SVLTA: Benchmarking Vision-Language Temporal Alignment via Synthetic Video Situation

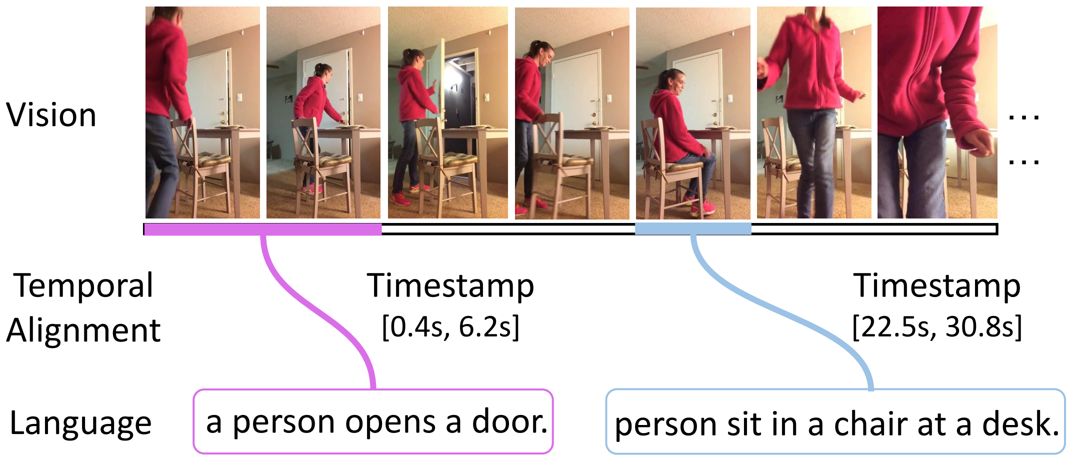
Hao Du*, Bo Wu*, Yan Lu, Zhendong Mao

Project : http://svlta.csail.mit.edu

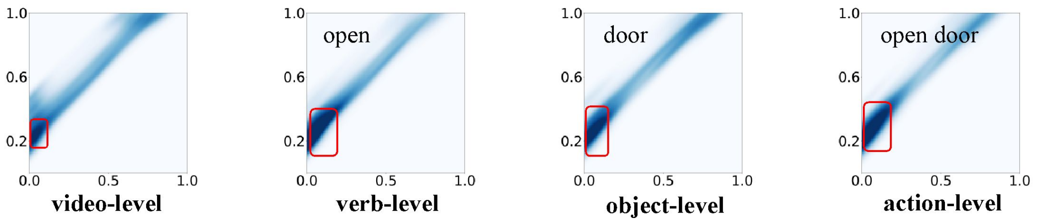
MIT-IBM Watson AI Lab

Project QR

Problem Formulation



Limitations of Existing Benchmarks



Limitations

- Biased temporal distributions
- Imprecise annotations
- Insufficient compositionally

Reasons

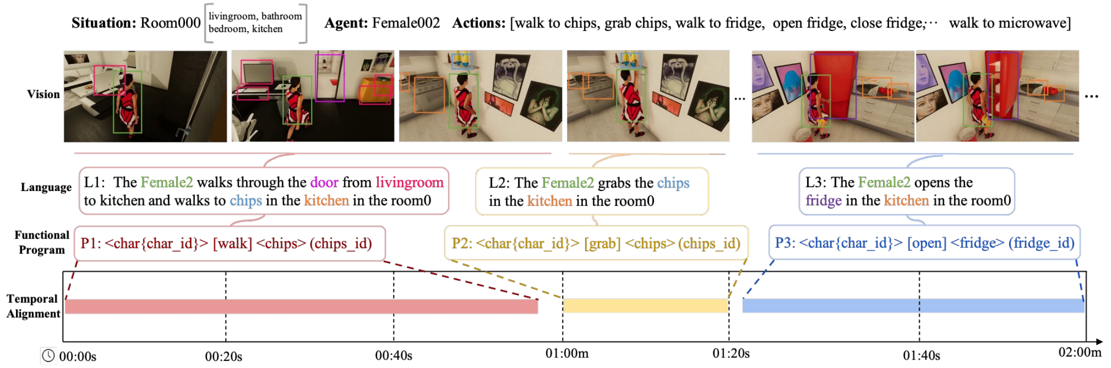
- Inherent properties of natural videos
- Human Annotation Challenges

TJSD: a diagnostic tool to analyze and quantify video temporal imbalance.

Question

Can we design a temporally fair benchmark to evaluate MLLMs?

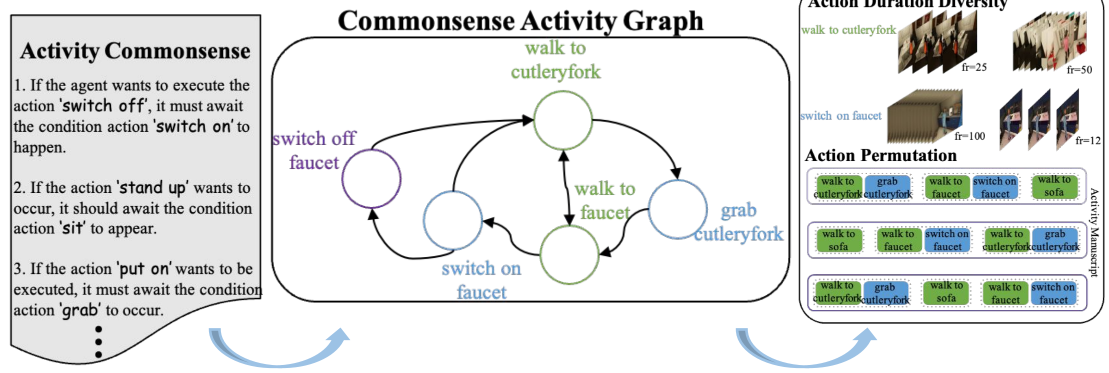
SVLTA Overview


Situation: Room000
livingroom, bathroom
bedroom, kitchen
Agent: Female002
Actions: [walk to chips, grab chips, walk to fridge, open fridge, close fridge,··· walk to microwave]
Vision
...
...
Language
L1: The Female2 walks through the door from livingroom to kitchen and walks to chips in the kitchen in the room0
L2: The Female2 grabs the chips in the kitchen in the room0
L3: The Female2 opens the fridge in the kitchen in the room0
Functional Program
P1: <char{char_id}> [walk] <chips> (chips_id)
P2: <char{char_id}> [grab] <chips> (chips_id)
P3: <char{char_id}> [open] <fridge> (fridge_id)
Temporal Alignment
00:00s
00:20s
00:40s
01:00m
01:20s
01:40s
02:00m

SVLTA includes synthetic videos, language, and high-quality temporal alignment.

Benchmark	Dataset Statistics			Dataset Characteristics				
	# Videos / # Annotations	# Actions	Avg. Video / Moment Duration (s)	Scalable	Controllable	Synthetic	Compositional	Unbiased
TACoS	0.1K / 18.8K	60	287.1 / 27.9	✗	✗	✗	✗	✗
ActivityNet Captions	14.9K / 54.9K	N/A	117.6 / 37.1	✗	✗	✗	✗	✗
Charades-STA	6.7K / 16.1K	157	30.0 / 8.1	✗	✗	✗	✗	✗
DiDeMo	10.5K / 40.5K	N/A	30.0 / 6.5	✗	✗	✗	✗	✗
TVR	21.8K / 109K	N/A	76.1 / 9.1	✗	✗	✗	✗	✗
MAD	0.7K / 384.6K	N/A	6646.2 / 4.1	✓	✗	✗	✗	✗
Ego4D	1.6K / 19.2K	N/A	495.3 / 11.2	✗	✗	✗	✗	✗
Ego4D Goal-Step	0.8K / 48K	N/A	1560.0 / 32.5	✗	✗	✗	✗	✗
E.T.Bench	7K / 7.3K	N/A	129.0 / 11.0	✗	✗	✗	✗	✗
SVLTA (ours)	25.3K / 77.1K	96	134.1 / 24.3	✓	✓	✓	✓	✓

SVLTA Generation



Activity Commonsense

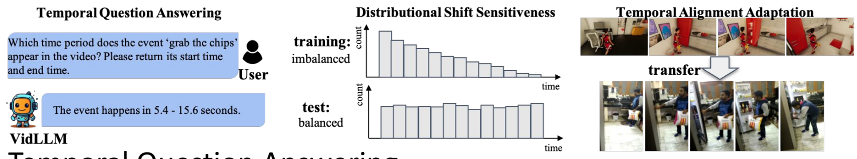
1. If the agent wants to execute the action 'switch off', it must await the condition action 'switch on' to happen.

2. If the action 'stand up' wants to occur, it should await the condition action 'sit' to appear.

3. If the action 'put on' wants to be executed, it must await the condition action 'grab' to occur.

⋮

Holistic Temporal Alignment Evaluations


Temporal Question Answering
User
VidLLM
Distributional Shift Sensitiveness
training: imbalanced
test: balanced
count
time
Temporal Alignment Adaptation
transfer

Which time period does the event 'grab the chips' appear in the video? Please return its start time and end time.

The event happens in 5.4 - 15.6 seconds.

Temporal Question Answering

Method	# Frames	Size	Visual Encoder	LLM	R@1				mIoU
					IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.7	
General Open-sourced Models: All models use their default setting. Except LLaVA-Video, due to the GPU memory limits.									
LLaVA-Video	16	7B	SIGLIP-SO400M	Qwen2	2.52	0.89	0.40	0.27	0.84
Videochat2	16	7B	UMT-L/16	Vicuna-0	2.93	0.87	0.32	0.13	0.87
Video-LLaVA	8	7B	LanguageBind-ViT-L/14	Vicuna-1.5	8.22	3.19	0.96	0.23	2.59
Video-ChatGPT	100	7B	CLIP-ViT-L/14	Vicuna-1.1	10.68	3.17	0.90	0.21	2.94
Video-LLaMA2	16	7B	CLIP-ViT-L/14	Mistral-7B	35.48	16.02	6.64	2.28	12.33
Time-aware Open-sourced Models: All models utilize their default configuration.									
E.T.Chat	1FPS	3.8B	EVA-ViT-G/14	Phi-3-Mini	17.86	8.07	3.48	1.36	6.29
TimeChat	96	7B	EVA-ViT-G/14	Llama-2	23.29	13.58	6.96	3.25	9.61
VTimeLLM	100	7B	CLIP-ViT-L/14	Vicuna-1.5	29.97	13.29	5.26	1.71	10.29
Close-sourced Models: Evaluated on a subset with 2000 samples.									
GPT-4o-mini	32	—	—	—	24.79	6.49	1.57	0.42	6.70
Gemini 1.5 Pro	1FPS	—	—	—	32.30	17.45	7.45	3.15	12.48
GPT-4o	32	—	—	—	49.54	27.38	11.69	5.62	18.90

Distributional Shift Sensitiveness

	Method	Test set	R@1				mIoU	RC ↓
			IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.9		
Biased Models	2D-TAN	high bias	93.82	87.08	72.55	35.06	76.41	10.85
		low bias	84.40$^{(-9.42)}$	76.10$^{(-10.98)}$	60.75$^{(-11.8)}$	22.75$^{(-12.31)}$	66.66$^{(-9.75)}$	
	VSLNet	high bias	98.14	97.03	95.26	83.40	92.63	14.31
		low bias	85.59$^{(-12.55)}$	83.22$^{(-13.81)}$	79.60$^{(-15.66)}$	67.34$^{(-16.06)}$	79.16$^{(-13.47)}$	
	LGI	high bias	97.02	94.26	87.38	56.36	85.25	14.94
		low bias	89.70$^{(-7.32)}$	82.98$^{(-11.28)}$	68.74$^{(-18.64)}$	31.49$^{(-24.87)}$	72.67$^{(-12.58)}$	
	QD-DETR	high bias	98.96	98.35	96.46	82.61	93.05	5.92
		low bias	95.59$^{(-3.37)}$	93.93$^{(-4.42)}$	90.17$^{(-6.29)}$	72.43$^{(-10.18)}$	87.72$^{(-5.33)}$	
Debiased Models	DCM	high bias	92.89	85.72	69.75	32.29	74.85	17.86
		low bias	79.55$^{(-13.34)}$	68.11$^{(-17.61)}$	46.15$^{(-23.6)}$	13.49$^{(-18.8)}$	58.88$^{(-15.97)}$	
	Shuffling	high bias	93.78	89.43	82.25	49.63	81.62	1.04
		low bias	93.26$^{(-0.52)}$	88.61$^{(-0.82)}$	80.23$^{(-2.02)}$	49.04$^{(-0.59)}$	80.36$^{(-1.26)}$	

Temporal Alignment Adaptation

Method	R@1			mIoU
	IoU=0.3	IoU=0.5	IoU=0.7	
2D-TAN	15.81	5.03	1.94	11.8
VSLNet	28.33	8.52	3.87	19.66
LGI	33.96	12.52	3.30	22.24
QD-DETR	33.74	18.39	7.55	22.32

Temporal Bias Comparison

Benchmark	Process	Entity		Composition
		Verb	Object	
TACoS	0.243	0.786	0.787	0.899
ActivityNet Captions	0.107	0.764	0.827	0.921
Charades-STA	0.287	0.739	0.877	0.881
TVR	0.229	0.779	0.84	0.914
MAD	0.628	0.842	0.869	0.926
SVLTA (ours)	**0.073**	**0.266**	**0.101**	**0.322**

References

[1] Puig, Xavier, et al. "Virtualhome: Simulating household activities via programs." CVPR. 2018.

[2] Otani, Mayu, et al. "Uncovering Hidden Challenges in Query-Based Video Moment Retrieval.". BMVC. 2020.