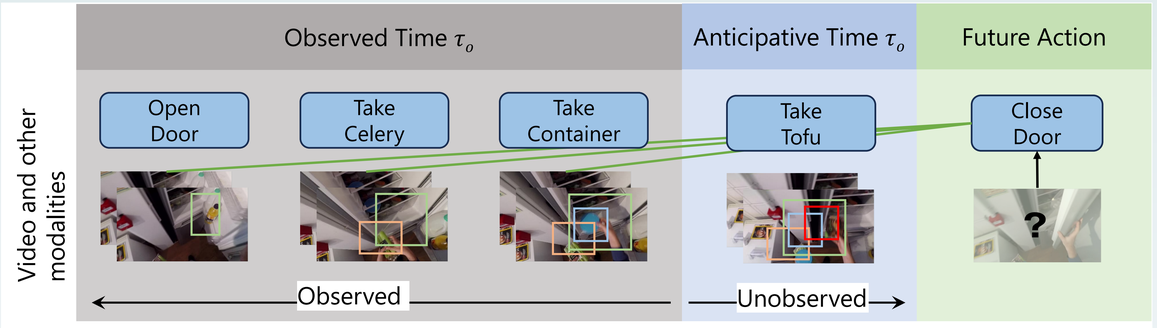
Text Descriptions of Actions and Objects Improve Action Anticipation

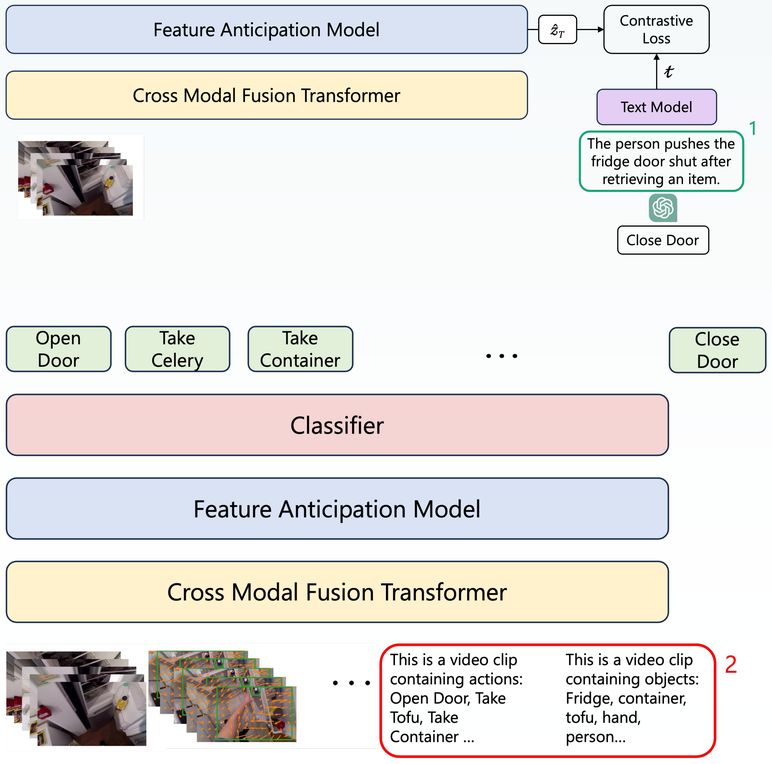
Apoorva Beedu, Harish Haresamudram, Irfan Essa
Georgia Institute of Technology

Motivation



- Task: predicting future actions based on video input and other modalities.
- No limit on observation period.
- Anticipation starts after a gap of τ_a seconds.
- Observed actions are not available during inference.

Pipeline



Pretrain Stage:

- With RGB as input, visual embedding is contrasted with text embeddings from GPT descriptions of actions

Finetune Stage:

- All modalities are introduced
- Observed actions and objects are introduced as inputs, in text form.
- Trained to predict actions

EGTEA

Table 1: **EGTEA Gaze+:** Model performance for Split=1 at $\tau_a = 0.5s$. $\rightarrow$ denotes the modalities used for pre-training and fine-tuning.

Model	Top-1			Class mean acc		
	Verb	Noun	Act.	Verb	Noun	Act.
AVT(TSN) [10]	51.7	50.3	39.8	**41.2**	41.4	28.3
AFFT* [34]	**52.1**	**50.7**	**41.4**	38.4	43.7	31.8
M-CAT (R $\rightarrow$ RF)	51.5	50.1	41.3	40.7	**45.9**	**33.5**

- Comparable performance to baseline.
- Effective contrastive learning requires large datasets and training in large batches.

EpicKitchens-55

Table 2: **EK55:** comparing performance of M-CAT against state-of-the-art methods on the val set of EK55.

Method	Verb		Noun		Action	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
RULSTM	32.4	79.6	23.5	51.8	15.3	35.3
ActionBanks	**35.8**	80.0	23.4	52.8	15.1	35.6
AVT	-	-	-	-	14.4	31.7
AVT+	32.5	79.9	24.4	54	16.6	37.6
AFFT	34.9	78.7	26.2	53.9	17.0	34.3
M-CAT (R $\rightarrow$ R)	32.4	80.1	28	56.4	16	36.5
M-CAT (ROFA $\rightarrow$ ROFA)	33	79.4	26	55.5	14.9	35.9
M-CAT (R $\rightarrow$ ROFA)	32.5	80.4	27.8	57	16.5	38.1
M-CAT* (R $\rightarrow$ ROFA+UV)	34.3	**80.6**	**29.7**	**58.8**	**17.9**	**39.8**

- Pretraining outperforms single modality training
- Action and object info. outperforms by ~5%.

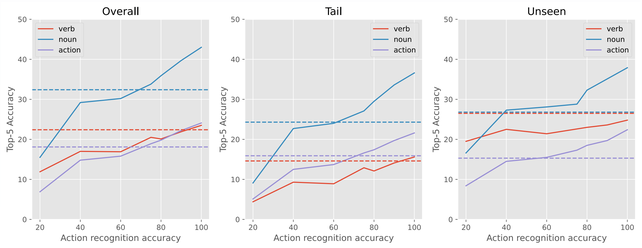
EpicKitchens-100

Table 4: **EK100**: comparison of M-CAT against state-of-the-art methods on the val set of EK100 using modalities provided by [8]. MeMViT uses only RGB data, while the rest use multiple modalities. $\rightarrow$ denotes the modalities used for pre-training and fine-tuning.

Method	Overall			Unseen			Tail		
	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
RULSTM	27.8	30.8	14.0	28.8	27.2	14.2	19.8	22.0	11.1
TempAgg	23.2	31.4	14.7	28	26.2	14.5	14.5	22.5	11.8
AVT	30.2	31.7	14.9	-	-	-	-	-	-
AVT+	28.2	32.0	15.9	29.5	23.9	11.9	21.2	25.8	14.1
MeMViT	**32.3**	37.0	17.7	28.6	27.4	15.2	25.3	31.0	15.5
AFFT(Swin+)	22.8	34.6	18.5	24.8	26.4	15.5	15.0	27.7	16.2
M-CAT (R $\rightarrow$ R)	30.1	32	16	32.7	28.4	15.3	23.4	25.3	13.8
M-CAT (R $\rightarrow$ ROFA)	31.9	35.9	17.3	32.5	30.2	14.5	**25.9**	30.3	15.4
M-CAT* (R $\rightarrow$ ROFA+UV)	31.3	**47.8**	**23.8**	**34.5**	**42.8**	**24**	23.8	**41.9**	**20.3**
InAViT	29.7	37.6	26.4	-	-	-	-	-	-
M-CAT (InAViT + UV)	**30.2**	**38.4**	**26.5**	-	-	-	-	-	-

- Using action and objects as input improves performance!

Impact of Action Recognition Accuracy



- Action recognition accuracy needs to be ~70% to outperform baselines.

Ablation Studies

(a) EK55 Ablations

Method	Action		
	Top-1	Top-5	Recall@5
M-CAT (w/ gpt)	15.6	**36.8**	16.1
M-CAT (w/o gpt)	14.8	36.7	16
M-CAT (w/o Aug)	14.8	36.3	14.8
M-CAT (w/ L_{v2v})	**16**	36.5	**17.5**

Removal of augmentations and GPT descriptions during pre-training reduce performance!

(b) EK100 Ablations

Method	Action		
	Overall	Unseen	Tail
M-CAT (ROFA$\rightarrow$ROFA)	15.8	16.6	13.3
M-CAT (R $\rightarrow$ R)	16	15.3	13.8
M-CAT (R $\rightarrow$ RV)	21.8	23.3	18.6
M-CAT (R $\rightarrow$ RAV)	22.4	21.9	19.2
M-CAT (R $\rightarrow$ ROFA+UV)	**23.8**	**24**	**20.3**

Additional modalities help improve performance!

Conclusion

- Action and object information is highly useful.
- Incorporating action/object via text descriptions is a viable solution, especially when training an LLM is infeasible.
- Accurate action recognition is important for action prediction.