ViLCo

Video-Language Continual learning Benchmark

Tianqi Tang[1], *Shohreh Deldari*[1], *Hao Xue*[1], *Celso M de Melo*[2], *Flora D. Salim*[1]

1 School of Computer Science and Engineering, UNSW, Australia
2 Army Research Laboratory, USA

This work supported by the International Technology Center Indo-Pacific (ITC-IPAC)

Why Multimodal Continual Learning

- One Crucial challenge in multimodal learning is **continuous adaptation.**

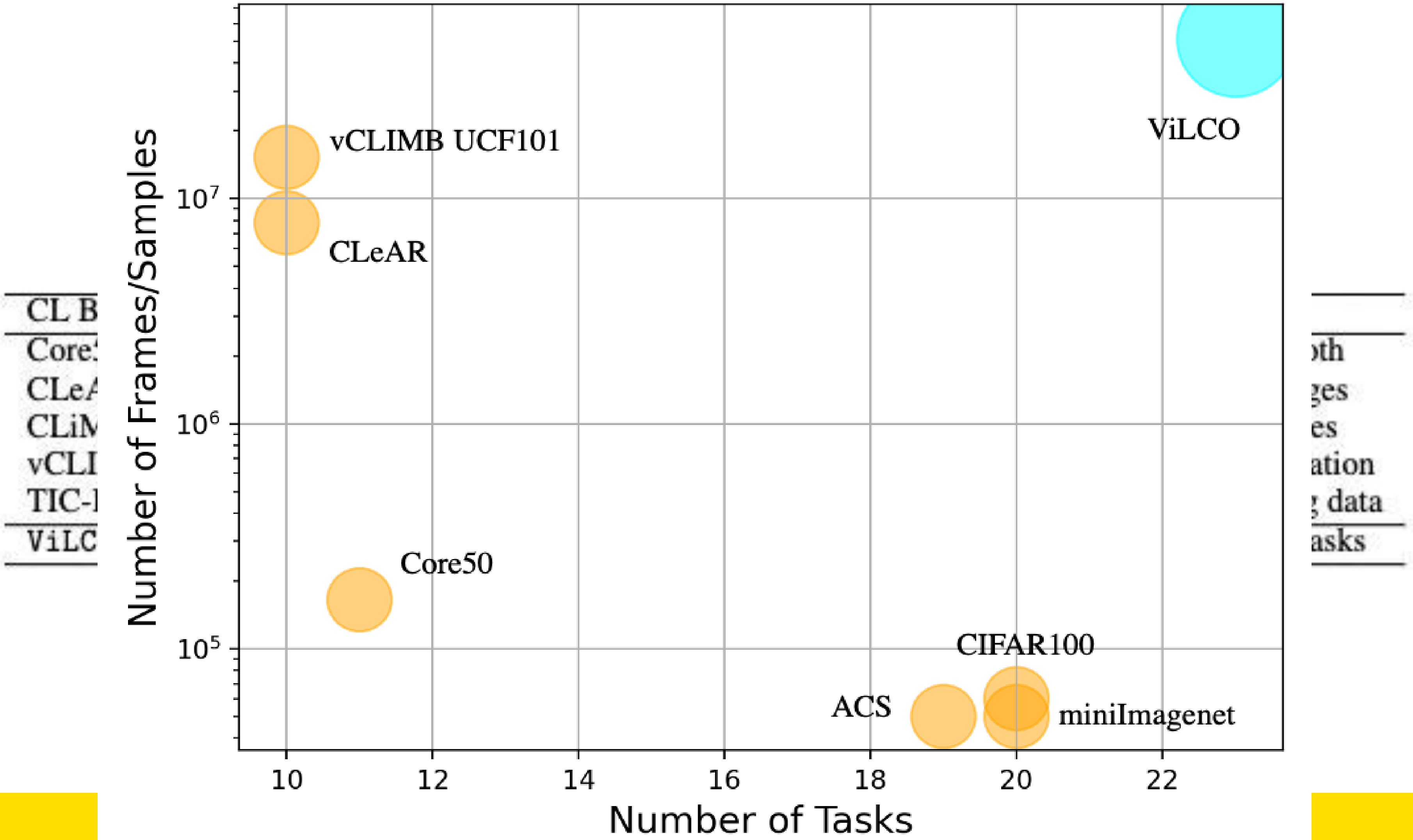
- There is always new data, new tasks, new query types...

- In multimodal scenarios, each mode of data can evolve together OR separately through
    - Emergence of new tasks
    - New data distribution
    - …

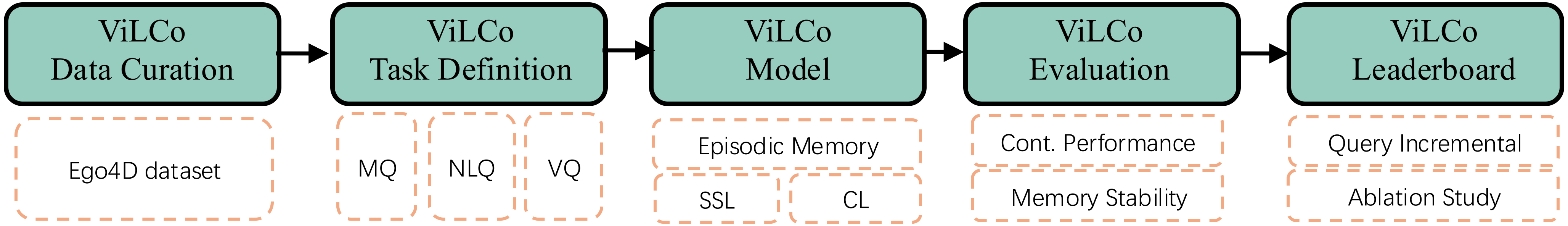
Q: When did I put the wood log?

A: from 45:12 to 45:51

Existing benchmark



ViLCo-Bench Pipeline



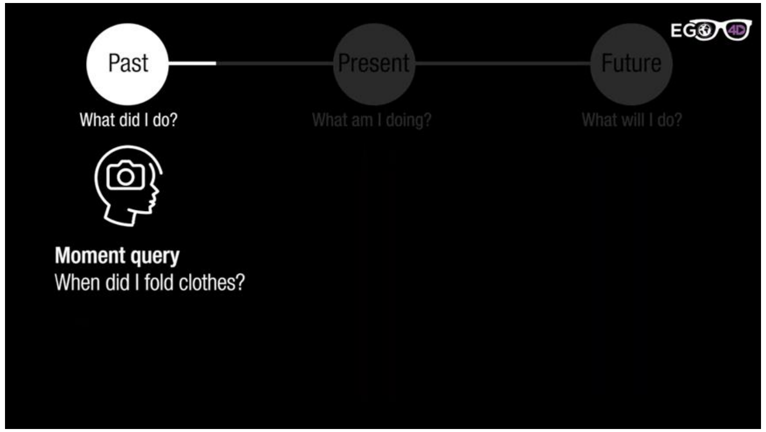
Data curation

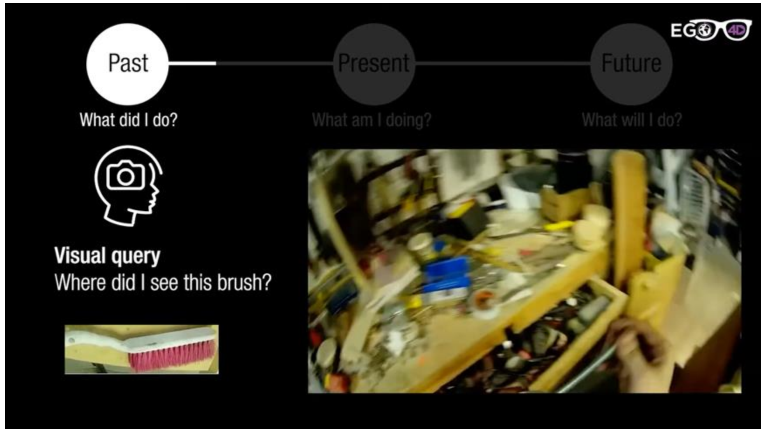
Curated Samples from Ego4D [1] dataset.

Moments Queries (MQ):
Inputs: video & names of activities;
Outputs: all temporal windows;
Includes a taxonomy of **110** activities

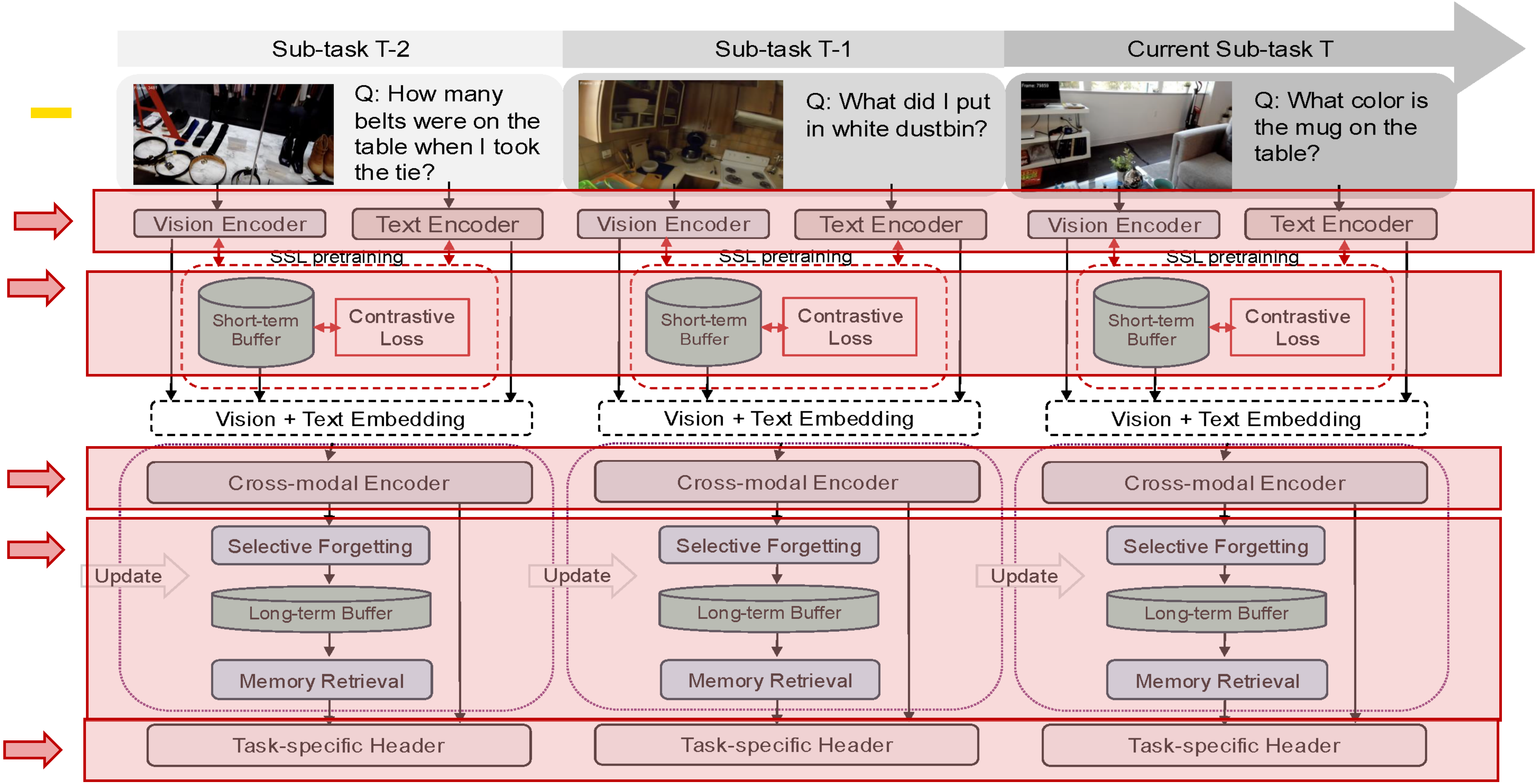
Natural Language Queries (NLQ):
Inputs: video & text query;
Outputs: temporal window where the answer is visible;
Includes **13** question template.

Visual Queries (VQ):
Inputs: video & image query;
Outputs: 2D bounding box
Includes **2000** classes.





[1] https://ego4d-data.org/


Sub-task T-2
Sub-task T-1
Current Sub-task T
Q: How many belts were on the table when I took the tie?
Q: What did I put in white dustbin?
Q: What color is the mug on the table?
Vision Encoder
Text Encoder
SSL pretraining
Short-term Buffer
Contrastive Loss
Vision + Text Embedding
Cross-modal Encoder
Selective Forgetting
Update
Long-term Buffer
Memory Retrieval
Task-specific Header

ViLCo

Leaderboard and Experiments

- Separate experiments for each type of task (MQ, NLQ, VQ)
- Impact of visual encoders
- Visual features
- Impact of each episodic memory and SSL module

Table 5: The impact of various visual features

Visual Backbone	BwF↓	Avg R@1 (%)↑ IoU=0.3	Avg R@1 (%)↑ IoU=0.5	Avg R@5 (%)↑ IoU=0.3	Avg R@5 (%)↑ IoU=0.5
Timersformer [4]	2.4	30.80	22.82	51.93	40.64
X3D [9]	1.4	31.50	23.01	48.09	36.59
ViViT [3]	1.2	40.0	35.82	56.05	47.40
EgoVLP-v2 [31]	2.9	33.58	26.24	53.75	42.30

Table 7: Comparing EM (episodic memory) and SSL (self-supervised learning) modules.

Method	BwF↓	Avg R@1 (%)↑ IoU=0.3	Avg R@1 (%)↑ IoU=0.5
Naive	18.8	22.74	17.58
ViLCo w/o EM	4.4	32.61	25.86
ViLCo w/o SSL	5.3	**33.70**	24.49
ViLCo	**2.9**	33.58	**26.24**

Table 6: The impact of various visual features. SF(Slowfast [10]) and OV(Omnivore [15])

Method	Vision Backbone	BwF↓	Avg R@1 (%)↑ IoU=0.3	Avg R@1 (%)↑ IoU=0.5	Avg R@1 (%)↑ mean	Avg R@5 (%)↑ IoU=0.3	Avg R@5 (%)↑ IoU=0.5	Avg R@5 (%)↑ mean
ViLCo	EgoVLP-v2	2.9	33.58	26.24	29.91	53.75	42.70	48.23
ViLCo	EgoVLP-v2 + InternVideo	2.8	42.73	33.53	38.13	62.97	50.50	56.74
ViLCo	EgoVLP-v2 + InternVideo + SF	4.3	38.33	29.75	34.04	56.95	46.69	51.82
ViLCo	EgoVLP-v2 + InternVideo + SF + OV	5.59	37.79	28.18	32.99	60.94	50.24	55.59

Thank you!

DEVCOM ARMY RESEARCH LABORATORY

UNSW SYDNEY

ViLCo-Bench

Paper

GitHub Repo

Flora.salim@unsw.edu.au