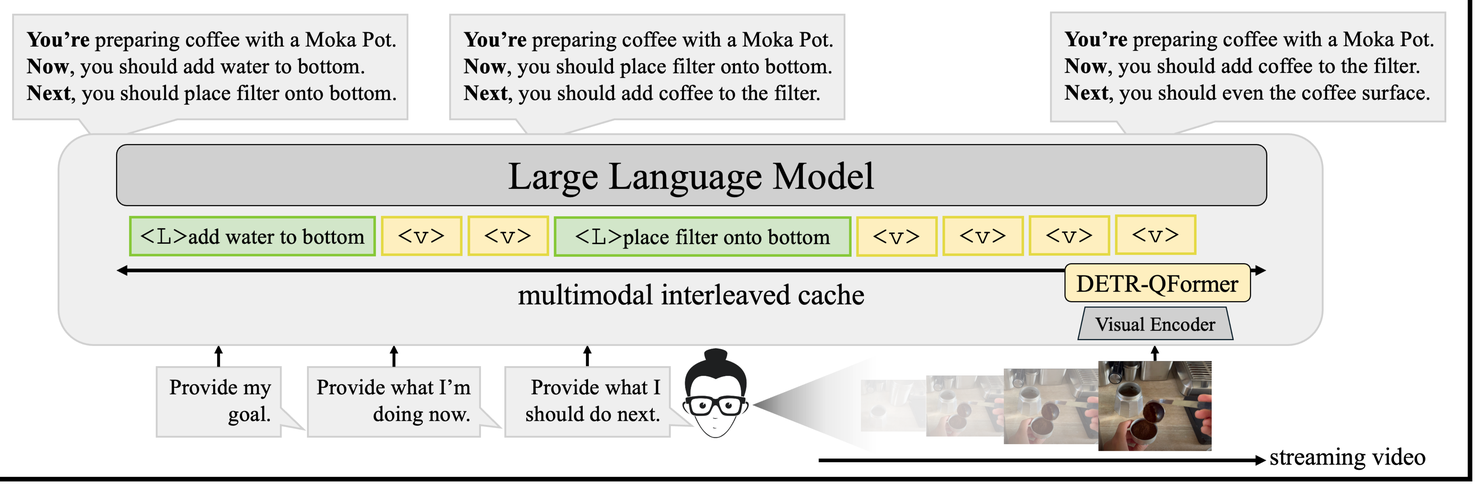
Streaming VideoLLMs for Real-Time Procedural Video Understanding

Dibyadip Chatterjee[1,2*], Edoardo Remelli[1], Yale Song[2], Bugra Tekin[1], Abhay Mittal[1], Bharat Bhatnagar[1], Necati Cihan Camgöz[1], Shreyas Hampali[1], Eric Sauser[1], Shugao Ma[1], Angela Yao[3], Fadime Sener[1]

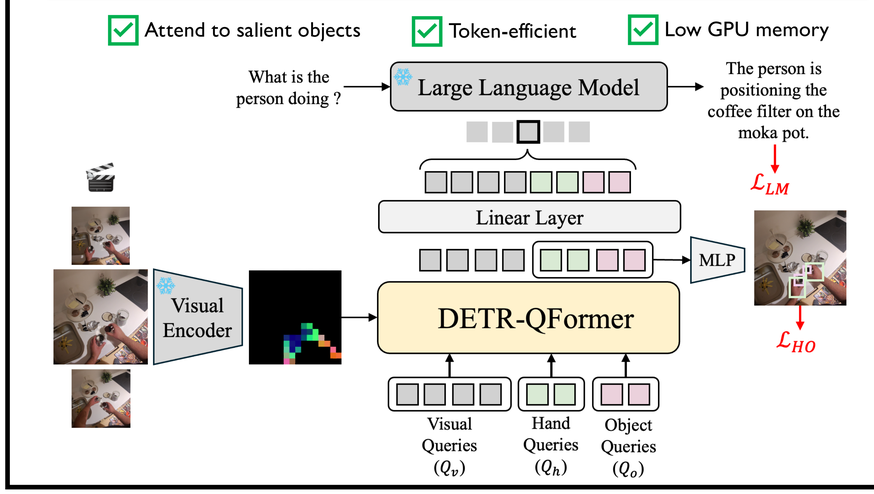
[1]National University of Singapore, [2]FAIR Meta, [3]Meta Reality Labs Research

[*]Work done during an internship at Meta

ProVideLLM: A state-of-the-art, real-time Procedural Video LLM

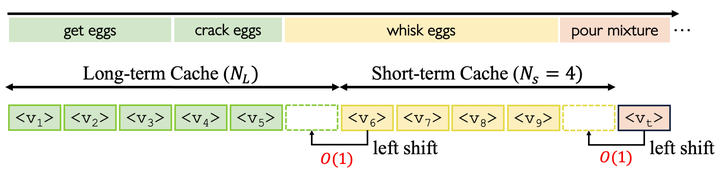


Tokenizing Short-Term Observations



Tokenizing Long-Term Observations

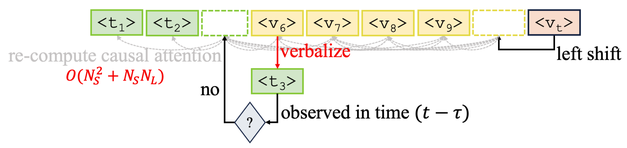
Long-Term Observations typically span a few-minutes to an hour



- Linear compute scaling
- Memory-intensive
- High Redundancy

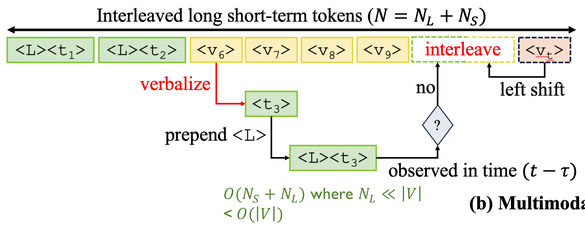
(a) Traditional streaming cache (visual-only)

Verbalize



- Low-memory
- Temporal Semantics
- Quadratic compute scaling

& Interleave



- Low-memory
- Temporal Semantics
- Sub-linear compute scaling

(b) Multimodal Interleaved Cache (vision-text)

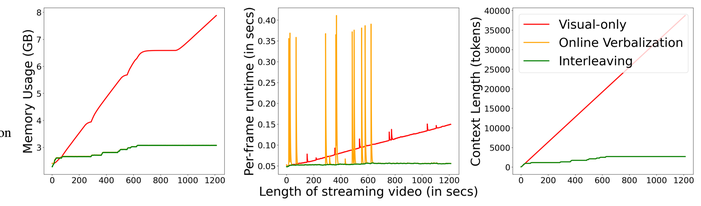
Ablations

DETR-QFormer on Ego4D

[CLS] Token	Patch Tokens	#Tokens /frame	Ego Val./ Acc. (%) SigLIP	Ego Val./ Acc. (%) DINOv2
w/ 2-layer MLP [45]				
✓	✗	1	31.8±0.4	28.7±0.2
✓	16×16 → 2×2	5	32.1±0.6	32.4±0.5
✓	16×16 → 4×4	17	33.3±0.6	36.6±0.5
w/ QFormer [39]				
✓	16×16 ⇒ 4	5	32.9±0.6	33.6±0.5
✓	16×16 ⇒ 16	17	33.2±0.4	38.1±0.7
w/ DETR-QFormer (2 hands, 2 objects)				
✓	16×16 ⇒ 0	5	30.9±0.3	34.7±0.5
✓	16×16 ⇒ 12	17	33.6±0.3	40.7±0.6

→ means avg-pooling
⇒ means Q-Former compression

Verbalize & Interleave on Ego4D-GoalStep



Only 3GB of GPU memory required to cache 20 mins of streaming video !!!

Results

Model	Training Views	Ego Acc. (%) Val / Test
TimeSFormer [9]	ego	35.13 / 35.93
EgoVLPv2 [54]	ego+exo	39.10 / 38.76
Viewpoint Distillation	ego+exo	38.19 / 39.49
View Invariant Encoder [50]	ego+exo	40.34 / 41.53
VideoLLM-MoD† [78]	ego	42.62 / —
ProVideLLM-8B/11	ego	**44.36 / 50.74**

State-of-the-art on EgoExo4D Keystep recognition (1st in Leaderboard)

Model	COIN Benchmark Top-1 Accuracy (%) Step	Task	Next	Proc.	Proc.+
VideoTF† [49]	56.5	91.0	42.4	40.2	46.4
VideoLLM-online [13]	63.1	92.7	49.1	49.8	54.1
VideoLLM-MoD [78]	63.4	92.8	49.7	49.8	53.3
ProVideLLM-8B/11+	**66.9**	**95.0**	**50.5**	**51.0**	**55.9**

State-of-the-art on COIN dataset across all tasks

Model	per-frame mAP Val	Test
LSTR [81] (only short-term)	8.8	-
LSTR [81]	8.9	8.1
EgoOnly [73]	10.3	10.9
ProVideLLM-1B/5 (no verbalization)	12.1	12.2
ProVideLLM-1B/5 (verbalization & interleaving)	**13.0**	**12.9**

State-of-the-art on Ego4D-Goalstep online step detection

Streaming Predictions	Model	FPS (↑) Vision Enc.	Connector + LLM	Full Model	Memory (GB) (↓)
Per-frame	ProVideLLM-1B/5	74.8	10.4	**9.1**	**2.0**
	ProVideLLM-8B/11	38.0	4.2	3.5	16.2
Dialogue [13]	ProVideLLM-1B/5	74.8	34.2	**24.6**	**2.2**
	ProVideLLM-8B/11	38.0	23.7	17.2	16.9

Real-time performance (>24fps) requiring ~2GB GPU memory

References

Y. Tang et al., COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, CVPR 2019
D. Shan et al., Understanding Human Hands in Contact at Internet Scale, CVPR 2020
M. Xu et al., Long Short-Term Transformer for Online Action Detection, NeurIPS 2021
C. Zhang et al., Helping Hands: An Object-Aware Ego-Centric Video Recognition Model, ICCV 2023
Y. Song et al., Ego4D Goal-Step: Toward Hierarchical Understanding of Procedural Activities, NeurIPS 2023 D&B
J. Chen et al., VideoLLM-online: Online Video Large Language Model for Streaming Video, CVPR 2024
K. Grauman et al., Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives, CVPR 2024

ICCV OCT 19-23, 2025 HONOLULU HAWAII