TOGA: Temporally Grounded Open-Ended Video QA with Weak Supervision

Ayush Gupta[1,2], Anirban Roy[1], Rama Chellappa[2], Nathaniel D. Bastian[3], Alvaro Velasquez[4], Susmit Jha[1]

[1]SRI [2]Johns Hopkins University [3]United States Military Academy [4]University of Colorado Boulder

ICCV OCT 19-23, 2025 HONOLULU HAWAII

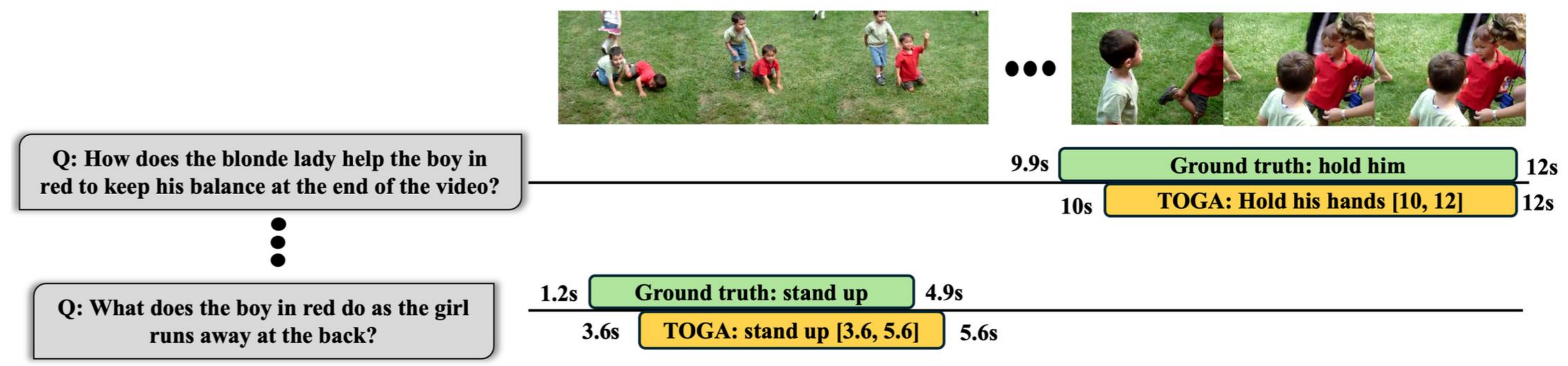
Introduction

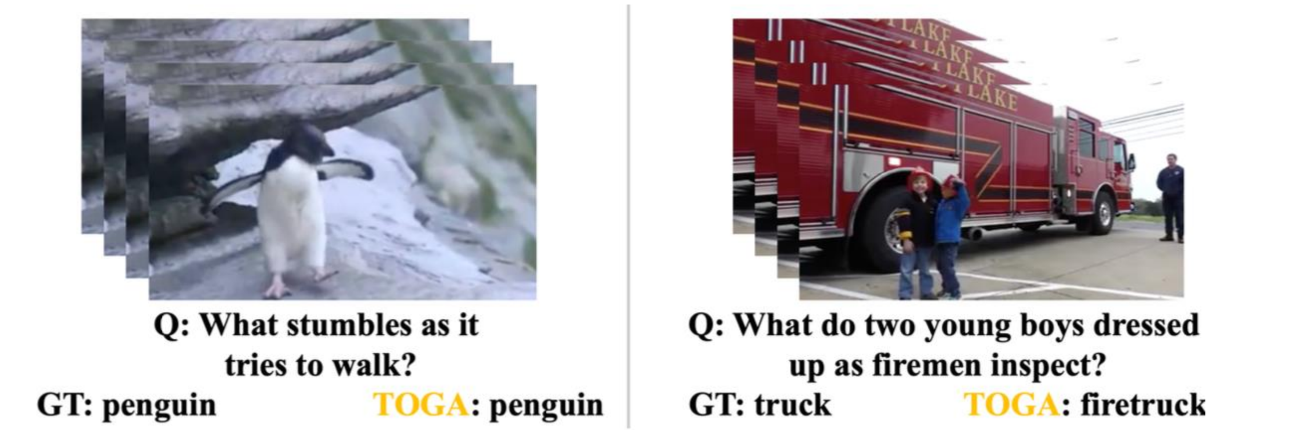
Limitation of existing works:

- **Less trustworthy:** Missing evidence for answers
- **Difficult to collect data:** Need temporal annotations
- **Less practical for deployment:** Choose answer from predefined options

We focus on the problem of **weakly supervised, grounded** video question answering (VQA).

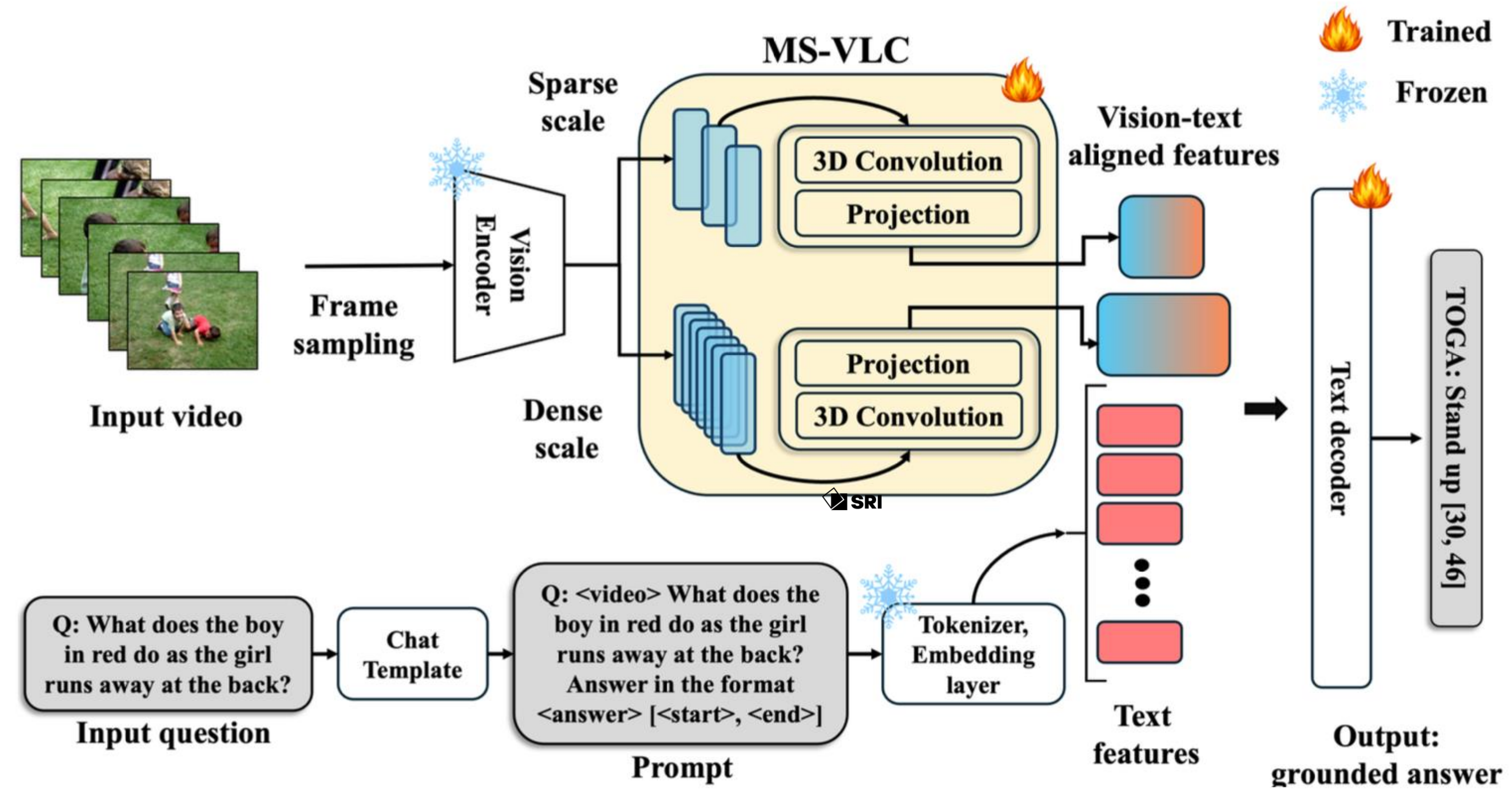
We propose a **vision language model (VLM)** for this task, to generate **open-ended** answers.





Architecture

- A CLIP vision encoder to encode vision features
- A **multi-scale vision language connector (MS-VLC)** to extract global-local features
- An LLM decoder to generate the answer and the grounding



Results

NExT-GQA dataset

	Open-ended evaluation	mIoU	mIoP	IoU@0.5	IoP@0.5	Acc@GQA
IGV [24] *(MM '22)*	✗	14	21.4	9.6	18.9	10.2
Temp[CLIP](NG+) [34, 49] *(CVPR '24)*	✗	12.1	25.7	8.9	25.5	16
FrozenBiLM (NG+) [49, 52] *(CVPR '24)*	✗	9.6	24.2	6.1	23.7	17.5
SeViLA [54] *(NeurIPS '23)*	✗	21.7	29.5	13.8	22.9	16.6
LLoVi [56] *(arXiv '24)*	✗	20	37.3	15.3	36.9	24.3
Grounded-VideoLLM [41] *(arXiv '24)*	✗	21.1	34.5	18	34.4	**26.7**
VideoStreaming [33] *(NeurIPS '24)*	✗	19.3	32.2	13.3	31	17.8
TOGA (Ours)	✓	**24.4**	**40.5**	**21.1**	**40.6**	24.6

ReXTime dataset

	Model	mIoU	R@1 (IoU=0.3)	R@1 (IoU=0.5)
Non Generative	UniVTG	28.17	41.34	26.88
	CG-DETR	23.87	31.31	16.67
LLM based	VTimeLLM	20.14	28.84	17.41
	TimeChat	11.65	14.42	7.61
	LITA	21.49	29.49	16.29
	Ours	**25.53**	**29.91**	**19.79**

Ablations and Analysis

Effect of MS-VLC

Model type	All	short	medium	long
		Query Type		
Sparse only	20.0	16.2	28.9	47.5
Dense Only	22.1	18.3	32.2	32.1
Multi-Scale (MS-VLC)	**24.4**	**20.5**	**34.7**	**49.3**

Question types: *before, after* clauses v/s *why, how* questions

Connector type	Question type: Causal		Temporal			All
	Why	How	Present	Past	Future	
Sparse only	21.7	21.1	18	17.9	12.8	19.55
Dense only	21.8	22.2	18.8	18	13.6	20.12
MS-VLC	26.1	27.4	23.4	18	18.1	24.6

Approach: Multi-stage training and consistency constraint

Three stages:

1. **Vision-text alignment**
2. **Instruction tuning for grounding/referring:** Learning initial timestamp representations
3. **Consistency tuning:** Refining timestamps

How to train without temporal annotations?

- **Stage 1**: Model can generate captions
- **Pseudo labels for stage 2**: Randomly crop sections of videos , caption them
- **Filtering noisy labels for stage 3**: Using consistency between model responses and GT

<Cropped video>: What is happening?
Model: A dog runs away from the water.

Q: What happens in [65, 79]?
A: A dog runs away from the water.

Q: Locate the scene where a dog runs away from the water.
A: [65, 79].

Q: How did the dog react to the coming waves?
GT Answer: Run away

Model Response: Run Away [60, 72]

Consistency

Q: What happens in [60, 72]?
Model Response: The dog run away.

Acknowledgements

This work was supported in part by the U.S. Air Force and DARPA under Contract No. FA8750-23-C-0519, the U.S. Army Research Laboratory Cooperative Research Agree- ment W911NF-17-2-0196, the U.S. Army Combat Capabil- ities Development Command (DEVCOM) Army Research Laboratory under Support Agreement No. USMA 21050, and the DARPA under Support Agreement No. USMA 23004. The opinions, findings, and conclusions expressed in this paper are those of the authors and do not reflect the position of the United States Department of Defense or the United States Government.

Contact Information

Ayush Gupta, final year PhD student

Looking for internship and full-time opportunities!

agupt120@jh.edu