CVPR Nashville JUNE 11-15, 2025

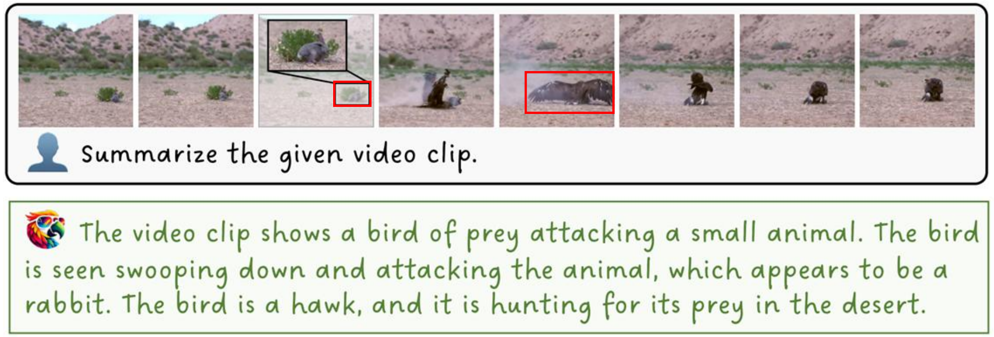
VideoEspresso: A Large-Scale Chain-of-Thought Dataset for Fine-Grained Video Reasoning via Core Frame Selection

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou,

Xiaojuan Qi, Yue Liao[†‡], Si Liu[†]

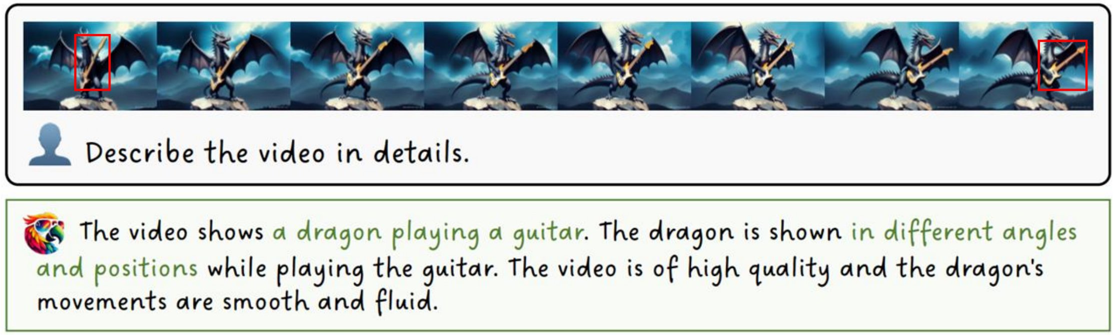
†Corresponding Author, ‡Project Leader

Background of Video Understanding



Summarize the given video clip.

The video clip shows a bird of prey attacking a small animal. The bird is seen swooping down and attacking the animal, which appears to be a rabbit. The bird is a hawk, and it is hunting for its prey in the desert.



Describe the video in details.

The video shows a dragon playing a guitar. The dragon is shown in different angles and positions while playing the guitar. The video is of high quality and the dragon's movements are smooth and fluid.

- Most existing video understanding focuses on **summarization** or detailed **description**
- Video reasoning often requires understanding **temporal logic** and **object interactions**
- Videos are **highly redundant —** many frames contribute little new information

[MVBench, CVPR 2024]

Insight of VideoEspresso

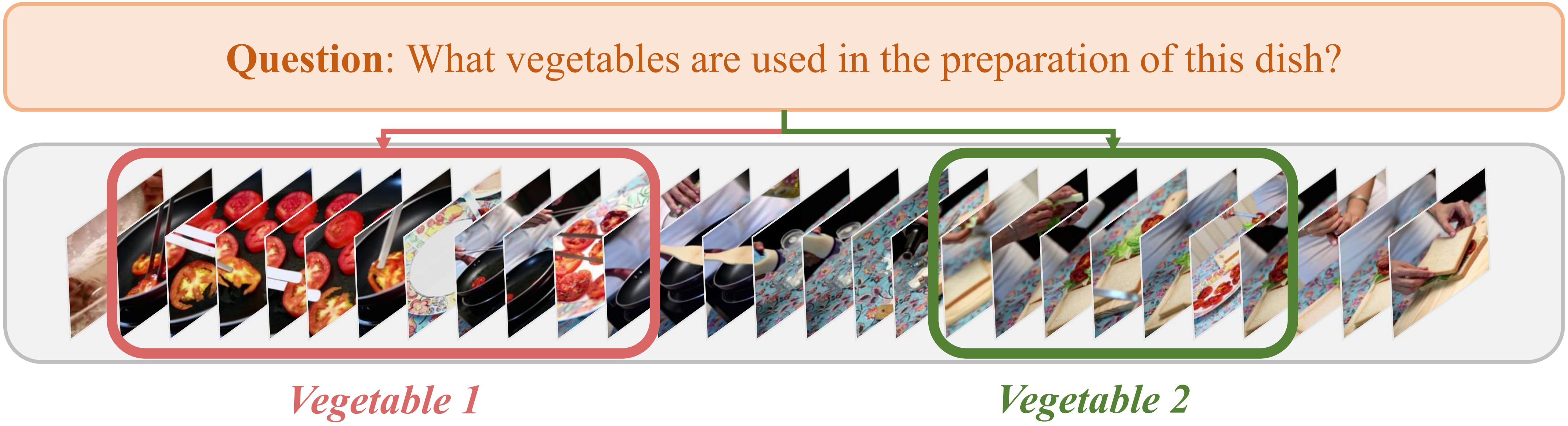
Question: What vegetables are used in the preparation of this dish?

200+ Frames Sampled at FPS=1

- Answering the question doesn’t require processing the entire sequence

[Youcook2, AAAI 2018]

Insight of VideoEspresso



Question: What vegetables are used in the preparation of this dish?

Vegetable 1

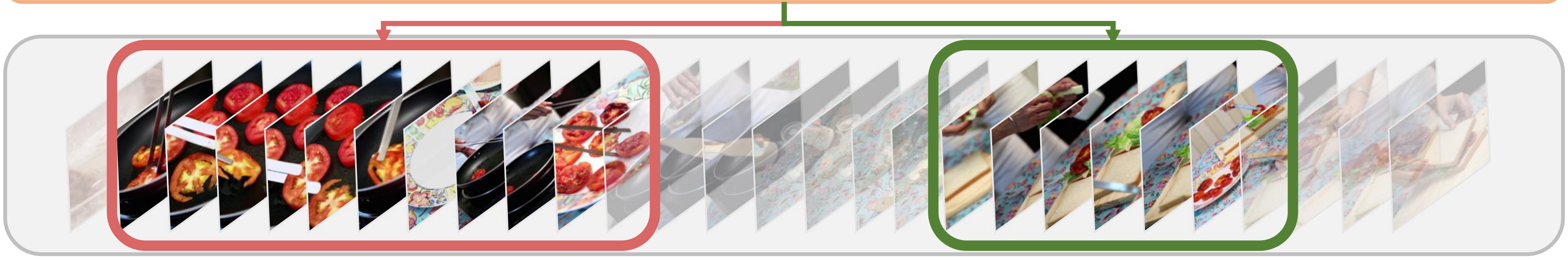
Vegetable 2

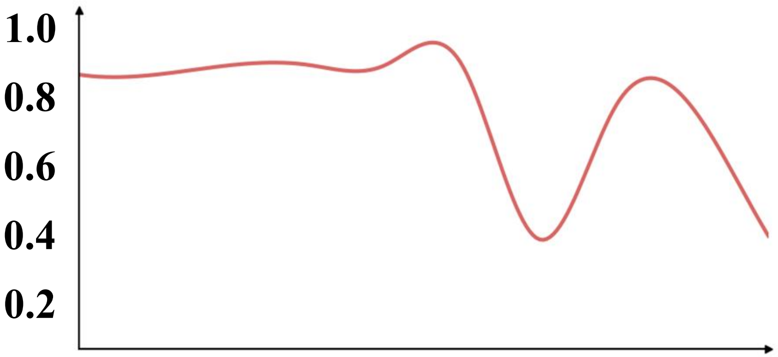
Still High Redundancy!

- Even key segments are not frame-efficient

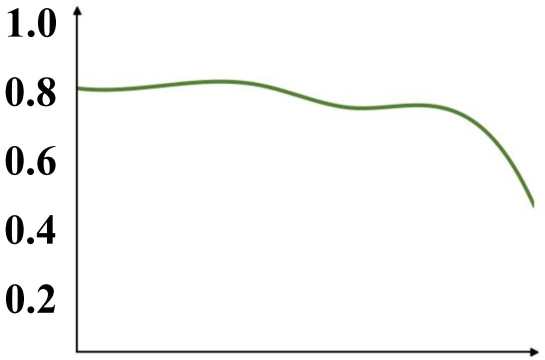
Insight of VideoEspresso

Question: What vegetables are used in the preparation of this dish?





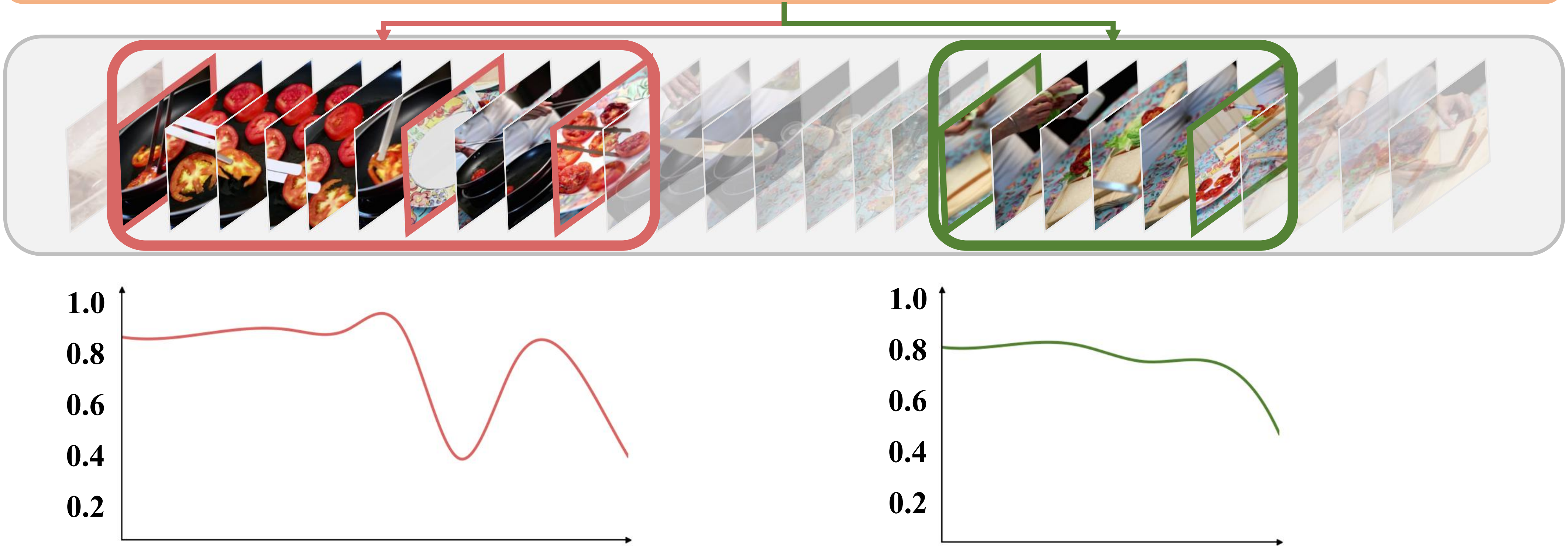
Intra-frame Similarity of segment-1



Intra-frame Similarity of segment-2

Insight of VideoEspresso

Question: What vegetables are used in the preparation of this dish?

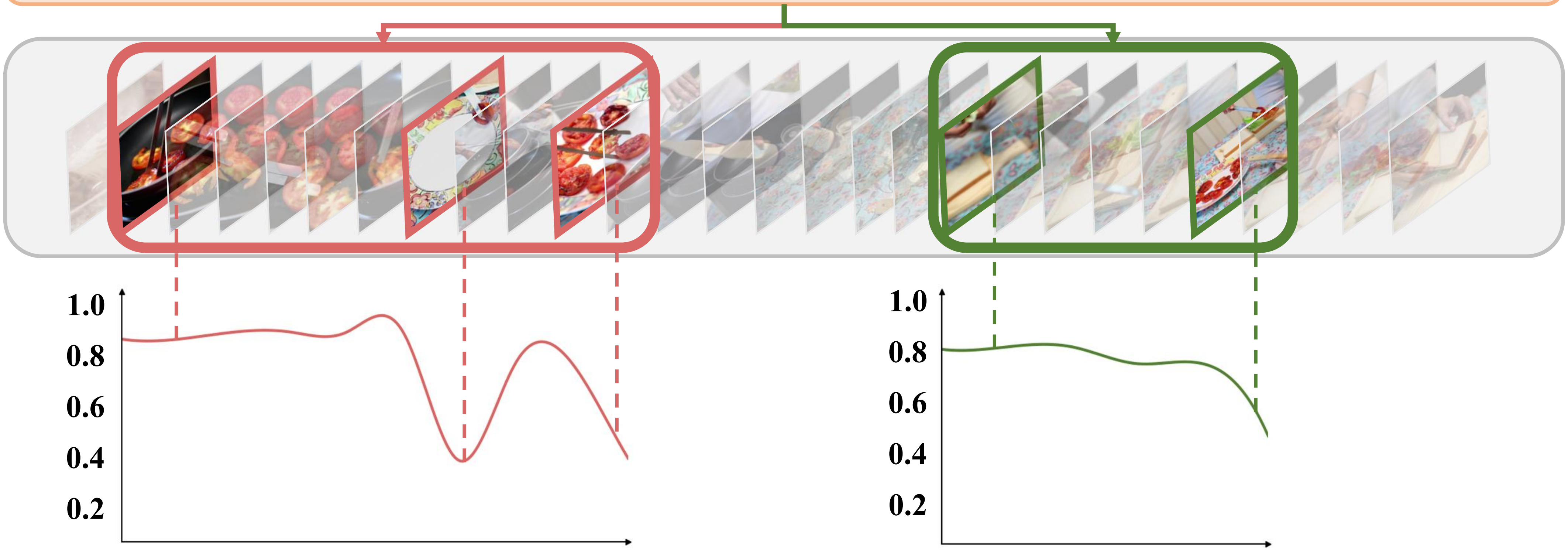


Intra-frame Similarity of segment-1

Intra-frame Similarity of segment-2

Insight of VideoEspresso

Question: What vegetables are used in the preparation of this dish?

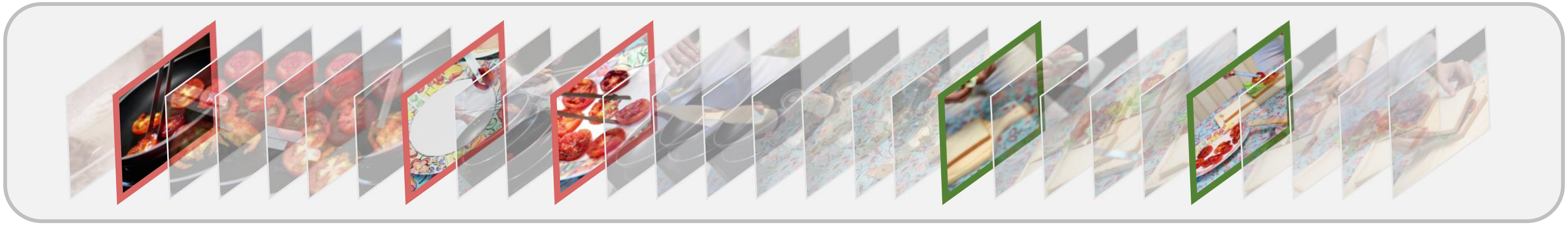


Intra-frame Similarity of segment-1

Intra-frame Similarity of segment-2

Insight of VideoEspresso

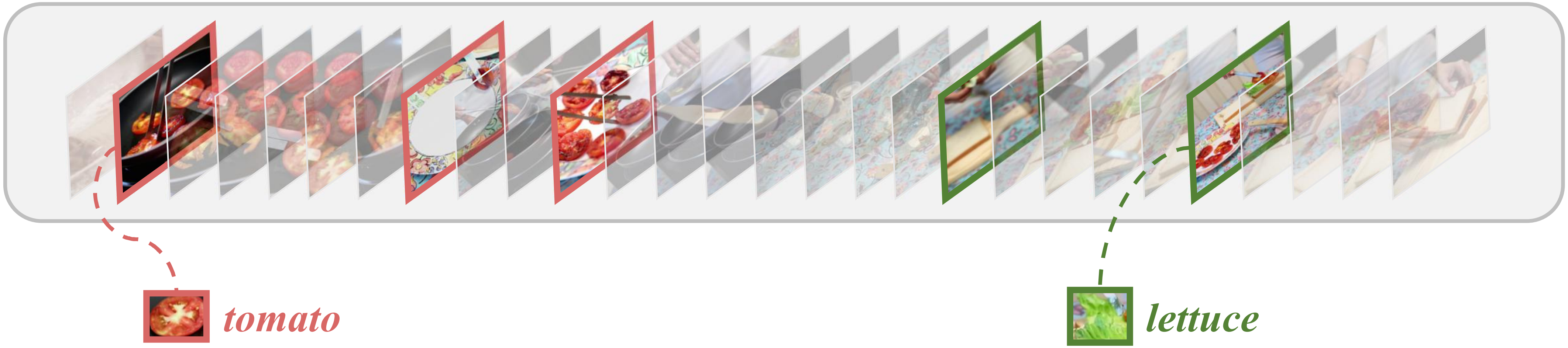
Question: What vegetables are used in the preparation of this dish?



***A Few Frames* are Enough!**

Insight of VideoEspresso

Question: What vegetables are used in the preparation of this dish?

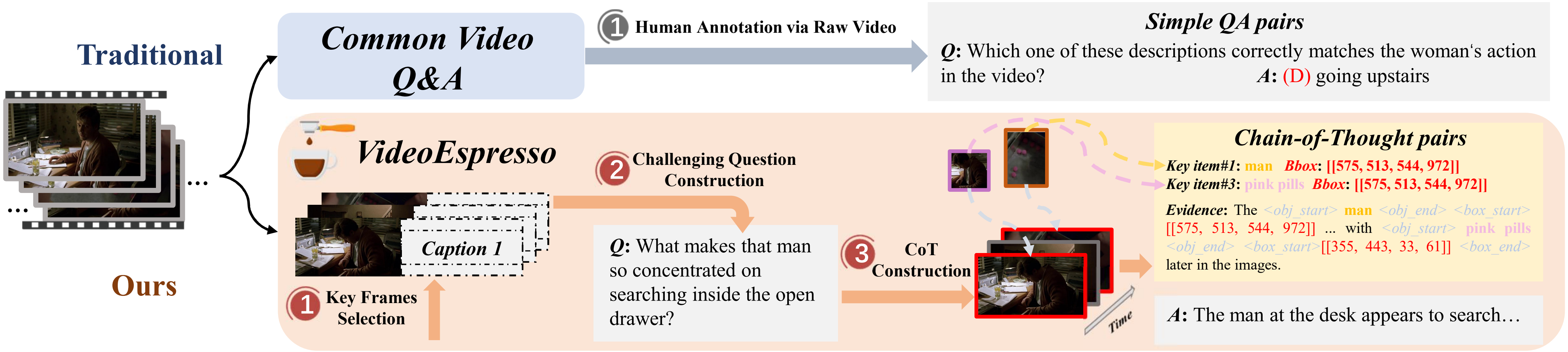


Spatial Location of Key Items helps Reason Better

But how to scale up data automatically?

Leveraging Core Frames and Key Items is key to reasoning

Rethinking VideoQA Construction



Traditional

Common Video Q&A

① Human Annotation via Raw Video

Simple QA pairs

***Q*:** Which one of these descriptions correctly matches the woman's action in the video? ***A*:** (D) going upstairs

Ours

VideoEspresso

① Key Frames Selection

② Challenging Question Construction

***Q*:** What makes that man so concentrated on searching inside the open drawer?

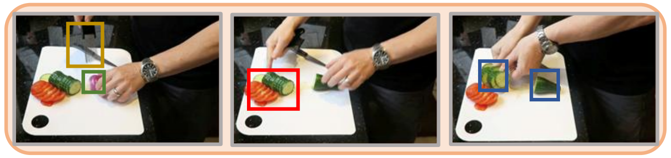
③ CoT Construction

Chain-of-Thought pairs

***Key item#1*:** **man** ***Bbox*:** **[[575, 513, 544, 972]]**
***Key item#3*:** **pink pills** ***Bbox*:** **[[575, 513, 544, 972]]**

***Evidence*:** The <obj_start> **man** <obj_end> <box_start> [[575, 513, 544, 972]] ... with <obj_start> **pink pills** <obj_end> <box_start>[[355, 443, 33, 61]] <box_end> later in the images.

***A*:** The man at the desk appears to search…

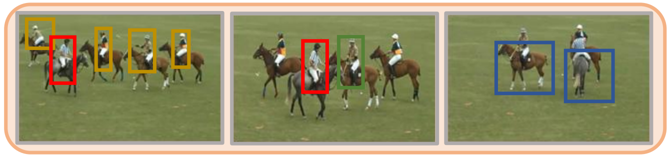


Question: Based on the sequence of actions described, what logical steps can be inferred about the purpose and functionality of the culinary tool used by the person?

***Key Items #1*:** **tool**, **onion**	***Bbox*:** **[[477, 517, 125, 137]…]**
***Key Items #2*:** **food**	***Bbox*:** **[[314, 489, 286, 980]]**
***Key Items #3*:** **cucumber**	***Bbox*:** **[[384, 498, 418, 977]…]**

Evidence: The sequence of actions demonstrates that the tool in **key frame 1** [[477, 517, 125, 137]] …

Answer: The culinary tool is designed to hold vegetables securely in place while they are being sliced, ensuring...

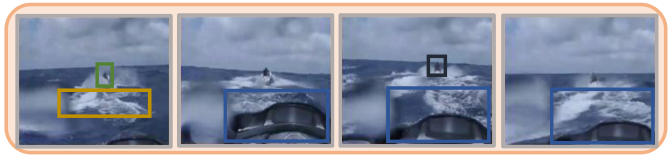


Question: Describe the interactions between the horse jockeys or players and the officials as observed in multiple sequences of the video.

***Key Items #1*:** **jockeys**,**referee**	***Bbox*:** **[[427, 257, 154, 274]…]**
***Key Items #2*:** **referee**, **men**	***Bbox*:** **[[712, 274, 56, 243]…]**
***Key Items #3*:** **horse**	***Bbox*:** **[[611, 399, 68, 288]…]**

Evidence: The horse jockeys in **key frame 1** [[427, 257, 54, 274]] interact with an…

Answer: The horse jockeys or players interact with officials or a referee, as one jockey says something to an official…



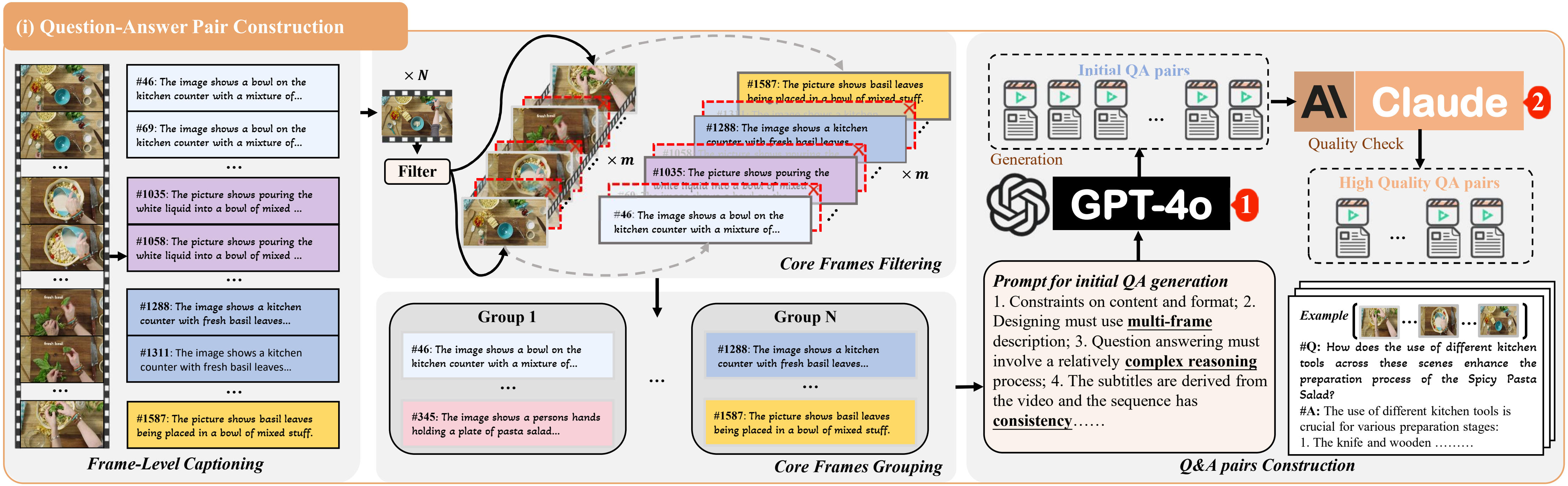
Question: Please infer the type of sports the person in the video is performing, and infer the relationship between the two objects in front and behind.

***Key Items #1*:** **spray**, **launch**	***Bbox*:** **[[293, 394, 203, 440]…]**
***Key Items #2*:** **dashboard**	***Bbox*:** **[[290, 385, 209, 399]]**
***Key Items #3*:** **man**, **dashboard**	***Bbox*:** **[[295, 392, 221, 422]…]**
***Key Items #4*:** **dashboard**	***Bbox*:** **[[184, 284, 938, 692]]**

Evidence: The person in **key frame 1** is riding a launch [[293,394,203,440]] and splashing huge waves…

Answer: Two people are riding a speedboat on the sea and creating huge waves. The person behind is taking a photo…

Detailed Pipeline for Multi-Frame Reasoning QA Construction



#46: The image shows a bowl on the kitchen counter with a mixture of...

#69: The image shows a bowl on the kitchen counter with a mixture of...

#1035: The picture shows pouring the white liquid into a bowl of mixed ...

#1058: The picture shows pouring the white liquid into a bowl of mixed ...

#1288: The image shows a kitchen counter with fresh basil leaves...

#1311: The image shows a kitchen counter with fresh basil leaves...

#1587: The picture shows basil leaves being placed in a bowl of mixed stuff.

#345: The image shows a persons hands holding a plate of pasta salad...

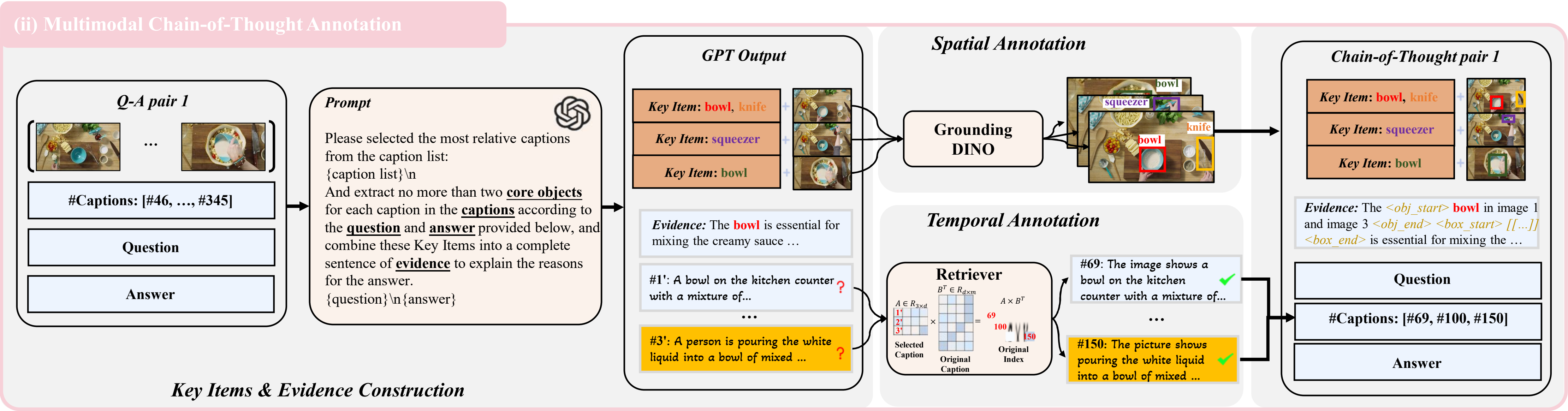
Prompt for initial QA generation
1. Constraints on content and format; 2. Designing must use **multi-frame** description; 3. Question answering must involve a relatively **complex reasoning** process; 4. The subtitles are derived from the video and the sequence has **consistency**……

#Q: How does the use of different kitchen tools across these scenes enhance the preparation process of the Spicy Pasta Salad?
#A: The use of different kitchen tools is crucial for various preparation stages:
1. The knife and wooden ………

- Select & group **core frames** for reasoning
- Use GPT-4o to generate **multi-frame** QA pairs
- Claude checks for consistency and quality

CVPR Nashville JUNE 11-15, 2025

Detailed Pipeline for Multimodal Chain-of-Thought Annotation



(ii) Multimodal Chain-of-Thought Annotation

Q-A pair 1

…

#Captions: [#46, …, #345]

Question

Answer

Prompt

Please selected the most relative captions from the caption list:
{caption list}\n
And extract no more than two **core objects** for each caption in the **captions** according to the **question** and **answer** provided below, and combine these Key Items into a complete sentence of **evidence** to explain the reasons for the answer.
{question}\n{answer}

Key Items & Evidence Construction

GPT Output

Key Item: bowl, knife

Key Item: squeezer

Key Item: bowl

Evidence: The **bowl** is essential for mixing the creamy sauce …

#1': A bowl on the kitchen counter with a mixture of... ?

…

#3': A person is pouring the white liquid into a bowl of mixed … ?

Spatial Annotation

Grounding DINO

bowl

squeezer

knife

bowl

Temporal Annotation

Retriever

$A \in R_{3\times d}$

1'

2'

3'

Selected Caption

×

$B^T \in R_{d\times m}$

Original Caption

=

$A \times B^T$

69

100

150

Original Index

#69: The image shows a bowl on the kitchen counter with a mixture of... ✓

…

#150: The picture shows pouring the white liquid into a bowl of mixed … ✓

Chain-of-Thought pair 1

Key Item: bowl, knife

Key Item: squeezer

Key Item: bowl

Evidence: The *<obj_start>* **bowl** in image 1 and image 3 *<obj_end> <box_start> [[...]] <box_end>* is essential for mixing the …

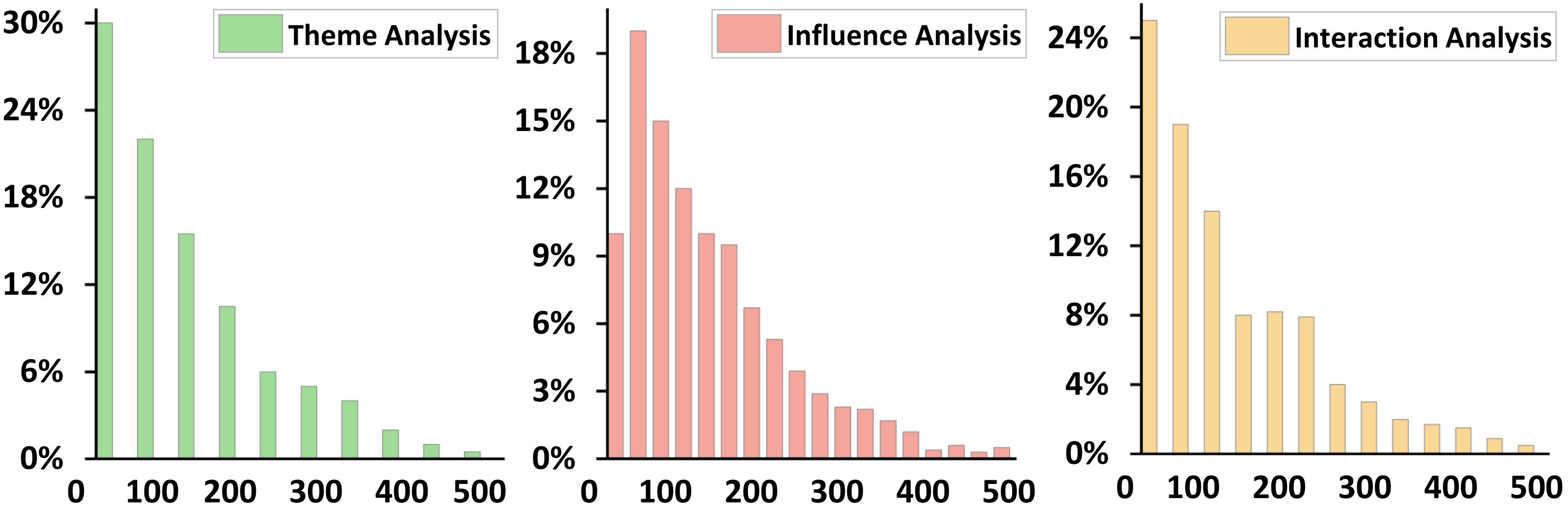
Question

#Captions: [#69, #100, #150]

Answer

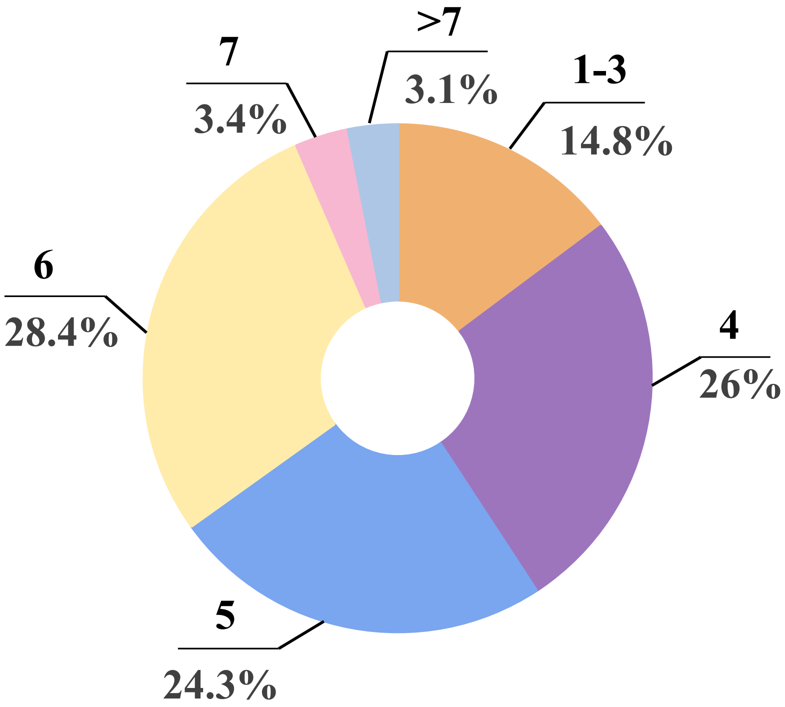
- GPT extracts **key items** & generates reasoning **evidence** with given QA pairs.
- **Spatial & temporal** annotation ensure multimodal alignment.
- Outputs interpretable, high-quality CoT annotations.

Distribution of Distance between adjacent core frames

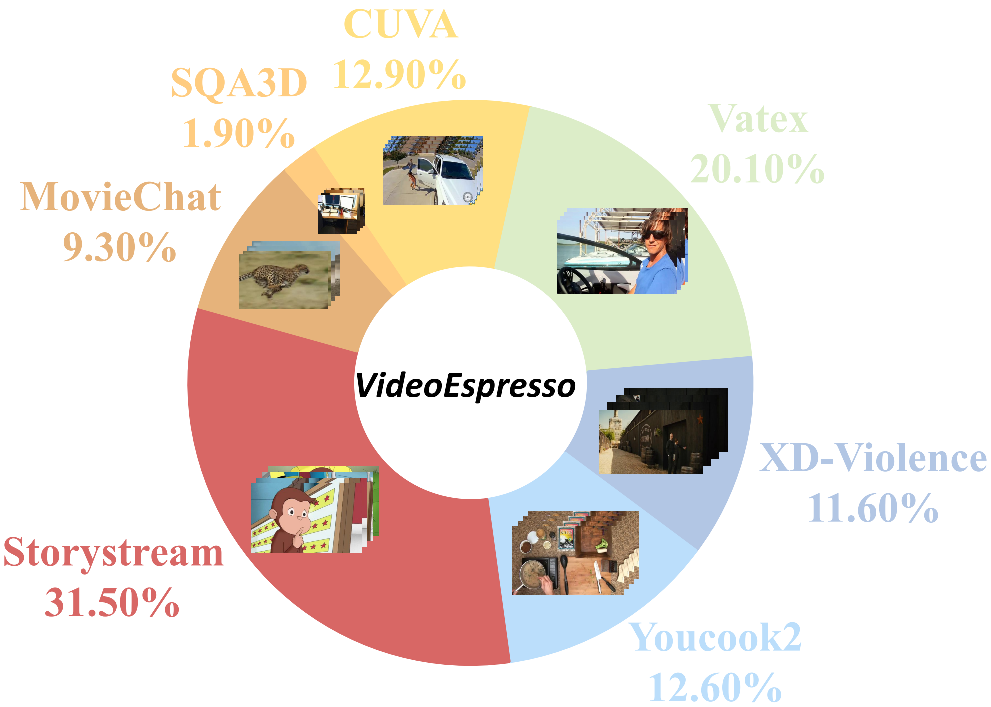


- Core frame **spacing differs** across reasoning types
- Sampling strategy should be **task-aware**
- A uniform frame sampling strategy is suboptimal

Details Data of VideoEspresso



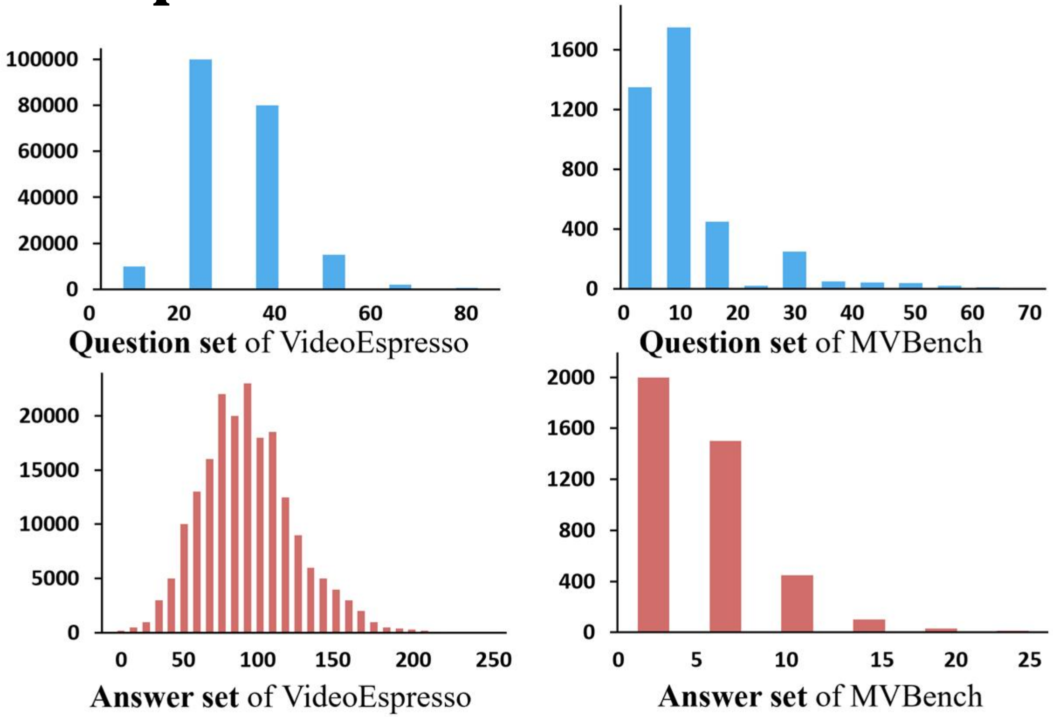
Number of key items



Data Source

- More than **200K** QA pairs with rich CoT annotation
- Most questions involve **4–6** key items, enabling focused reasoning
- The dataset is built from **diverse** sources, covering **various** domains

Comparison with MVBench



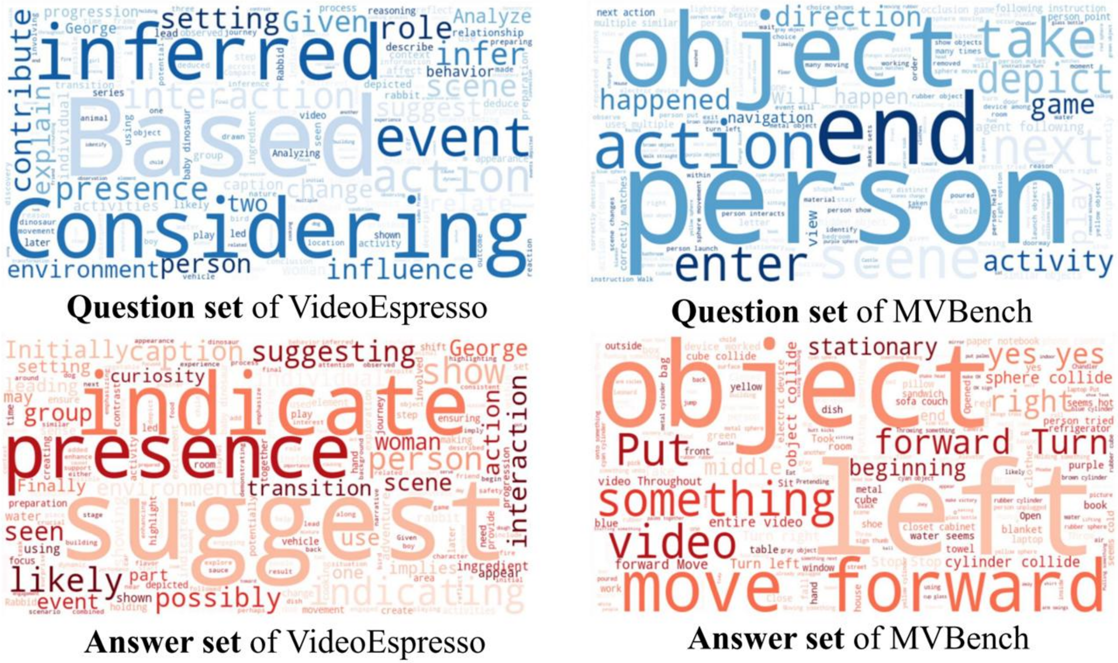
Question set of VideoEspresso

Question set of MVBench

Answer set of VideoEspresso

Answer set of MVBench

Comparison of **Token Length**



Question set of VideoEspresso

Question set of MVBench

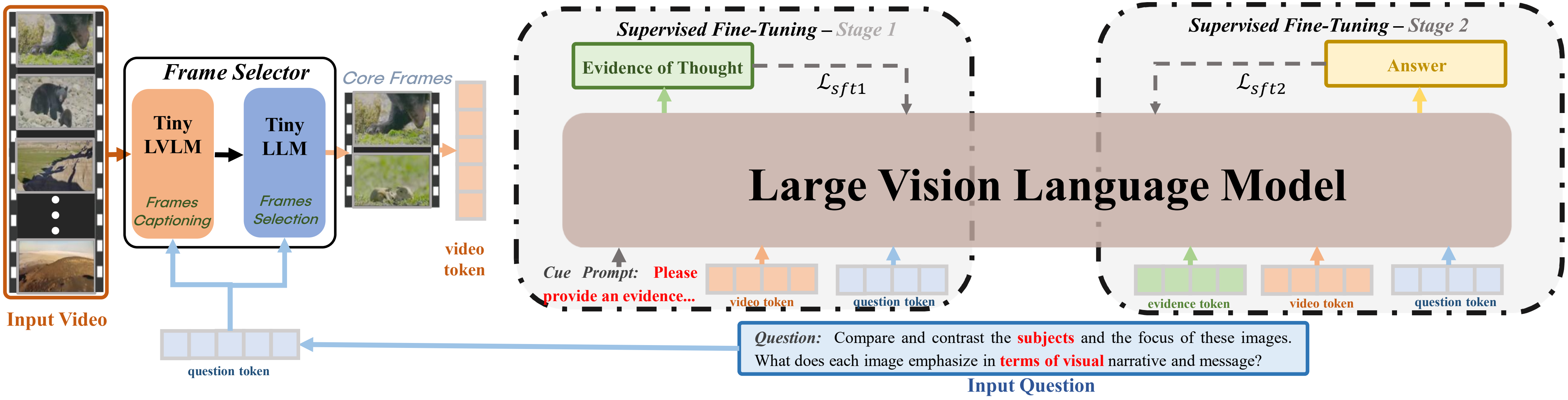
Answer set of VideoEspresso

Answer set of MVBench

Comparison of **Word Cloud**

- VideoEspresso uses **longer**, more **descriptive** questions/answers
- Focuses more on reasoning verbs like *consider, indicate, suggest*
- MVBench centers on **simpler** object/action terms

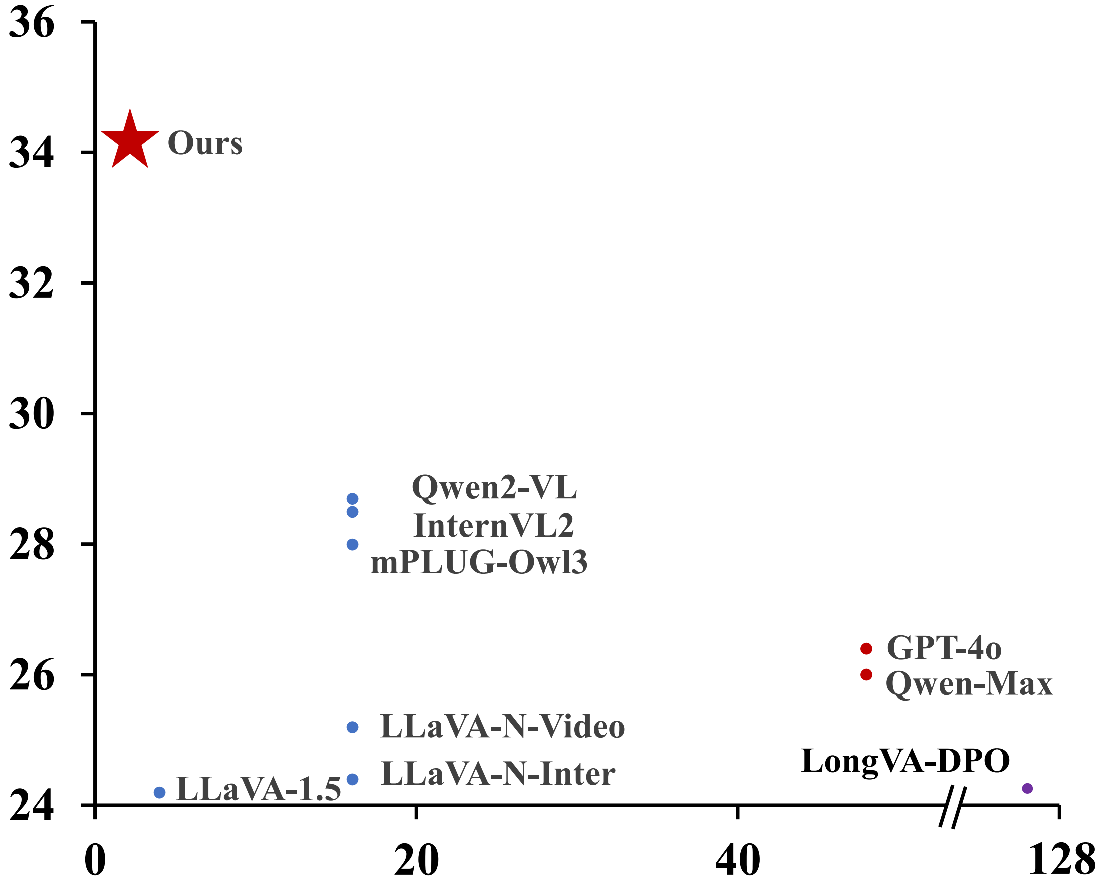
Hybrid LVLMs Collaboration for VideoQA



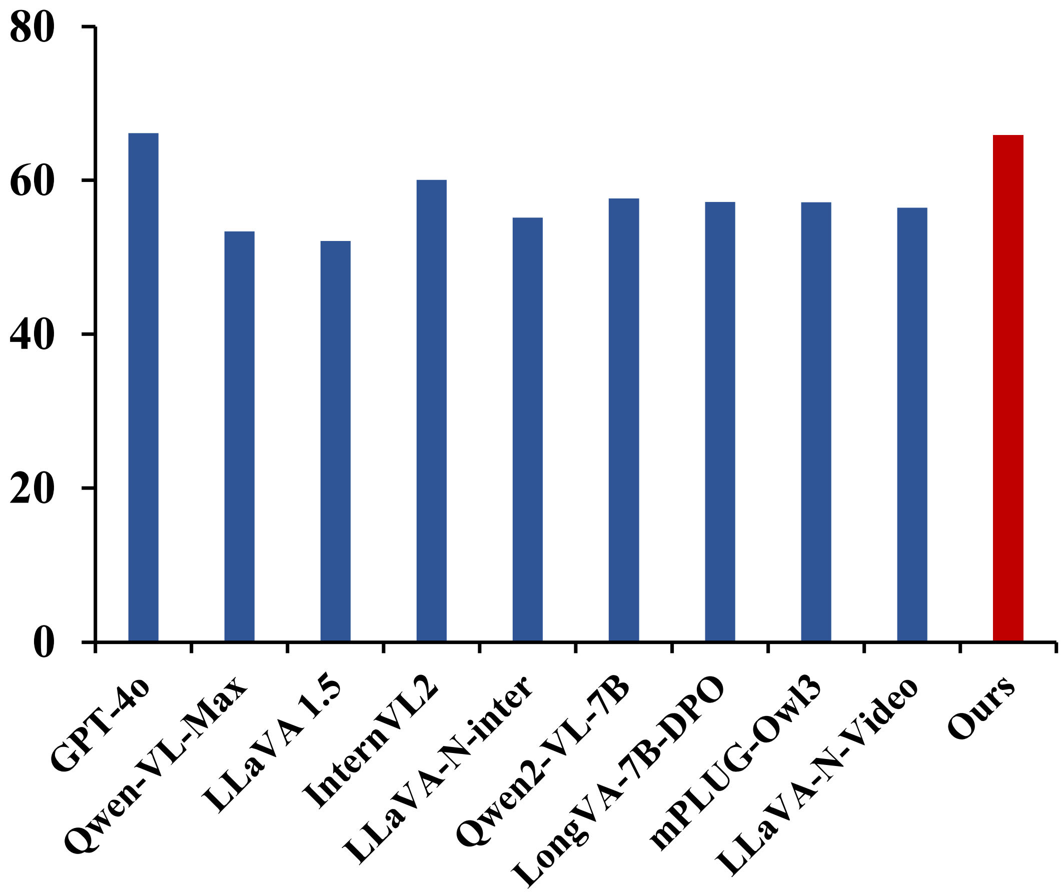
Input Video
Frame Selector
Tiny LVLM
Frames Captioning
Tiny LLM
Frames Selection
Core Frames
video token
question token
Supervised Fine-Tuning – Stage 1
Evidence of Thought
$\mathcal{L}_{sft1}$
Cue Prompt: Please provide an evidence...
video token
question token
Large Vision Language Model
Supervised Fine-Tuning – Stage 2
Answer
$\mathcal{L}_{sft2}$
evidence token
video token
question token
Question: Compare and contrast the subjects and the focus of these images. What does each image emphasize in terms of visual narrative and message?
Input Question

- Tiny model selects **informative frames** from video
- Large model focuses on **reasoning** with selected evidence
- Efficient pipeline without sacrificing answer quality

Main Results

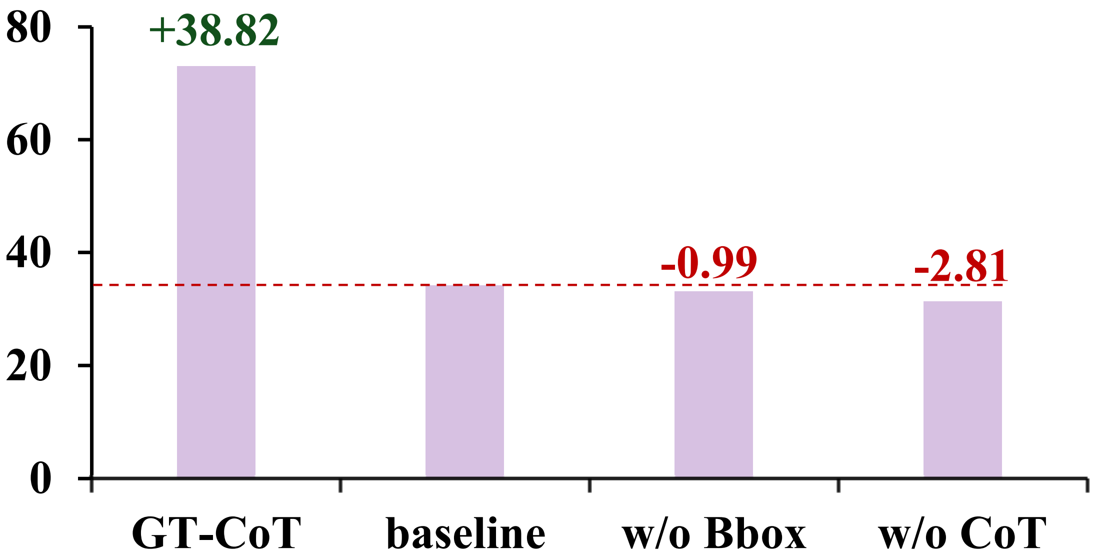


Result on objective Benchmark

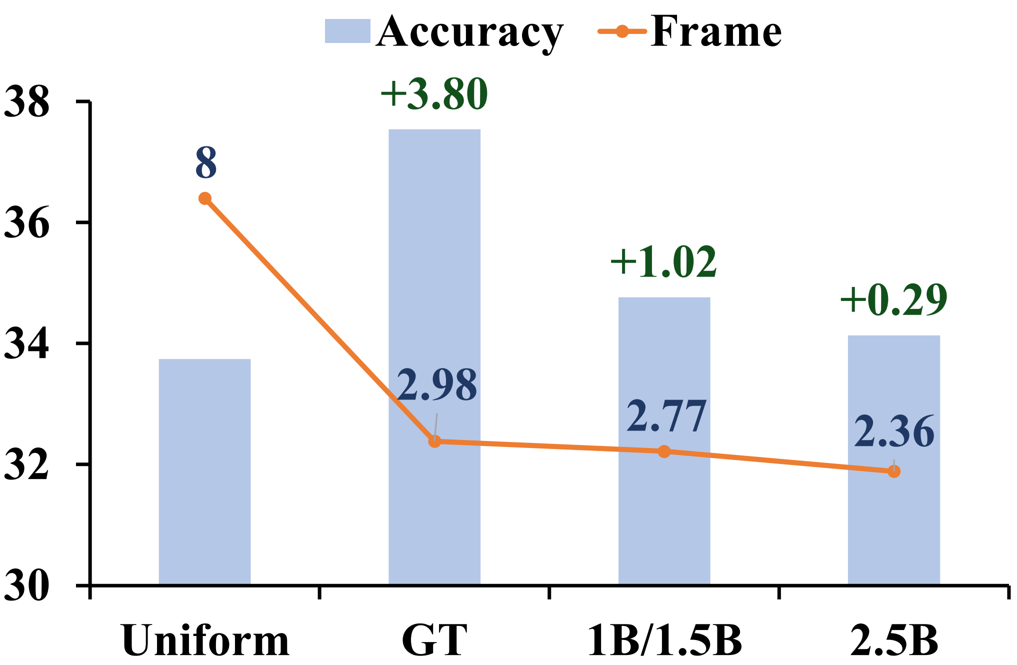


Result on subjective Benchmark

Ablation Study



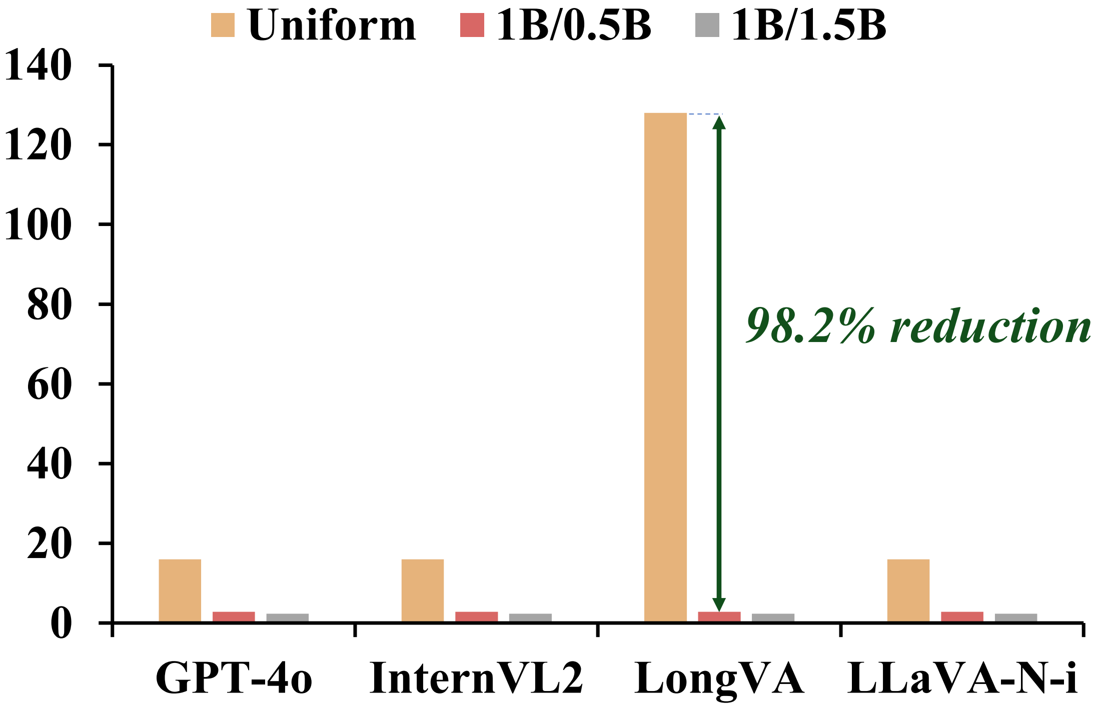
Ablation Studies on different **CoT strategies**



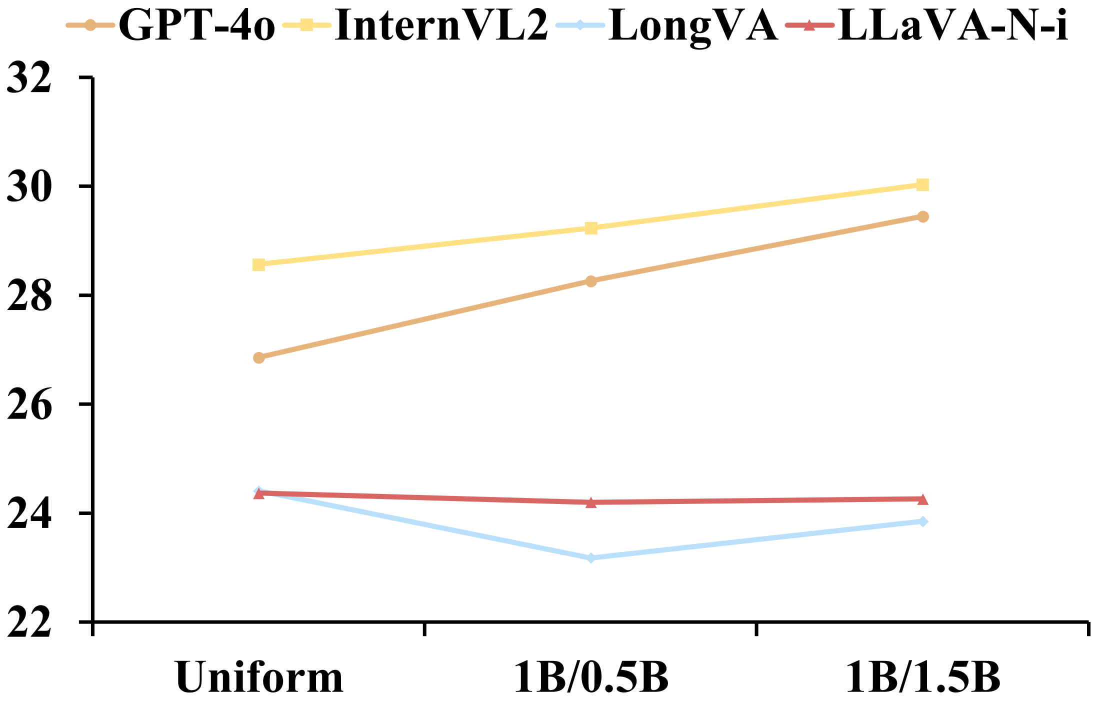
Ablation studies on **Selector**

- CoT reasoning significantly **boosts** performance
- Spatial grounding (Bbox) **contributes** essential object-level cues.
- Learned selectors reduce frame count while maintaining accuracy.

Evaluations Results with Selector Adoption



Impact of Selector on **Frame Utilization** across Models



Impact of Selector on **Accuracy Performance** across Models

- **Significantly reduces input frames** (up to 98.2% fewer) with lightweight selector.
- **Maintains or even improves accuracy**, despite fewer frames being used.

VideoEspresso: A Large-Scale Chain-of-Thought Dataset for Fine-Grained Video Reasoning via Core Frame Selection

VideoEspresso = Large-Scale + CoT + Reasoning + Core Frame

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou,

Xiaojuan Qi, Yue Liao†‡, Si Liu†

†Corresponding Author, ‡Project Leader

https://github.com/hshjerry/VideoEspresso

paper

code