**
EGTEA Gaze+ [28]	Baseline	81.2	71.7	60.4
	ETM(Ours)	**83.4**	**72.9**	**62.5**

Cross Model

Results on EK100 across a wide variety of models

Model	w/o ETM	w/ ETM
MoviNet A0 [24]	36.8	**40.2**
MoviNet A2 [24]	41.2	**43.4**
X3D-XS [11]	35.5	**38.1**
X3D-S [11]	40.5	**42.2**
ConvNeXt-S 224 [31]	20.1	**32.4**
LambdaResNet-50 [4]	26.6	**27.1**
EfficientNet-B0 [57]	25.3	**26.3**
EfficientNet-B4 [11]	29.2	**29.4**
AVT-b [14]	30.4	**30.7**

Experimental results - Action Anticipation

Cross Dataset

Results with MoViNet A0

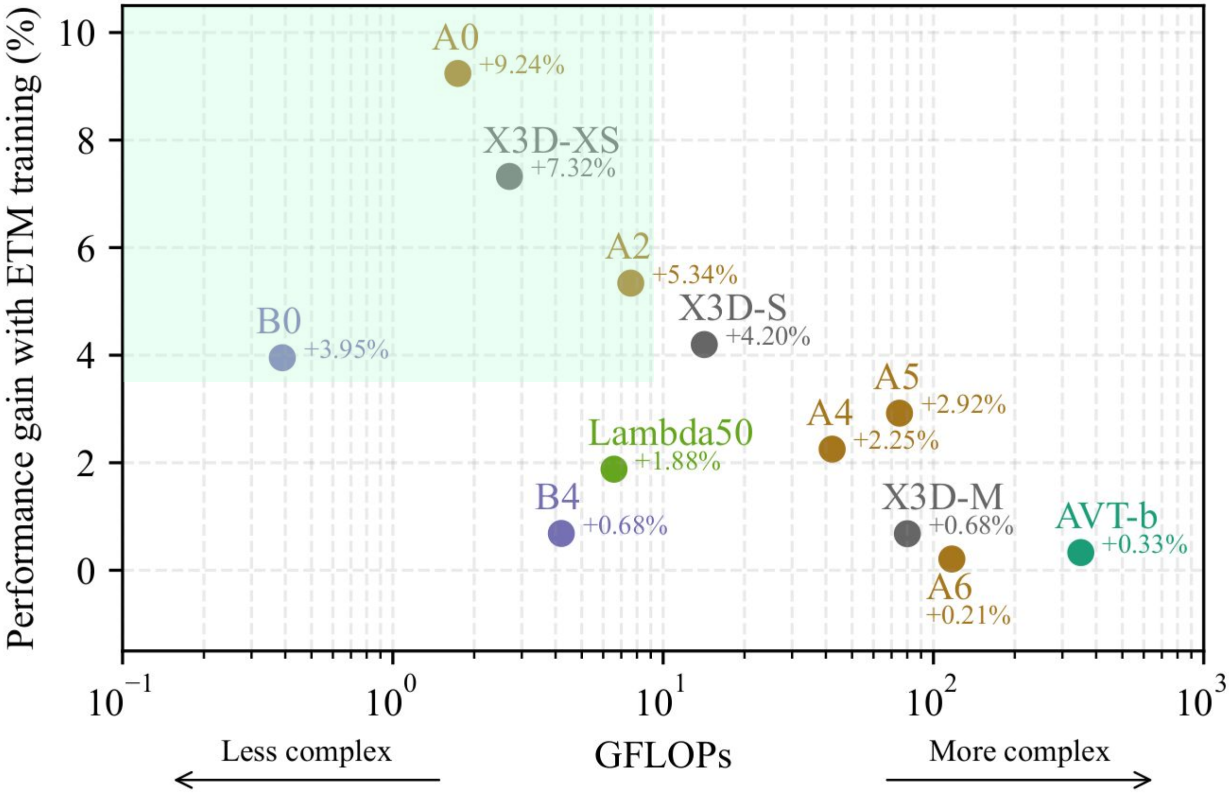
Dataset	Frozen Encoder?	Baseline			ETM (Ours)		
		Verb ↑	Noun ↑	Action ↑	Verb ↑	Noun ↑	Action ↑
EK100	✓	19.9	20.4	7.2	**21.5**	**20.5**	**8.1**
		20.8	21.3	8.0	**22.4**	**22.7**	**9.1**
EGO4D LTA	✓	17.1	16.6	10.3	**18.1**	**17.8**	**11.4**
		18.2	17.5	11.1	**19.9**	**19.1**	**12.9**
EGTEA Gaze+	✓	42.1	37.6	28.9	**43.4**	**38.9**	**31.3**
		43.5	38.5	30.3	**46.5**	**40.7**	**34.1**

Cross Models

Results on EK100

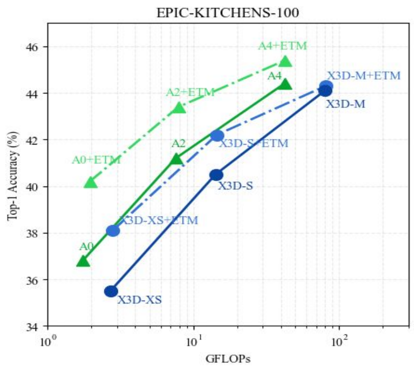
Encoder	w/o ETM	with ETM
MoViNet A0 [24]	8.0	**9.1**
MoViNet A2 [24]	10.2	**10.8**
X3D-XS [11]	6.3	**7.4**
X3D-S [11]	9.4	**9.9**
ConvNeXt-S 224 [31]	4.1	**5.0**
EfficientNet B0 [57]	7.2	**8.0**
EfficientNet B4 [57]	9.4	**10.1**
AVT-b [14]	13.4	**13.5**

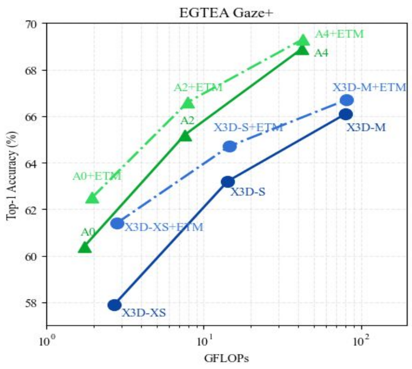
Larger gains on smaller models!



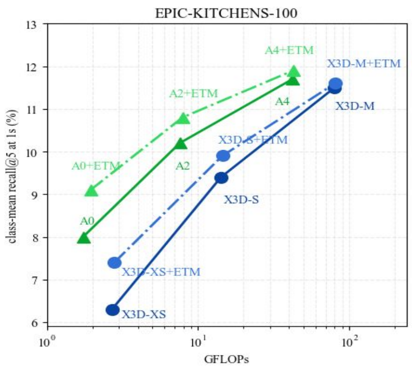
Performance on architecture families

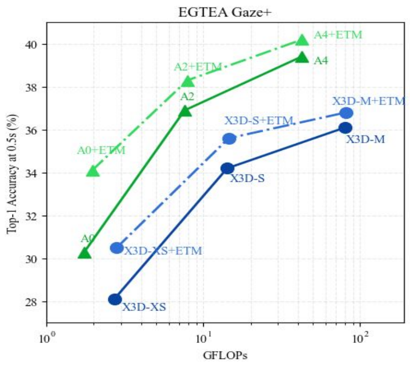
Action Recognition





Action Anticipation

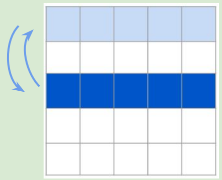




MoViNet+ETM (— -▲); MoViNet (——▲); X3D+ETM (— -●); X3D (——●).

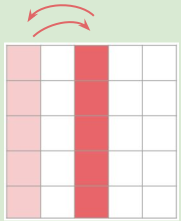
Baseline tests and ablations

Shuffle Rows



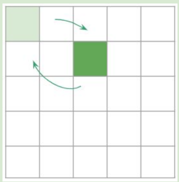
Row shuffling in ETM matrix: Distribution of **future actions** doesn’t match the action index at a given **row**.

Shuffle Columns



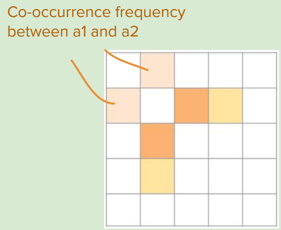
Row shuffling in ETM matrix: Distribution of **past actions** doesn’t match the action index at a given **column**.

Full Shuffle



Full shuffling in ETM matrix: no transition probability estimation matches its original action pair.

Co-occurrence



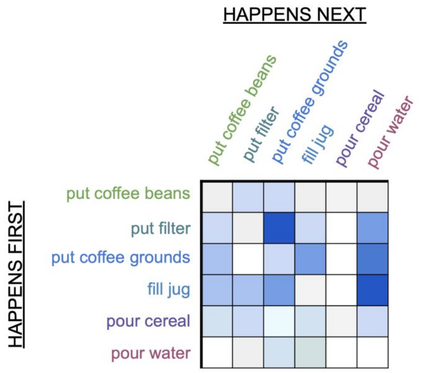
Using a co-occurrence matrix: cells correspond to **co-occurrence frequencies** instead of transition probabilities

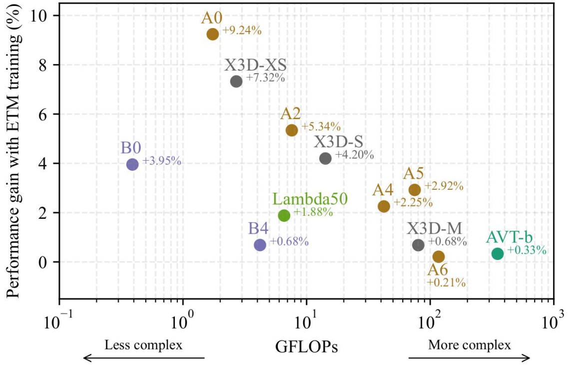
Baseline tests and ablations

Model	Present			MAE on	MAE on
	Verb ↑	Noun ↑	Action ↑	**Past** ↓	**Future** ↓
Baseline	64.8	47.4	36.8	-	-
Full shuffle	64.1	47.2	36.3	4.117	4.012
Columns/rows shuffle	64.7	47.6	36.7	3.254	3.101
Co-occurrence	65.3	49.0	37.9	1.211	1.115
Only past vector	65.7	49.3	38.2	0.901	-
Only future vector	65.5	49.8	38.3	-	0.898
ETM (Ours)	**67.9**	**51.2**	**40.2**	**0.882**	**0.859**

Conclusions

- We introduce a new training regime that uses **external temporal regularities** to boost video understanding.
- Using our ETM as a training target enhances action recognition and anticipation, particularly on **low representational power models.**
- Our ETM protocol's key benefits: flexibility, simplicity, cost-effectiveness, and easy integration.





Thank you!

Leveraging Temporal Context in Low Representational Power Regimes

Camilo Fosco, SouYoung Jin, Emilie Josephs, Aude Oliva

Project page:
camilofosco.com/etm_website

Contact: camilolu@mit.edu

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA