Multi-level and Multi-modal Action Anticipation

Seulgi Kim, Ghazal Kaviani, Mohit Prabhushankar, Ghassan AlRegib
Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
{ seulgi.kim, gkaviani, mohit.p, alregib } @ gatech.edu

Project DARai Dataset

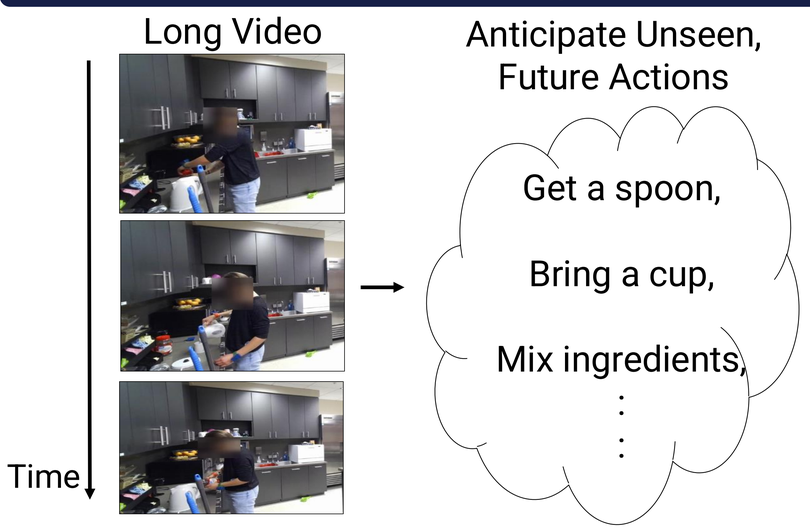
Contribution

✓ Multi-level (Hierarchical) & Multi-modal Modeling

✓ Fine-grained Action Text Generator

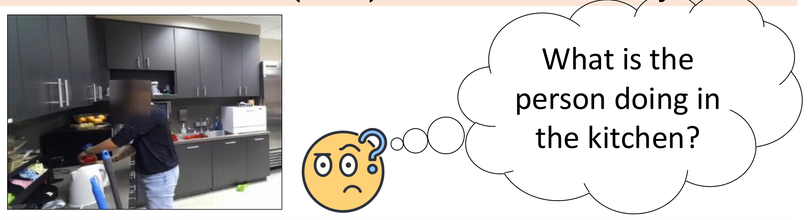
✓ Benchmark on our newly released *DARai Dataset*

What is Action Anticipation?

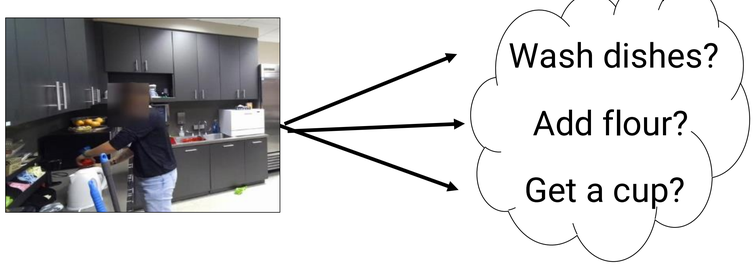


Challenge of Action Anticipation

1. Visual information (RGB) alone lacks clarity



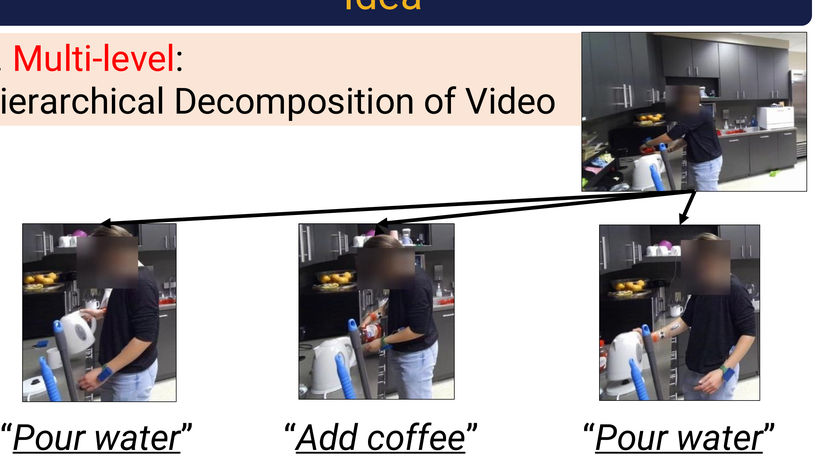
2. Ambiguity leads to multiple plausible futures



3. Ambiguity increases uncertainty of future actions

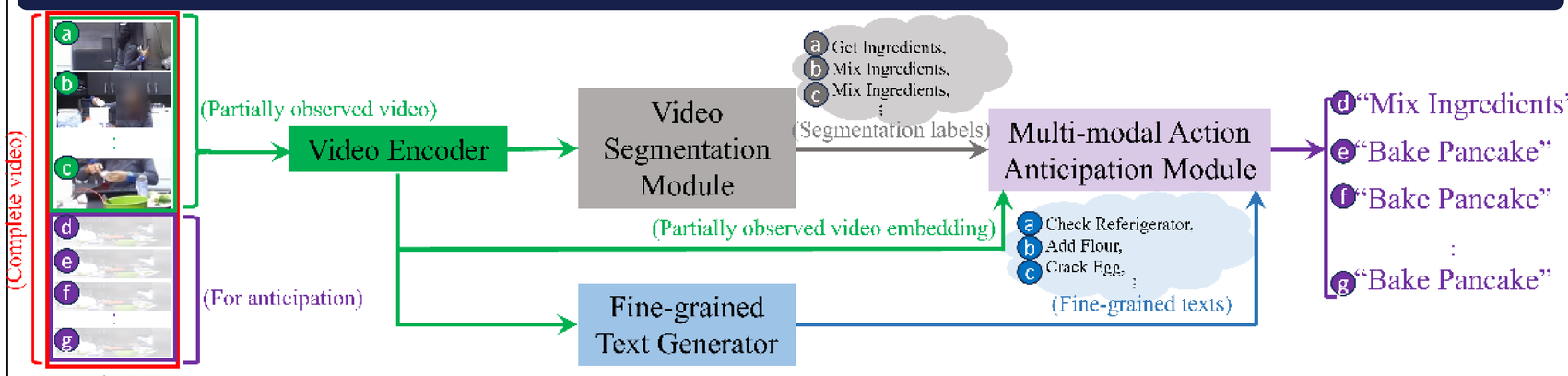
Idea

1. Multi-level:
Hierarchical Decomposition of Video



2. Multi-modal: Generate fine-grained texts

Architecture



1). Video Segmentation Module
- Transforms each video frame into action labels.

2). Fine-grained Text Generator
- Transforms each video frame into fine-grained texts.

3). Multi-modal Action Anticipation Module
- Aligns RGB with Text modality and anticipate next actions.

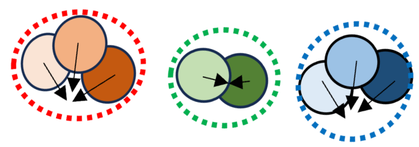
Fine-grained text generator

Hierarchical Decomposition of Video

x_1 t=1, Pour water
x_2 t=2, Pour water
x_3 t=3, Pour water
x_4 t=4, Add coffee
x_5 t=5, Add coffee
x_6 t=6, Pour water
x_7 t=7, Pour water
x_8 t=8, Pour water

$$Loss = L_{Cross\ Entropy} + \alpha L_{intra} + \beta L_{inter}$$

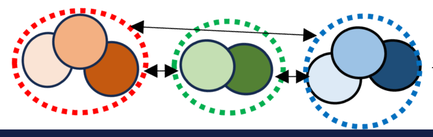
1. L_{intra} : Pull Frames within the same temporal interval



$$L_{intra} = \sum_{k} \sum_{x_i \in X(k)} ||x_i - \mu_k||^2 ,$$

$$\mu_k = \frac{1}{X(k)} \sum_{x_i \in X(k)} x_i$$

2. L_{inter} : Push Frames that are temporally distant



$$L_{inter} = \sum_{k,k' \neq k} \frac{1}{||\mu_k - \mu_{k'}||}$$

Results (Evaluation Metric: Mean of Classes)

Dataset	Methods	$\alpha = 0.2$				$\alpha = 0.3$			
		$\beta(0.1)$	$\beta(0.2)$	$\beta(0.3)$	$\beta(0.5)$	$\beta(0.1)$	$\beta(0.2)$	$\beta(0.3)$	$\beta(0.5)$
Breakfast	RNN	18.11	17.20	15.94	15.81	21.64	20.02	19.73	19.21
	CNN	17.90	16.35	15.37	14.54	22.44	20.12	19.69	18.76
	FUTR	47.06	47.08	47.11	47.11	65.55	65.55	65.54	65.51
	GTD	49.86	49.75	49.65	49.58	65.00	64.39	63.95	63.79
	Ours	**50.55**	**50.53**	**50.52**	**50.56**	**66.49**	**66.53**	**66.63**	**66.68**
50Salads	RNN	30.06	25.43	18.74	13.49	30.77	17.19	14.79	09.77
	CNN	21.24	19.03	15.98	09.87	29.14	20.14	17.46	10.86
	FUTR	54.83	54.33	52.45	**51.03**	100.00	**100.00**	98.62	**53.00**
	Ours	**69.61**	**69.27**	**69.73**	48.35	**100.00**	99.88	**98.70**	39.61

α	Methods	$\beta(0.1)$	$\beta(0.2)$	$\beta(0.3)$	$\beta(0.5)$
0.1	FUTR	24.26	24.25	24.74	23.46
	AFFT	20.25	23.13	23.63	23.42
	Ours	**26.53**	**25.82**	**25.98**	**24.71**
0.2	FUTR	25.05	25.11	**24.48**	**23.18**
	AFFT	23.14	24.78	23.62	21.02
	Ours	**25.75**	**25.70**	24.24	23.12
0.3	FUTR	40.71	33.57	33.42	30.79
	AFFT	33.82	29.25	28.33	25.45
	Ours	**42.00**	**34.71**	**34.49**	**31.34**

▲ Comparison with the State-of-the-arts across 3 Different Human Activity Datasets

Dataset	Methods	Observation Rate (α)			
		0.1	0.2	0.3	0.4
Breakfast	w/o multi-level modeling	25.51	47.10	65.54	**68.24**
	w/ multi-level modeling	**30.05**	**50.54**	**66.59**	67.20
50Salads	w/o multi-level modeling	31.25	53.16	**87.91**	68.80
	w/ multi-level modeling	**39.33**	**64.24**	84.55	**79.82**
DARai	w/o multi-level modeling	24.18	24.45	34.63	34.16
	w/ multi-level modeling	**25.76**	**24.70**	**35.64**	**38.36**

▲ Hierarchical (Multi-level) Modeling

Dataset	Methods	Observation Rate (Input)			
		0.1	0.2	0.3	0.4
Breakfast	Uni-modal	26.86	47.21	62.43	66.64
	Multi-modal	**30.05**	**50.54**	**66.59**	**67.20**
50Salads	Uni-modal	29.76	53.76	**90.92**	73.92
	Multi-modal	**39.33**	**64.24**	84.55	**79.82**
DARai	Uni-modal	**25.87**	22.02	28.92	27.08
	Multi-modal	25.76	**24.70**	**35.64**	**38.36**

▲ Multi-modal Modeling

α: Observation Rate
β: Prediction Rate