Leveraging Temporal Context in Low Representational Power Regimes

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

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

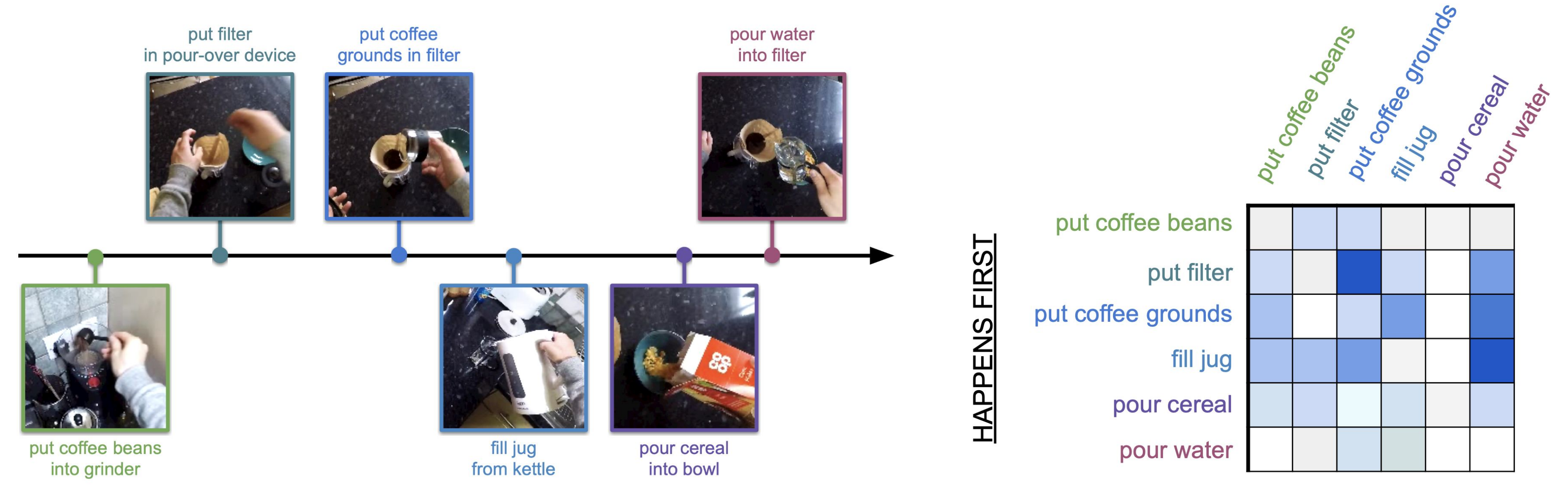
Problem

Can the performance of **low-complexity video understanding models** be improved by supplementing training with additional information about temporal regularities in a dataset?

We explore this in the domains of action recognition and action anticipation, over egocentric videos.

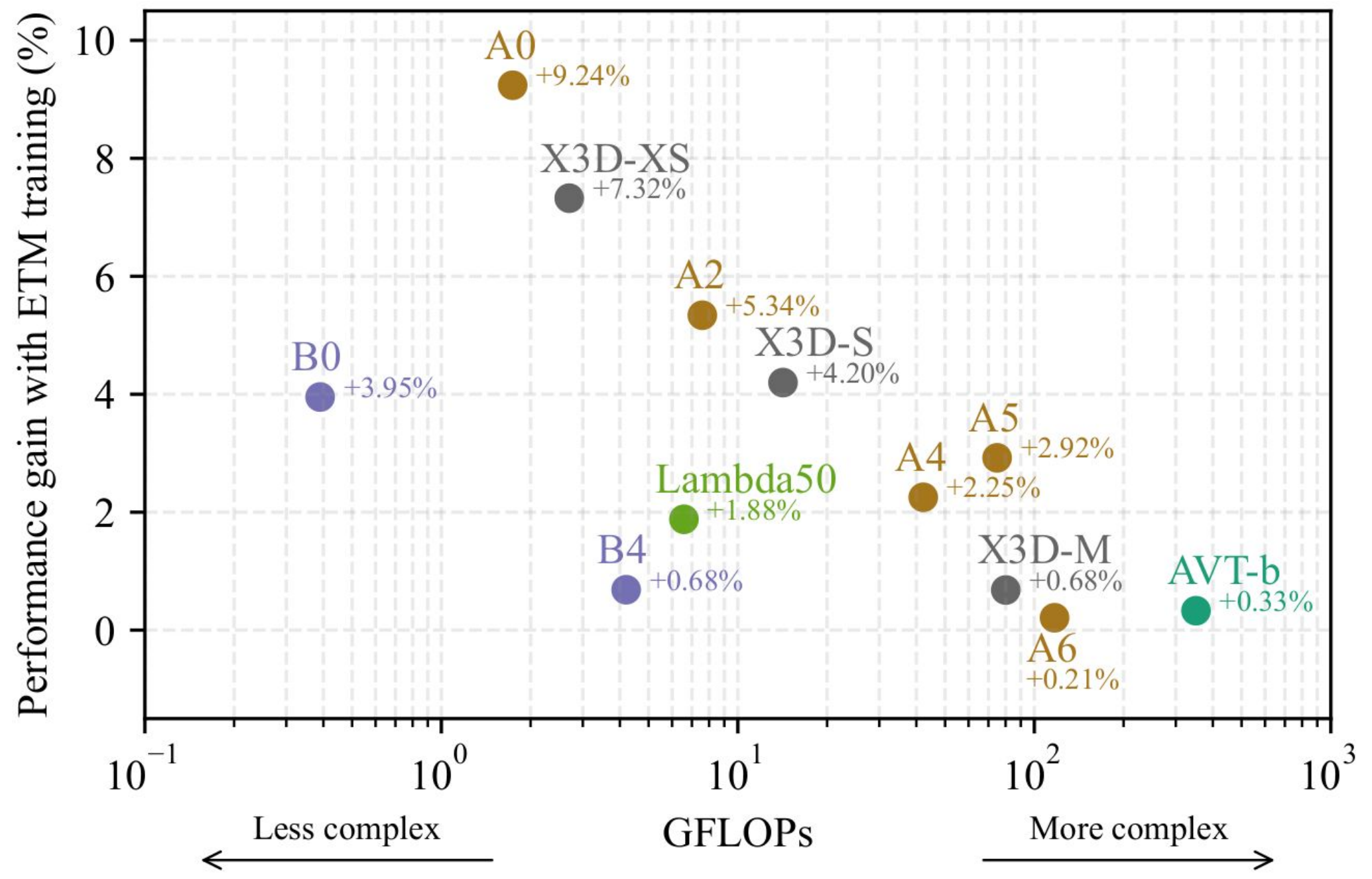
- We introduce **The Event Transition Matrix (ETM),** computed from action labels in an untrimmed video dataset, which captures the temporal context of a given action.
- We show that **including ETM information during training** improves action recognition and anticipation performance on various egocentric video datasets.

The Event Transition Matrix



The advantage

Training with the ETM approach increases performance on action recognition. Interestingly, lower complexity models appear to benefit more from this approach.

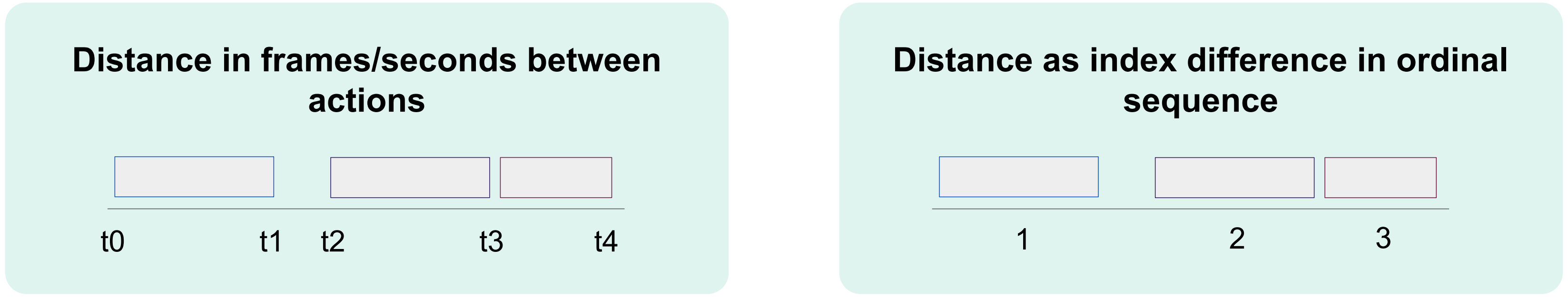


Contact: camilolu@mit.edu

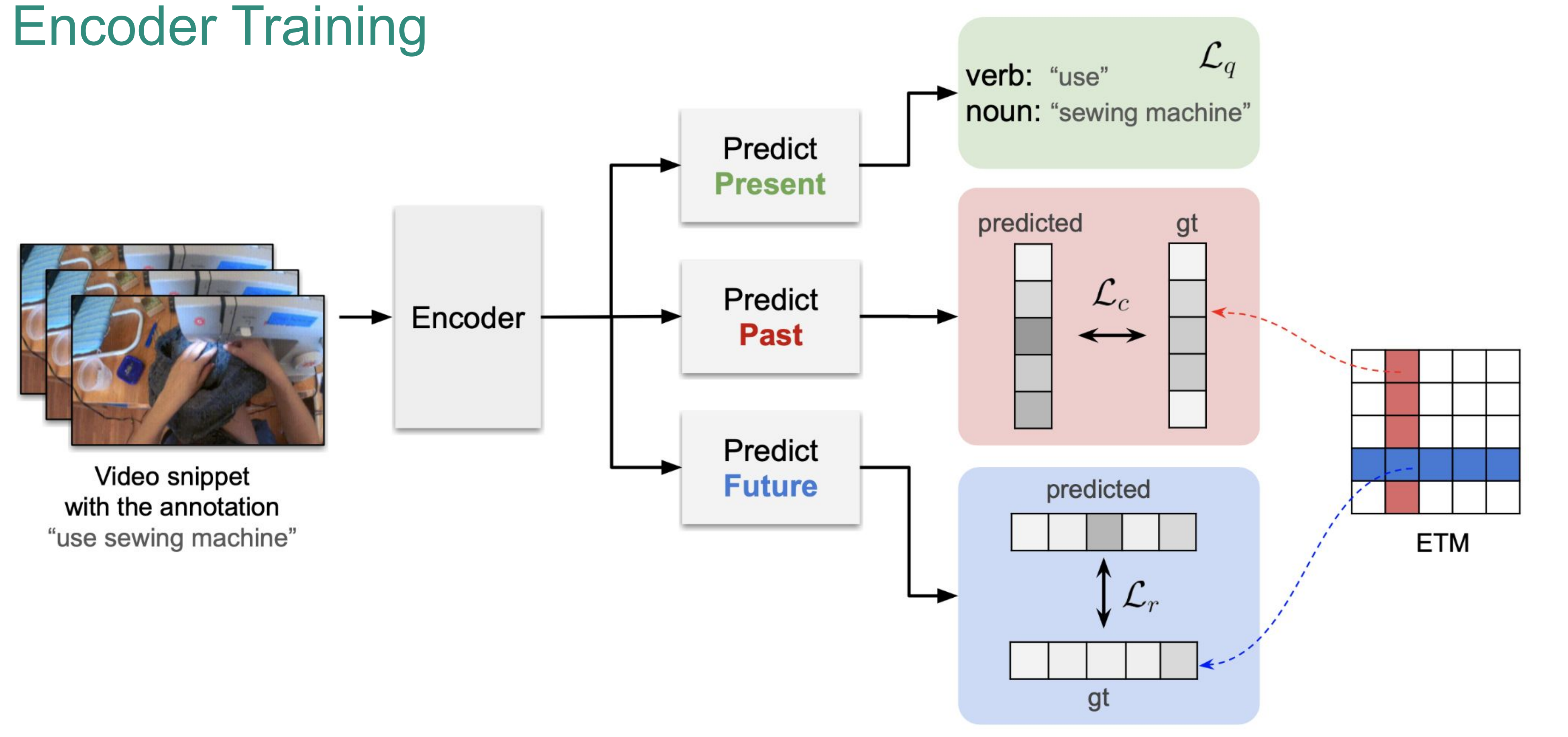
Proposed Approach

How do we build the ETM?

We combine information from all previous and subsequent events, weighted by their temporal distance from the queried action, capturing long-range relationships among events. This temporal distance can be measured in two ways: **time** or **number of events,** with different tradeoffs.



Encoder Training



Ablation studies

To show that these effects are due to our specific protocol, we train models with simple baselines and compare AR performance.

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**

Results

A MoViNet A0 pre-trained with ETM supervision shows **increased performance over 3 different datasets**. In AA, We observe improvements of 12.5%, 10.7% and 8.3% on EK100, EGO4D and EGTEA Gaze+, respectively.

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

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**

Performance on different architecture families