Latency Matters: Real-Time Action Forecasting Transformer




Harshayu Girase* | Nakul Agarwal | Chiho Choi | Karttikeya Mangalam*

*denotes equal technical contribution

Poster Session THU-AM-218

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA

Problem Formulation

Input

Video consisting of past frames
(without action labels)

Output

Predicted action at a predetermined time t_f after the present

F_0

...

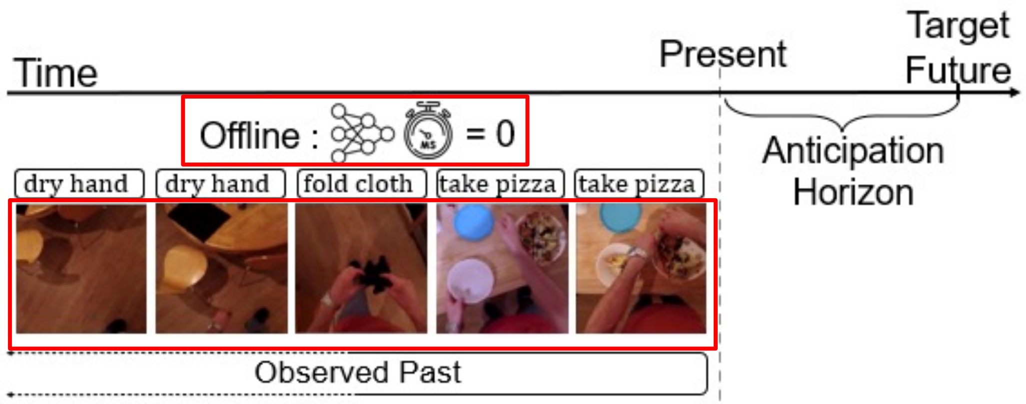
F_t

wash knife

Real-time Action Forecasting Evaluation

Offline Evaluation

- Ignores model latency
- Uses all video data up to present time T to predict the future



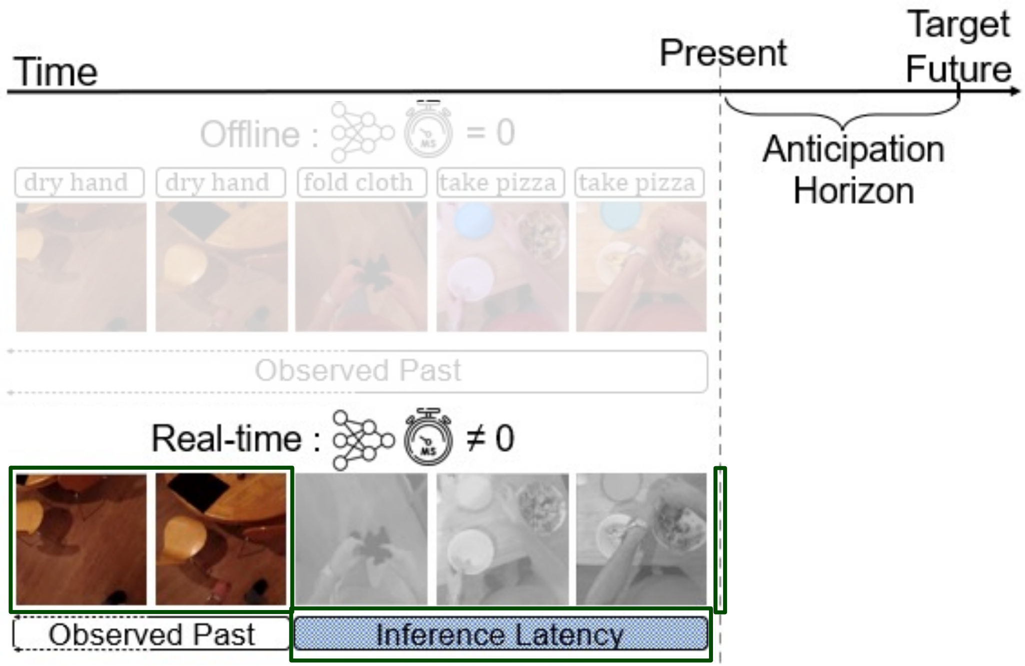
Real-time Action Forecasting Evaluation

Offline Evaluation

- Ignores model latency
- Uses all video data up to present time T to predict the future

Real-time Evaluation

- Takes model latency into account
- Model can only use video data up to time $T - t_{latency}$ to predict
- Forces prediction to arrive at time T



BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA

Real-time Action Forecasting Evaluation

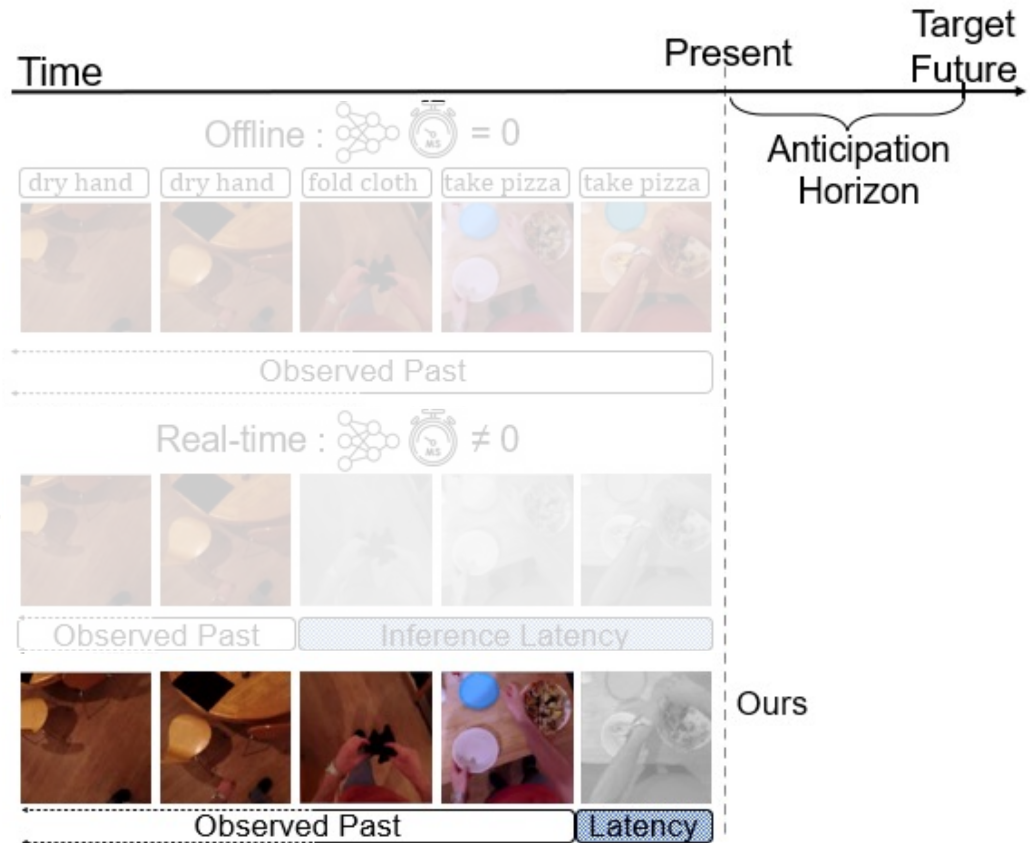
Offline Evaluation

- Ignores model latency
- Uses all video data up to present time T to predict the future

Real-time Evaluation

- Takes model latency into account
- Model can only use video data up to time $T - t_{latency}$ to predict
- Forces prediction to arrive at time T

We propose RAFTformer, a novel action anticipation transformer that balances high performance with low latency



Results: Offline Setting

All past frames up to the present time T
are used to predict the action at time T + t_f

Split	Method	Addl. Modality	Init	Epic Boxes	Top-5 Recall Verb	Top-5 Recall Noun	Top-5 Recall Action	Parameters ($\times 10^6$)	GPU Hours	Inference Latency (ms)
Val	TempAgg [70]	None	IN1K		24.2	29.8	13.0	-	-	-
	RULSTM [19]	None	IN1K		-	-	13.3	-	-	-
	RULSTM [19]	Obj+Flow	IN1K	✓	30.8	27.8	14.0	-	-	-
	TempAgg [70]	Obj+Flow+ROI	IN1K	✓	23.2	31.4	14.7	-	-	-
	AVT [25]	None	IN21K		30.2	31.7	14.9	378	-	420
	AVT+ [25]	Obj	IN21K	✓	28.2	32.0	15.9	-	-	-
	TSN-AVT+ [25]	Obj	IN21K	✓	31.8	25.5	14.8	-	-	-
	MeMVit [80]	None	K400		32.8	33.2	15.1	59	-	160
	MeMVit [80]	None	K700		32.2	37.0	17.7	212	368	350
	RAFTformer	None	K400 + IN1K		33.3	35.5	17.6	**26**	**23**	**40**
	RAFTformer	None	K700		33.7	37.1	18.0	**26**	27	110
	RAFTformer-2B	None	K700 + IN1K		**33.8**	**37.9**	**19.1**	52	50	160

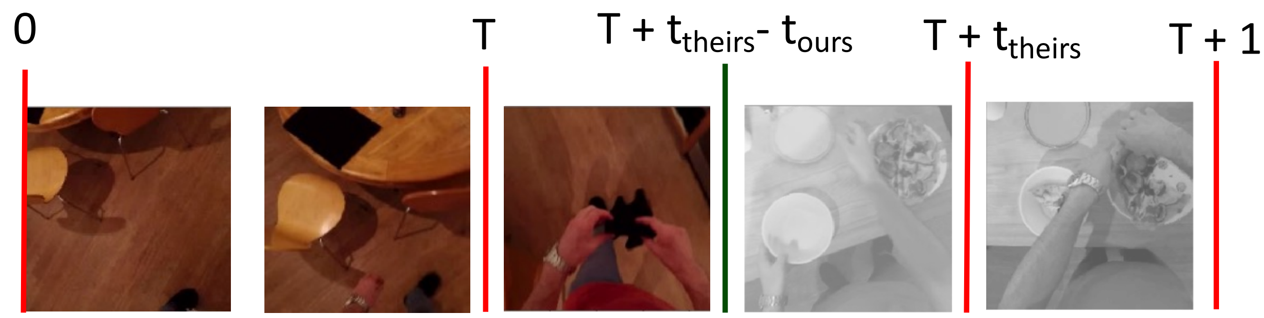
~9x less latency

State-of-the-art results

~8x less parameters

~94% less GPU hours

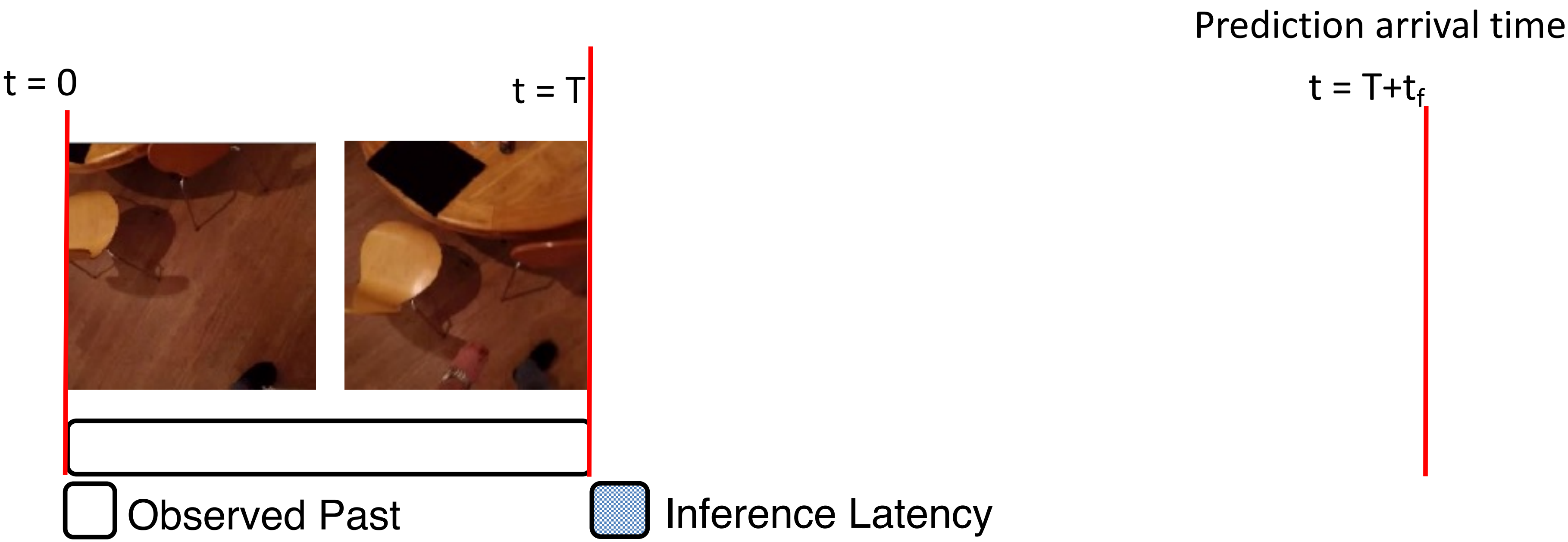
Results: Online Setting



Model	Init	Latency (t_l ms)	Inference Start Time Stamp	Inference End Time Stamp	Target Time Stamp	Top-5 Recall Verb	Top-5 Recall Noun	Top-5 Recall Action
AVT [25]	IN21K	$t_{\text{avt}} = 420$	T	$T + t_{\text{avt}}$	$T+1$	30.2	31.7	14.9
RAFTformer	K400 + IN1k	$t_{\text{ours}} = 40$	$T + t_{\text{avt}} - t_{\text{ours}}$	$T + t_{\text{avt}}$	$T+1$	34.1	38.2	**19.3** (+4.4)
MemViT [80]	K400	$t_{\text{vit}} = 160$	T	$T + t_{\text{vit}}$	$T+1$	32.8	33.2	15.1
RAFTformer	K400 + IN1k	$t_{\text{ours}} = 40$	$T + t_{\text{vit}} - t_{\text{ours}}$	$T + t_{\text{vit}}$	$T+1$	33.8	37.1	**18.1** (+3.0)
MemViT [80]	K700	$t_{\text{vit}} = 350$	T	$T + t_{\text{vit}}$	$T+1$	32.2	37.0	17.7
RAFTformer	K400 + IN1k	$t_{\text{ours}} = 40$	$T + t_{\text{vit}} - t_{\text{ours}}$	$T + t_{\text{vit}}$	$T+1$	33.7	37.9	**19.0** (+1.3)

Offline Forecasting Evaluation: Shortcomings

Forecasting models often must be done in real time.
But current good forecasting models have high latency.

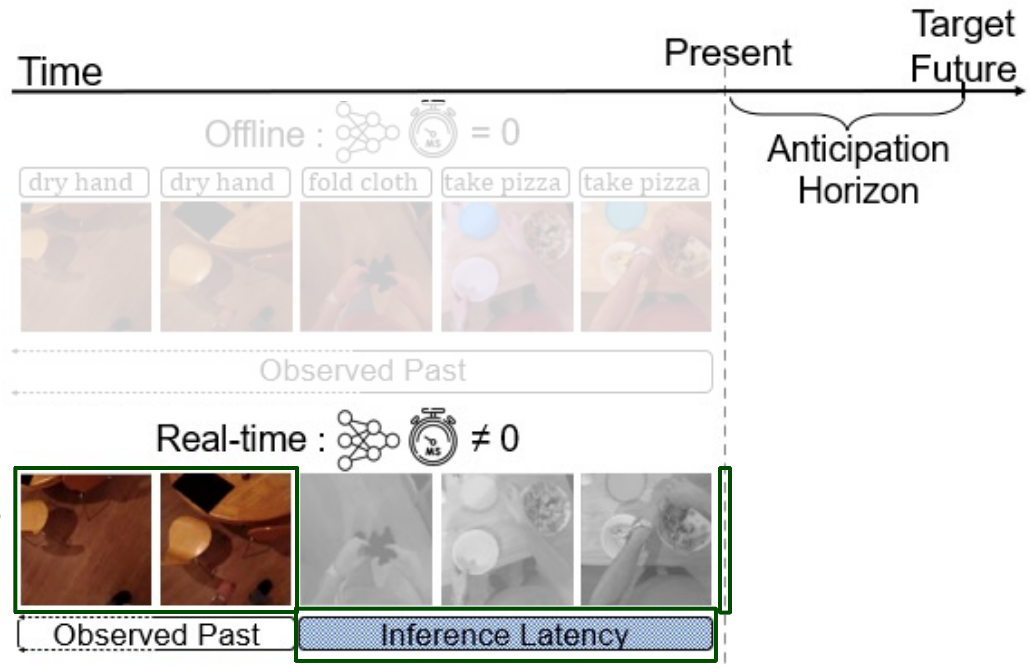


The latency can even be so high that the model furnishes predictions after the future action has already happened

Real-time Action Forecasting Evaluation

Offline Evaluation

- Ignores model latency
- Uses all video data up to present time T to predict the future



Real-time Evaluation

- Takes model latency into account
- Model can only use video data up to time T – $t_{latency}$ to predict
- Forces prediction to arrive at time T

- Tradeoff between latency and real-time performance. Larger models can lead to poorer real-time performance.

Bigger is not necessarily better in real-time!

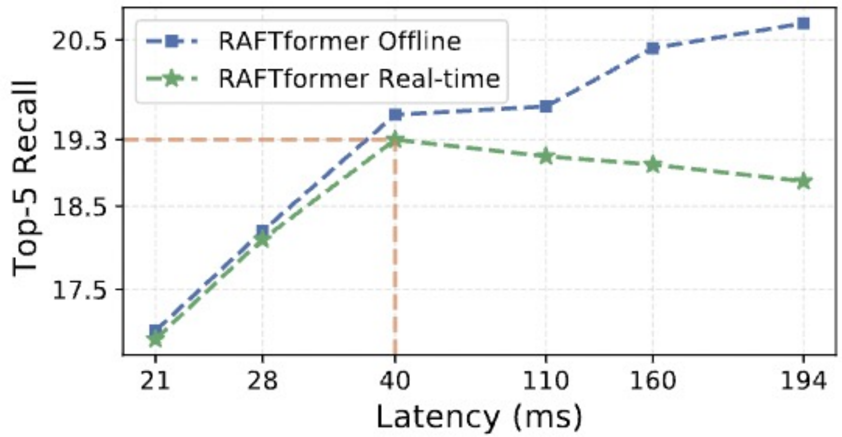
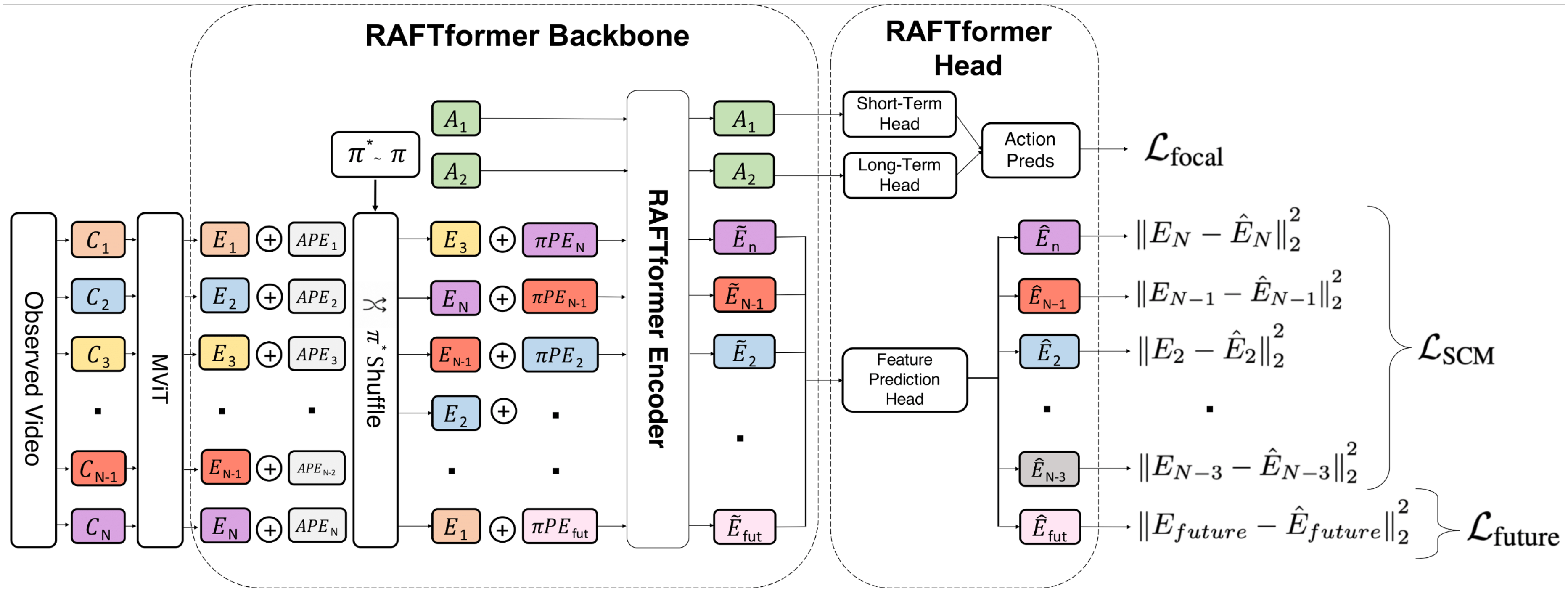
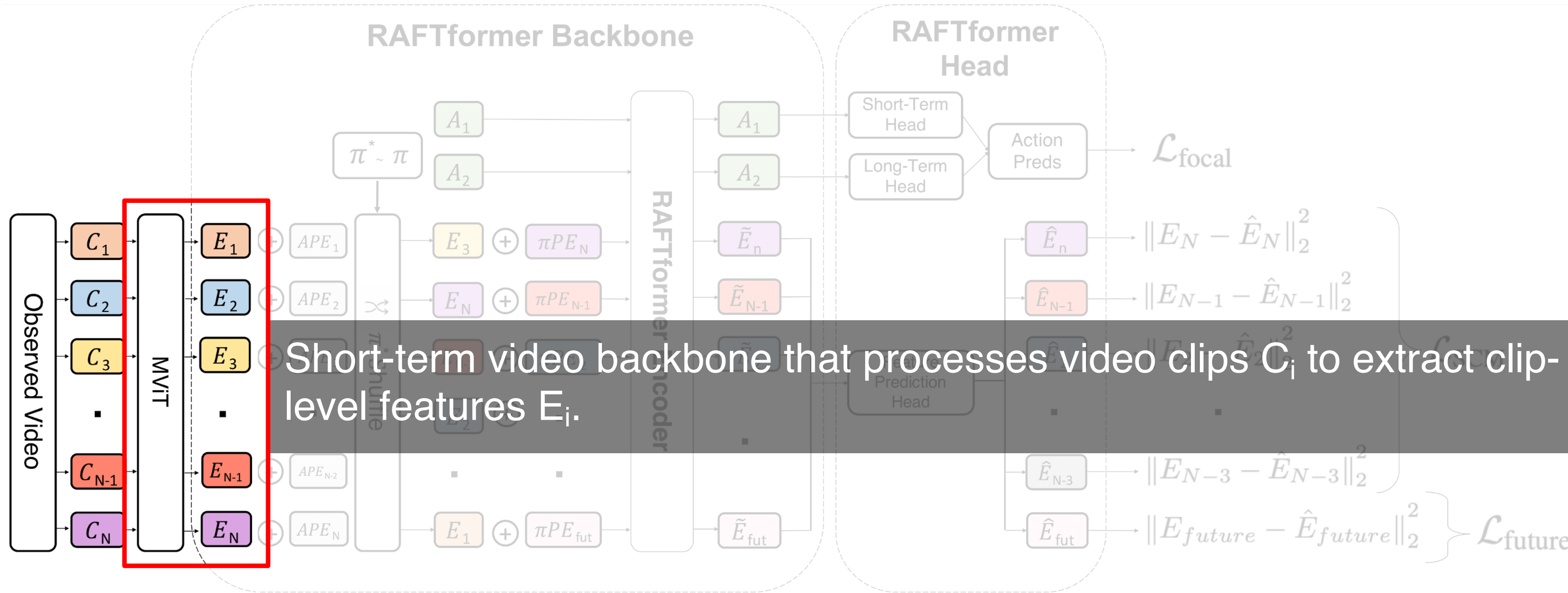


Figure 2. **Evaluation Performance *vs*. Latency.** Bigger models perform better in latency agnostic offline settings. In the real-time evaluation setting, we observe that, beyond a limit, bigger models with higher latency cause a drop in forecasting performance. In practical deployment, there exists a trade-off between latency and high-fidelity forecasts. See §4.3.1 for details.

RAFTformer Architecture

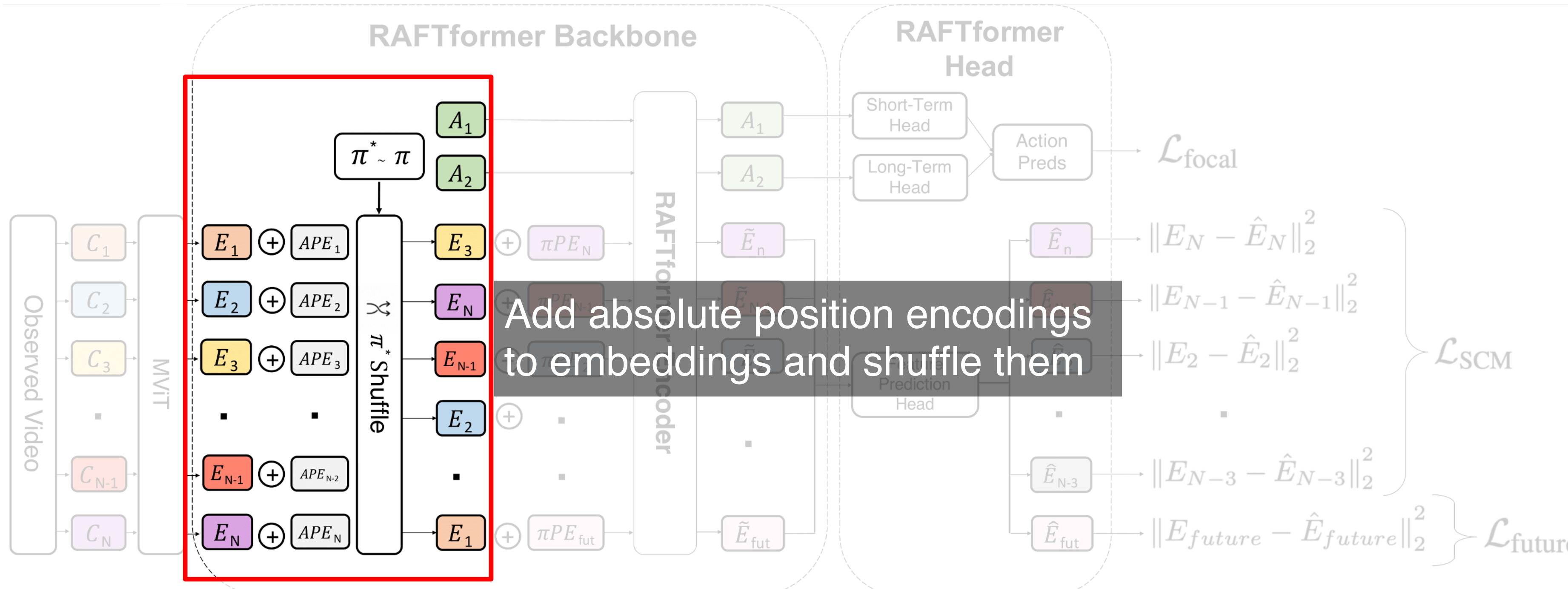


RAFTformer Architecture



Short-term video backbone that processes video clips C_i to extract clip-level features E_i.

RAFTformer Architecture

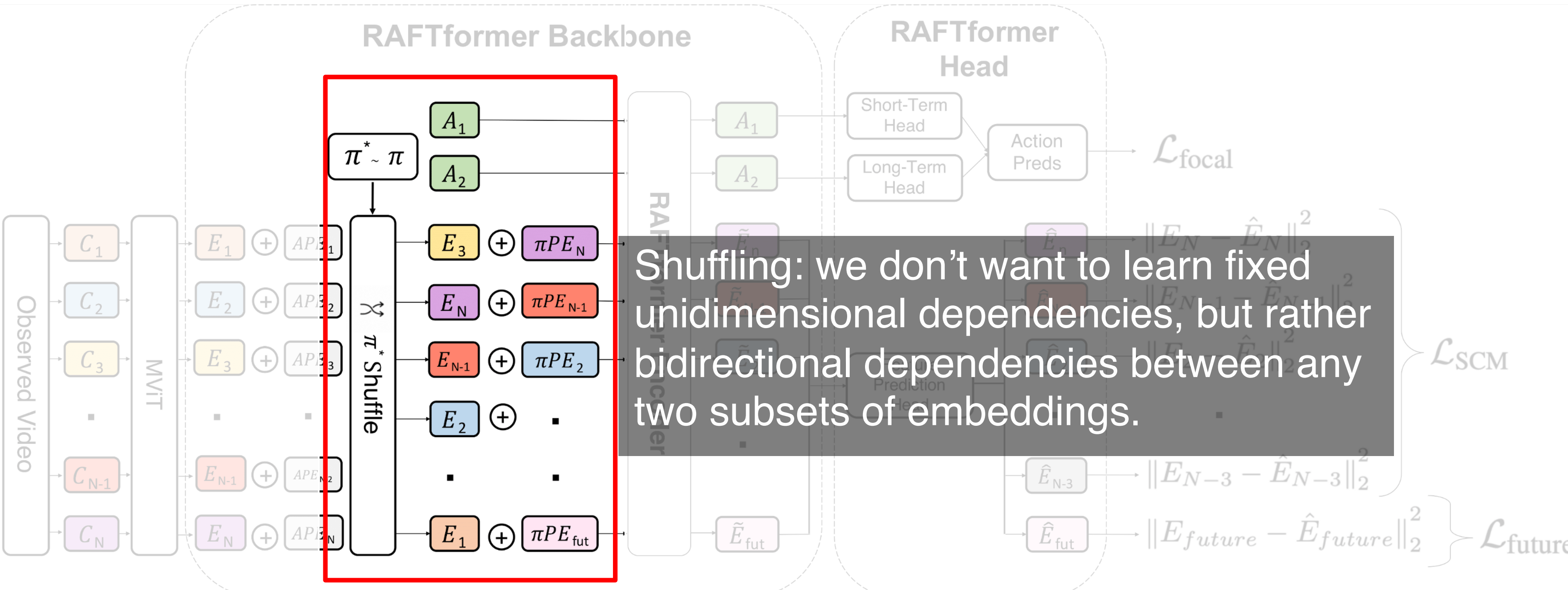


Add absolute position encodings to embeddings and shuffle them

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA

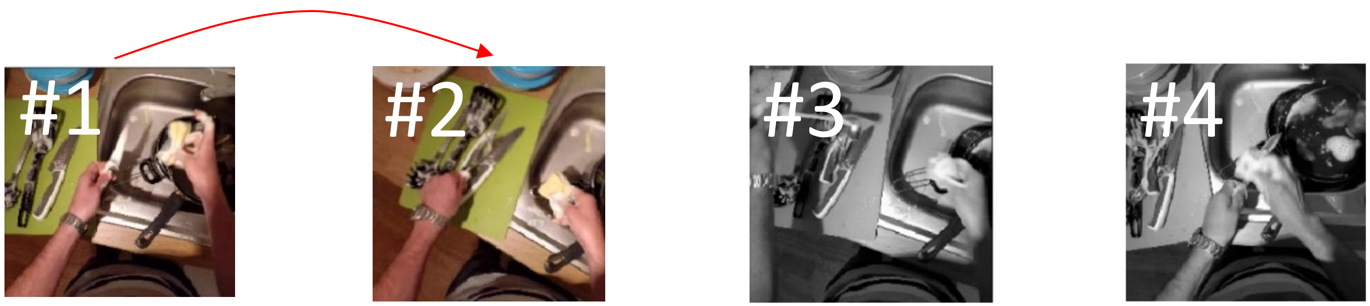
RAFTformer Architecture



Shuffling: we don't want to learn fixed unidimensional dependencies, but rather bidirectional dependencies between any two subsets of embeddings.

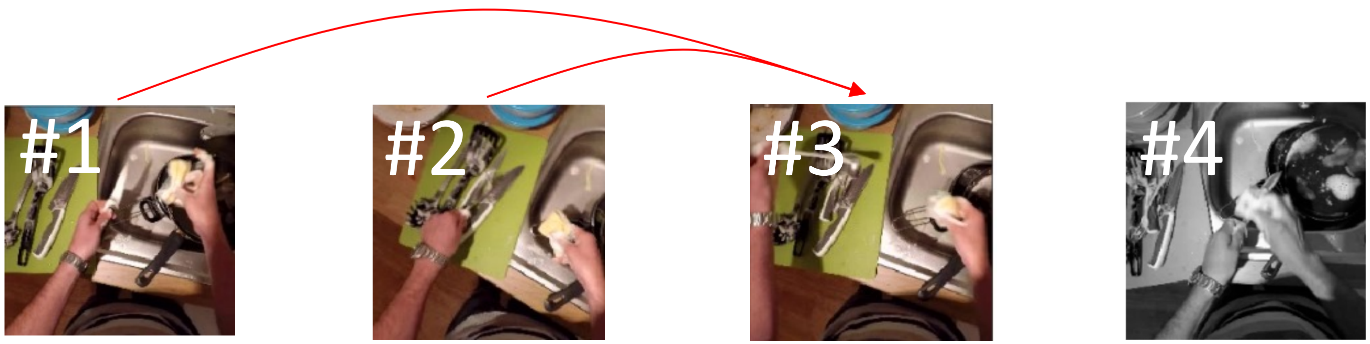
Why Do Shuffling?

Without shuffling, a model using causal attention masking can only learn sequential unidimensional dependencies.



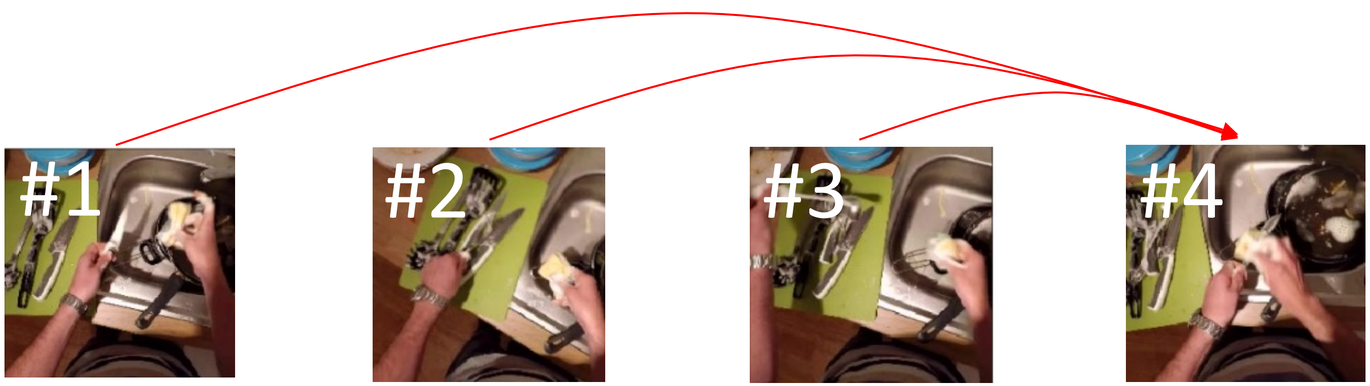
Why Do Shuffling?

Without shuffling, a model using causal attention masking can only learn sequential unidimensional dependencies.



Why Do Shuffling?

Without shuffling, a model using causal attention masking can only learn sequential unidimensional dependencies.



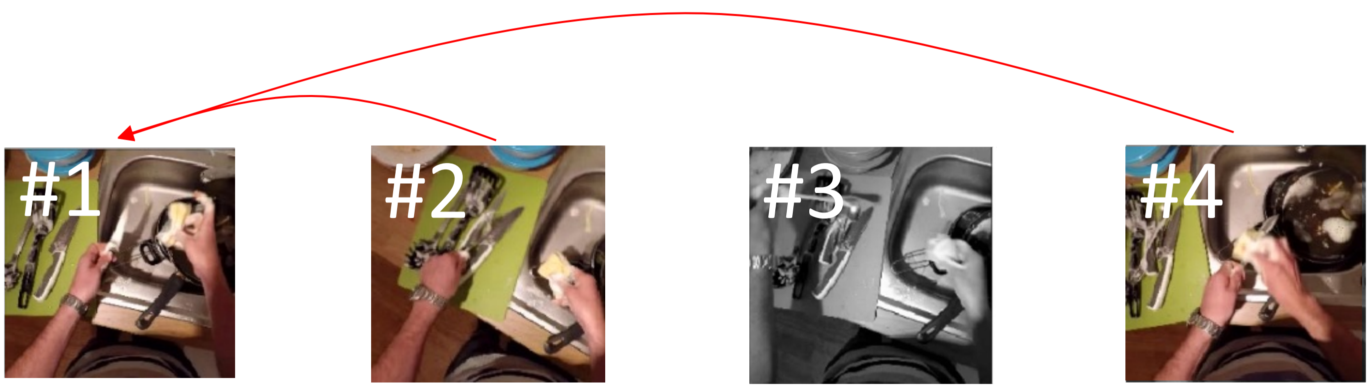
Why Do Shuffling?

Shuffling tokens carefully* allows the model to learn bidirectional dependencies between any two token subsets.

*requires careful changes to masked attention to ensure no multi-hop information leakage through self-attention layers. Please see paper for details.

Why Do Shuffling?

Shuffling tokens carefully* allows the model to learn bidirectional dependencies between any two token subsets.



*requires careful changes to masked attention to ensure no multi-hop information leakage through self-attention layers. Please see paper for details.

Why Do Shuffling?

Shuffling tokens carefully* allows the model to learn bidirectional dependencies between any two token subsets.

*requires careful changes to masked attention to ensure no multi-hop information leakage through self-attention layers. Please see paper for details.

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

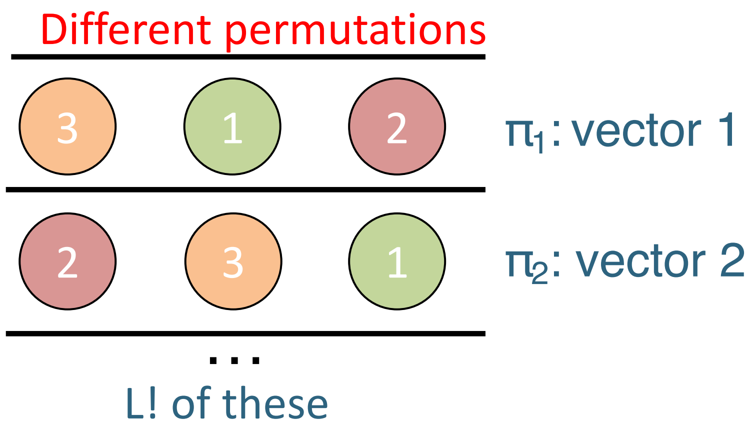
Berkeley
UNIVERSITY OF CALIFORNIA

Encoding Permutation π

Goal: We want RAFTformer encoder to “know” sampled permutation π^*, so we want to embed π^* vectorially for use in the encoder.

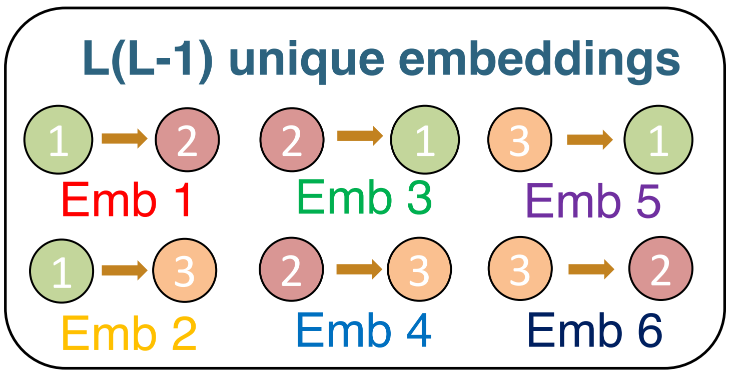
Encoding Permutation π
Naïve Method

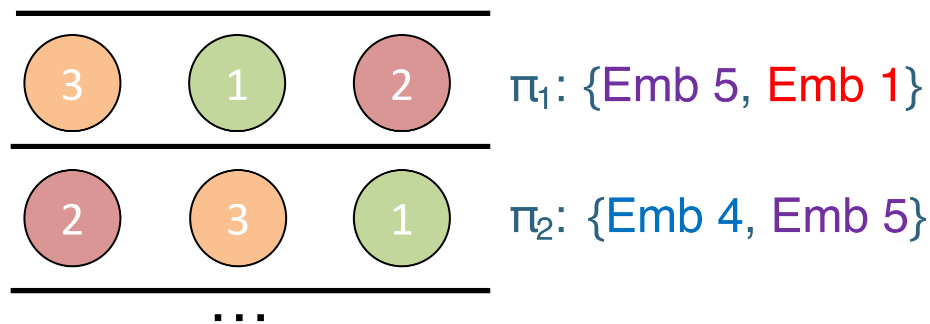
Method: Assign each π a single learnable vector
L! unique embeddings needed to encode all possible permutations π (one for each permutation)



Encoding Permutation π
Predecessor Successor Method

Method: Encode each π as a set of predecessor → successor relationships. L(L-1) unique embeddings needed to encode all possible permutations π (one for each pair i→j)





Encoding Permutation π
πPE Method

Method: Encode only the successor PE in the permuted π^* (predecessor is already encoded through Absolute PE)

L(L-1) unique embeddings needed to encode all possible permutations π (one for each i)

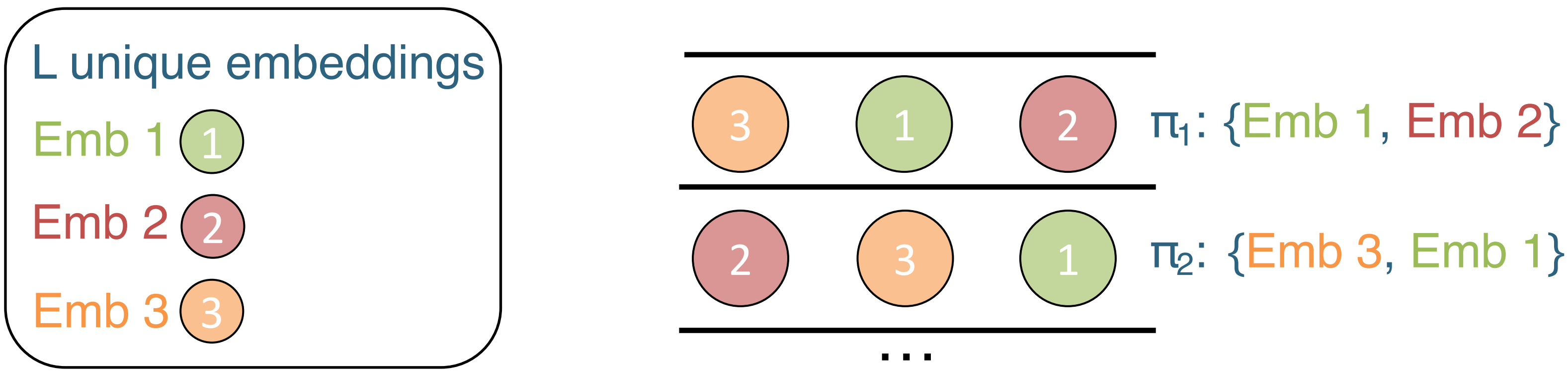
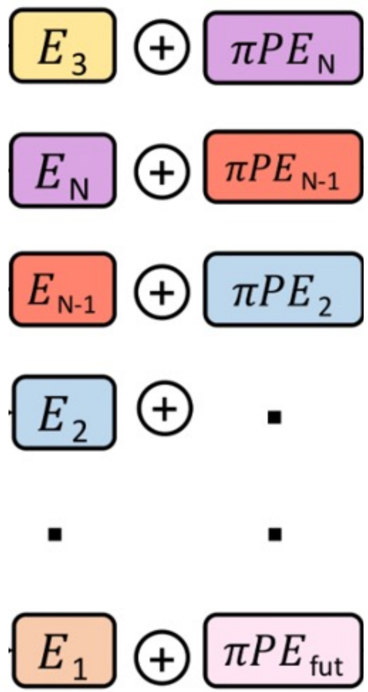


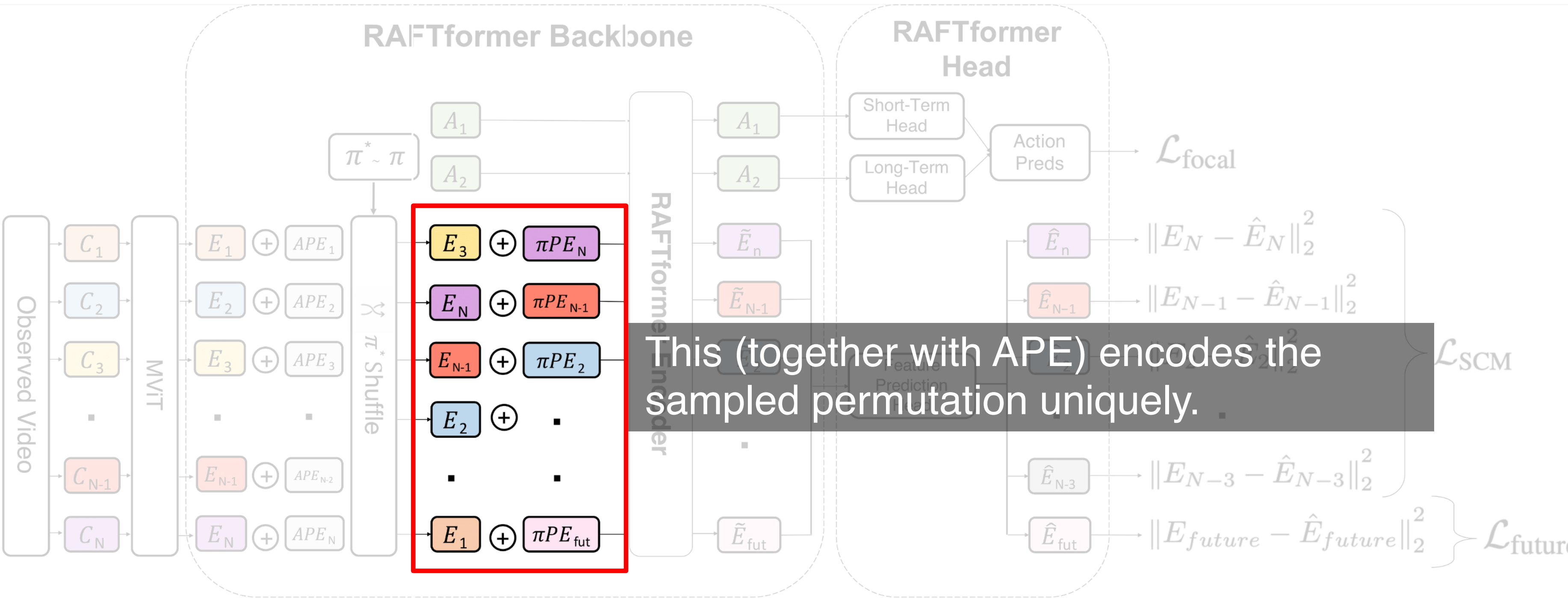
Illustration of πPE



πPE is the encoding of the original temporal position of the successor in the permuted sequence

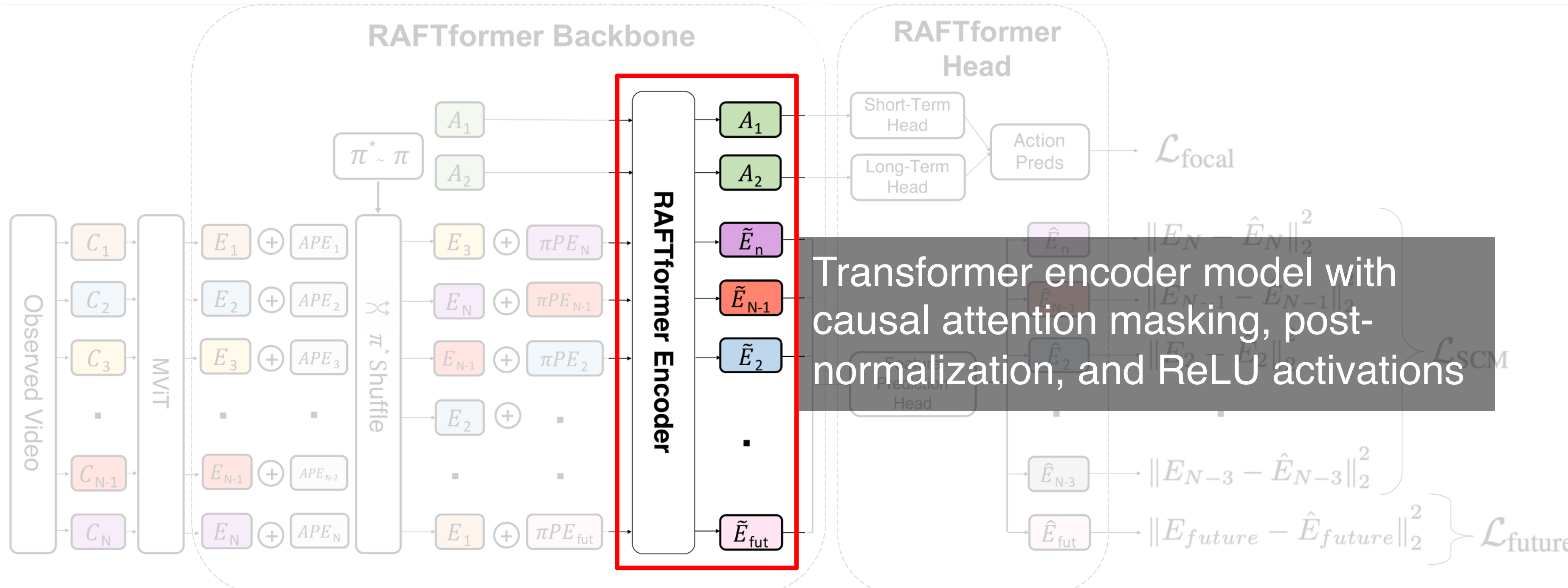
The last token adds πPE_{fut}, which is used to help generate the "future token"

RAFTformer Architecture



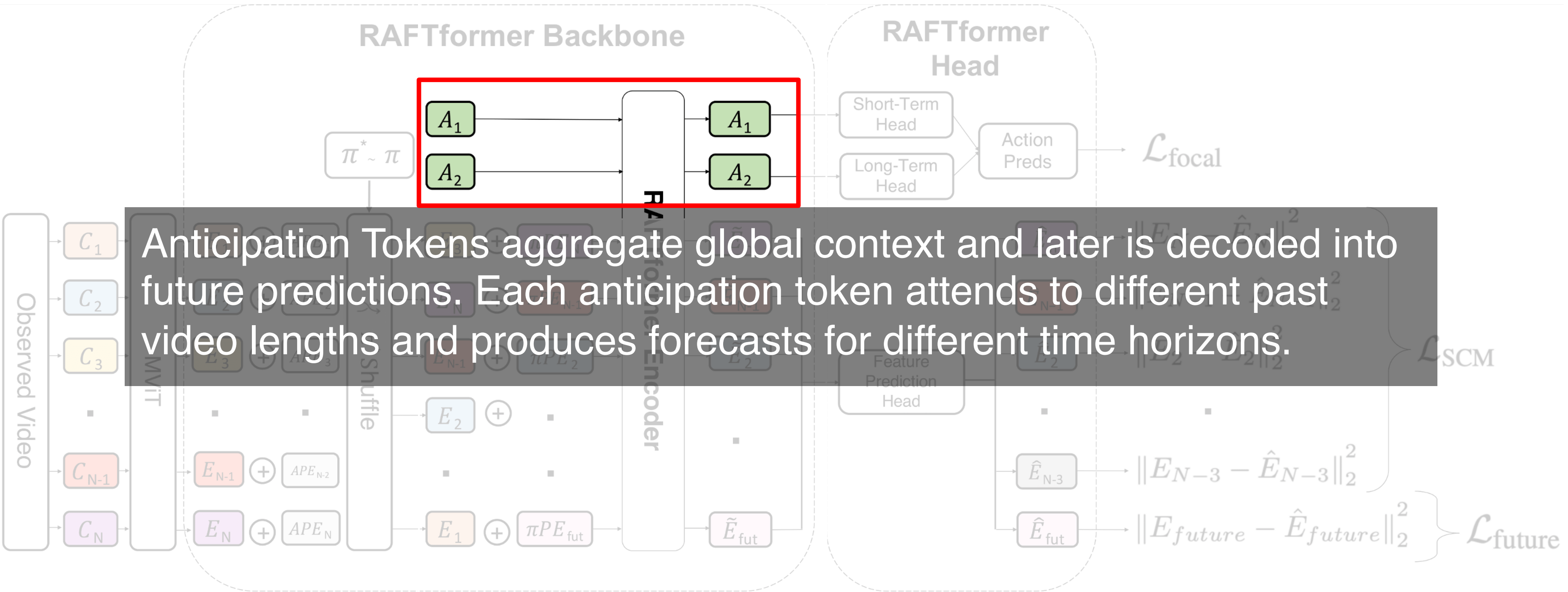
This (together with APE) encodes the sampled permutation uniquely.

RAFTformer Architecture



*RAFTformer encoder has a special form of masked attention that prevents information leakage under shuffling. Please see paper for details.

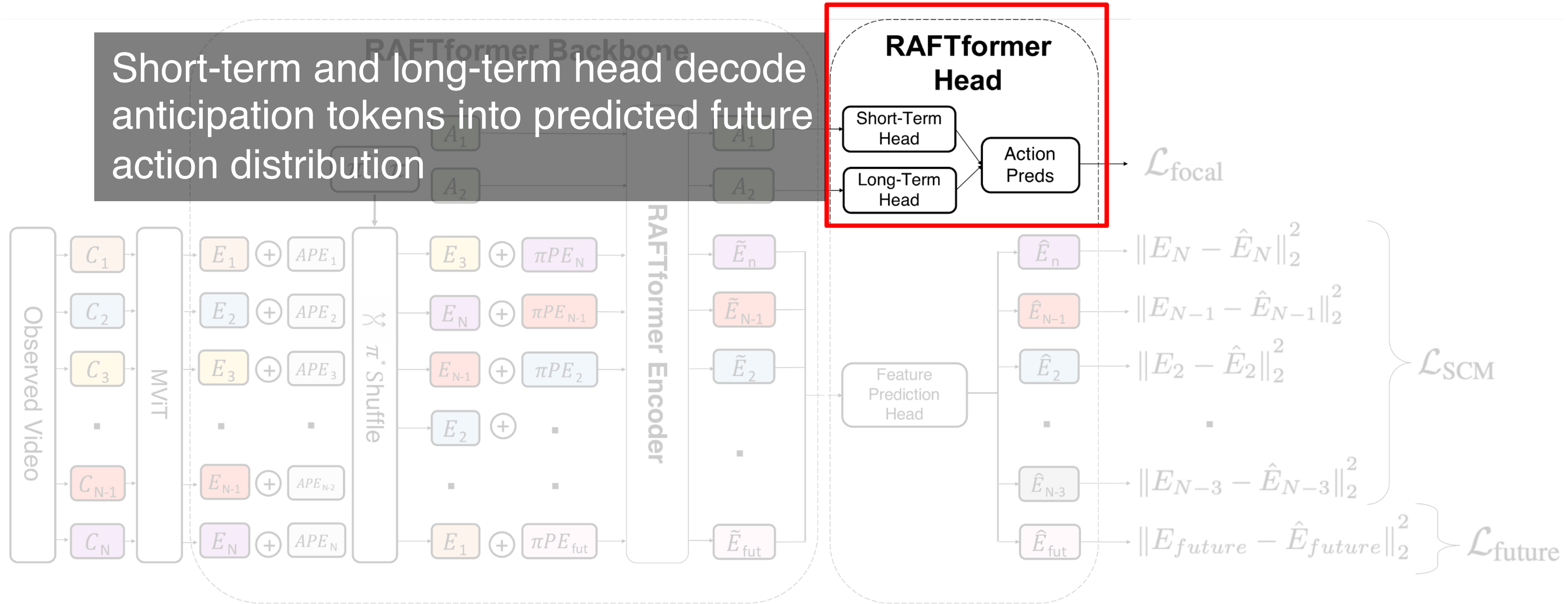
RAFTformer Architecture



Anticipation Tokens aggregate global context and later is decoded into future predictions. Each anticipation token attends to different past video lengths and produces forecasts for different time horizons.

RAFTformer Architecture

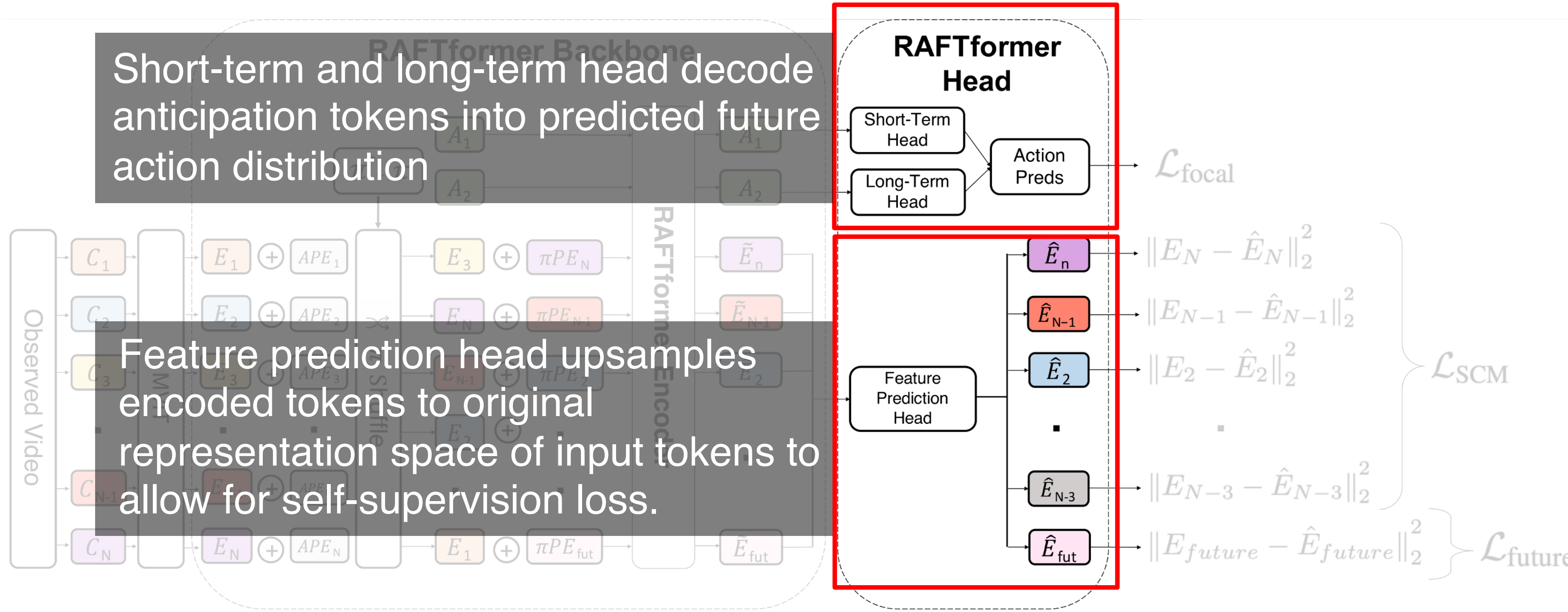
Short-term and long-term head decode anticipation tokens into predicted future action distribution



RAFTformer Architecture

Short-term and long-term head decode anticipation tokens into predicted future action distribution

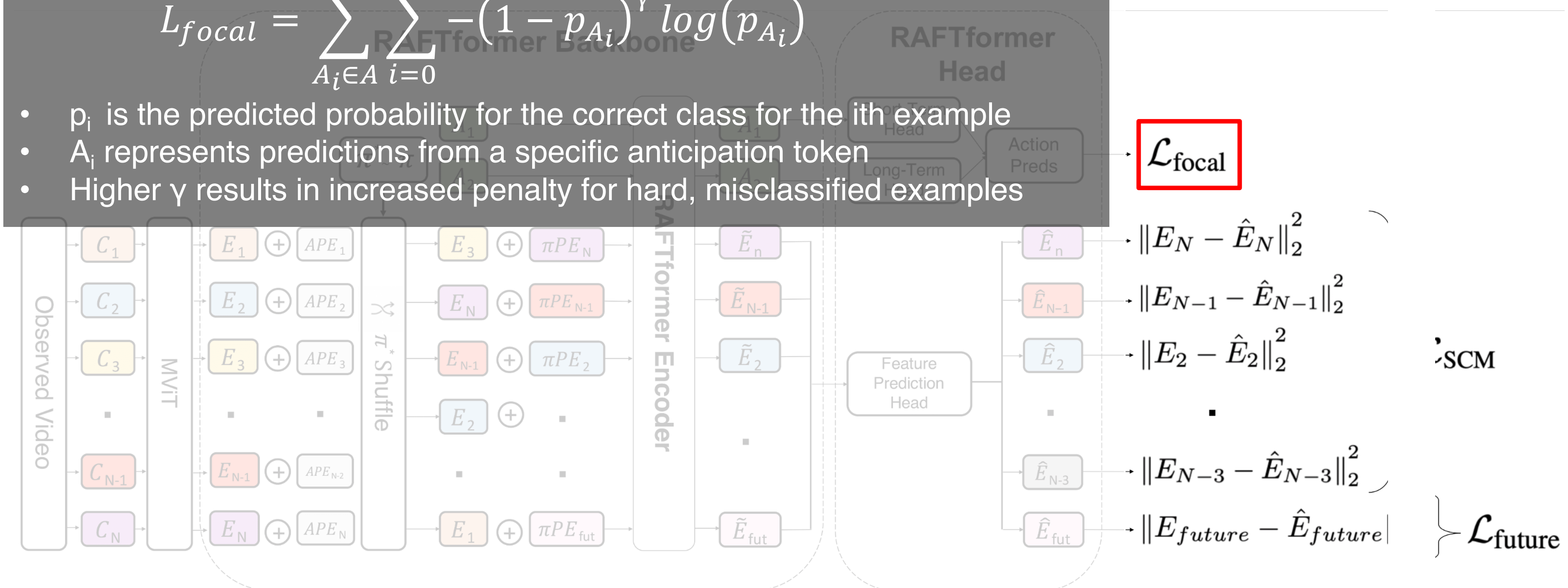
Feature prediction head upsamples encoded tokens to original representation space of input tokens to allow for self-supervision loss.



RAFTformer Architecture

$$L_{focal} = \sum_{A_i \in A} \sum_{i=0}^{n} -(1 - p_{A_i})^{\gamma} \log(p_{A_i})$$

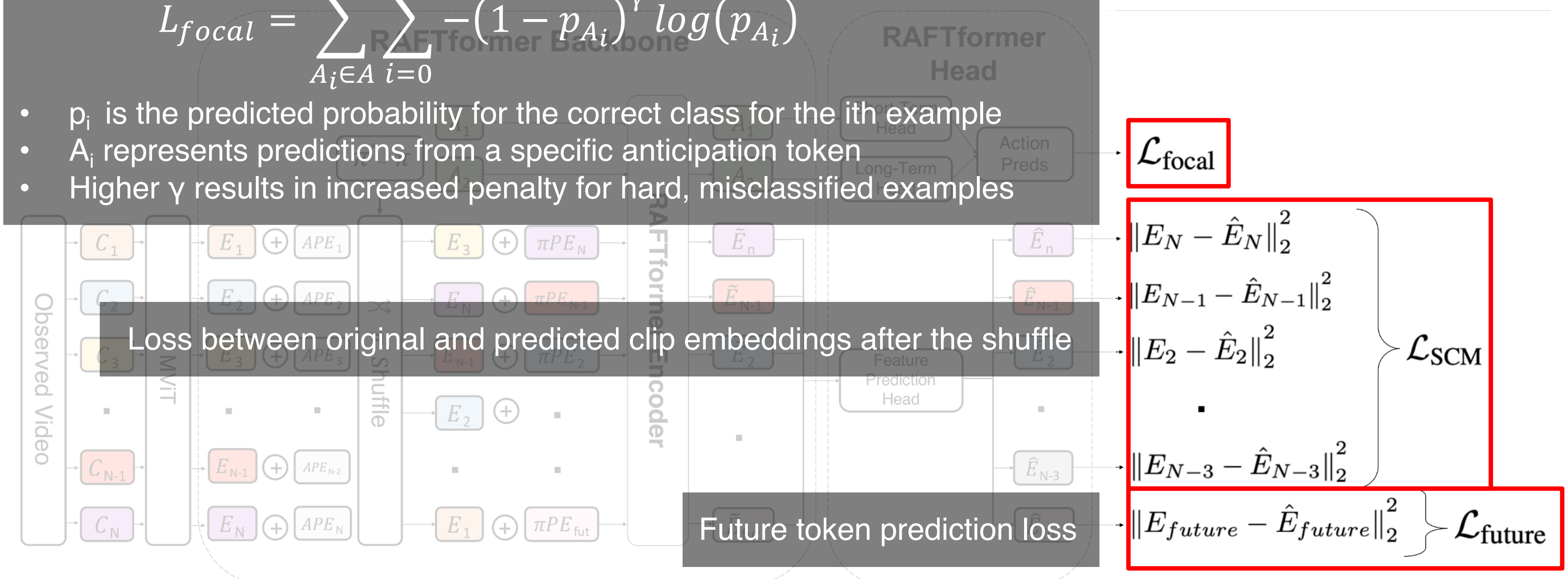
- p_i is the predicted probability for the correct class for the ith example
- A_i represents predictions from a specific anticipation token
- Higher γ results in increased penalty for hard, misclassified examples



RAFTformer Architecture

$$L_{focal} = \sum_{A_i \in A} \sum_{i=0}^{n} -(1 - p_{A_i})^{\gamma} \log(p_{A_i})$$

- p_i is the predicted probability for the correct class for the ith example
- A_i represents predictions from a specific anticipation token
- Higher γ results in increased penalty for hard, misclassified examples



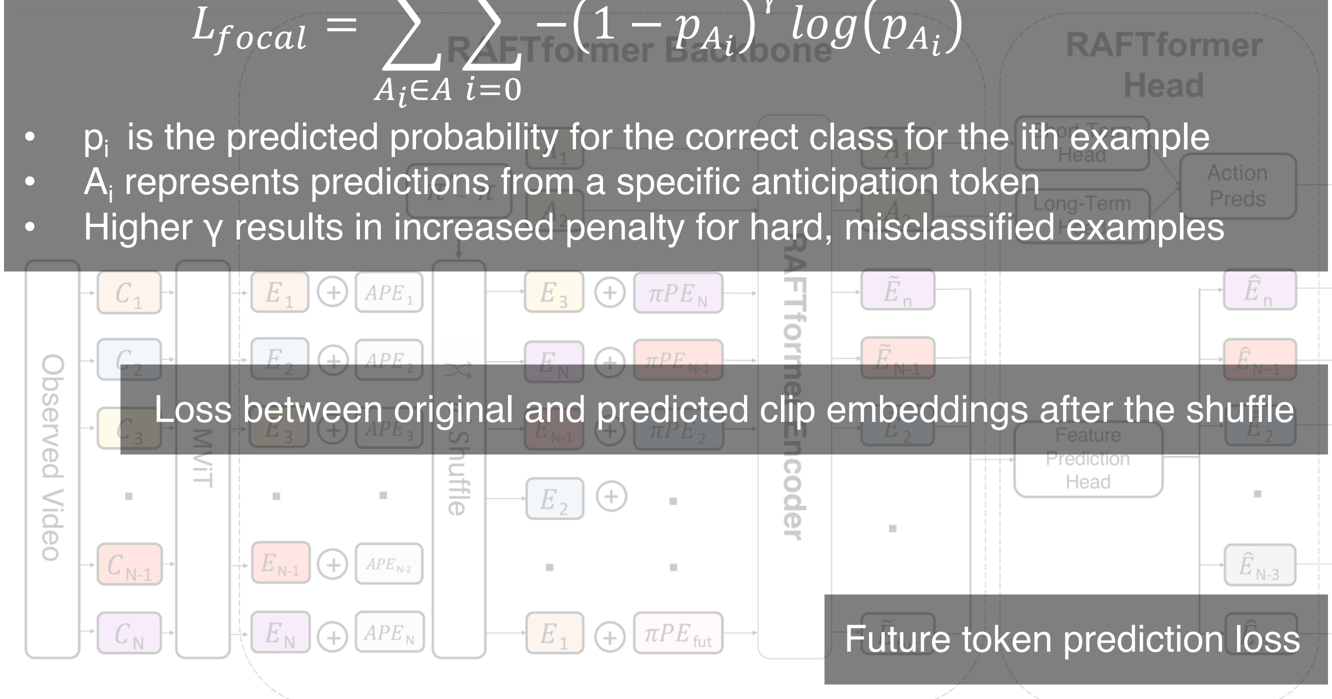
Loss between original and predicted clip embeddings after the shuffle

Future token prediction loss

RAFTformer Architecture

$$L_{focal} = \sum_{A_i \in A} \sum_{i=0}^{n} -(1 - p_{A_i})^{\gamma} \log(p_{A_i})$$

- p_i is the predicted probability for the correct class for the ith example
- A_i represents predictions from a specific anticipation token
- Higher γ results in increased penalty for hard, misclassified examples



Loss between original and predicted clip embeddings after the shuffle

Future token prediction loss

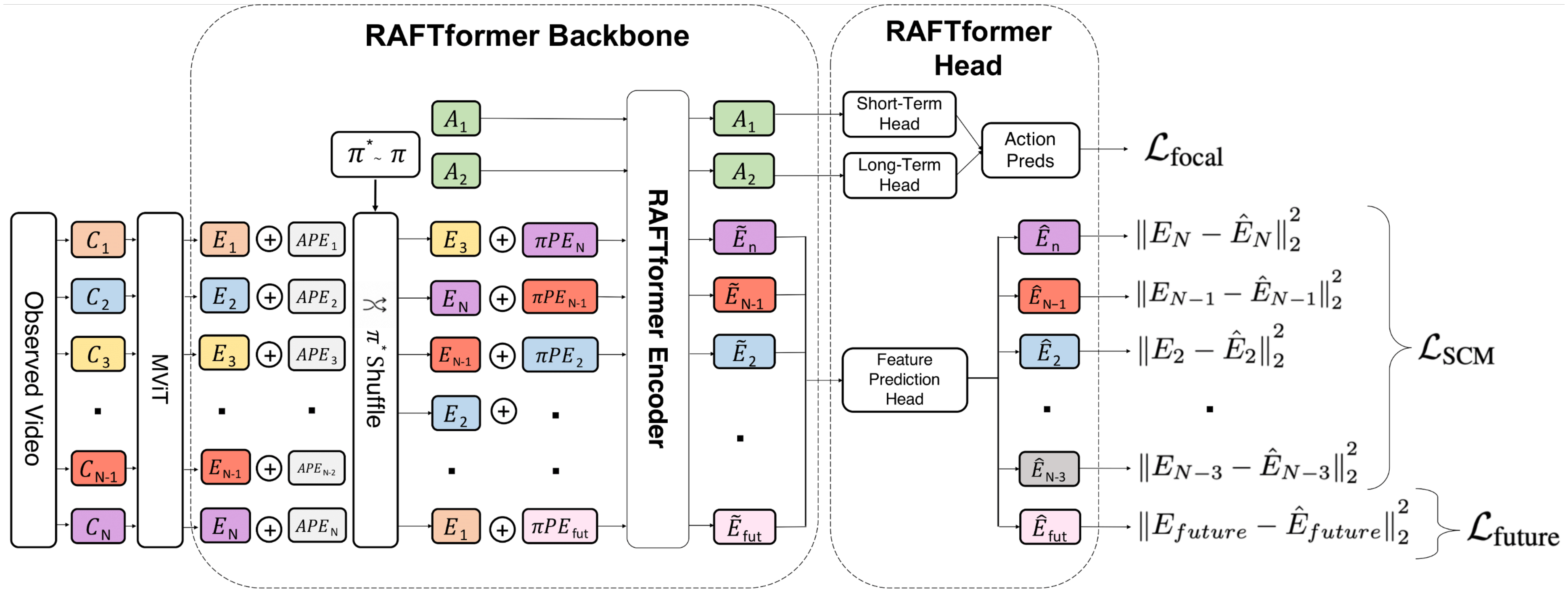
$\mathcal{L}_{\text{focal}}$

$$\mathcal{L}_{\text{SCM}} = \left\{ \begin{array}{l} \|E_N - \hat{E}_N\|_2^2 \\ \|E_{N-1} - \hat{E}_{N-1}\|_2^2 \\ \|E_2 - \hat{E}_2\|_2^2 \\ \cdot \\ \|E_{N-3} - \hat{E}_{N-3}\|_2^2 \end{array} \right.$$

$$\mathcal{L}_{\text{future}} = \|E_{future} - \hat{E}_{future}\|_2^2$$

$$L = L_{focal} + \lambda_1 L_{SCM} + \lambda_2 L_{future}$$

RAFTformer Architecture



Results: Offline Setting

All past frames up to the present time T
are used to predict the action at time T + t_f

Split	Method	Addl. Modality	Init	Epic Boxes	Top-5 Recall			Parameters ($\times 10^6$)	GPU Hours	Inference Latency (ms)
					Verb	Noun	Action			
Val	TempAgg [70]	None	IN1K		24.2	29.8	13.0	-	-	-
	RULSTM [19]	None	IN1K		-	-	13.3	-	-	-
	RULSTM [19]	Obj+Flow	IN1K	✓	30.8	27.8	14.0	-	-	-
	TempAgg [70]	Obj+Flow+ROI	IN1K	✓	23.2	31.4	14.7	-	-	-
	AVT [25]	None	IN21K		30.2	31.7	14.9	378	-	420
	AVT+ [25]	Obj	IN21K	✓	28.2	32.0	15.9	-	-	-
	TSN-AVT+ [25]	Obj	IN21K	✓	31.8	25.5	14.8	-	-	-
	MeMVit [80]	None	K400		32.8	33.2	15.1	59	-	160
	MeMVit [80]	None	K700		32.2	37.0	17.7	212	368	350
	RAFTformer	None	K400 + IN1K		33.3	35.5	17.6	**26**	**23**	**40**
	RAFTformer	None	K700		33.7	37.1	18.0	**26**	27	110
	RAFTformer-2B	None	K700 + IN1K		**33.8**	**37.9**	**19.1**	52	50	160

Results: Offline Setting

All past frames up to the present time T
are used to predict the action at time T + t_f

Split	Method	Addl. Modality	Init	Epic Boxes	Top-5 Recall Verb	Top-5 Recall Noun	Top-5 Recall Action	Parameters ($\times 10^6$)	GPU Hours	Inference Latency (ms)
Val	TempAgg [70]	None	IN1K		24.2	29.8	13.0	-	-	-
	RULSTM [19]	None	IN1K		-	-	13.3	-	-	-
	RULSTM [19]	Obj+Flow	IN1K	✓	30.8	27.8	14.0	-	-	-
	TempAgg [70]	Obj+Flow+ROI	IN1K	✓	23.2	31.4	14.7	-	-	-
	AVT [25]	None	IN21K		30.2	31.7	14.9	378	-	420
	AVT+ [25]	Obj	IN21K	✓	28.2	32.0	15.9	-	-	-
	TSN-AVT+ [25]	Obj	IN21K	✓	31.8	25.5	14.8	-	-	-
	MeMVit [80]	None	K400		32.8	33.2	15.1	59	-	160
	MeMVit [80]	None	K700		32.2	37.0	17.7	212	368	350
	RAFTformer	None	K400 + IN1K		33.3	35.5	17.6	**26**	**23**	**40**
	RAFTformer	None	K700		33.7	37.1	18.0	**26**	27	110
	RAFTformer-2B	None	K700 + IN1K		**33.8**	**37.9**	**19.1**	52	50	160

State-of-the-art
results

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA

Results: Offline Setting

All past frames up to the present time T
are used to predict the action at time $T + t_f$

Split	Method	Addl. Modality	Init	Epic Boxes	Top-5 Recall Verb	Top-5 Recall Noun	Top-5 Recall Action	Parameters ($\times 10^6$)	GPU Hours	Inference Latency (ms)
Val	TempAgg [70]	None	IN1K		24.2	29.8	13.0	-	-	-
	RULSTM [19]	None	IN1K		-	-	13.3	-	-	-
	RULSTM [19]	Obj+Flow	IN1K	✓	30.8	27.8	14.0	-	-	-
	TempAgg [70]	Obj+Flow+ROI	IN1K	✓	23.2	31.4	14.7	-	-	-
	AVT [25]	None	IN21K		30.2	31.7	14.9	378	-	420
	AVT+ [25]	Obj	IN21K	✓	28.2	32.0	15.9	-	-	-
	TSN-AVT+ [25]	Obj	IN21K	✓	31.8	25.5	14.8	-	-	-
	MeMVit [80]	None	K400		32.8	33.2	15.1	59	-	160
	MeMVit [80]	None	K700		32.2	37.0	17.7	212	368	350
	RAFTformer	None	K400 + IN1K		33.3	35.5	17.6	**26**	**23**	**40**
	RAFTformer	None	K700		33.7	37.1	18.0	**26**	27	110
	RAFTformer-2B	None	K700 + IN1K		**33.8**	**37.9**	**19.1**	52	50	160

State-of-the-art results

~8x less parameters

Results: Offline Setting

All past frames up to the present time T
are used to predict the action at time $T + t_f$

Split	Method	Addl. Modality	Init	Epic Boxes	Top-5 Recall Verb	Noun	Action	Parameters ($\times 10^6$)	GPU Hours	Inference Latency (ms)
Val	TempAgg [70]	None	IN1K		24.2	29.8	13.0	-	-	-
	RULSTM [19]	None	IN1K		-	-	13.3	-	-	-
	RULSTM [19]	Obj+Flow	IN1K	✓	30.8	27.8	14.0	-	-	-
	TempAgg [70]	Obj+Flow+ROI	IN1K	✓	23.2	31.4	14.7	-	-	-
	AVT [25]	None	IN21K		30.2	31.7	14.9	378	-	420
	AVT+ [25]	Obj	IN21K	✓	28.2	32.0	15.9	-	-	-
	TSN-AVT+ [25]	Obj	IN21K	✓	31.8	25.5	14.8	-	-	-
	MeMVit [80]	None	K400		32.8	33.2	15.1	59	-	160
	MeMVit [80]	None	K700		32.2	37.0	17.7	212	368	350
	RAFTformer	None	K400 + IN1K		33.3	35.5	17.6	**26**	**23**	**40**
	RAFTformer	None	K700		33.7	37.1	18.0	**26**	27	110
	RAFTformer-2B	None	K700 + IN1K		**33.8**	**37.9**	**19.1**	52	50	160

State-of-the-art results

~8x less parameters

~94% less GPU hours

Results: Offline Setting

All past frames up to the present time T
are used to predict the action at time T + t_f

Split	Method	Addl. Modality	Init	Epic Boxes	Top-5 Recall			Parameters	GPU	Inference
					Verb	Noun	Action	($\times 10^6$)	Hours	Latency (ms)
Val	TempAgg [70]	None	IN1K		24.2	29.8	13.0	-	-	-
	RULSTM [19]	None	IN1K		-	-	13.3	-	-	-
	RULSTM [19]	Obj+Flow	IN1K	✓	30.8	27.8	14.0	-	-	-
	TempAgg [70]	Obj+Flow+ROI	IN1K	✓	23.2	31.4	14.7	-	-	-
	AVT [25]	None	IN21K		30.2	31.7	14.9	378	-	420
	AVT+ [25]	Obj	IN21K	✓	28.2	32.0	15.9	-	-	-
	TSN-AVT+ [25]	Obj	IN21K	✓	31.8	25.5	14.8	-	-	-
	MeMVit [80]	None	K400		32.8	33.2	15.1	59	-	160
	MeMVit [80]	None	K700		32.2	37.0	17.7	212	368	350
	RAFTformer	None	K400 + IN1K		33.3	35.5	17.6	**26**	**23**	**40**
	RAFTformer	None	K700		33.7	37.1	18.0	**26**	27	110
	RAFTformer-2B	None	K700 + IN1K		**33.8**	**37.9**	**19.1**	52	50	160

~9x less latency

State-of-the-art results

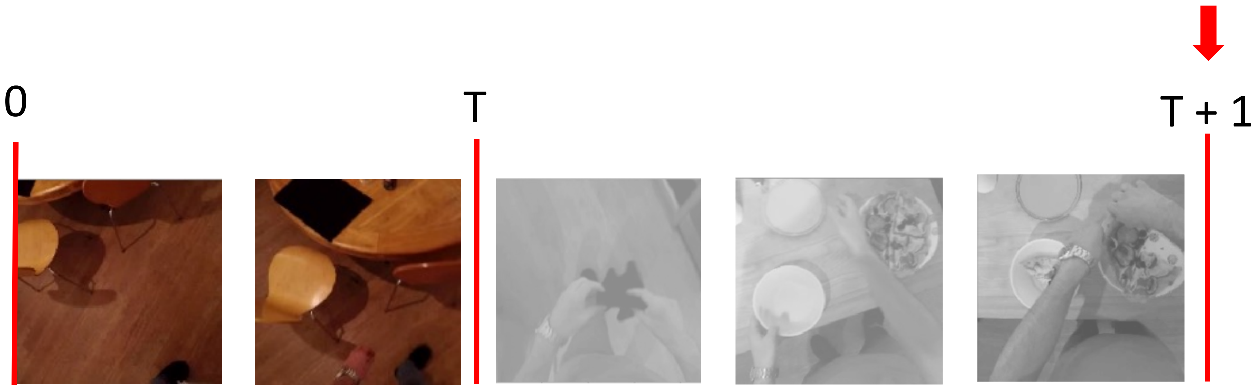
~8x less parameters

~94% less GPU hours

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA

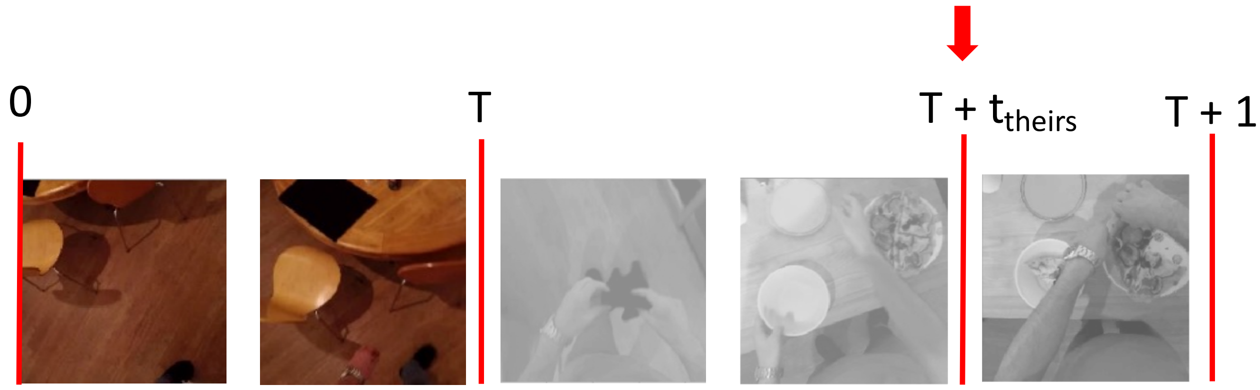
Results: Online Setting



Base models are given a video up to time T, which they used to predict the action at time T+1

Model	Init	Latency (t_l ms)	Inference Start Time Stamp	Inference End Time Stamp	Target Time Stamp	Top-5 Recall		
						Verb	Noun	Action

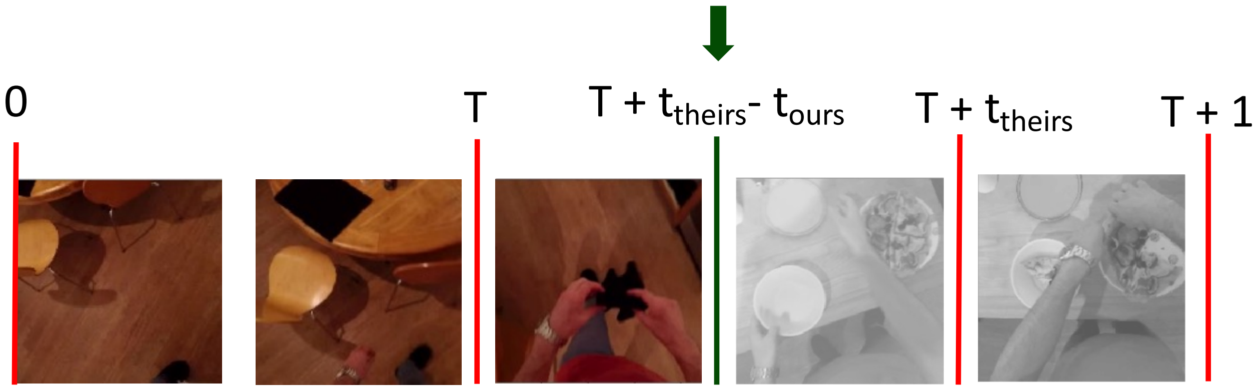
Results: Online Setting



With a latency of t_{theirs}, the prediction will arrive at time $T + t_{theirs}$

Model	Init	Latency (t_l ms)	Inference Start Time Stamp	Inference End Time Stamp	Target Time Stamp	Top-5 Recall Verb	Top-5 Recall Noun	Top-5 Recall Action

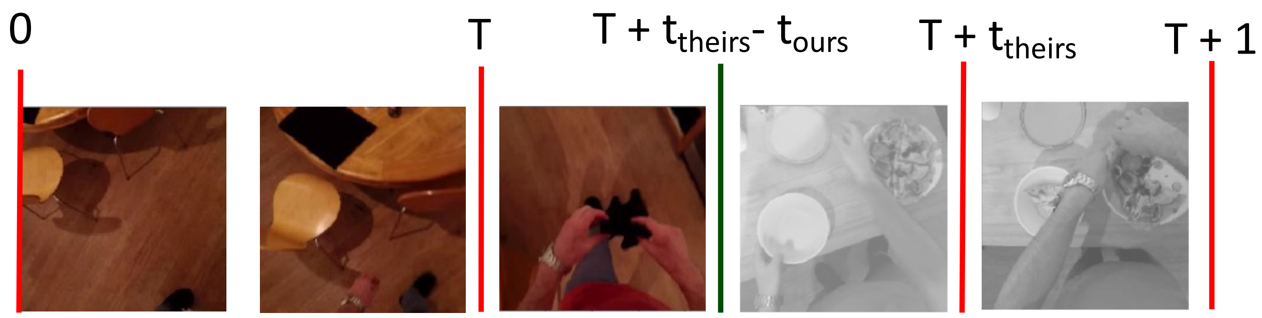
Results: Online Setting



When comparing two models, the prediction arrival times should be the same. With a latency of t_{ours}, RAFTFormer must start prediction at time $T + t_{theirs} - t_{ours}$ so that the prediction arrives at time $T + t_{theirs}$

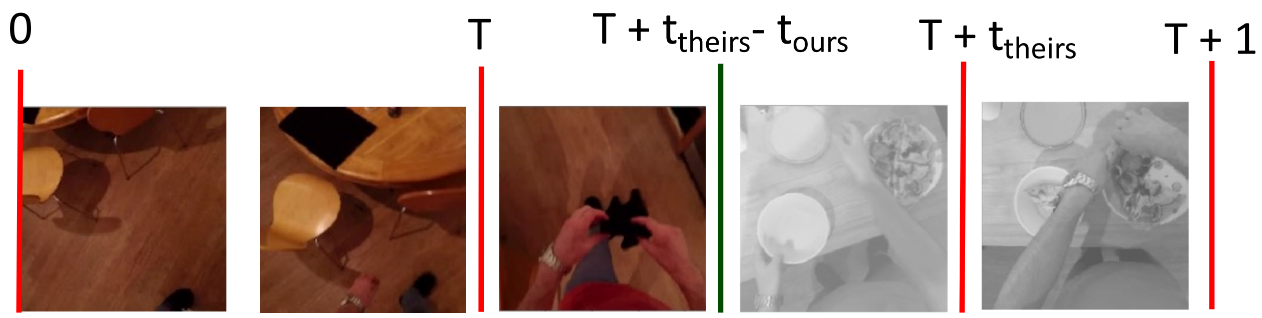
Model	Init	Latency (t_l ms)	Inference Start Time Stamp	Inference End Time Stamp	Target Time Stamp	Top-5 Recall		
						Verb	Noun	Action

Results: Online Setting



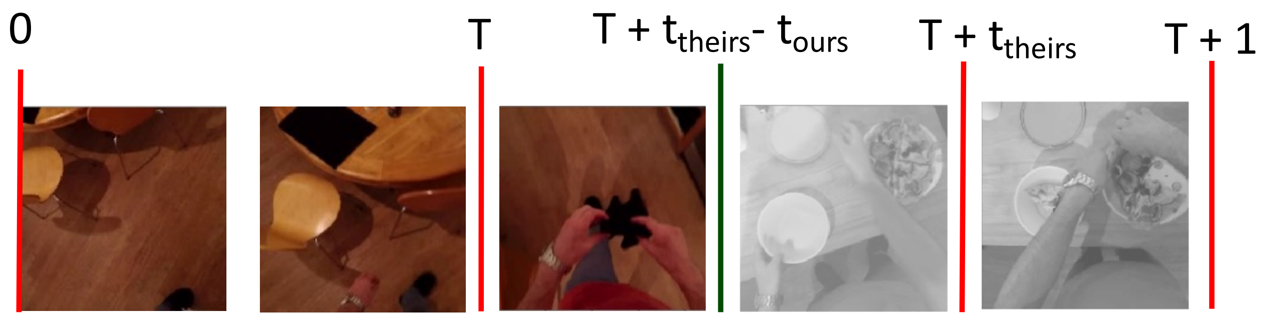
Model	Init	Latency (t_l ms)	Inference Start Time Stamp	Inference End Time Stamp	Target Time Stamp	Top-5 Recall Verb	Top-5 Recall Noun	Top-5 Recall Action
AVT [25]	IN21K	$t_{\text{avt}} = 420$	T	$T + t_{\text{avt}}$	$T + 1$	30.2	31.7	14.9
RAFTformer	K400 + IN1k	$t_{\text{ours}} = 40$	$T + t_{\text{avt}} - t_{\text{ours}}$	$T + t_{\text{avt}}$	$T + 1$	34.1	38.2	**19.3** (+4.4)

Results: Online Setting



Model	Init	Latency (t_l ms)	Inference Start Time Stamp	Inference End Time Stamp	Target Time Stamp	Top-5 Recall Verb	Top-5 Recall Noun	Top-5 Recall Action
AVT [25]	IN21K	$t_{\text{avt}} = 420$	T	$T + t_{\text{avt}}$	$T + 1$	30.2	31.7	14.9
RAFTformer	K400 + IN1k	$t_{\text{ours}} = 40$	$T + t_{\text{avt}} - t_{\text{ours}}$	$T + t_{\text{avt}}$	$T + 1$	34.1	38.2	**19.3** (+4.4)
MemViT [80]	K400	$t_{\text{vit}} = 160$	T	$T + t_{\text{vit}}$	$T + 1$	32.8	33.2	15.1
RAFTformer	K400 + IN1k	$t_{\text{ours}} = 40$	$T + t_{\text{vit}} - t_{\text{ours}}$	$T + t_{\text{vit}}$	$T + 1$	33.8	37.1	**18.1** (+3.0)

Results: Online Setting



Model	Init	Latency (t_l ms)	Inference Start Time Stamp	Inference End Time Stamp	Target Time Stamp	Top-5 Recall Verb	Top-5 Recall Noun	Top-5 Recall Action
AVT [25]	IN21K	$t_{\text{avt}} = 420$	T	$T + t_{\text{avt}}$	$T + 1$	30.2	31.7	14.9
RAFTformer	K400 + IN1k	$t_{\text{ours}} = 40$	$T + t_{\text{avt}} - t_{\text{ours}}$	$T + t_{\text{avt}}$	$T + 1$	34.1	38.2	**19.3** (+4.4)
MemViT [80]	K400	$t_{\text{vit}} = 160$	T	$T + t_{\text{vit}}$	$T + 1$	32.8	33.2	15.1
RAFTformer	K400 + IN1k	$t_{\text{ours}} = 40$	$T + t_{\text{vit}} - t_{\text{ours}}$	$T + t_{\text{vit}}$	$T + 1$	33.8	37.1	**18.1** (+3.0)
MemViT [80]	K700	$t_{\text{vit}} = 350$	T	$T + t_{\text{vit}}$	$T + 1$	32.2	37.0	17.7
RAFTformer	K400 + IN1k	$t_{\text{ours}} = 40$	$T + t_{\text{vit}} - t_{\text{ours}}$	$T + t_{\text{vit}}$	$T + 1$	33.7	37.9	**19.0** (+1.3)

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA

Latency Matters: Real-Time Action Forecasting Transformer

Thank You!

For full results on all datasets, paper, code and further details please visit the project homepage

SCAN ME

https://karttikeya.github.io/publication/RAFTformer/

Berkeley
UNIVERSITY OF CALIFORNIA