清華大學
Tsinghua University

Research Proposal:
Enhancing MovieChat with Keyframe Extraction for Efficient Long-Term Video Understanding

Naifan Zhang

2024

Research Goal

Improving memory efficiency and accuracy in long video understanding by enhancing MovieChat's memory management through keyframe extraction.

Research Background

Video understanding

aiming at interpreting video content

Various tasks

- Video Question Answering
- Video Classification
- Video Scene Segmentation
- …

Limitation:

1. **LLMs max input length**
2. **GPU memory consumption**

Q#1 at 4500

Q#2 at 10000

Question #1 What is this clip talking about now?

Answer #1 It shows a young man standing in a dugout on a black background. He is then seen kneeling on the field during a football game, followed by two young men in a dugout. Next, we see a young man standing on the field. And finally, we see a group of young men standing on the field and watching the game. The video also features football players and fans, as well as a football field, field, and dugout.

Question #2 Has there been a black speaker in the video so far?

Answer #2 Yes, there is a man in a red shirt standing on a street with a dark blue sky in the background, and a man with a tie standing on a street with a red car in the background. There is also a man in a black suit and another man in a white jacket standing in front of a building with a street light.

Question #3 Which movie sequence lasts the longest?

Answer #3 The video shows the scene of a dark night with a football field and a group of people playing football.

Content

Long video understanding

Reduce the computing requirements needed to analyze long videos while maintaining good accuracy

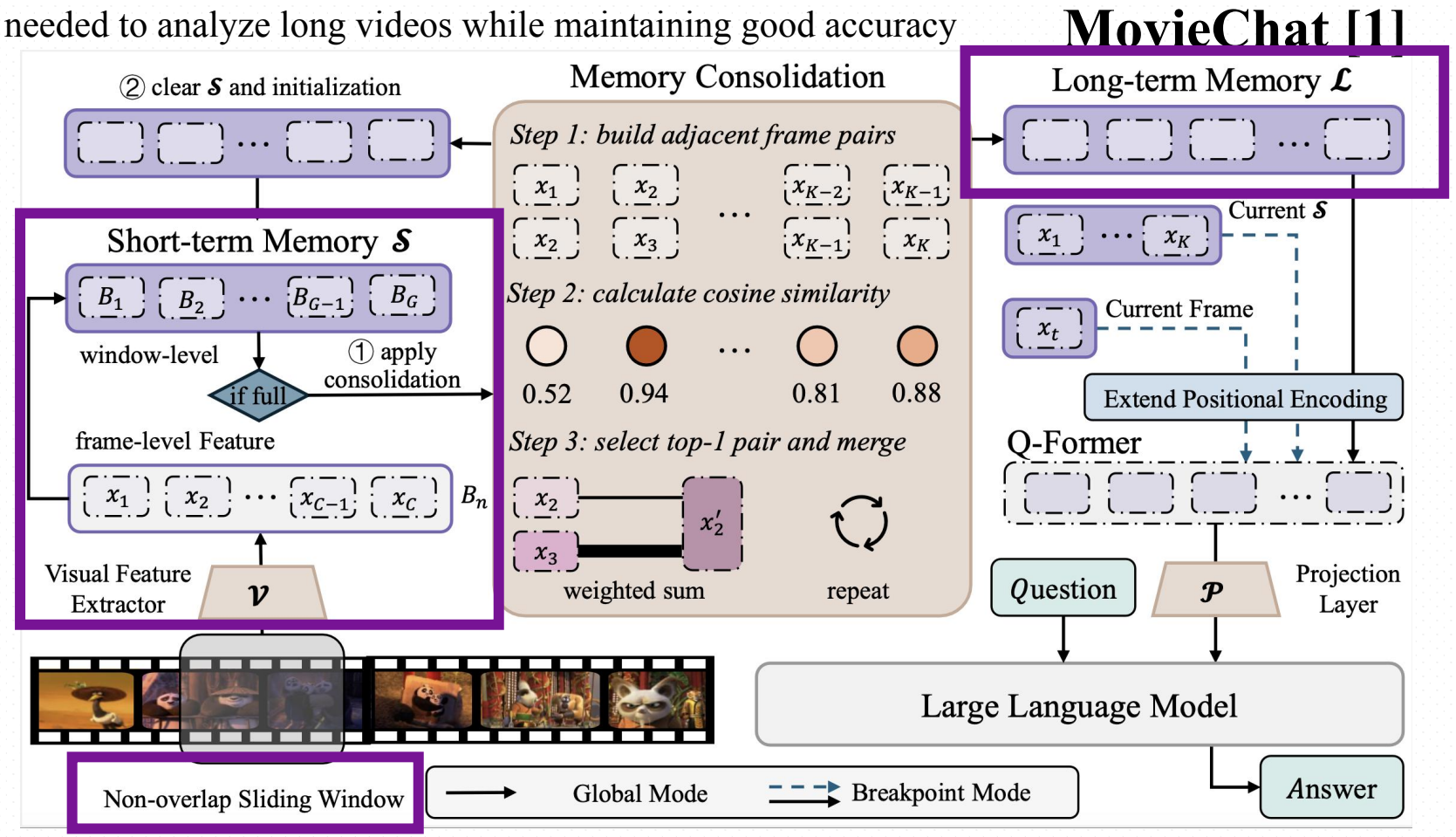
Methods

1. Sampling
2. Aggregation
3. Memory Bank

...

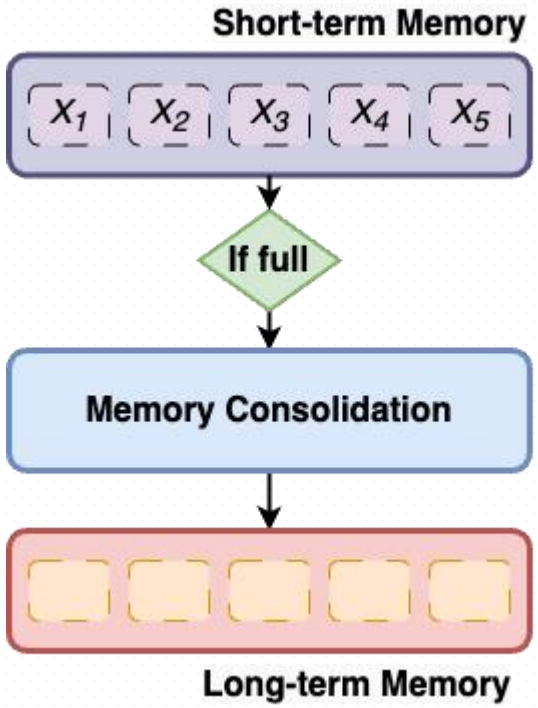
1. Short-term Memory
   Similar to computer's RAM

2. Long-term Memory
   Similar to computer's hard drive

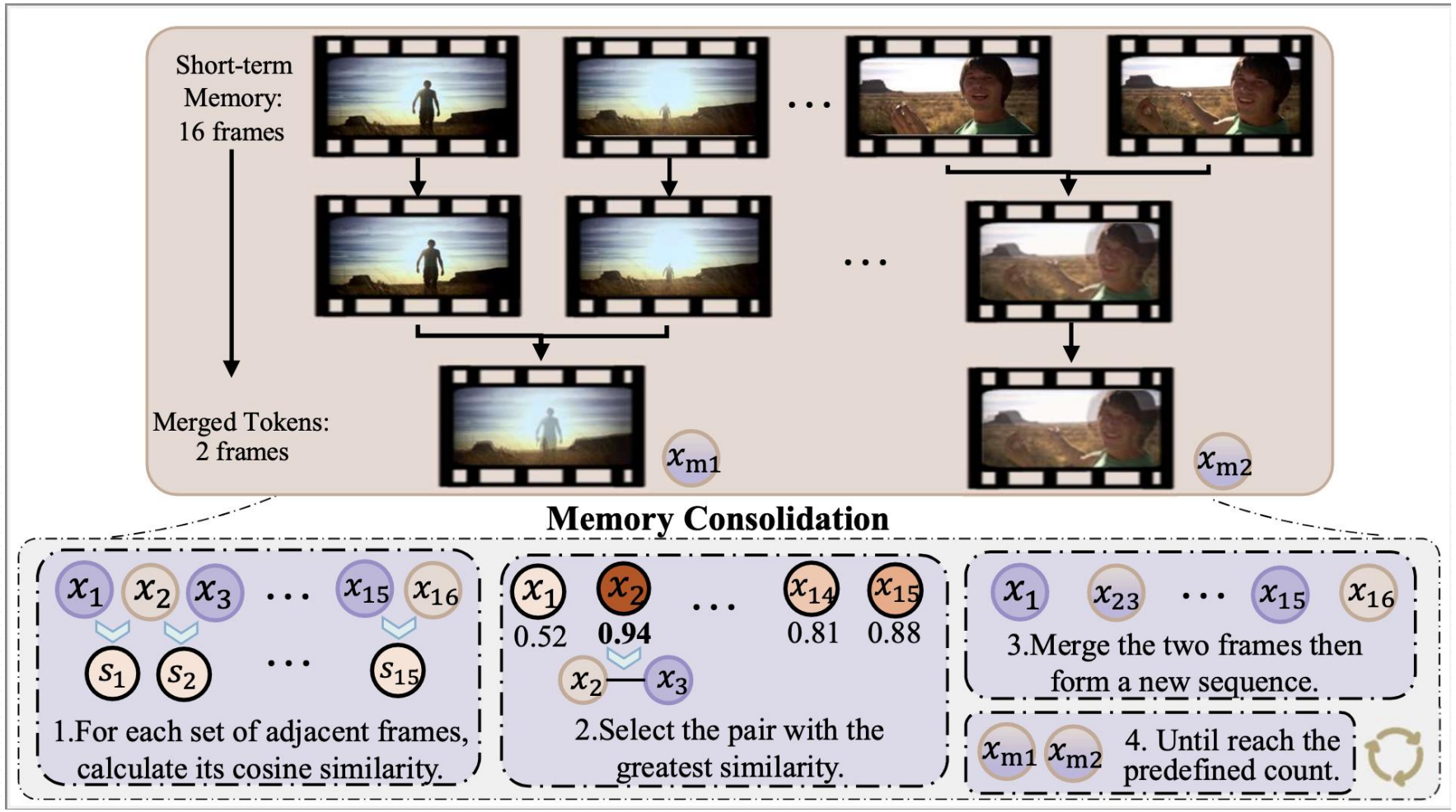


[1] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18221–18232, 2024.

MovieChat



Memory Consolidation



MovieChat

1. Breakpoint Mode

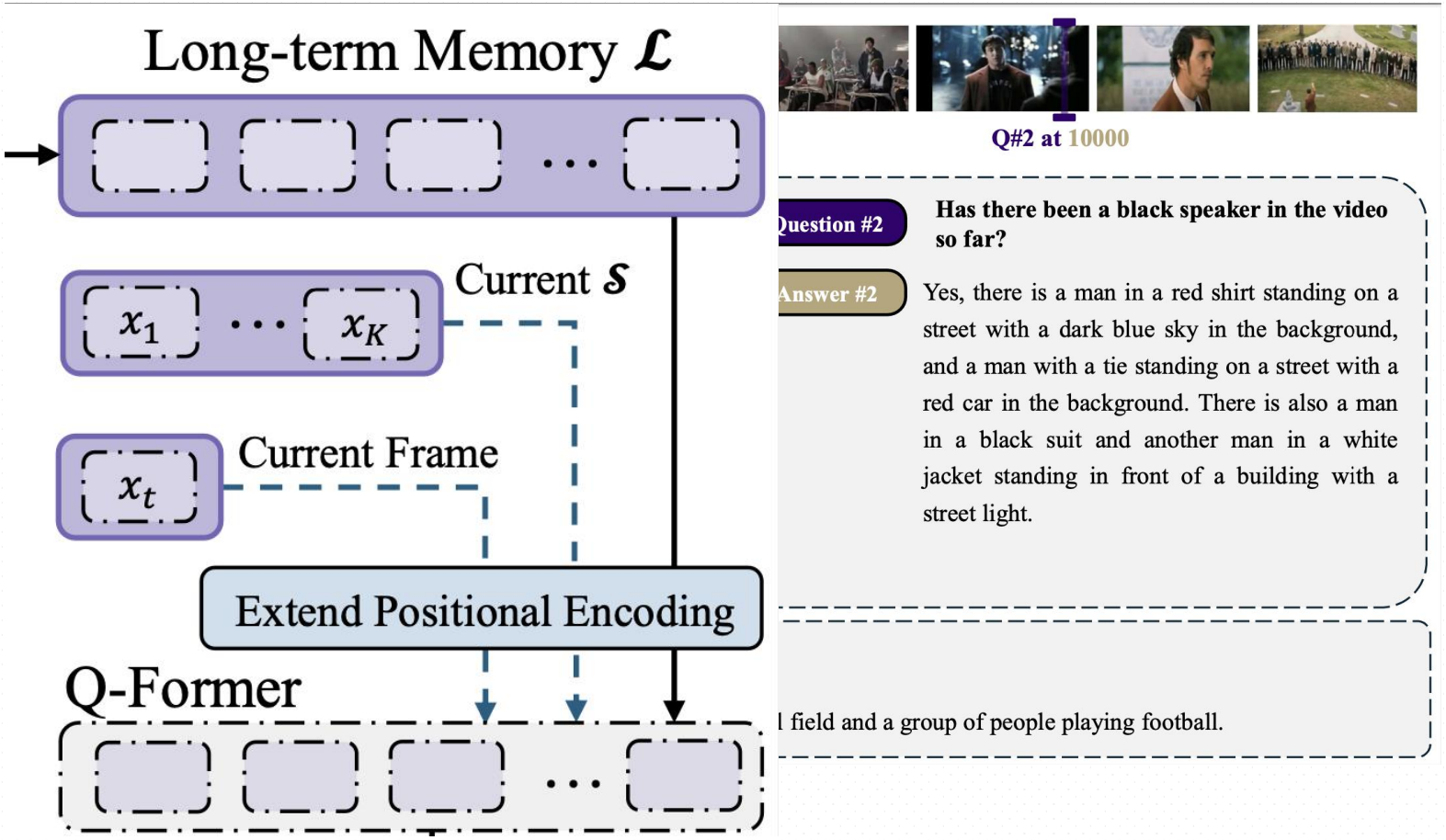
Short-term memory, Long-term memory, and Current video frames

2. Global Mode

Long-term memory

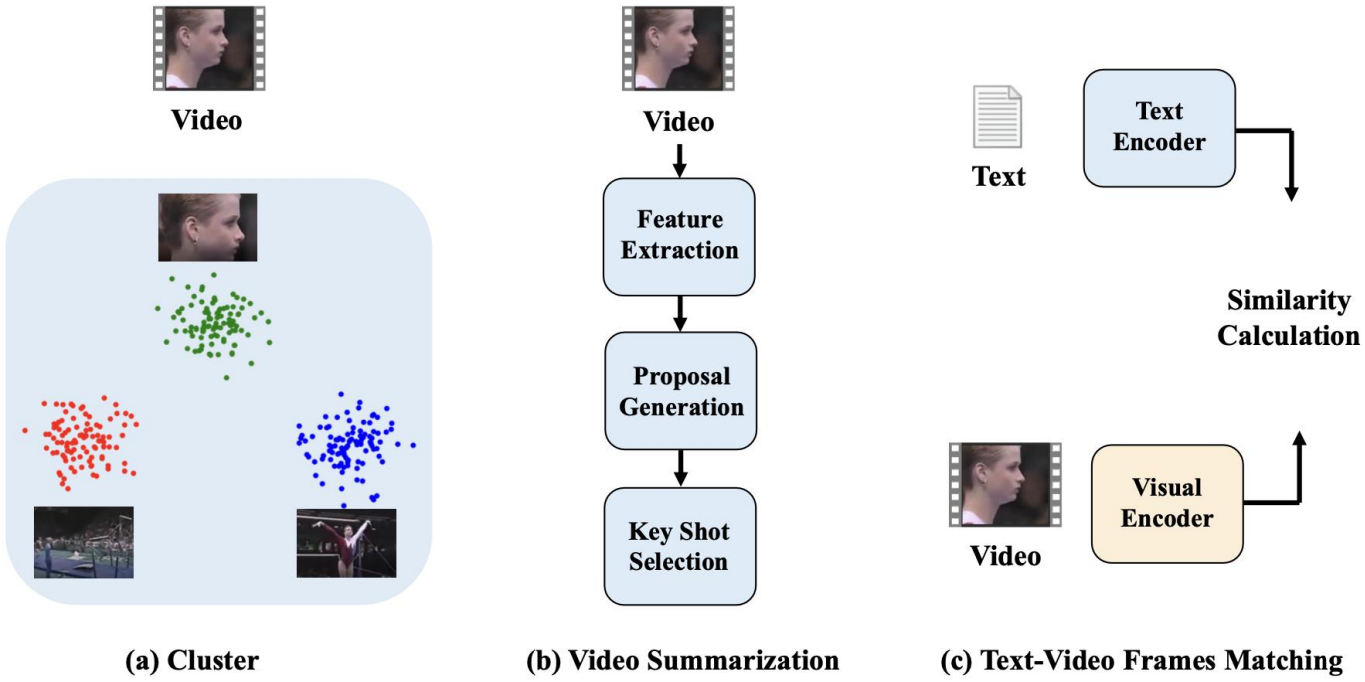
MovieChat offers an efficient approach to understanding long videos by combining memory management with multi-modal large language models.

Inference Mode



Keyframe Extraction

- **Uniform Sampling**: Selecting frames at regular intervals.
- **Boundary-Based Methods**: Detecting scene changes.
- **Activity-Based Methods**: Focusing on frames with significant motion.
- **Visual Content-Based Methods**: Analyzing visual features like color and texture.
- **Clustering-Based Methods**: Grouping similar frames and selecting representative ones.
- **Text-Video Similarity-Based Methods**: Choosing frames that match user queries.



It is recommended to conduct experiments to compare the performance, strengths, and limitations of various methods

Content

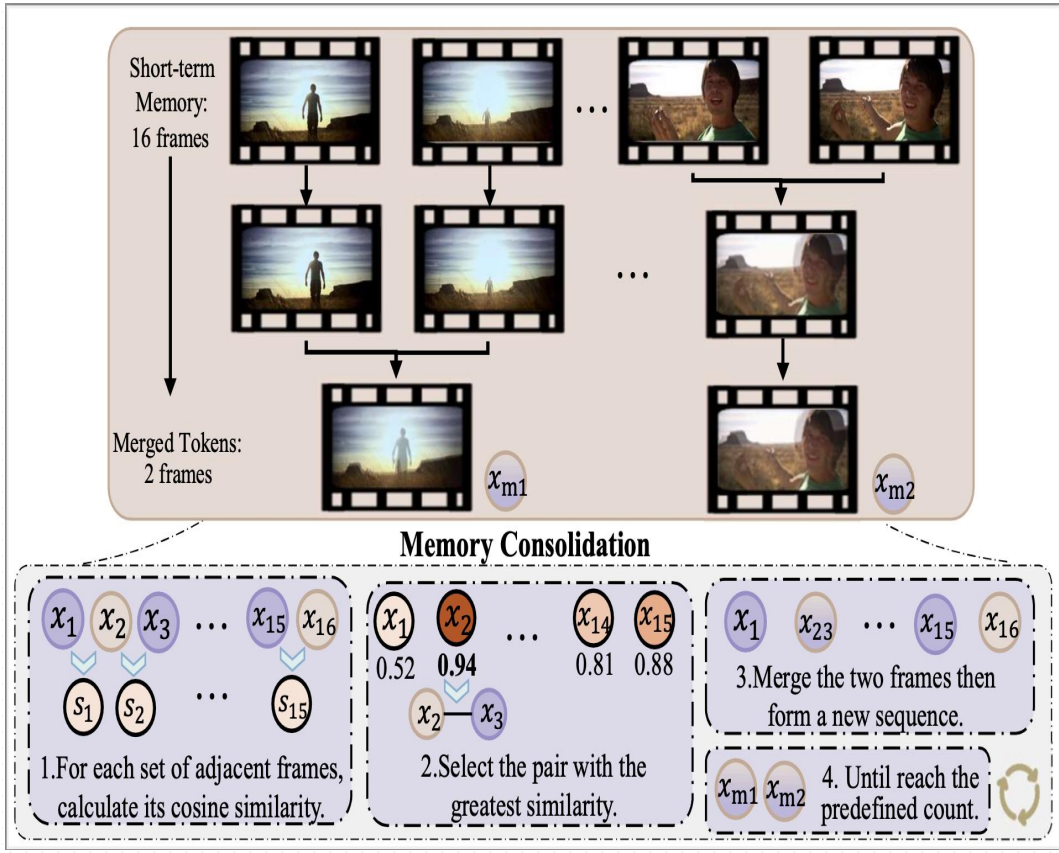
Motivation

Limitation: Loss of Critical Information During Memory Updates

Only **Token similarity** and not **Token importance**

Merged Keyframes with similar neighboring frames

Loss of critical information

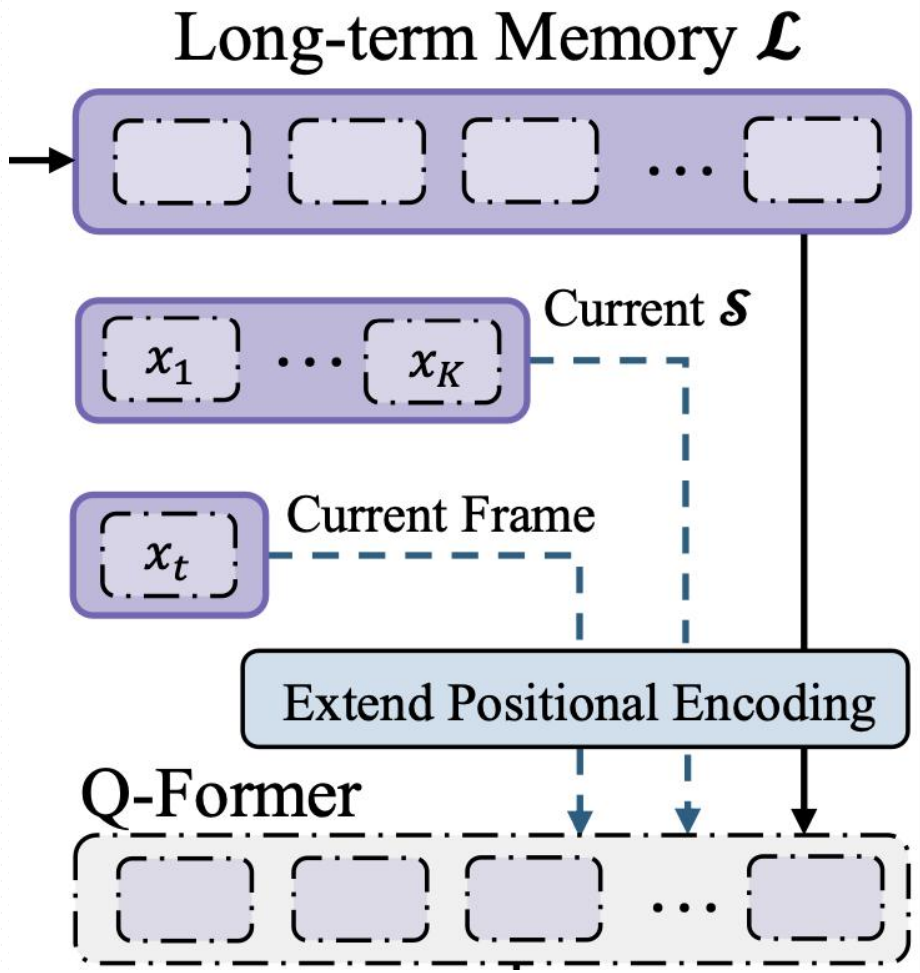


Motivation

Limitation: Inefficient Use of the Memory Bank

All memory bank as input without selection or filter

Prevent the memory bank for reducing computational load



Content

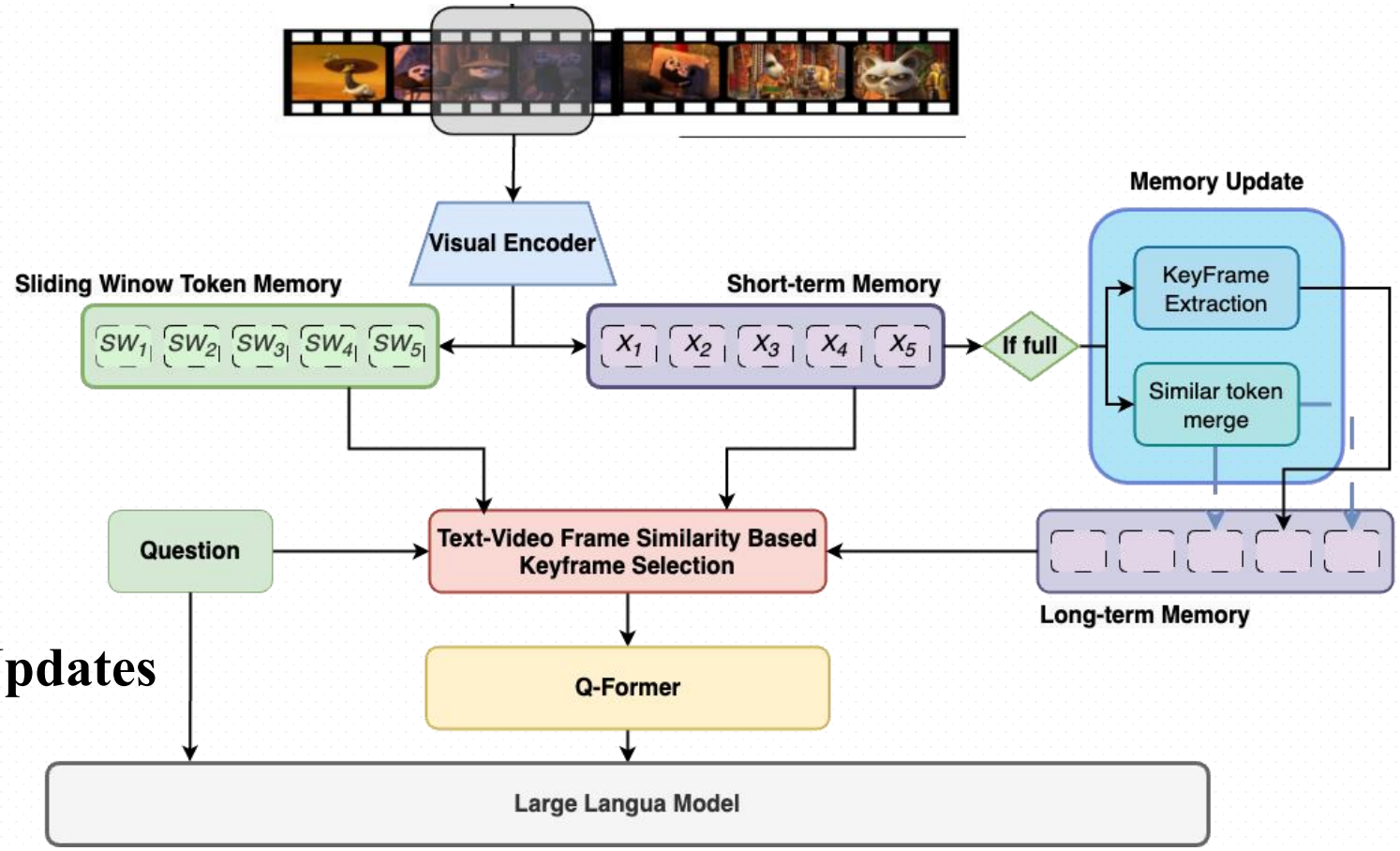
Methodology

Limitation:

1. **Loss of Critical Information During Memory Updates**
2. **Inefficient Use of the Memory Bank**

Key improvements:

1. **Keyframe Selection for Memory Updates**
2. **Text-Based Keyframe Selection**
3. **Sliding Window Memory Bank**

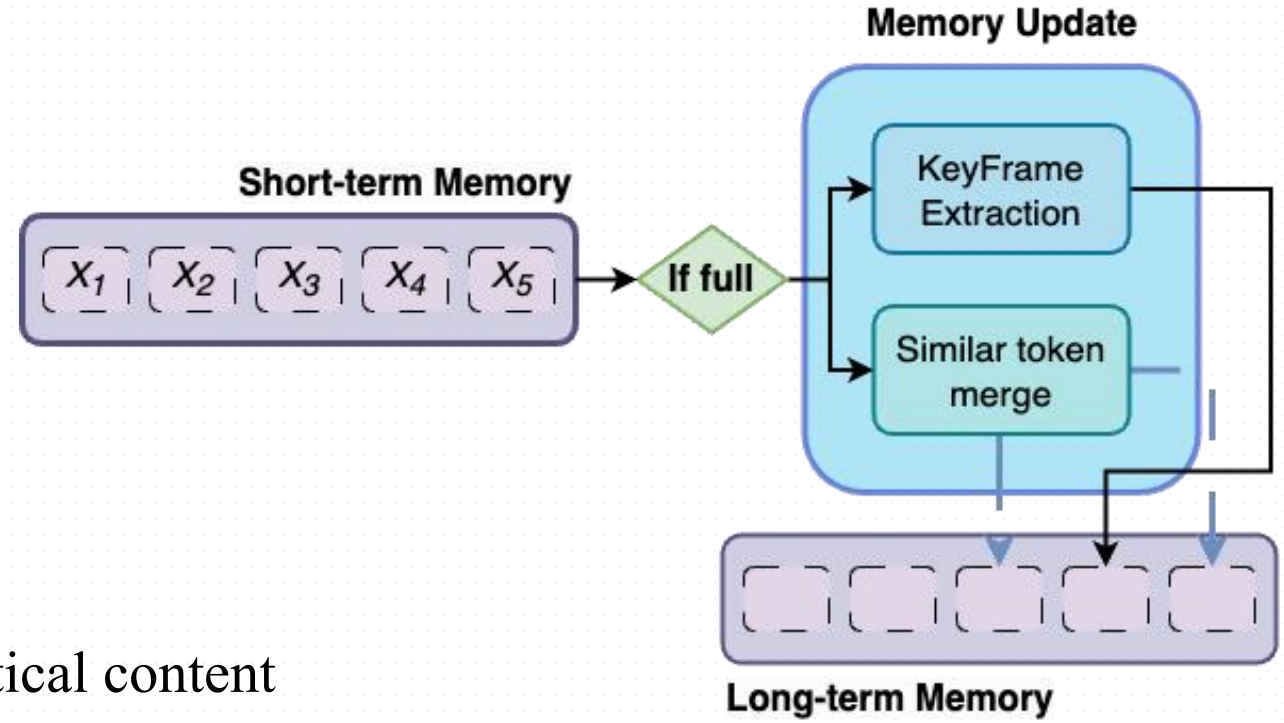


Methodology

Using Keyframe Selection for Long-term Memory Updates

1. **Keyframe Extraction**
   Extract multi Keyframe and then stored in Long-term Memory

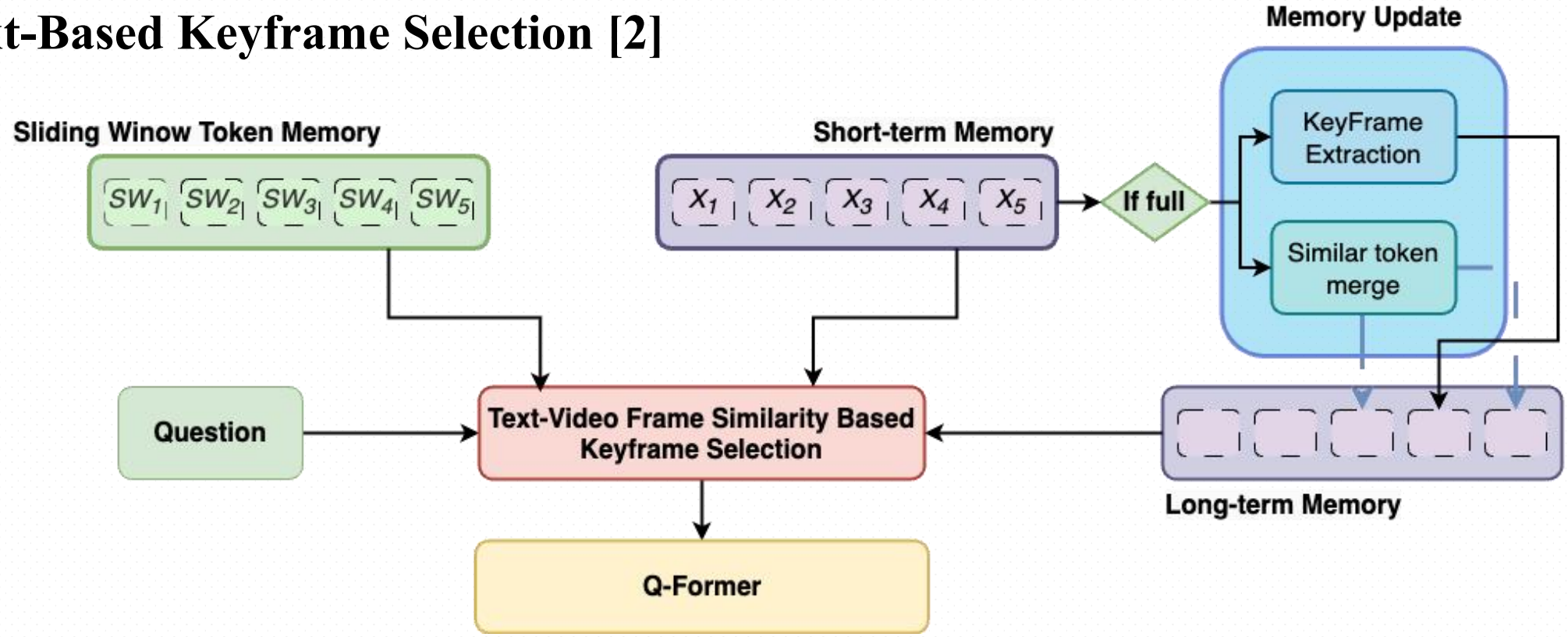
2. **Similarity-Based Fusion**
   Memory consolidation as MovieChat for the remaining tokens



This approach addresses the problem of critical content disappearing due to the current update process.

Methodology

Text-Based Keyframe Selection [2]



1. Calculates the similarity
2. Selecte the top-N relevant frames

This approach further reduces the computational cost, while still maintaining response accuracy.

[2] Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. Keyvideollm: Towards large-scale video keyframe selection, 2024.

Methodology

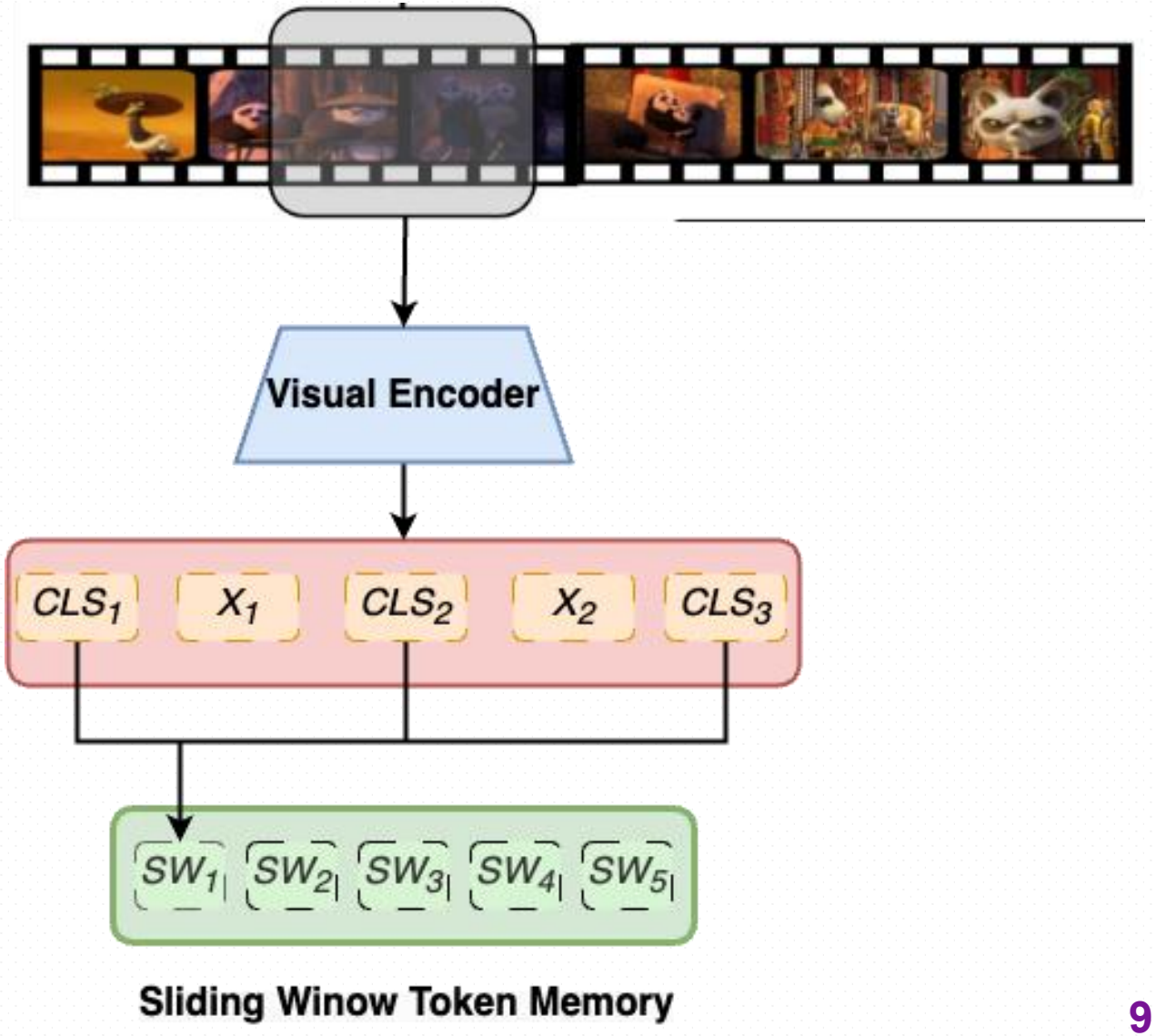
Sliding Window Memory Bank

Purpose:
Provide supplementary information with a longer temporal span

For N_{th} sliding window, with length L:

$$SW_N = \frac{1}{L}\sum_{i=1}^{L} CLS_i$$

Future work will explore more advanced methods



Content

Experiments

Text-Based Keyframe Selection method

Benchmark: MovieChat-1k
Test: 170 videos, durations 10k -- 12k frames

MovieChat ---- 256 frames
Ours ---- 128 framse

Table 1: Accuracy on MovieChat-1k

Model	Acc
MovieChat	0.527
Ours with Text-based Key Frame Selection	0.513

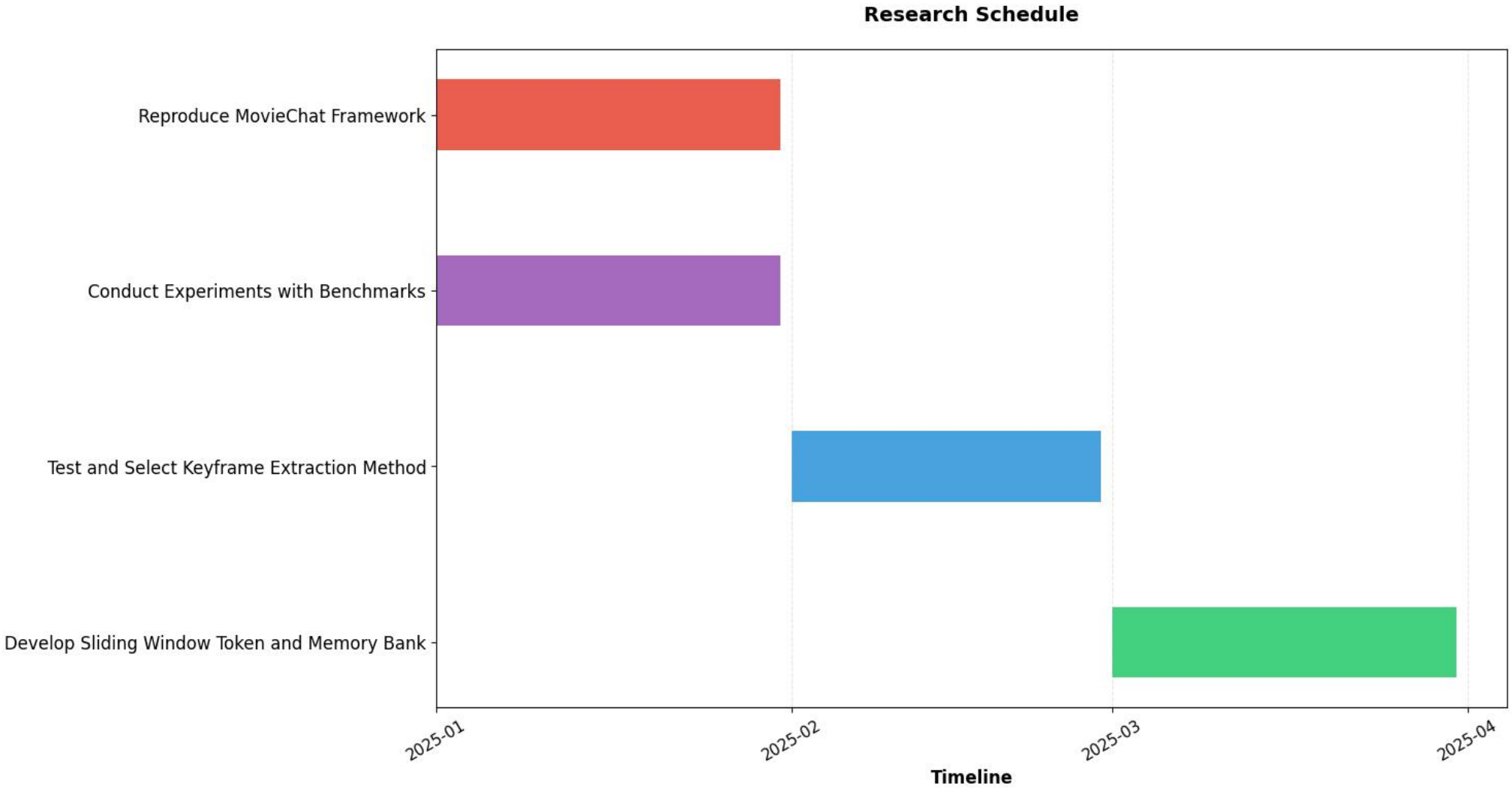
Result:

- **Only half of the frames, the accuracy is closed to the original MovieChat**

It is possible to reduce the number of frames required for processing, while maintaining similar response accuracy

Content

Further Plan



清華大學
Tsinghua University

Thanks for Listening.

Naifan Zhang

Dec 18th, 2024