Self-Chained Image-Language Model for Video Localization and Question Answering

Shoubin Yu, Jaemin Cho, Prateek Yadav, Mohit Bansal
UNC Chapel Hill

NeurIPS 2023

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- Recent studies have explored efficient training of video-LMs by leveraging pre-trained image-LMs.

- They typically concatenate uniformly/randomly sampled video frames as visual inputs **without explicit question-aware modeling.**

Motivation

- Recent studies have explored efficient training of video-LMs by leveraging pre-trained image-LMs.

- They typically concatenate uniformly/randomly sampled video frames as visual inputs **without explicit question-aware modeling.**

- Such a simple sampling can lead to losing important visual cues, resulting in the video-LMs focusing on frames that are unimportant to language.

Motivation

UNC NLP

- Recent studies have explored efficient training of video-LMs by leveraging pre-trained image-LMs.

- They typically concatenate uniformly/randomly sampled video frames as visual inputs **without explicit question-aware modeling.**

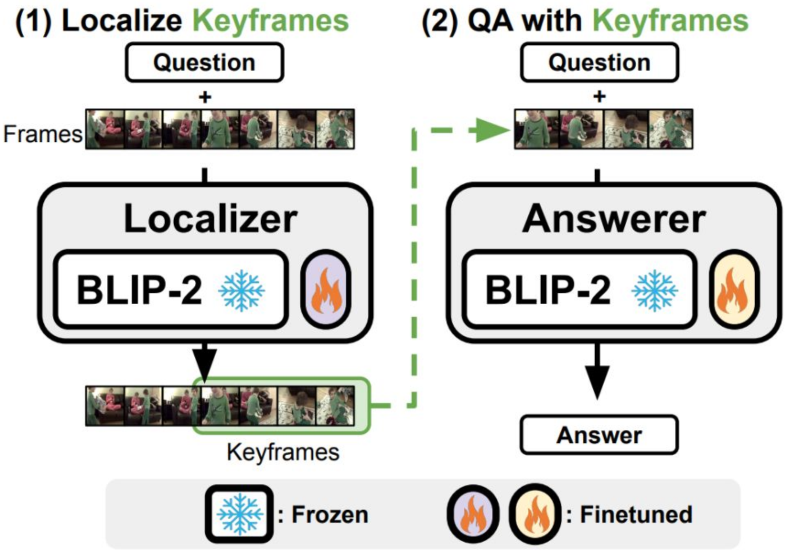
- Such a simple sampling can lead to losing important visual cues, resulting in the video-LMs focusing on frames that are unimportant to language.

- we introduce a novel video-language framework where we adopt **a single image-LM** to handle both **temporal localization** and **question answering** on videos, while avoiding expensive language-aware grounding annotations

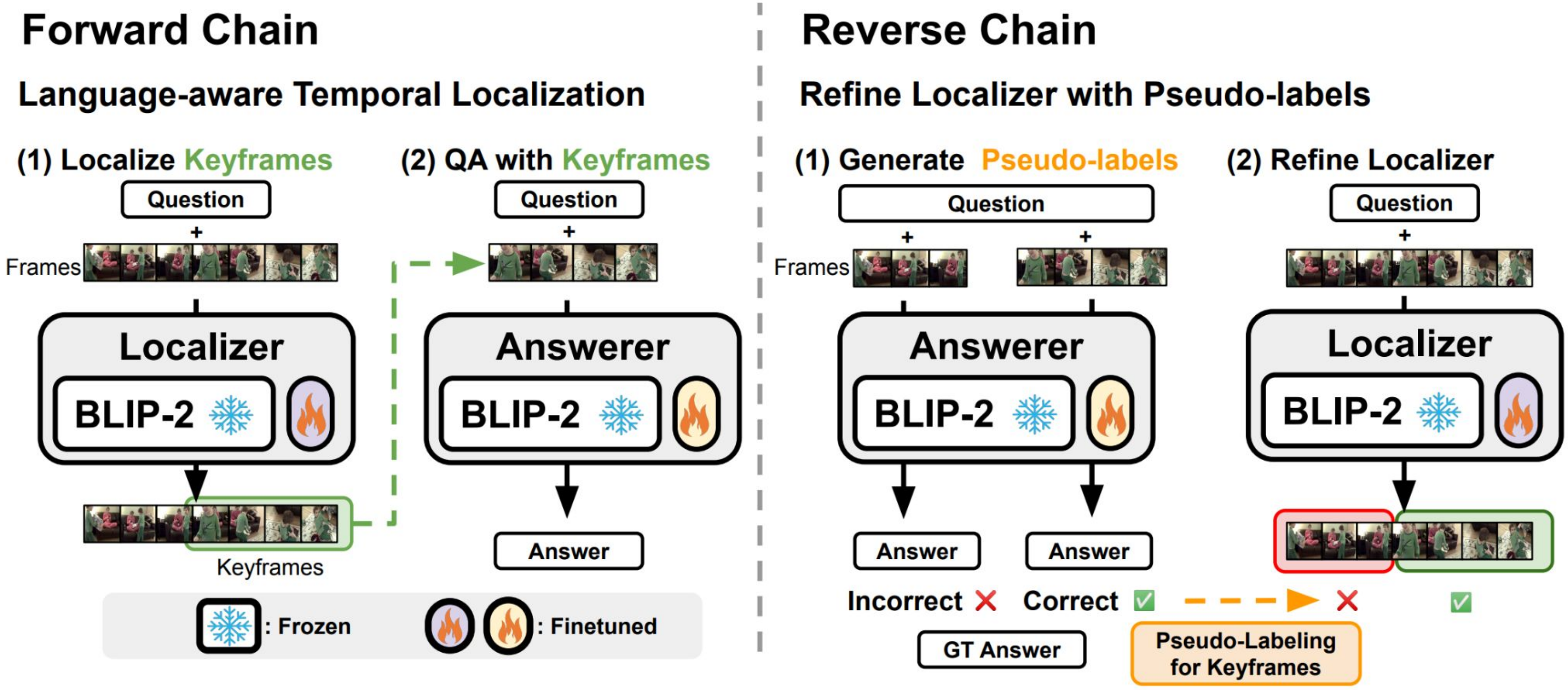
SeViLA: Self-chained Video Localization-Answering

Forward Chain

Language-aware Temporal Localization



SeViLA: Self-chained Video Localization-Answering



SeViLA: Self-chained Video Localization-Answering

Localizer

[Question]
what does the boy do after unwrapping the present?

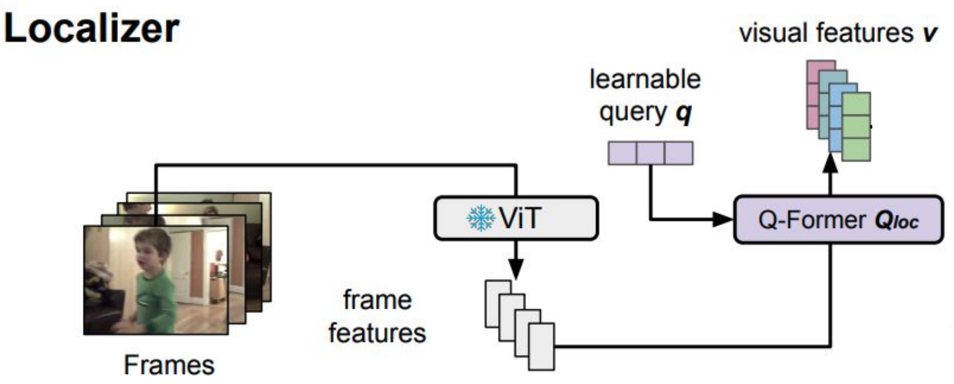
[Options]
A: Show to the camera. B: Cry.
C: Relax his fingers. D: Walk alone.

[Loc Prompt]
Does the information within the frame provide necessary details to accurately answer the question?

[QA Prompt]
Considering information in frames, select the correct answer from the options.

SeViLA: Self-chained Video Localization-Answering

Localizer



[Question]
what does the boy do after unwrapping the present?

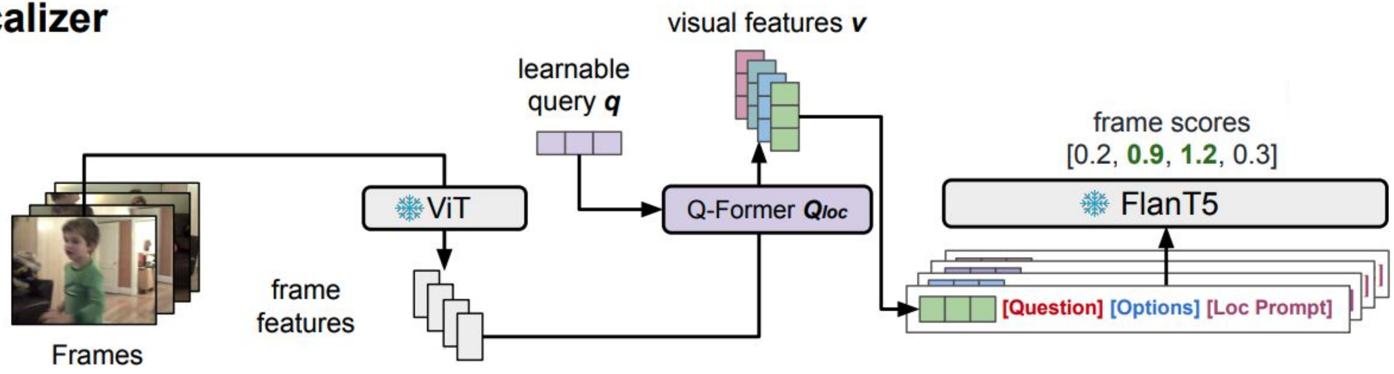
[Options]
A: Show to the camera. B: Cry.
C: Relax his fingers. D: Walk alone.

[Loc Prompt]
Does the information within the frame provide necessary details to accurately answer the question?

[QA Prompt]
Considering information in frames, select the correct answer from the options.

SeViLA: Self-chained Video Localization-Answering

Localizer



[Question]
what does the boy do after unwrapping the present?

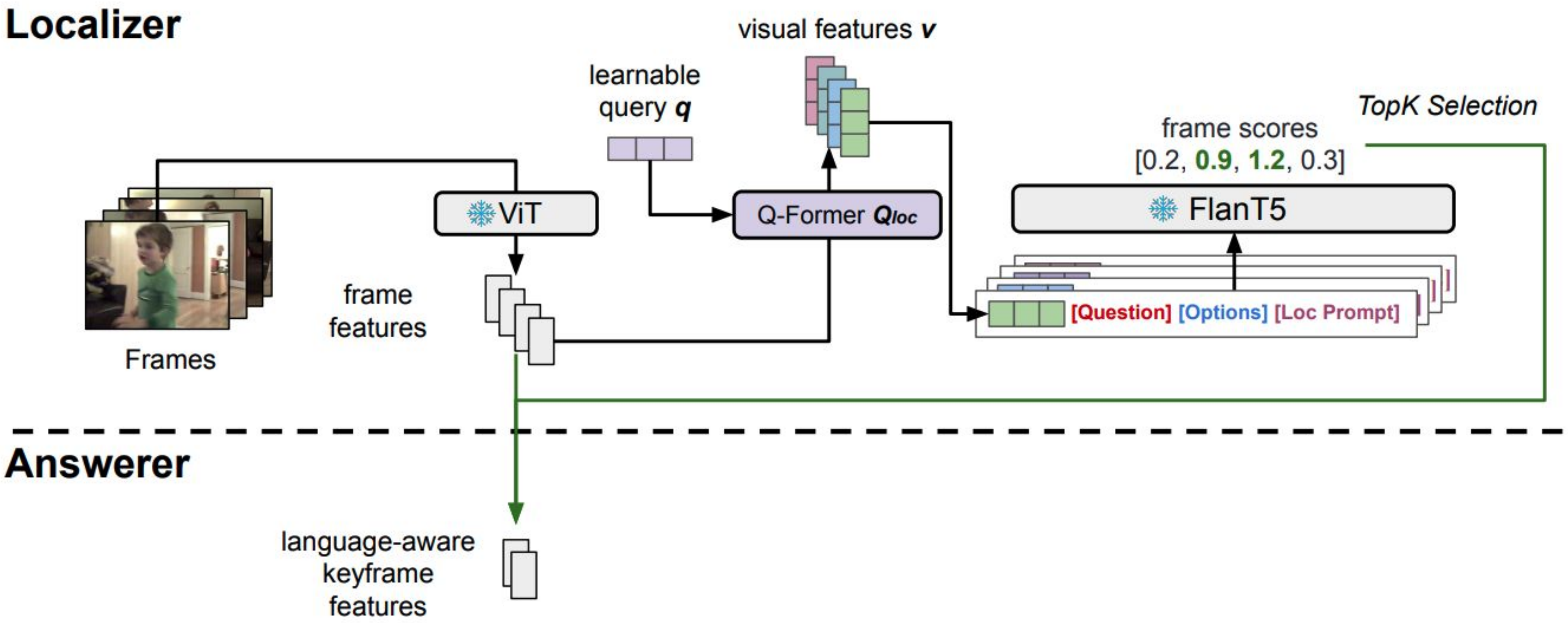
[Options]
A: Show to the camera. B: Cry.
C: Relax his fingers. D: Walk alone.

[Loc Prompt]
Does the information within the frame provide necessary details to accurately answer the question?

[QA Prompt]
Considering information in frames, select the correct answer from the options.

SeViLA: Self-chained Video Localization-Answering

UNC NLP



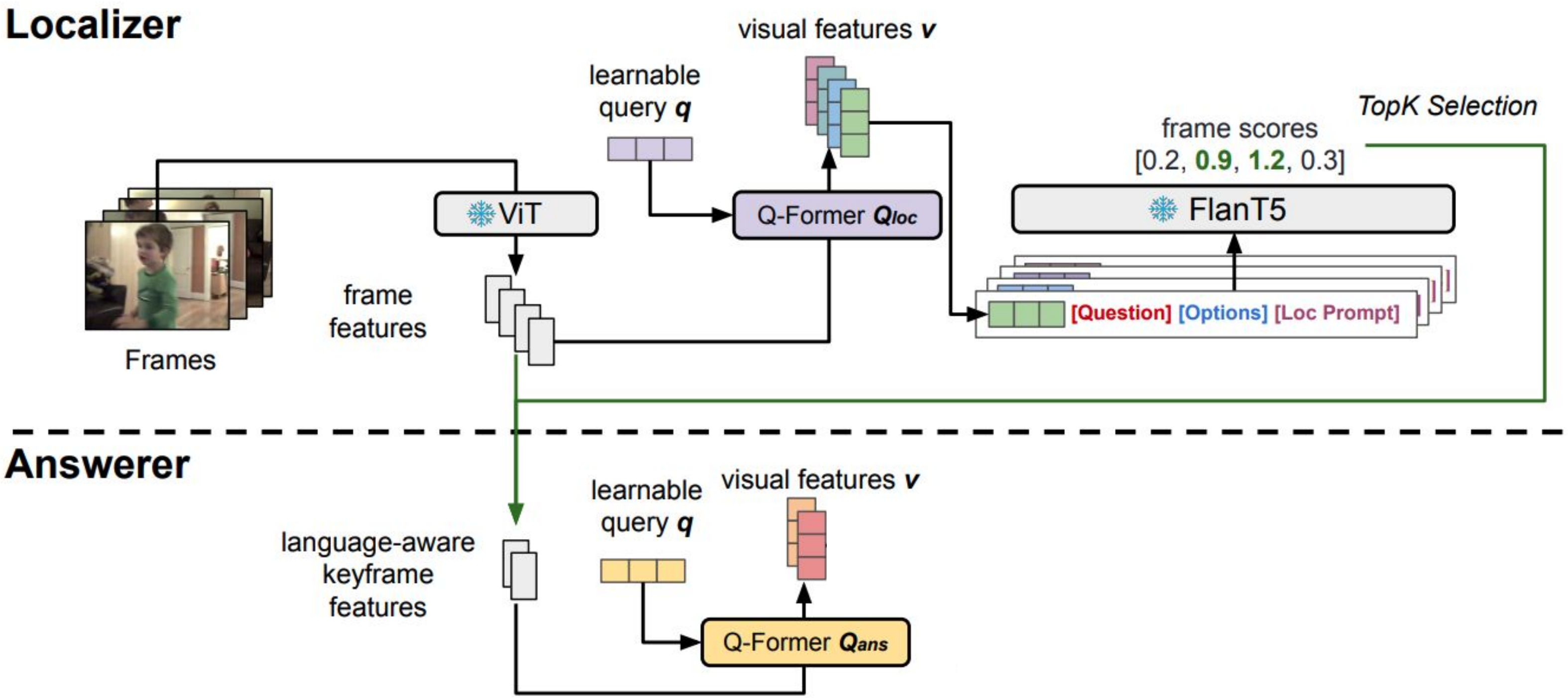
[Question]
what does the boy do after unwrapping the present?

[Options]
A: Show to the camera. B: Cry.
C: Relax his fingers. D: Walk alone.

[Loc Prompt]
Does the information within the frame provide necessary details to accurately answer the question?

[QA Prompt]
Considering information in frames, select the correct answer from the options.

SeViLA: Self-chained Video Localization-Answering



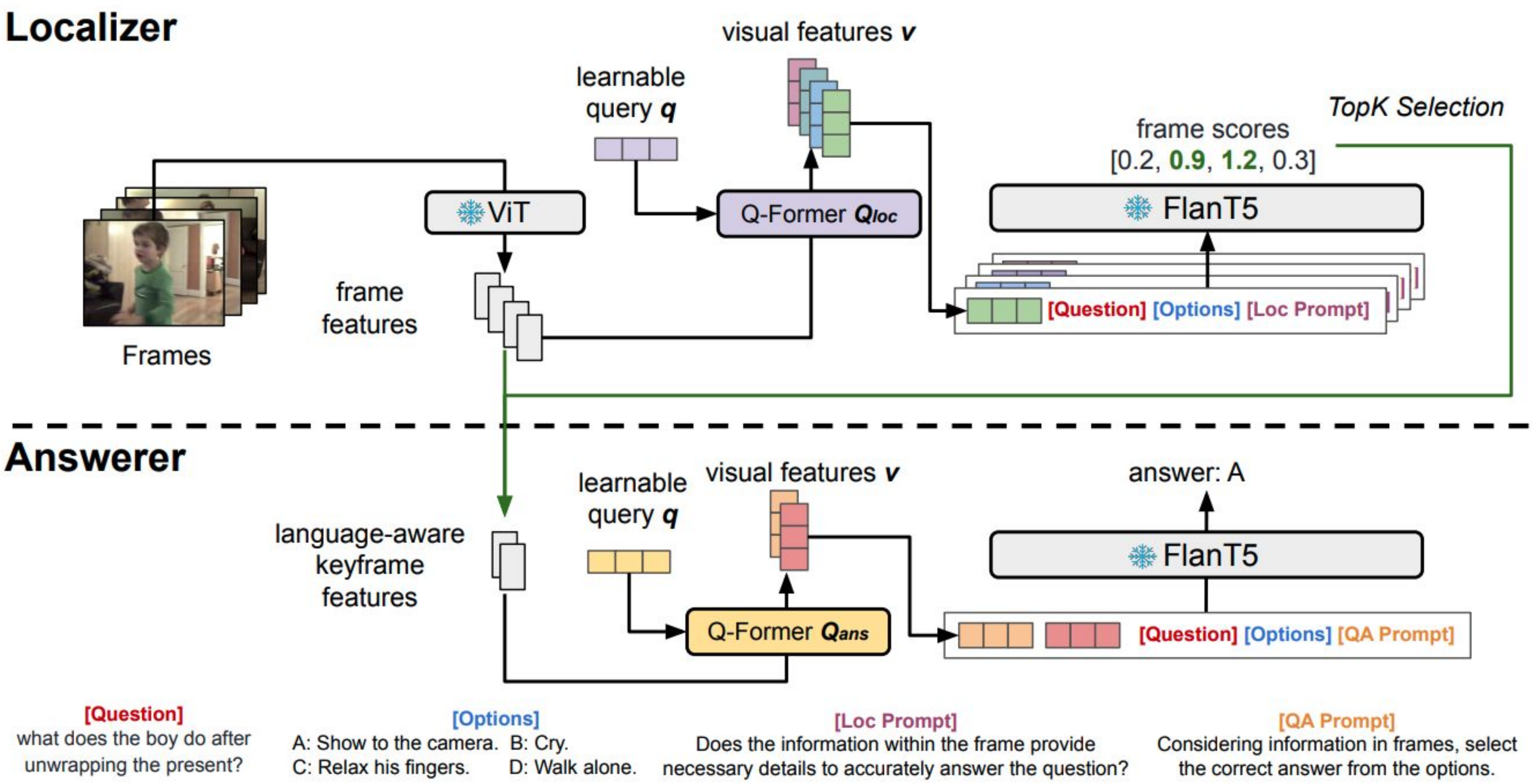
[Question]
what does the boy do after unwrapping the present?

[Options]
A: Show to the camera. B: Cry.
C: Relax his fingers. D: Walk alone.

[Loc Prompt]
Does the information within the frame provide necessary details to accurately answer the question?

[QA Prompt]
Considering information in frames, select the correct answer from the options.

SeViLA: Self-chained Video Localization-Answering



Training SeViLA

- Pre-training Localizer on Video Moment Retrieval dataset
- Fine-tuning Answerer on downstream tasks with keyframes from Localizer
- Self-refining Localizer with pseudo labels from Answerer

Datasets and Evaluation Metrics

- Taks & Dataset:
  - Video Question Answering (QA)
    - **NeXT-QA**: 52K questions with an an average length of 44s video
    - **STAR**: 60K questions with an average length of 12s video
    - **How2QA**: 44k questions with an average length of 60s video
    - **TVQA**: 152K questions with an average length of 76s video
  - Video Event Prediction (EP)
    - **VLEP**: 28K questions along with10K diverse video from TV Shows and YouTube Lifestyle Vlog video clips
  - Video Moment Retrieval
    - **QVHighlights**: 10K videos with a duration of 150s, 18K moments, and 10K queries

Fine-tuning Results on Video QA & Event Prediction

Model (# Frames)	NExT-QA				STAR					How2QA	TVQA	VLEP
	Tem.	Cau.	Des.	Avg.	Int.	Seq.	Pre.	Fea.	Avg.			
(w/ speech input or use dense frames)												
HERO (dense/1fps) [36]	-	-	-	-	-	-	-	-	-	73.8	73.6	-
JustAsk (20) [84]	51.4	49.6	63.1	52.3	-	-	-	-	-	84.4	-	-
FrozenBiLM (10) [85]	-	-	-	-	-	-	-	-	-	86.7	82.0	-
VidIL 4-shot (12) [72]	-	-	-	-	-	-	-	-	-	-	-	72.0
T+T (dense/1fps) [40]	-	-	-	-	-	-	-	-	-	92.4	-	-
T+T (+ASR, dense/1fps) [40]	-	-	-	-	-	-	-	-	-	93.2	-	-
Flamingo-80B 32-shot (30) [1]	-	-	-	-	-	-	-	-	42.2	-	-	-
FrozenBiLM (10) [85]	-	-	-	-	-	-	-	-	-	81.5	57.5	-
All-in-One (32) [67]	48.6	48.0	63.2	50.6	47.5	50.8	47.7	44.0	47.5	-	-	-
Temp[ATP] (32) [3]	49.3	48.6	65.0	51.5	50.6	52.8	49.3	40.6	48.3	-	-	-
VGT (32) [78]	55.0	52.2	64.0	55.0	-	-	-	-	44.2	-	-	-
MIST (32) [18]	56.6	54.6	66.9	57.1	55.5	54.2	54.2	44.4	51.1	-	-	-
VFC (32) [50]	53.3	57.6	72.8	58.6	-	-	-	-	-	-	-	-
CoVGT (32) [79]	57.4	58.8	69.3	60.0	-	-	-	-	45.9	-	-	-
SeViT$_{\text{FiD}}$ (10) [24]	-	-	-	60.6	-	-	-	-	-	-	-	-
HiTeA (16) [87]	58.3	62.4	75.6	63.1	-	-	-	-	-	-	-	-
InternVideo* (8) [71]	58.5	62.5	75.8	63.2	62.7	65.6	54.9	51.9	58.7	79.0	57.2	63.9
BLIP-2$^{\text{voting}}$ (4)	65.2	70.1	80.1	70.1	52.3	54.8	49.0	51.2	51.8	79.6	54.5	67.0
BLIP-2$^{\text{concat}}$ (ANSWERER) (4)	68.1	72.9	81.2	72.6	**65.4**	69.0	59.7	54.2	62.0	82.2	59.8	68.6
SEVILA† (32 → 4)	68.8	73.4	**83.5**	73.4	63.2	66.6	61.3	60.0	62.7	**83.7**	59.7	**69.0**
SEVILA (32 → 4)	**69.4**	**74.2**	81.3	**73.8**	63.7	**70.4**	**63.1**	**62.4**	**64.9**	83.6	**61.6**	68.9

+6.1%

+5.9%

Zero-shot Results on Video QA & Event Prediction

Model (# Frames)	NExT-QA Tem.	NExT-QA Cau.	NExT-QA Des.	NExT-QA Avg.	STAR Int.	STAR Seq.	STAR Pre.	STAR Fea.	STAR Avg.	How2QA	TVQA	VLEP
(w/ speech input or use dense frames)												
JustAsk (20) [84]	-	-	-	-	-	-	-	-	-	51.1	-	-
FrozenBiLM (10) [85]	-	-	-	-	-	-	-	-	-	58.4	59.2	-
ViperGPT (dense/1fps) [63]	-	-	-	60.0	-	-	-	-	-	-	-	-
Flamingo-80B (30) [1]	-	-	-	-	-	-	-	-	39.7	-	-	-
FrozenBiLM (10) [85]	-	-	-	-	-	-	-	-	-	41.9	29.7	-
VFC (32) [50]	45.4	51.6	64.1	51.5	-	-	-	-	-	-	-	-
InternVideo* (8) [71]	43.4	48.0	65.1	49.1	43.8	43.2	42.3	37.4	41.6	62.2	35.9	58.7
BLIP-2$^{\text{voting}}$ (4)	59.1	61.3	74.9	62.7	41.8	39.7	40.2	39.5	40.3	69.8	35.7	63.8
BLIP-2$^{\text{concat}}$ (ANSWERER) (4)	59.7	60.8	73.8	62.4	45.5	41.8	41.8	40.0	42.2	70.8	36.6	64.0
SEVILA† (32→4)	**61.3**	**61.5**	**75.6**	**63.6**	**48.3**	**45.0**	**44.4**	**40.8**	**44.6**	**72.3**	**38.2**	**64.4**

+2.1%

- SeViLA achieves the state-of-the-art in both fine-tuning and zero-shot setting on multiple datasets.

Video Moment Retrieval Results

Model	R1@0.5	R1@0.7	mAP
CAL [12]	25.4	11.5	9.8
XML [28]	41.8	30.3	32.1
Moment-DETR [29]	52.8	33.0	30.7
QD-DETR [47]	62.4	44.9	39.8
LOCALIZER (Ours)	54.5	36.5	32.3

- SeViLA-Localizer can also work as a standalone model for video moment retrieval task.

Qualitative Results for SeViLA

Thank you!

We Introduce **SeViLA**, a self-chained video-language framework, to handle temporal localization and QA in video with SoTA performance.

Paper: https://arxiv.org/abs/2305.06988

Demo Website: https://huggingface.co/spaces/SeViLA/SeViLA

Code: https://github.com/Yui010206/SeViLA

If you have any questions, please contact shoubin@cs.unc.edu