AntGPT: Can Large Language Models Help Long-term Action Anticipation from Videos? 

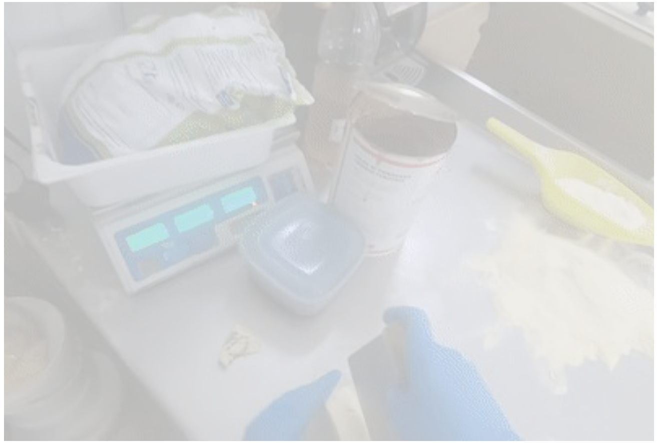
ICLR 2024

Qi Zhao (Kevin)*[1], Shijie Wang*[1], Ce Zhang[1], Changcheng Fu[1],
Minh Quan Do[1], Nakul Agarwal[2], Kwonjoon Lee[2], Chen Sun[1]

1: Brown University, 2: Honda Research Institute

Task Definition: Long-term Action Anticipation (LTA)

Observed Video

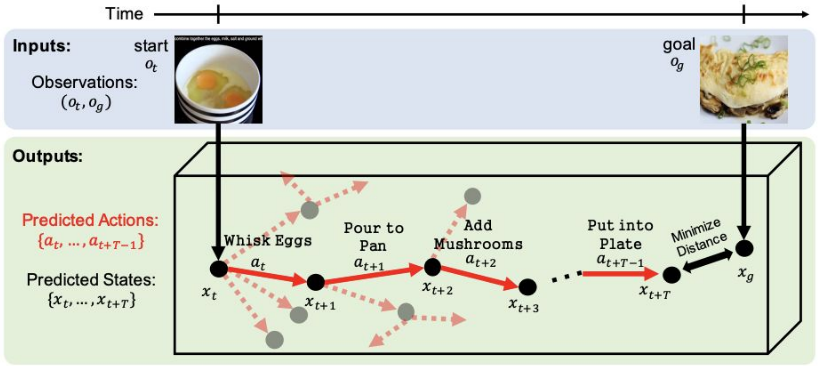


Future Actions

{cut cheese, …, put cheese}

- Given video observations, the LTA task aims to predict future actions of the person in long time spans.
- Different benchmarks has different task setup and metrics:
  - Ego4D LTA (order-specific): edit distance
  - EK-55/GAZE (order-agnostic): mean average precision

Question 1: Can we infer goals from LLMs that are helpful for LTA?



1. Can LLMs infer reasonable goals from observations?
2. Does inferring goals from LLM improve models' ability to predict future actions?

- Bottom-up LTA: Predict the next actions auto-regressively from previous actions
  E.g. predicting next action such as "mix eggs" from history actions "crack eggs".

- Top-down LTA: Infer the **goal** of the actor, then predict future actions to accomplish the **goal**.
  E.g. Predict next actions based on history actions and a long-term goal "making egg fried rice".

Question 2: Can LLMs help model temporal dynamics?

Vision

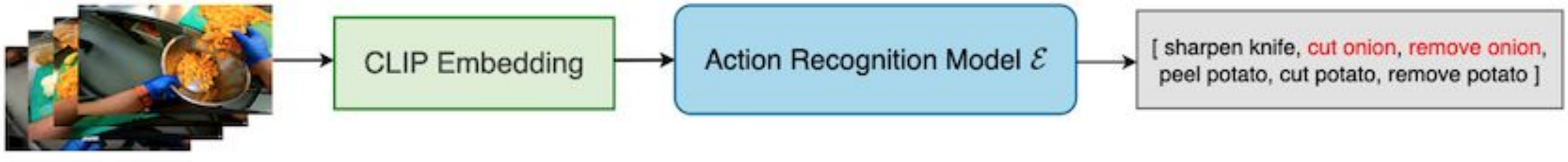
Text

Stretch dough *Sprinkle flour* *Roll rolling-pin*

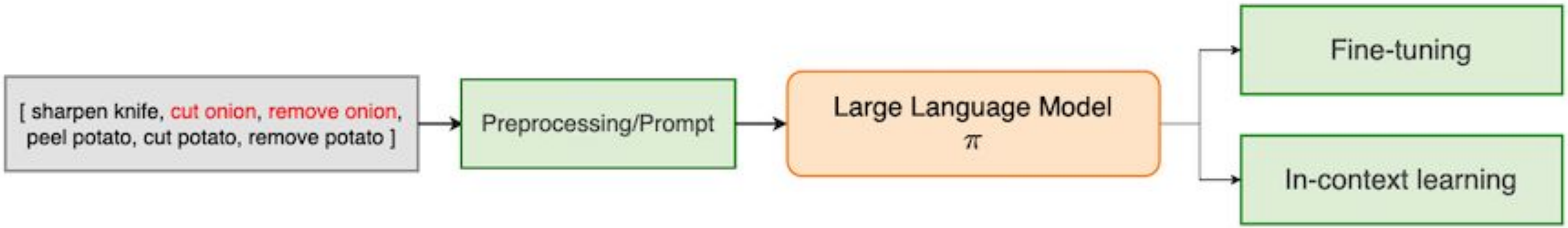
Pre-trained Large Language Model

LLMs demonstrated strong ability for sequence modeling and generation. How can we utilize it for LTA?

AntGPT: what is a good interface to interact with LLMs?



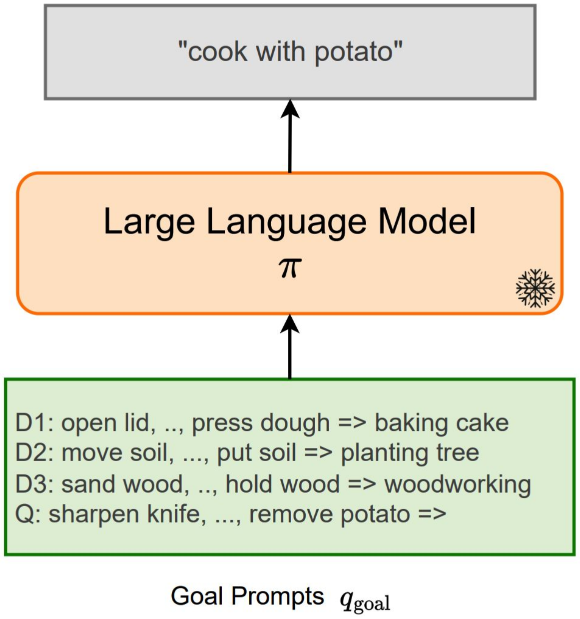
We use CLIP embeddings to train a transformer-based action recognition model to output action labels



We then preprocess the action labels into text tokens to perform fine-tuning or build prompt for ICL.

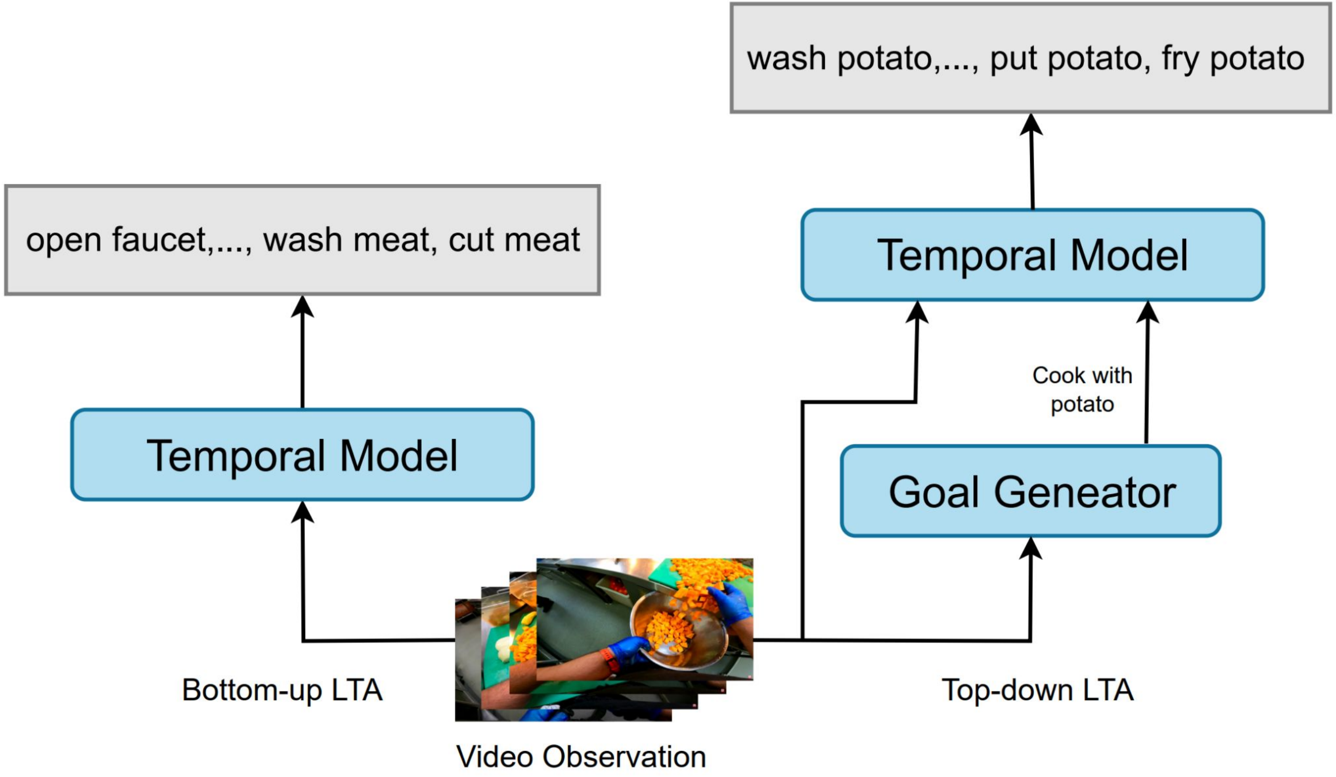
Can LLM generate goals from action observations?

- Bottom-up LTA: Predict the next actions solely from previous actions
- Top-down LTA: Infer the **goal** of the actor, then predict future actions to accomplish the **goal**.



(c) Few-shot Goal Generation with LLM

Do goals inferred by LLM benefit LTA?



(a) Overview of LTA Paradigms

Do goals reasoned by LLM benefit LTA?

Method	Ego4d v1 (ED)		EK-55 (mAP)			EGTEA (mAP)		
	Verb ↓	Noun ↓	ALL ↑	Freq ↑	Rare ↑	ALL ↑	Freq ↑	Rare ↑
image features	0.735	0.753	38.2	**59.3**	29.0	78.7	84.7	68.3
image features + Llama2 inferred goals	0.728	0.747	**40.1**	58.1	**32.1**	80.0	84.6	70.0
image features + GPT-3.5 inferred goals	**0.724**	**0.744**	**40.1**	58.8	31.9	**80.2**	**84.8**	**72.9**
image features + oracle goals*	-	-	40.9	58.7	32.9	81.6	86.8	69.3

Table 1: **Impact of goal conditioning on LTA performance.** Goal-conditioned (top-down) models outperforms the bottom-up model in all three datasets. We report edit distance for Ego4D, mAP for EK-55 and EGTEA. All results are reported on the validation set.

Inferred goals lead to **consistent improvements** for the top-down approach, especially for the rare actions of EK-55 and Gaze.

Do goals reasoned by LLM benefit LTA?

Method	EK-55			GAZE		
	ALL	FREQ	RARE	ALL	FREQ	RARE
I3D [9]	32.7	53.3	23.0	72.1	79.3	53.3
ActionVLAD [24]	29.8	53.5	18.6	73.3	79.0	58.6
Timeception [28]	35.6	55.9	26.1	74.1	79.7	59.7
VideoGraph [29]	22.5	49.4	14.0	67.7	77.1	47.2
EGO-TOPO [38]	38.0	56.9	29.2	73.5	80.7	54.7
Anticipatr [40]	39.1	58.1	29.1	76.8	83.3	55.1
AntGPT (ours)	**40.2**	**58.8**	**32.0**	**80.2**	**84.5**	**74.0**

Table 5: **Comparison with SOTA methods on the EK-55 and GAZE Dataset in mAP.** ALL, FREQ and RARE represent the highest performances on all, frequent, and rare target actions respectively.

Can LLMs Model Temporal Dynamics?

Vision Model
(CNN/Transformer…)

wash potato,...,put potato, fry potato

a) LTA with vision models: classification

“sharpen knife”,...,“cut potato”, “wash potato”

LLM

“wash potato”,...,“put potato”, “fry potato”

b) LTA as text sequence completion

Can LLMs Model Temporal Dynamics?

Model	Goal	Input	Verb ↓	Noun ↓
Transformer	GPT-3.5	image features	0.724	0.744
GPT-3-curie	GPT-3.5	recog actions	**0.709**	**0.729**
Transformer	Llama2-13B	image features	0.728	0.747
Llama2-7B	Llama2-13B	recog actions	**0.700**	**0.717**

Table 2: **Comparison of temporal models for top-down LTA.** Results on Ego4D v1 val set.

LLM largely outperforms from-scratched transformers with vision inputs, indicating its advantages on temporal modeling and the effectiveness of using actions as discrete representations.

LLMs benefit from Language Priors

Action Labels:
take photo, …, open door

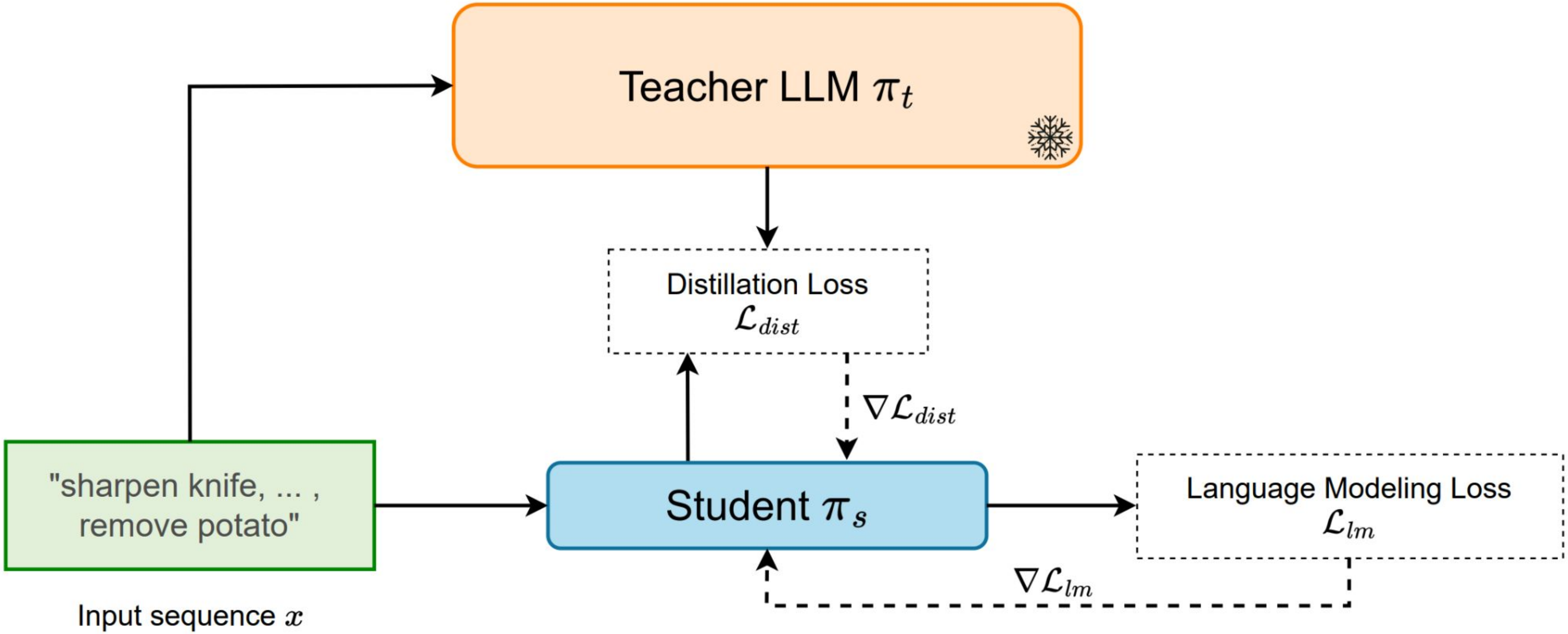
Shuffled Labels:
open potato, …, eat mask

Label Indices:
3 21, …, 15 7

Seq Type	Verb ↓	Noun ↓	Action ↓
Action Labels	**0.6794**	**0.6757**	**0.8912**
Shuffled Labels	0.6993	0.6972	0.9040
Label Indices	0.7249	0.6805	0.9070

Table 4: **Benefit of language prior.** Results on Ego4D v2 test set. We replace original action sequences to semantically nonsensical sequences.

Model Distillation



(d) Knowledge Distillation of LLM

Model Distillation

Model	Setting	Verb ↓	Noun ↓	Action ↓
7B	Pre-trained	0.6794	0.6757	0.8912
91M	From-scratch	0.7176	0.7191	0.9117
91M	Distilled	**0.6649**	**0.6752**	**0.8826**

Table 5: **LLM as temporal model.** Results on Ego4D v2 test set. Llama2-7B model is fine-tuned on Ego4D v2 training set. 91M models are randomly initialized.

Compare with SoTA models

Method	Version	Verb ↓	Noun ↓	Action ↓
HierVL [3]	v1	0.7239	0.7350	0.9276
ICVAE[35]	v1	0.7410	0.7396	0.9304
VCLIP [12]	v1	0.7389	0.7688	0.9412
Slowfast [23]	v1	0.7389	0.7800	0.9432
AntGPT (ours)	v1	**0.6584**±7.9e-3	**0.6546**±3.8e-3	**0.8814**±3.1e-3
Slowfast [23]	v2	0.7169	0.7359	0.9253
VideoLLM [10]	v2	0.721	0.725	0.921
PaMsEgoAI [29]	v2	0.6838	0.6785	0.8933
Palm [26]	v2	0.6956	0.6506	0.8856
AntGPT (ours)	v2	**0.6503**±3.6e-3	**0.6498**±3.4e-3	**0.8770**±1.2e-3

Table 6: **Comparison with SOTA methods on the Ego4D v1 and v2 test sets in ED@20.** Ego4d v1 and v2 share the same test set. V2 contains more training and validation examples than v1.

AntGPT: Can Large Language Models Help Long-term Action Anticipation from Videos?

ICLR 2024

Qi Zhao (Kevin)*[1], Shijie Wang*[1], Ce Zhang[1], Changcheng Fu[1], Minh Quan Do[1], Nakul Agarwal[2], Kwonjoon Lee[2], Chen Sun[1]

1: Brown University, 2: Honda Research Institute