HierarQ: Task-Aware Hierarchical Q-Former for Enhanced Video Understanding

Shehreen Azad

Vibhav Vineet

Yogesh S Rawat

Center for Research in Computer Vision, University of Central Florida;

Microsoft Research

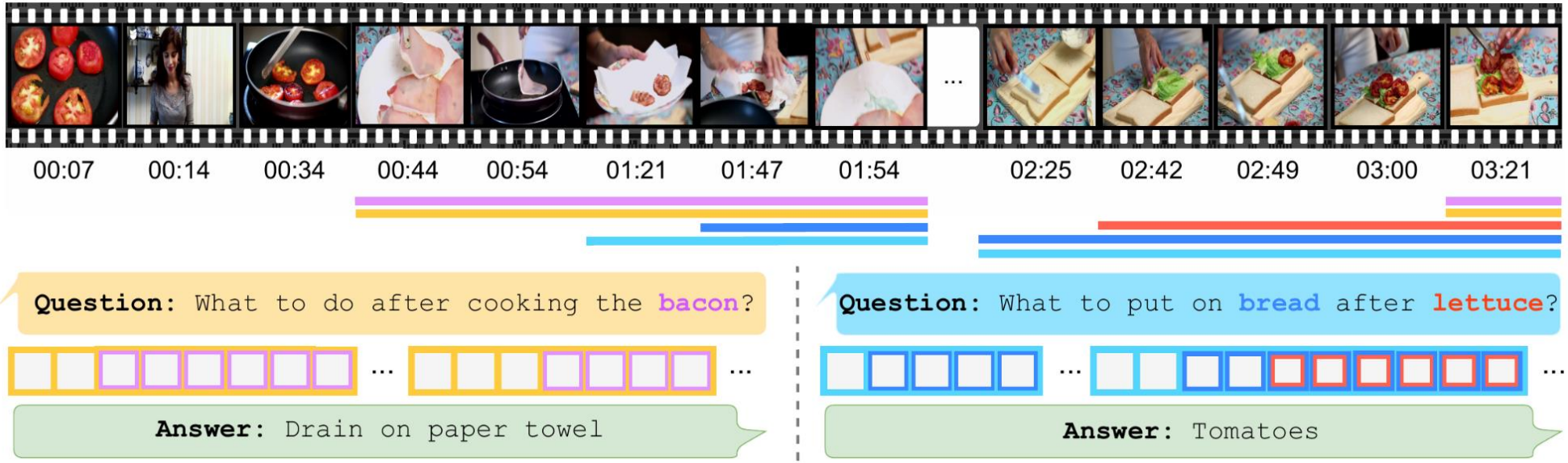
Introduction

Current limitation

- **Context length/ Input frames** constraint.
- Frame **sampling**, Coarse **compression**.

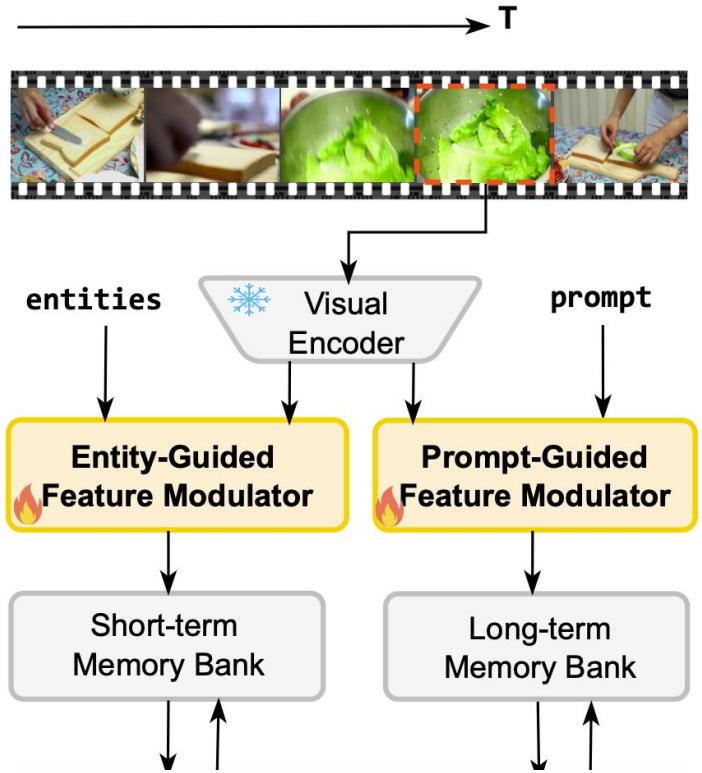
Our approach limitation

- **Auto-regressive** processing.
- **Task-focused** awareness.



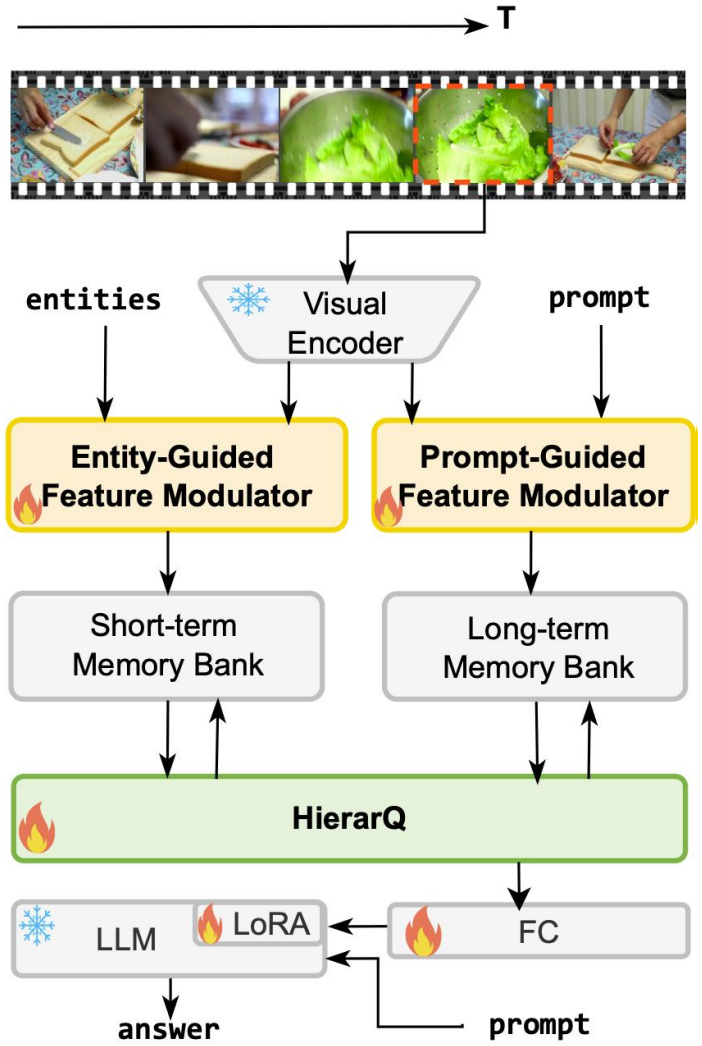
Framework Overview

- Our framework processes entire video auto-regressively,
- Using language-guided two-stream feature modulators to incorporate task-awareness.
- The visual features are cross-attended with entities or prompt to modulate them according to task-relevance.
- The modulated features are stored in dedicated memory banks.



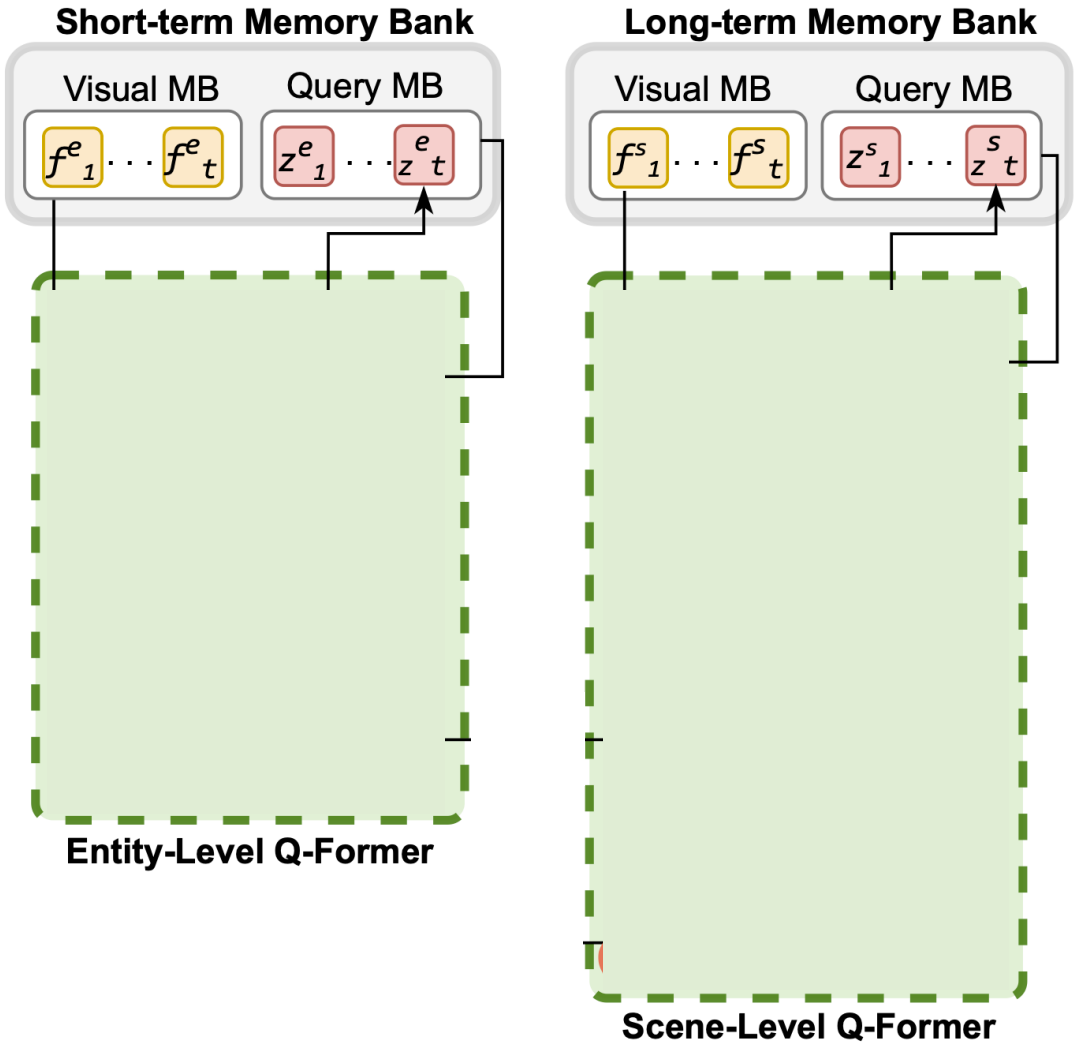
Framework Overview

- These memory banks enable our proposed Hierarchical Querying Transformer (HierarQ) to model long-term temporal relationships.
- Finally, the output from HierarQ is sent to an LLM to generate response text.



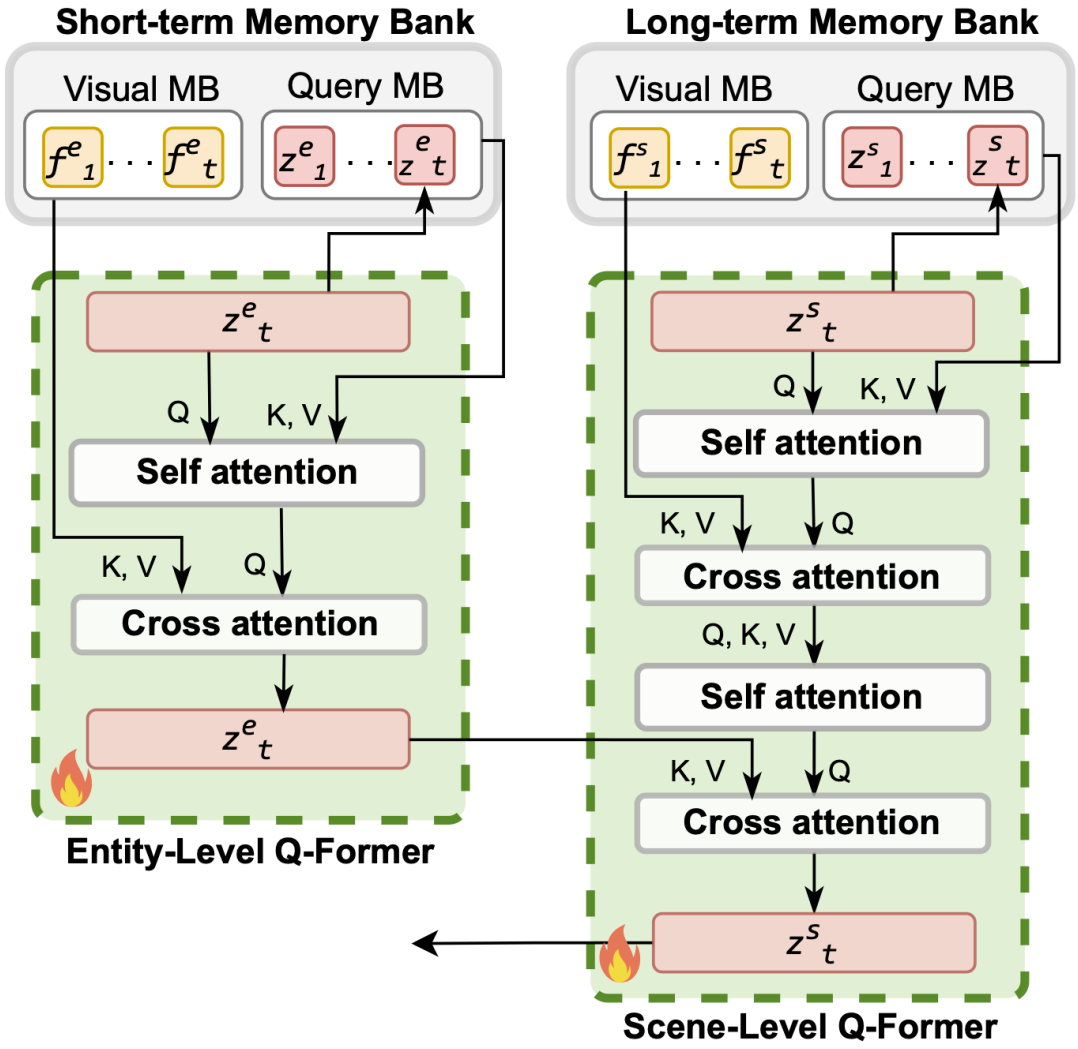
HierarQ

- Consists of two Q-Formers:
  - Entity level
  - Scene level
- Each Q-Former interacts with corresponding memory banks.
- Each memory bank contains two types of memory :
  - Visual
  - Query
- The short term memory bank is updated by FIFO
- The long term memory bank is updated by merging similar tokens



HierarQ

- Each Q-Former has its own learnable query,
- Which self refine itself through self attention and query memory banks.
- Then align themselves with visual features from the visual memory banks.
- The entity level Q-Former's learned query then interacts with the scene level Q-Former's learned query.
- The final learned query of the scene level Q-Former contains historical information of the scene complemented by the entity details.



Results

Long Video Understanding

Model	Relation ↑	Speak ↑	Scene ↑	Avg
VideoBERT [63]	52.8	37.9	54.9	48.5
Obj_T4mer[78]	54.8	33.2	52.9	47.0
Orthoformer [48]	50.0	38.3	66.3	51.5
VIS4mer [29]	57.1	40.8	67.4	55.1
LF-VILA [65]	61.5	41.3	68.0	56.9
TranS4mer [30]	59.5	39.2	70.9	56.5
S5 [72]	67.1	42.1	73.5	60.9
Movies2Scene [13]	**71.2**	42.2	68.2	60.5
VideoMamba [38]	62.5	40.4	70.4	57.8
MA-LMM [24]	58.2	44.8	80.3	61.1
HierarQ ‡	67.9	<u>48.7</u>	<u>83.8</u>	<u>66.8</u>
HierarQ	<u>69.4</u>	**49.3**	**85.1**	**67.9**

Short Video Understanding

Model	Breakfast	COIN
Timeception [27]	71.3	-
VideoGraph [28]	69.5	-
GHRM [92]	75.5	-
Dist-Sprv [42]	89.9	90.0
ViS4mer [29]	88.2	88.4
TranS4mer [30]	90.3	89.2
S5 [72]	90.7	90.8
FACT [44]	84.5	-
VideoMamba [38]	94.3	86.2
MA-LMM [24]	93.0	93.2
HierarQ‡	<u>96.1</u>	<u>94.6</u>
HierarQ	**97.4**	**96.0**

Results

Short Video Question Answering

Model	MSR-QA	MSVD-QA	ANet-QA
JustAsk [84]	41.5	46.3	38.9
FrozenBiLM [85]	47.0	54.4	43.2
SINGULARITY [33]	43.5	-	43.1
GIT [71]	43.2	56.8	-
VIOLETv2 [21]	44.5	54.7	-
mPLUG-2 [81]	48.0	58.1	-
UMT-L [37]	47.1	55.2	47.9
Mirasol3B [49]	50.4	-	51.1
MA-LMM [24]	48.5	60.6	49.8
HierarQ‡	53.4	64.4	56.8
HierarQ	**54.1**	**66.2**	**57.1**

Long Video Question Answering

Model	G. Acc.	G. Sc.	B. Acc.	B. Sc.
Video Chat [36]	57.8	3.0	46.1	2.3
Video LLaMA [88]	51.7	2.7	39.1	2.0
Video-ChatGPT [46]	47.6	2.6	48.0	2.5
MovieChat [61]	62.3	3.2	48.3	2.6
FVS+S3 [75]	84.0	4.6	73.5	4.0
MA-LMM † [24]	61.4	3.2	50.4	2.7
HierarQ‡	86.9	**4.7**	74.2	4.1
HierarQ	**87.5**	**4.7**	**76.4**	**4.2**

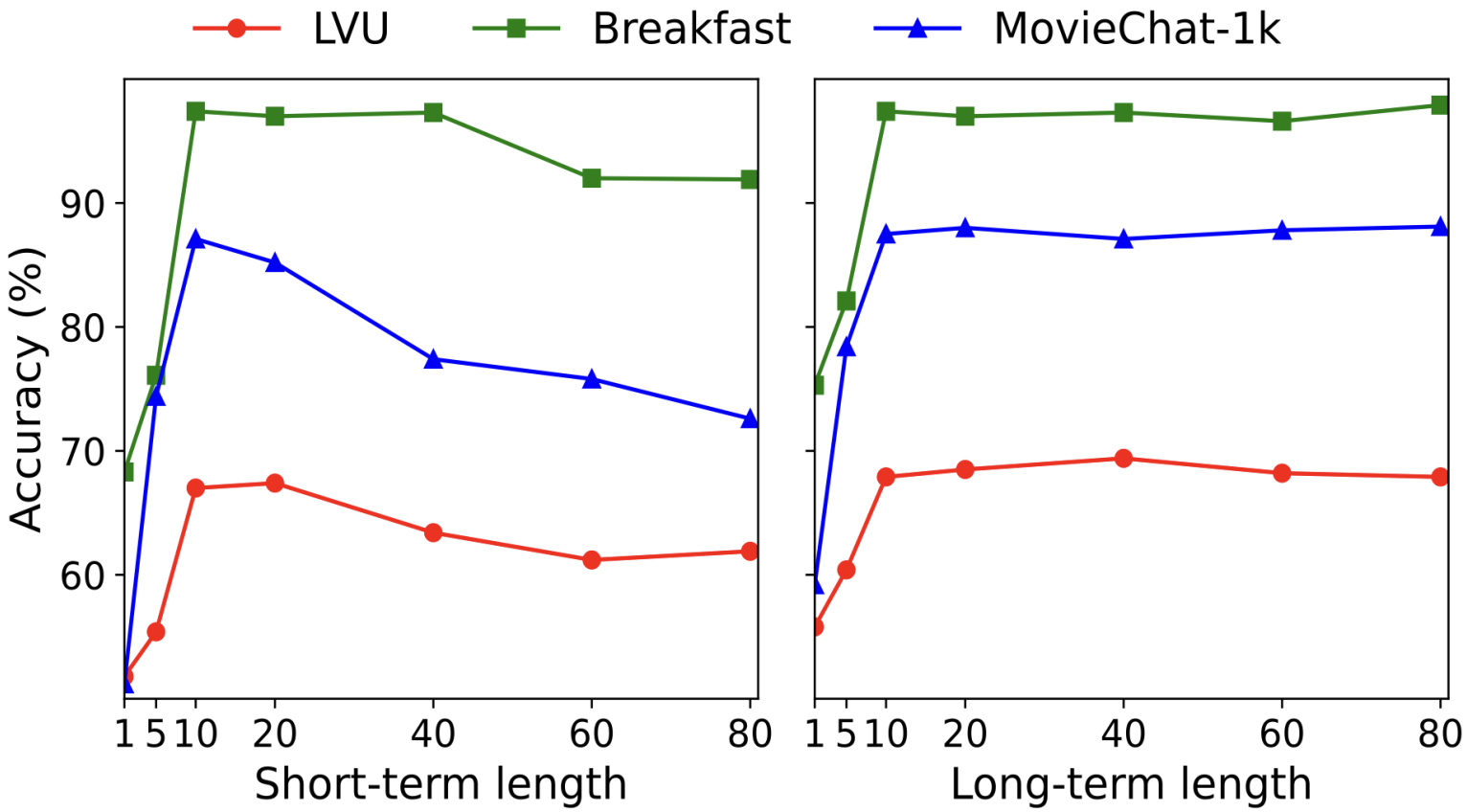
Results

Video Captioning

Model	MSRVTT	MSVD	YouCook2
SwinBERT [41]	55.9	149.4	109.0
GIT [71]	73.9	180.2	129.8
mPLUG-2 [81]	<u>80.3</u>	165.8	-
HowToCaption [60]	65.3	154.2	116.4
MA-LMM [24]	74.6	179.1	<u>131.2</u>
HierarQ‡	79.8	<u>182.6</u>	<u>134.4</u>
HierarQ	**80.5**	**183.1**	**136.1**

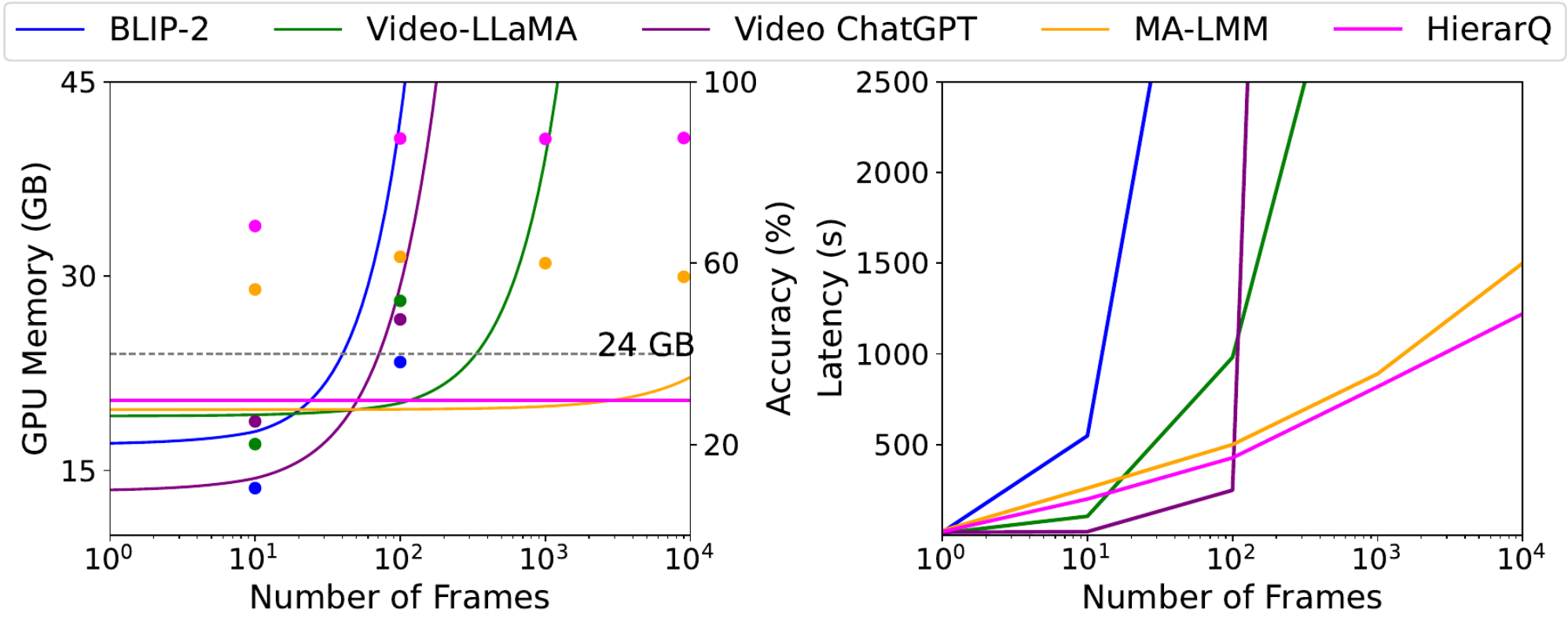
Ablation Studies

Memory Bank Size: 10 frames, 32 tokens each



Cost Analysis

HierarQ maintains constant token length (32) while supporting over 10k frames on a single 24G GPU.



Qualitative Analysis

Superior subtle event identification capability

Question: Are there any **animals**?

✓ **HierarQ:** Yes ✗ **MA-LMM:** No

Question: Do **stars** appear?

✓ **HierarQ:** No ✗ **MA-LMM:** Yes

Qualitative Analysis

Robust to implicit entities

Q: What does the video describe?

Ans: Two teams are playing a football game.

Qualitative Analysis

Robust to ambiguous entities

Thank You

Corresponding author:
Shehreen.Azad@ucf.edu