

Look, Remember and Reason: Grounded Reasoning in Videos with Language Models

Apratim Bhattacharyya*, Sunny Panchal, Mingu Lee, Reza Pourreza,
Pulkit Madan, Roland Memisevic

*Engineer, Senior; Qualcomm AI Research.
Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Video Language Models

- Great advances have been made recently in text-based reasoning tasks using LM (language models).
- However, on video data Video-LMs focus primarily on high-level question answering.

Q: What activity are the dog and the woman engaged in?



A: They are playing fetch ...

Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models, Maaz et. al., 2023.

Video Language Models

- Great advances have been made recently in text-based reasoning tasks using LM (language models).
- However, on video data Video-LMs focus primarily on high-level question answering.

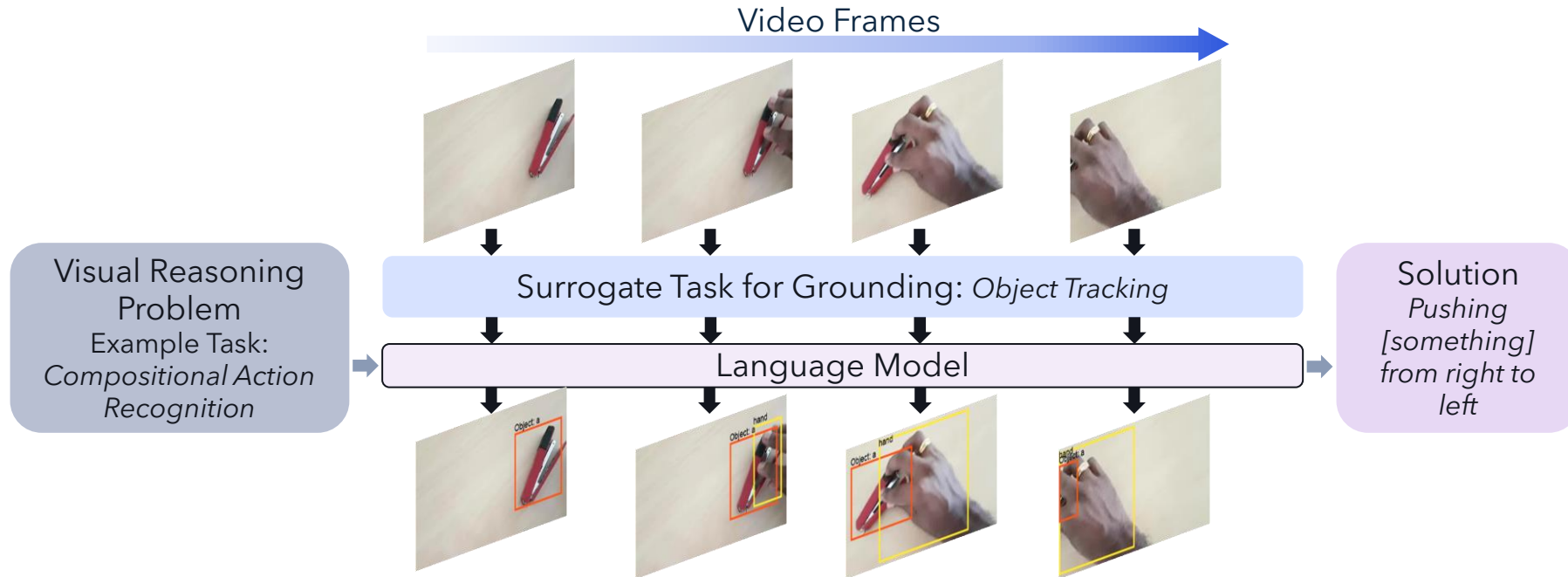
Q: What is the person doing with the stapler?



A: [Moving the stapler from right to left]

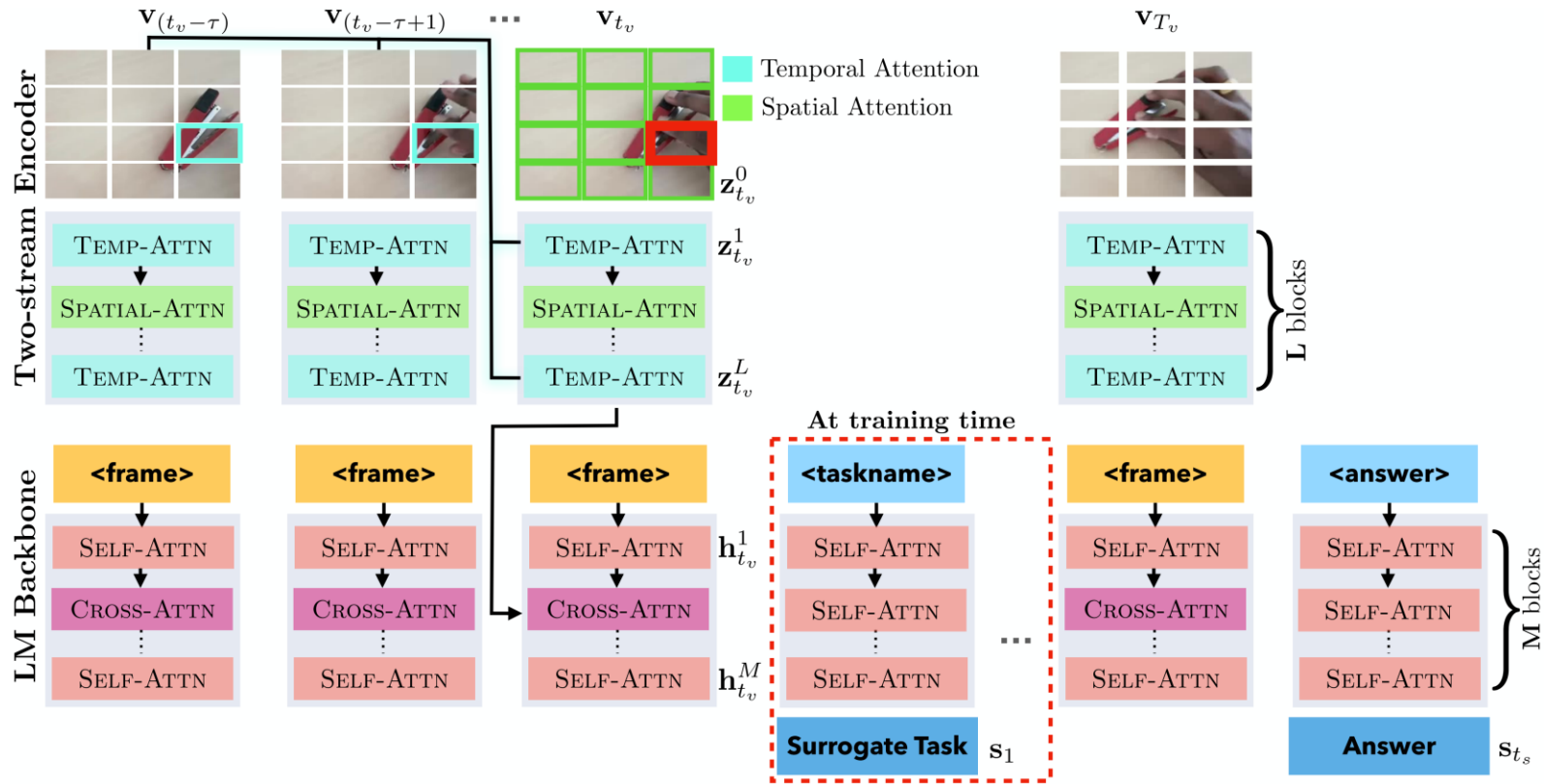
“Look, Remember and Reason” (LRR) Model

- Focus on reasoning tasks requires a fine-grained understanding of low-level details of motion and interactions.

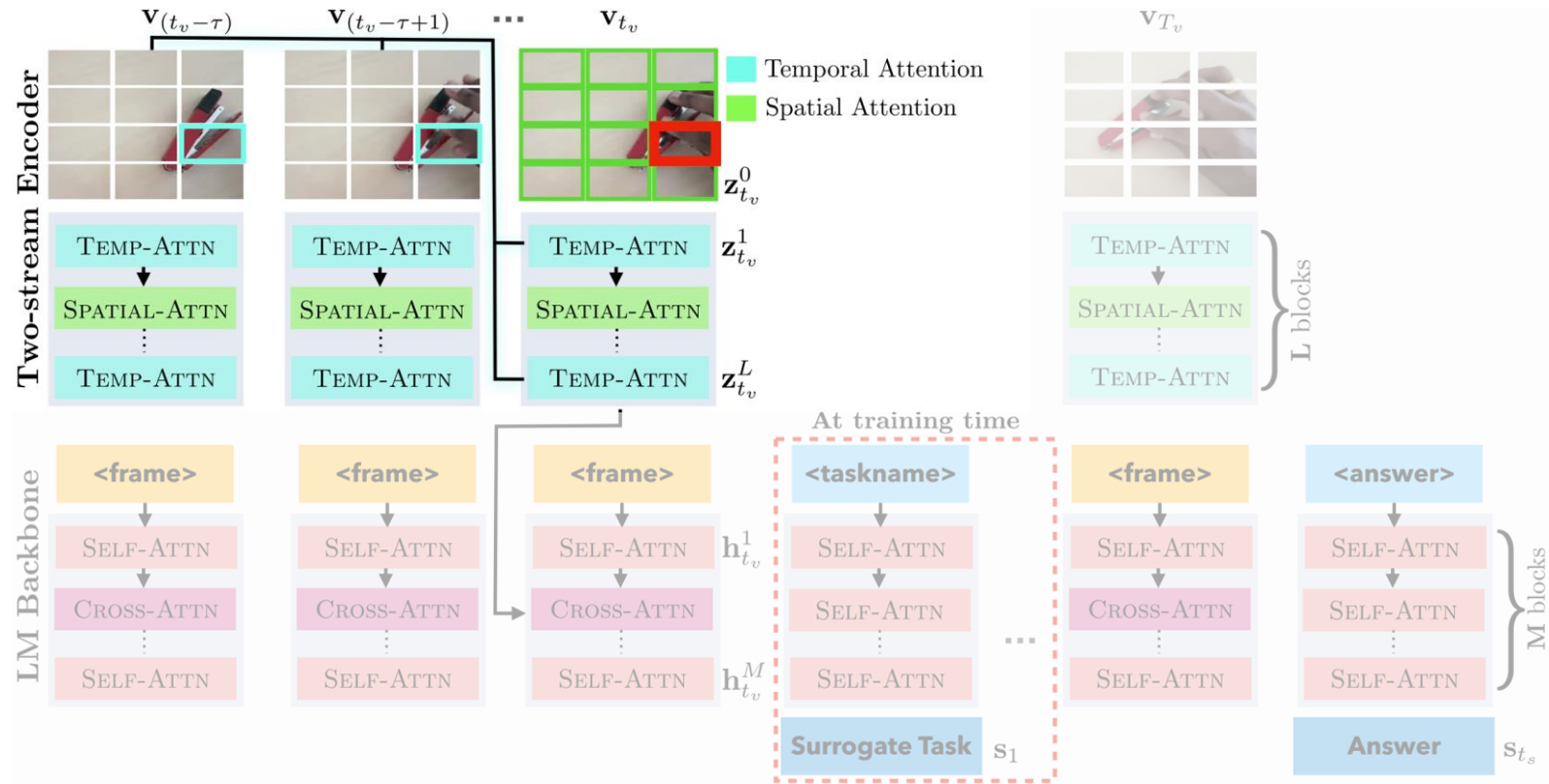


- Our proposed “LRR” model:
 - **Architecture:** “Two-stream” video encoder that captures scene structure and motion.
 - **Random Probes:** Low-level surrogate tasks such as object recognition, tracking and re-identification.

“Look, Remember and Reason” (LRR) Model

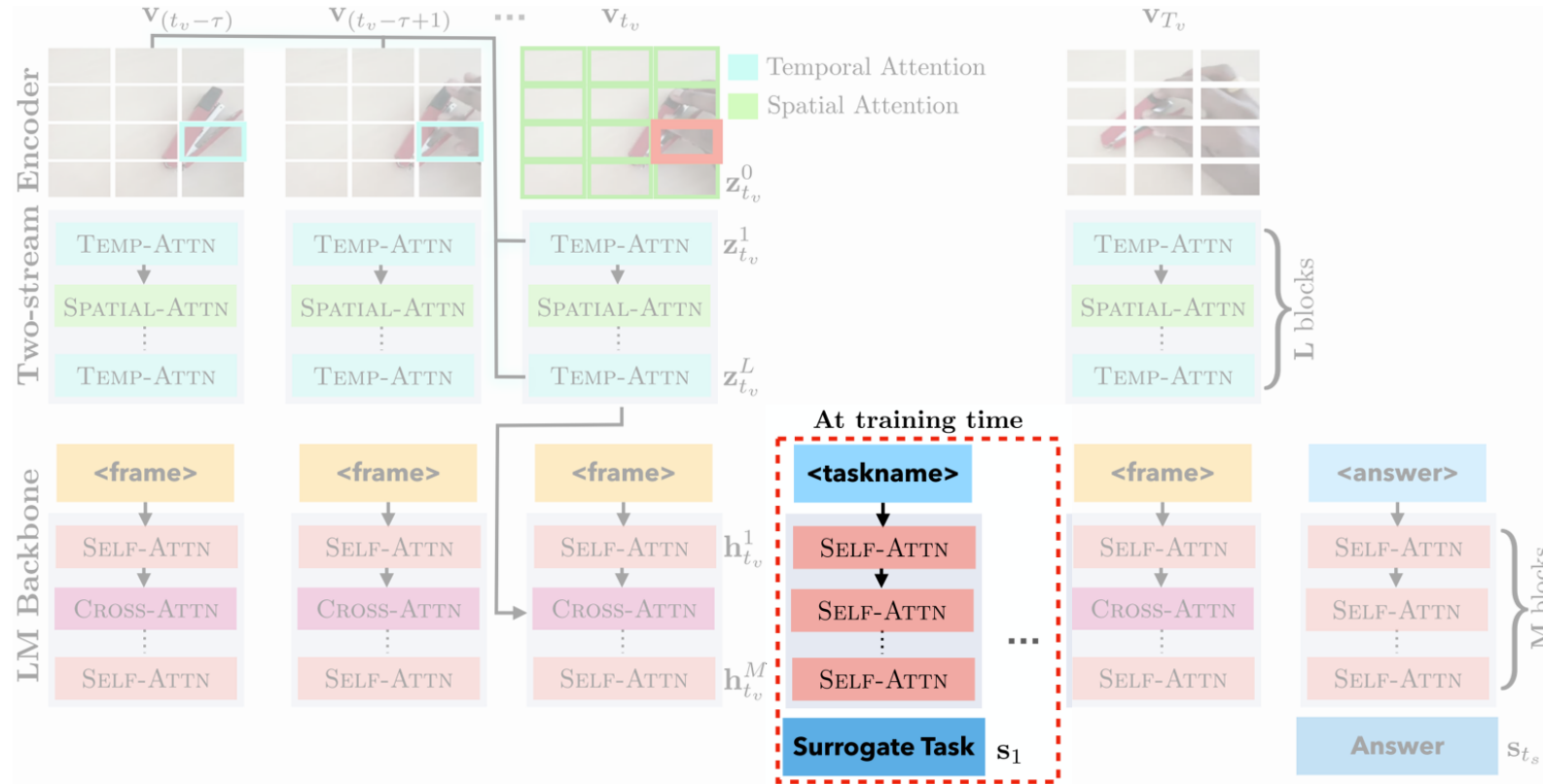


“Look, Remember and Reason” (LRR) Model



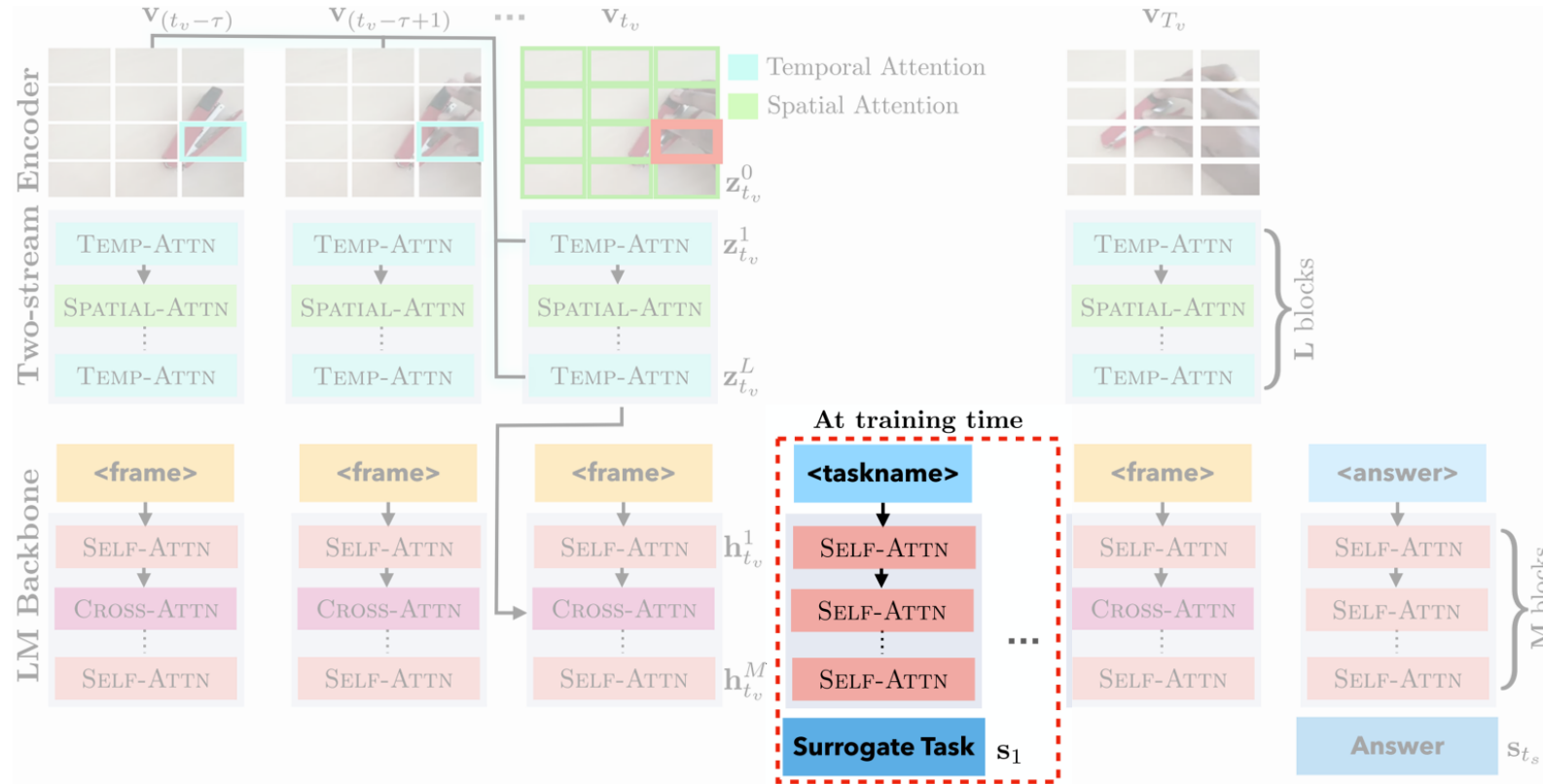
- Efficiently encodes both structure and motion in the input video.
 - **Structure:** Spatial attention on patches in the current frame.
 - **Motion:** Temporal attention on patches in previous τ frames.

“Look, Remember and Reason” (LRR) Model



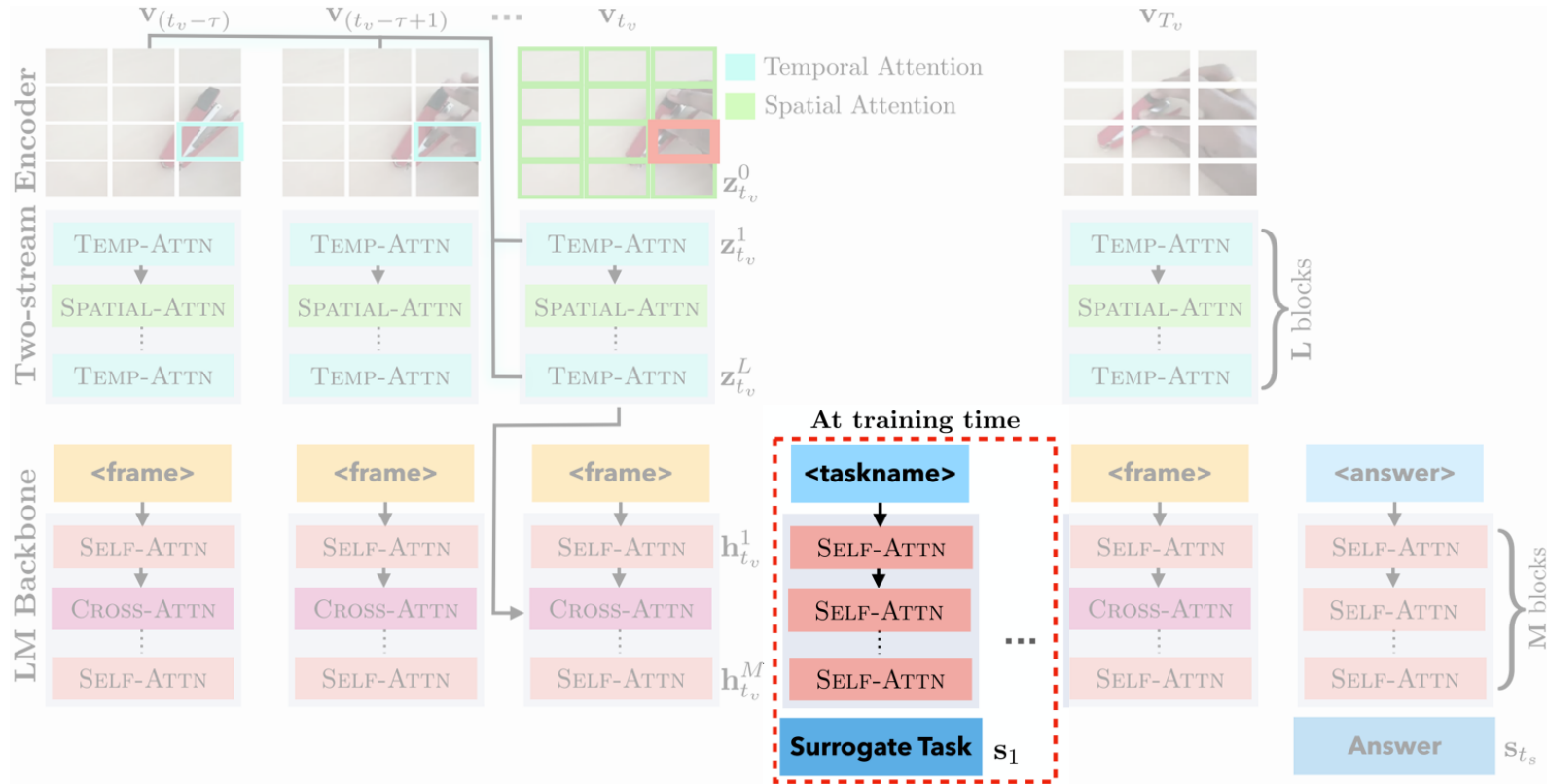
- **Random Probes:** We prompt our LRR model to solve certain low-level surrogate tasks at randomly selected time steps within the video.

“Look, Remember and Reason” (LRR) Model



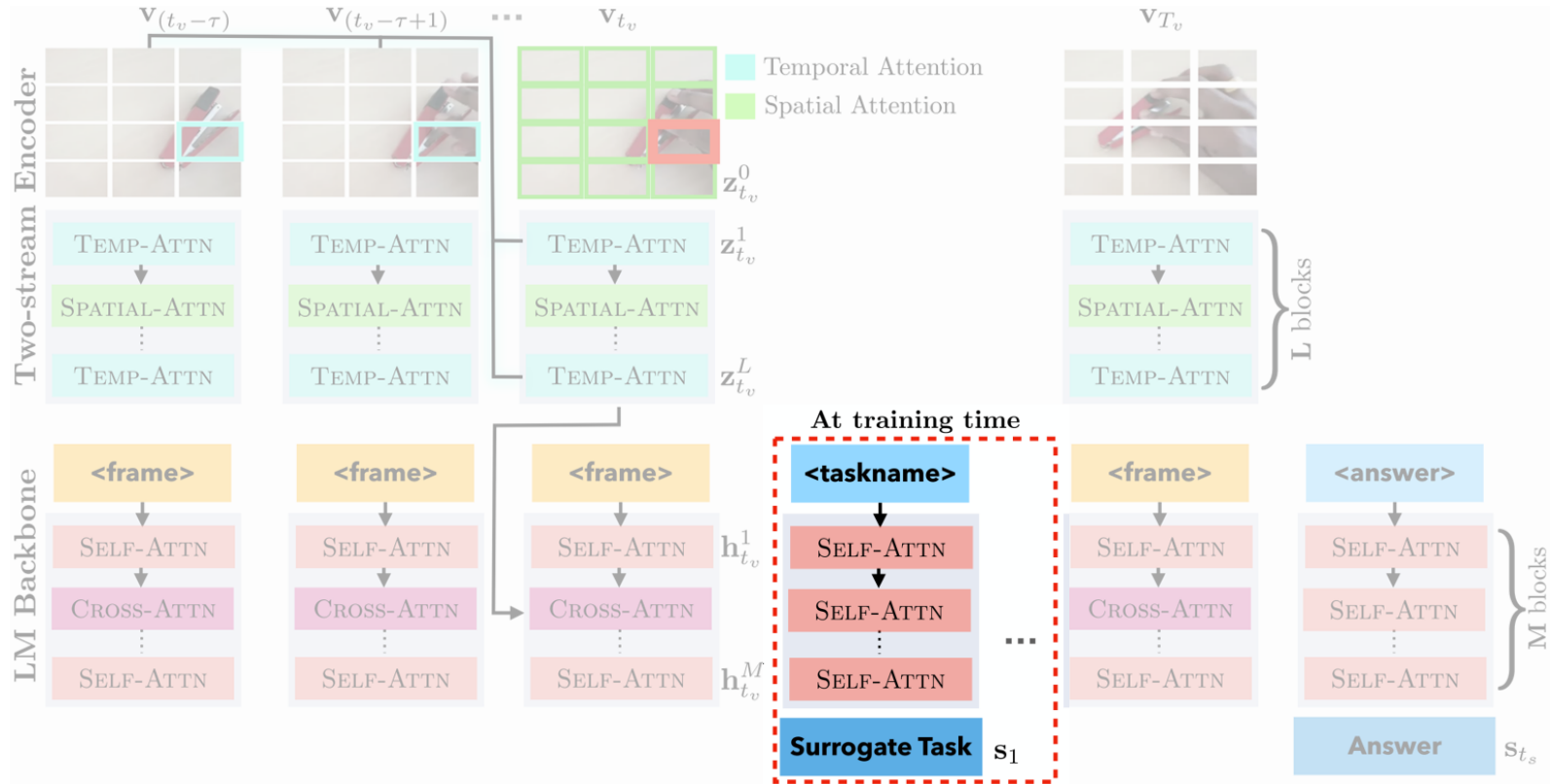
- **Random Probes:** Object recognition, localization, re-identification and tracking, fundamental to solving a range of visual reasoning problems.

“Look, Remember and Reason” (LRR) Model



- Our LRR model highly flexible and can be prompted to solve a wide range of low-level surrogate tasks.

“Look, Remember and Reason” (LRR) Model



- Ground-truth can be usually obtained using off-the-shelf vision models.

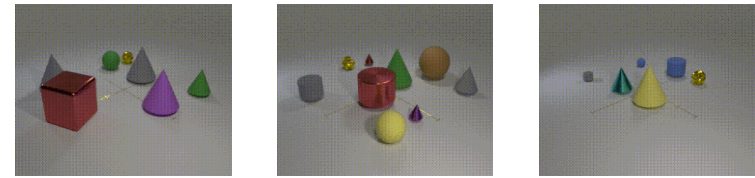
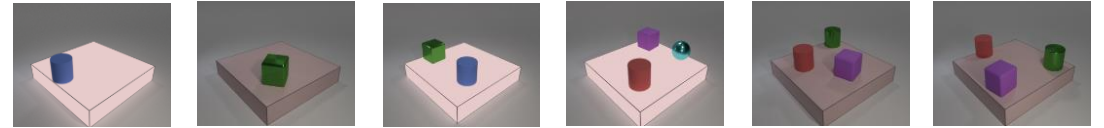
Evaluation

- **State-of-the-art results on:**

- ACRE (Zhang et al., 2021).
- CATER (Ding et al., 2021).
- Something-Else (Materzynska et al., 2020).
- STAR (Wu et al., 2021).

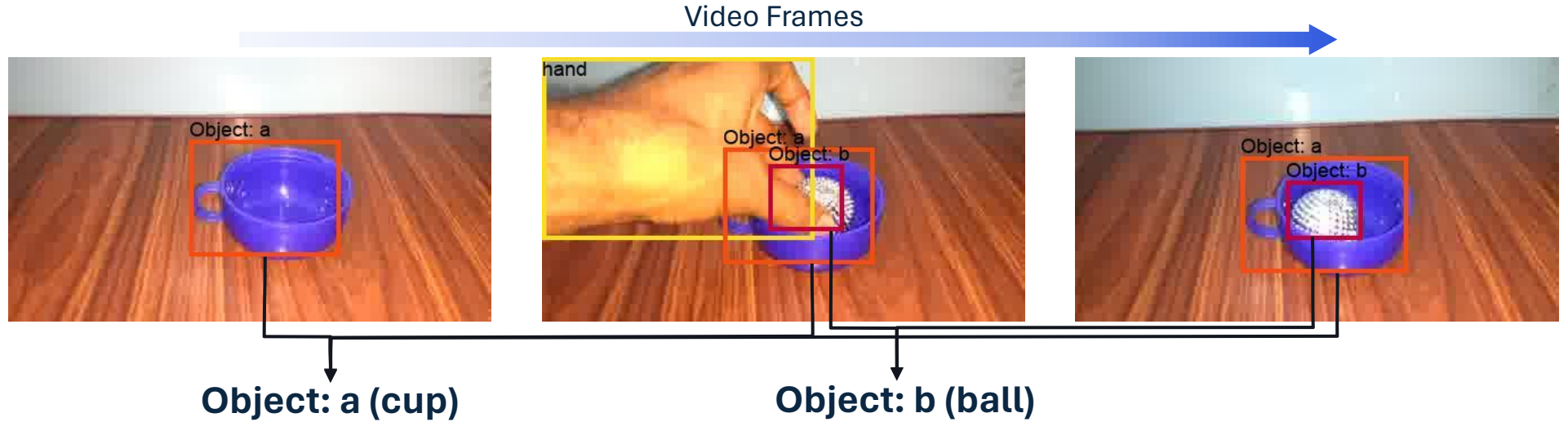
- **LRR model details:**

- LM backbone: OPT-125M/350M/1.3B.
- We fine-tune the video encoder and the LLM backbone.



Evaluation: Something-Else

Surrogate Task:
Multi-object tracking



Method	Base		Compositional	
	Top-1	Top-5	Top-1	Top-5
STIN + OIE + NL (Materzynska et al., 2020, MIT)	78.1	94.5	56.2	81.3
Video-ChatGPT (Maaz et al., 2023)	52.6	75.8	38.6	67.8
LRR (Ours)	80.2	96.1	62.0	86.3
LRR (w/o Two-stream Encoder)	73.2	90.4	53.6	76.1
LRR (w/o Surrogate Tasks)	52.6	75.8	50.1	70.8

Evaluation: STAR

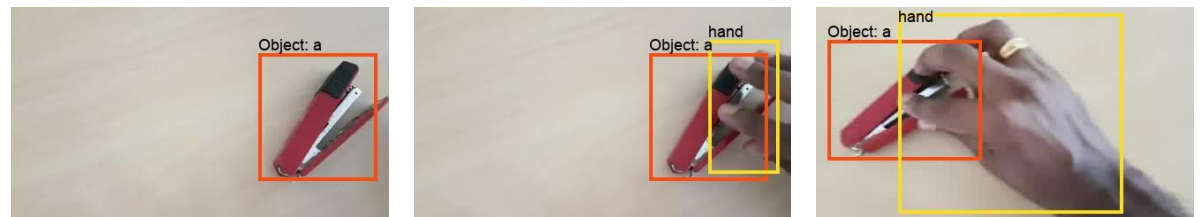
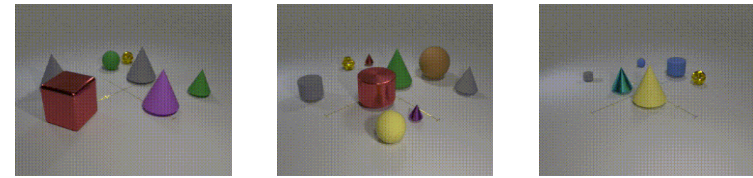
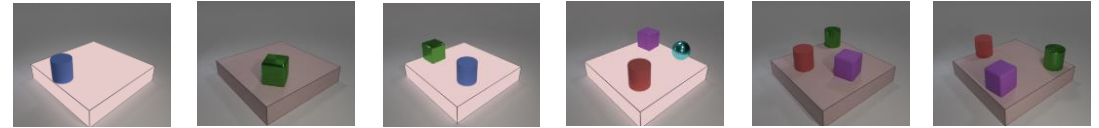
STAR Challenge Leaderboard on eval.ai (April 2024)

Rank ↕	Participant team ↕	Mean (↑) ↕	Last submission at ↕
1	SN3946 (Look, Remember and Reason)	71.89	9 days ago
2	sn12 (#7024-IPRM)	70.28	2 months ago
3	No22 (T816)	69.53	9 days ago
4	ThreeMissingOne	64.25	3 months ago
5	Fudan Nebula (Work in progress (submitted to))	62.67	6 months ago

- **Ranked 1st on the STAR leaderboard.**
- **Surrogate Tasks:**
 - Object recognition: STAR (Wu et al., 2021).
 - Action recognition: Kinetics (Kay et al., 2017) and Moments-in-Time (Monfort et al., 2020).
 - Tracking: Something-Else (Materzynska et al., 2020).
 - Regularize on text data.

Conclusion

- We show that off-the-shelf LMs can solve complex visual reasoning tasks on videos using our LRR framework.
- Surrogate tasks ensures that the LM can utilize relevant low-level visual cues.
- Grounding predictions to low-level visual cues combined with the high-level reasoning ability of the LM is the key to the success of the model.



Thank you

Qualcomm

Follow us on: [in](#) [X](#) [@](#) [▶](#) [f](#)

For more information, visit us at:

qualcomm.com & qualcomm.com/blog

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

© Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm is a trademark or registered trademark of Qualcomm Incorporated. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to "Qualcomm" may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.