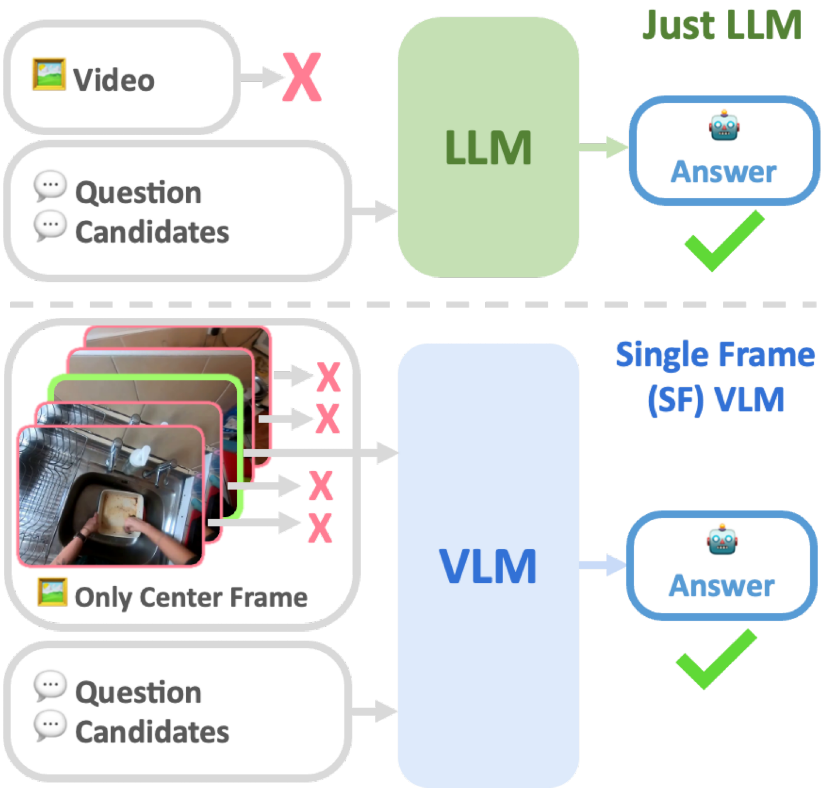
Stony Brook University

Understanding Long Videos with Multimodal Language Models

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, Michael S. Ryoo

Video Understanding with LLMs

- LLMs contain extensive world knowledge & reasoning skills
- How does this effect video tasks?
- Strong performance on video benchmarks maybe misleading!

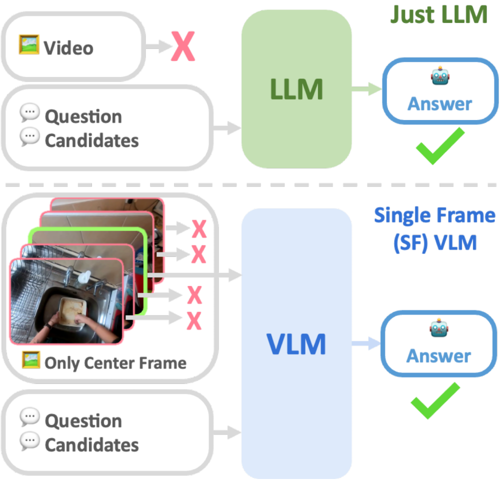


Video Understanding with LLMs

❖ LLMs solve some video QnA tasks significantly better than random with no video information! (similar findings in [2])

❖ Solve temporal tasks with single frame inputs…

Method	Param	Video Frames	ES-S		NextQA-T	
			Acc	Time	Acc	Time
Random	-	-	20.0	-	20.0	-
Just-LLM	7B	0	45.8	0.41	40.1	0.55
SF-VLM	13B	1	55.8	1.89	51.2	2.03
SOTA [1]	20B	180	50.8	381	54.3	207



[1] Zhang, Ce et al. "A Simple LLM Framework for Long-Range Video Question-Answering." EMNLP 2023.
[2] Min, Juhong et al. "MoreVQA: Exploring modular reasoning models for video question answering." CVPR 2024.

Why is this a problem?

- Strong performance on such benchmarks may not generalize
  - Possibly ignores important *video specific* information
  - Does not need these since LLM shortcuts with strong world knowledge & reasoning
- Spurious, unexpected performance on real world deployments

SOLUTION:

1) Explicit Motion Specific Information
2) Visual Grounding of Information
3) More Interpretable Framework

MVU: Multi-Modal Video Understanding Framework

Question x_t:

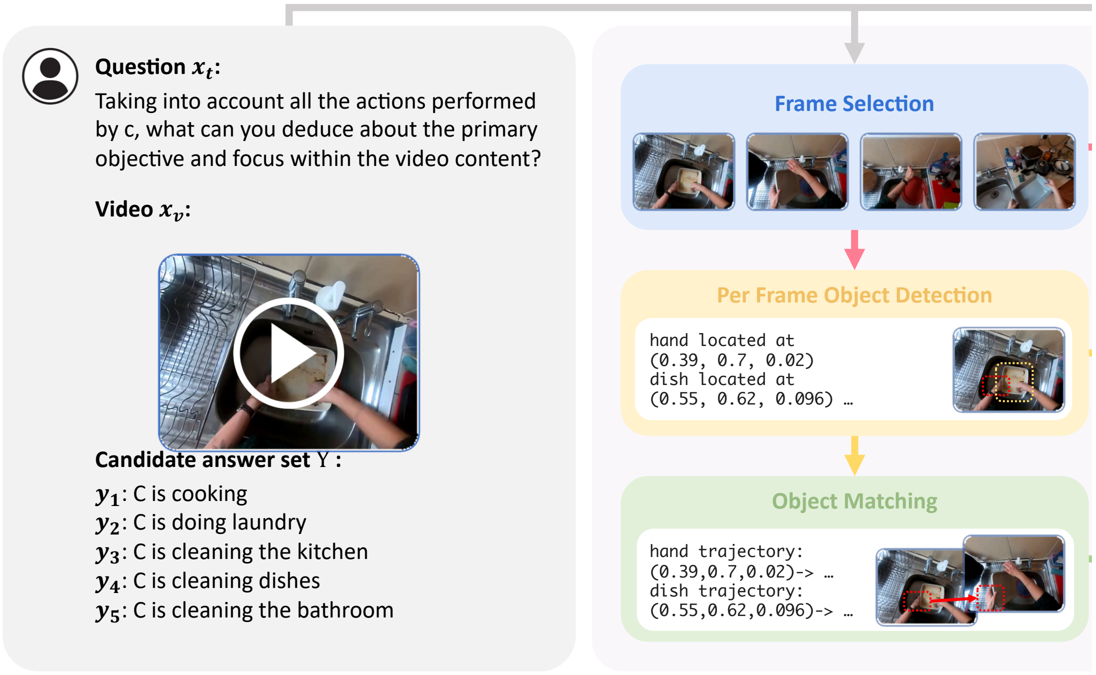
Taking into account all the actions performed by c, what can you deduce about the primary objective and focus within the video content?

Video x_v:

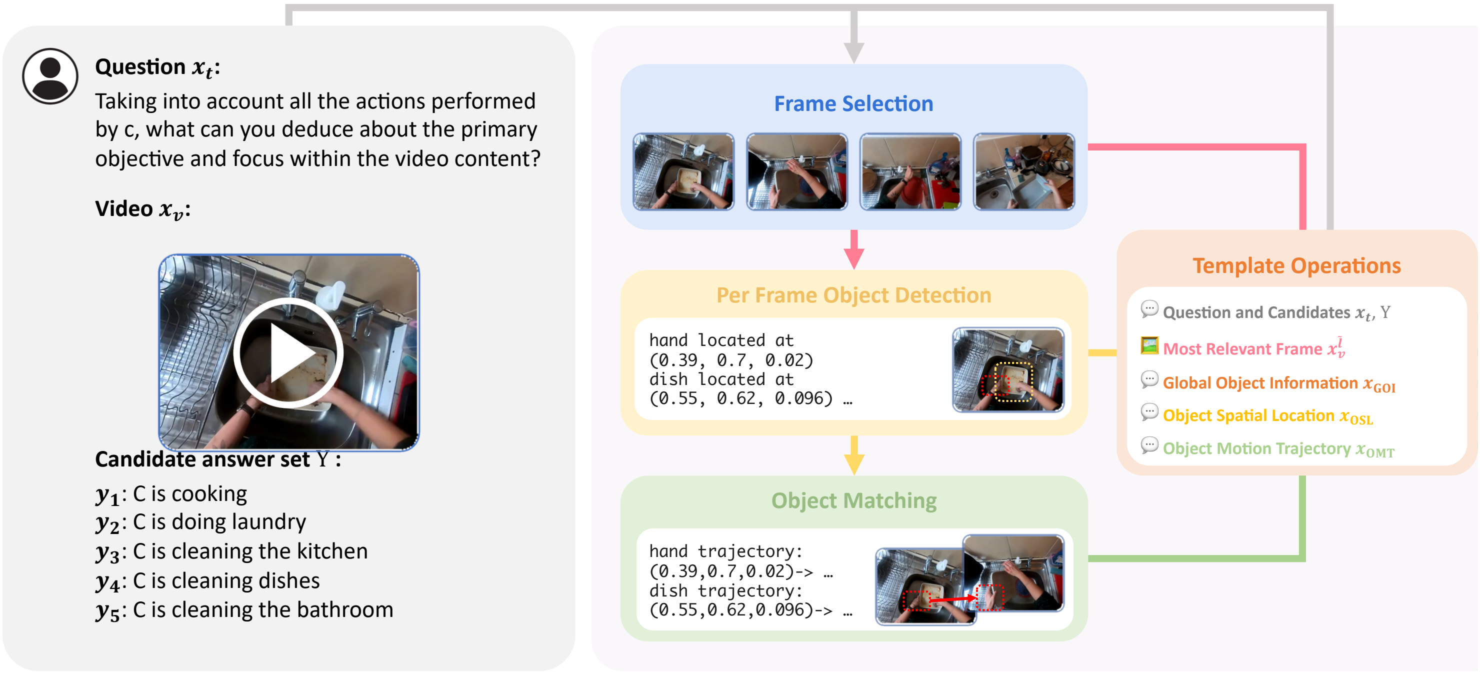
Candidate answer set Υ :

y_1: C is cooking
y_2: C is doing laundry
y_3: C is cleaning the kitchen
y_4: C is cleaning dishes
y_5: C is cleaning the bathroom

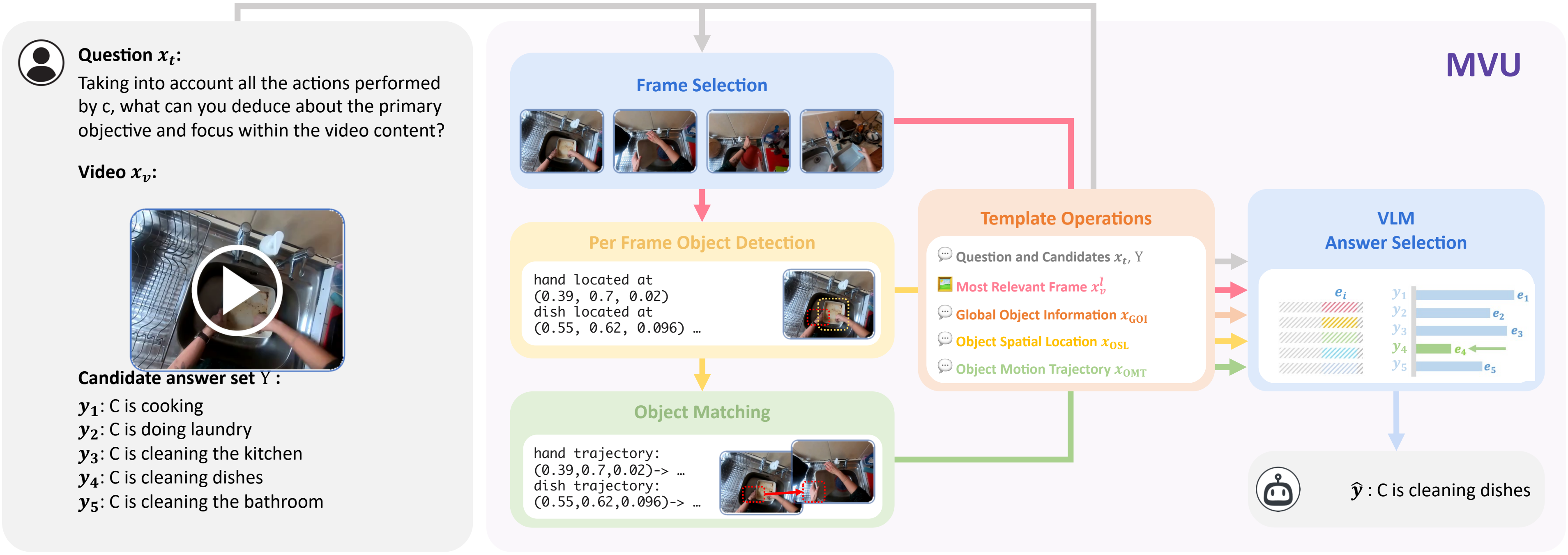
MVU: Multi-Modal Video Understanding Framework



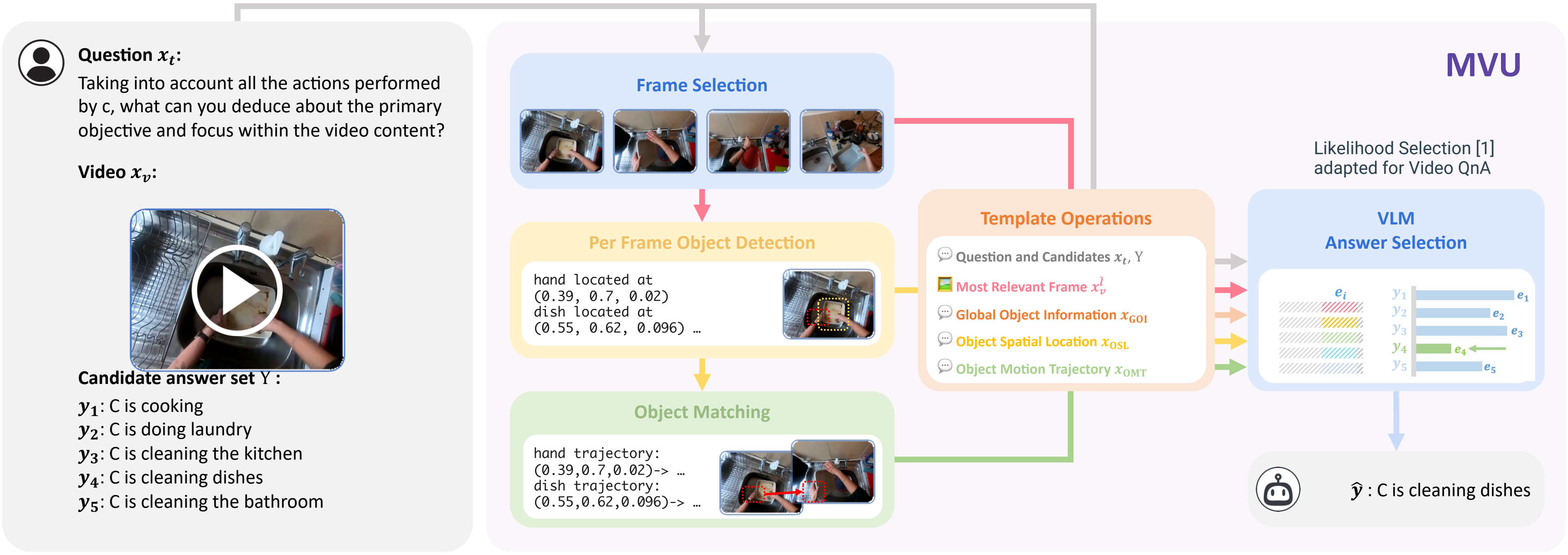
MVU: Multi-Modal Video Understanding Framework



MVU: Multi-Modal Video Understanding Framework

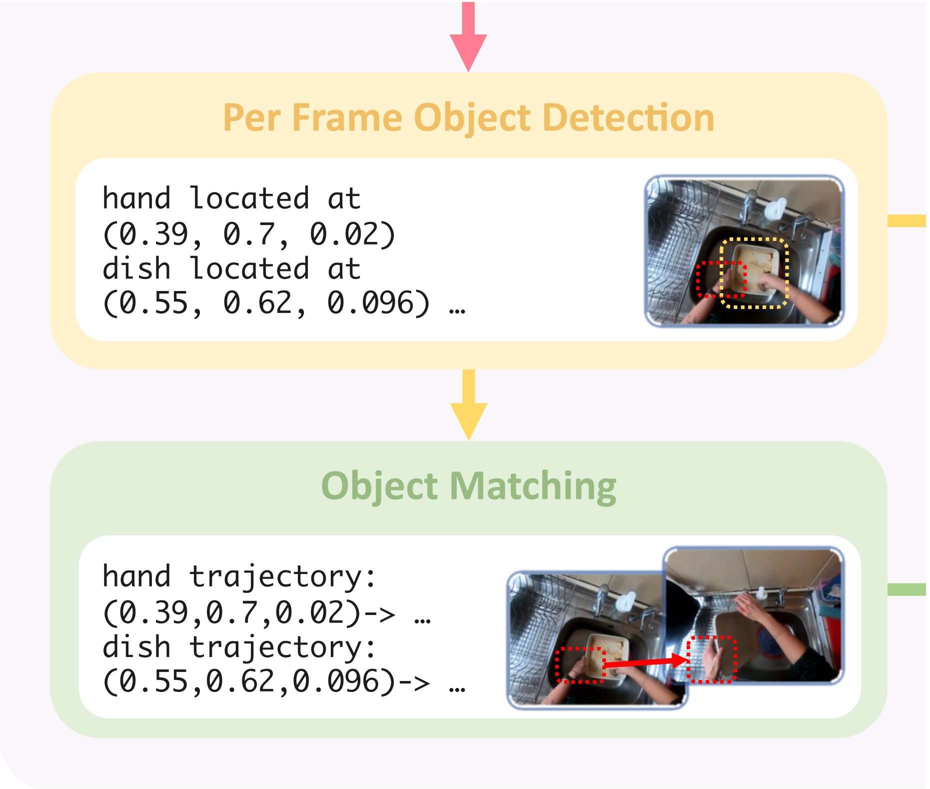


MVU: Multi-Modal Video Understanding Framework



[1] Robinson, Joshua et al. “Leveraging large language models for multiple choice question answering.” ICLR, 2023.

MVU: Multi-Modal Video Understanding Framework



- Unique objects across video
- Trajectories of each object

- Encode **object trajectories** explicitly in natural language

Object Trajectories in Natural Language

Object Trajectories in Natural Language

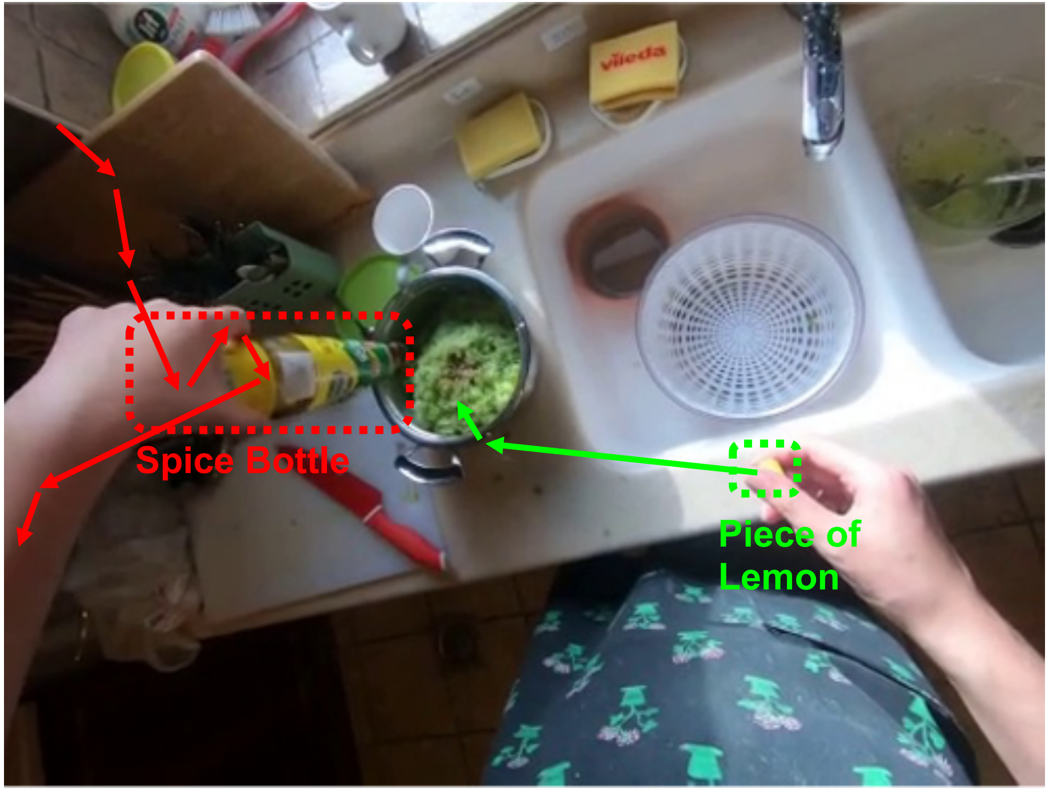
Represent motions as string of (x,y) sequence

"Spice Bottle Trajectory:
(0.13,0.20) -> (0.16,0.24)
->(0.17, 0.36)
.
.
.
->(0.02, 0.64)"

Object Trajectories in Natural Language

Represent motions as string of (x,y) sequence

"Spice Bottle Trajectory:
(0.13,0.20) -> (0.16,0.24)
->(0.17, 0.36)
.
.
.
->(0.02, 0.64)"



Represent motions as string of (x,y) sequence

"Piece of Lemon Trajectory:
(0.64,0.58) -> (0.48,0.52)
->(0.46, 0.49)"

Object Trajectories in Natural Language

Represent motions as string of (x,y) sequence

"Spice Bottle Trajectory:
(0.13,0.20) -> (0.16,0.24)
->(0.17, 0.36)
.
.
.
->(0.02, 0.64)"

Represent motions as string of (x,y) sequence

"Piece of Lemon Trajectory:
(0.64,0.58) -> (0.48,0.52)
->(0.46, 0.49)"

Represent motions as string of (x,y) sequence

"Cooking Pan Trajectory:
(0.46,0.47) -> (0.52,0.44)
->(0.47, 0.40)
->(0.45, 0.45)"

Object Trajectories in Natural Language

Sample Prompt

"Consider following objects moving along (x, y) trajectories in video to answer the question:" +
"Piece of Lemon Trajectory: (0.64,0.58) -> (0.48,0.52) ->(0.46, 0.49)" +
"Spice Bottle Trajectory: (0.13,0.20) -> 0.16,0.24) ->(0.17, 0.36) . . . ->(0.02, 0.64)" +
"Cooking Pan Trajectory: (0.46,0.47) -> (0.52,0.44) ->(0.47, 0.40) ->(0.45, 0.45)" +
". What does the person do after adding spice to the dish?"

Sample Response
"The person adds lime to the dish."

NOTE: Area / dimensions / frame index data omitted in example. The (x,y) can be replaced with (x,y,h,w,t).

[1] Zeng, Andy et al. "Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language." ICLR 2023.

Object Trajectories in Natural Language

"Piece of Lemon Trajectory: (0.64,0.58) -> (0.48,0.52) ->(0.46, 0.49)" +
"Spice Bottle Trajectory: (0.13,0.20) -> 0.16,0.24) ->(0.17, 0.36) . . . ->(0.02, 0.64)" +
"Cooking Pan Trajectory: (0.46,0.47) -> (0.52,0.44) ->(0.47, 0.40) ->(0.45, 0.45)" +

Obtaining Trajectories?

Use off-the-shelf object detector and tracker

- Much faster than generative VLM
- Can apply more densely on frames
- Spatial grounding gives interpretability
- Tracking reduces hallucinations

Evaluations

Video QnA: EgoSchema

Method	Zero Shot	Video Training	Closed Model	Params	Full
Random Selection	-	-	-	-	20.0
VIOLET (Fu et al., 2022)	✓	✓	✗	198M	19.9
FrozenBiLM (Yang et al., 2022)	✓	✓	✗	1.2B	26.9
SeViLA (Yu et al., 2024)	✓	✓	✗	4B	22.7
mPLUG-Owl (Ye et al., 2023b)	✓	✓	✗	7.2B	31.1
InternVideo (Wang et al., 2022)	✓	✓	✗	478M	32.1
ImageViT (Papalampidi et al., 2023)	✗	✓	✗	1B	30.9
SeViLA+ShortViViT (Papalampidi et al., 2023)	✗	✓	✗	5B	31.3
LongViViT (Papalampidi et al., 2023)	✗	✓	✗	1B	33.3
MC-ViT-L (Balavzevi'c et al., 2024)	✗	✓	✗	424M	44.4
InternVideo2 (Wang et al., 2024b)	✓	✓	✗	7B	55.8
Tarsier (Wang et al., 2024a)	✓	✓	✗	7B	49.9
Tarsier (Wang et al., 2024a)	✓	✓	✗	34B	61.7
Vamos (Wang et al., 2023a)	✓	✗	✗	13B	36.7
LLoVi (Zhang et al., 2023a)	✓	✗	✗	13B	33.5
LangRepo (Kahatapitiya et al., 2024)	✓	✗	✗	12B	41.2
Vamos (Wang et al., 2023a)	✓	✗	✓	1.8T	48.3
LLoVi (Zhang et al., 2023a)	✓	✗	✓	1.8T	50.3
LifelongMemory (Wang et al., 2023b)	✓	✗	✓	1.8T	62.4
MoreVQA (Min et al., 2024)	✓	✗	✓	-	51.7
VideoAgent (Wang et al., 2025)	✓	✗	✓	1.8T	54.1
VideoTree (Wang et al., 2024c)	✓	✗	✓	1.8T	61.1
LVNet (Park et al., 2024)	✓	✗	✓	1.8T	61.1
SF-VLM (ours)	✓	✗	✗	13B	36.4
SF-VLM + MVU (ours)	✓	✗	✗	13B	37.6
LVNet + MVU (ours)	✓	✗	✓	1.8T	61.3

- Strong results on EgoSchema dataset
- Easy integration with SOTA methods like LVNet [1]

[1] Park, Jongwoo et al. "Too Many Frames, not all Useful: Efficient Strategies for Long-Form Video QA." NeurIPS-W 2024.

Video QnA: NextQA

Method	ZS	VT	Params	Cau.	Tem.	Des.	All
Random Selection	-	-		20.0	20.0	20.0	20.0
CoVGT (Xiao et al., 2023)	✗	✓	149M	58.8	57.4	69.3	60.0
SeViT (Kim et al., 2023)	✗	✓	215M	-	-	-	60.6
HiTeA (Ye et al., 2023a)	✗	✓	297M	62.4	58.3	75.6	63.1
InternVideo (Wang et al., 2022)	✗	✓	478M	62.5	58.5	75.8	63.2
MC-ViT-L (Balavzevi'c et al., 2024)	✗	✓	424M	-	-	-	65.0
BLIP-2 (Li et al., 2023a)	✗	✓	4B	70.1	65.2	80.1	70.1
SeViLA (Yu et al., 2024)	✗	✓	4B	74.2	69.4	81.3	73.8
LLama-VQA-7B (Ko et al., 2023)	✗	✓	7B	72.7	69.2	75.8	72.0
Vamos (Wang et al., 2023a)	✗	✓	7B	72.6	69.6	78.0	72.5
Just-Ask (Yang et al., 2021)	✓	✓	66M	31.8	30.4	36.0	38.4
VFC (Momeni et al., 2023)	✓	✓	164M	45.4	51.6	64.1	51.5
InternVideo (Wang et al., 2022)	✓	✓	478M	43.4	48.0	65.1	49.1
SeViLA(Yu et al., 2024)	✓	✓	4B	61.3	61.5	75.6	63.6
CaKE-LM (Su et al., 2023)	✓	✗	2.7B	35.7	35.3	36.8	34.9
LLoVi (Zhang et al., 2023a)	✓	✗	13B	55.6	47.9	63.2	54.3
ViperGPT (Surís et al., 2023)	✓	✗	175B	-	-	-	60.0
LLoVi (Zhang et al., 2023a) (GPT-4)	✓	✗	1.8T	69.5	61.0	75.6	67.7
MoreVQA (Min et al., 2024)	✓	✗	1.7T	70.2	64.6	-	69.2
VideoAgent (Wang et al., 2025)	✓	✗	1.7T	72.7	64.5	81.1	71.3
VideoTree (Wang et al., 2024c)	✓	✗	1.7T	75.2	67.0	81.3	73.5
LVNet (Park et al., 2024)	✓	✗	1.8T	75.0	65.5	81.5	72.9
SF-VLM + MVU (ours)	✓	✗	13B	55.7	48.2	64.2	55.4
LVNet + MVU (ours)	✓	✗	1.8T	75.2	66.8	81.3	73.3

- Strong results on NextQA dataset
- Easy integration with SOTA methods like LVNet [1]

[1] Park, Jongwoo et al. “Too Many Frames, not all Useful: Efficient Strategies for Long-Form Video QA.” NeurIPS-W 2024.

More Evaluations

Method	Acc (%)
Phi-3-Vision-Instruct (Abdin et al., 2024)	49.7
Phi-3-Vision-Instruct + MVU	50.4

Longer Videos: MVU can improve performance on longer video benchmarks such as LongVideoBench.

Method	OMT	Accuracy
Random	-	0.6
CLIP (Radford et al., 2021)	-	4.0
MAXI (Lin et al., 2023b)	-	6.4
MVU (ours)	✗	3.6
MVU (ours)	✓	**7.2**

Motion Ablation on SSv2: We use the motion focused SSv2 dataset for a special ablation of our MVU's explicit motion information, establishing its clear usefulness in recognizing motion patterns.

Dataset	Obs.	Size	CC	Random	Baseline	MVU
ASU TableTop Manipulation	T	110	83	13.6	19.1	**20.9**
Berkeley MVP Data	F	480	6	20	26.0	**33.1**
Berkeley RPT Data	F	908	4	24.6	23.1	**26.2**
CMU Play Fusion	T	576	44	20.3	34.0	**35.6**
CMU Stretch	T	135	5	23	18.5	**24.4**
Furniture Bench	T	5100	9	20.2	24.8	**26.4**
Furniture Bench	F	5100	9	20.2	22.6	**24.9**
CMU Franka Pick-Insert Data	T	631	7	18.7	19.3	**21.2**
CMU Franka Pick-Insert Data	F	631	7	23.1	**57.8**	49.3
Imperial F Cam	T	170	17	20	22.9	**24.1**
Imperial F Cam	F	170	17	23.5	20.6	**24.7**
USC Jaco Play	T	1085	89	21.8	26.4	**30.6**
USC Jaco Play	F	1085	89	19.4	28.6	**32.4**
NYU ROT	T	14	12	21.4	**57.1**	**57.1**
Roboturk	T	1959	3	34.7	43.0	**44.2**
Stanford HYDRA	T	570	3	35.1	54.7	**68.2**
Stanford HYDRA	F	570	3	31.2	45.3	**48.9**
Freiburg Franka Play	F	3603	406	20.4	**32.2**	31.6
Freiburg Franka Play	T	3603	406	19.7	21.8	**24.0**
LSMO Dataset	T	50	2	34.0	68.0	**72.0**
UCSD Kitchen	T	150	8	19.3	32.0	**32.7**
Austin VIOLA	T	150	3	26.7	32.7	**33.3**
Austin VIOLA	F	150	3	30.0	33.3	**34.0**
Total	-	27000	-	22.1	28.5	**30.4**

OOD Generalization: MVU works on Robotics Domain tasks constructed from OpenX-Embodiment Dataset.

Summary of Contributions

1. Highlight issues of LLM based Video QnA
2. Build efficient setup for LLM-based video QnA
3. Propose Framework to use Video-Specific information
   a. Extraction of Object Centric information
   b. Language based fusion with VLM
4. Evaluation across established video QnA benchmarks

github.com/kahnchana/mvu

Our MVU framework performs from strong video QnA with better interpretability.