Code

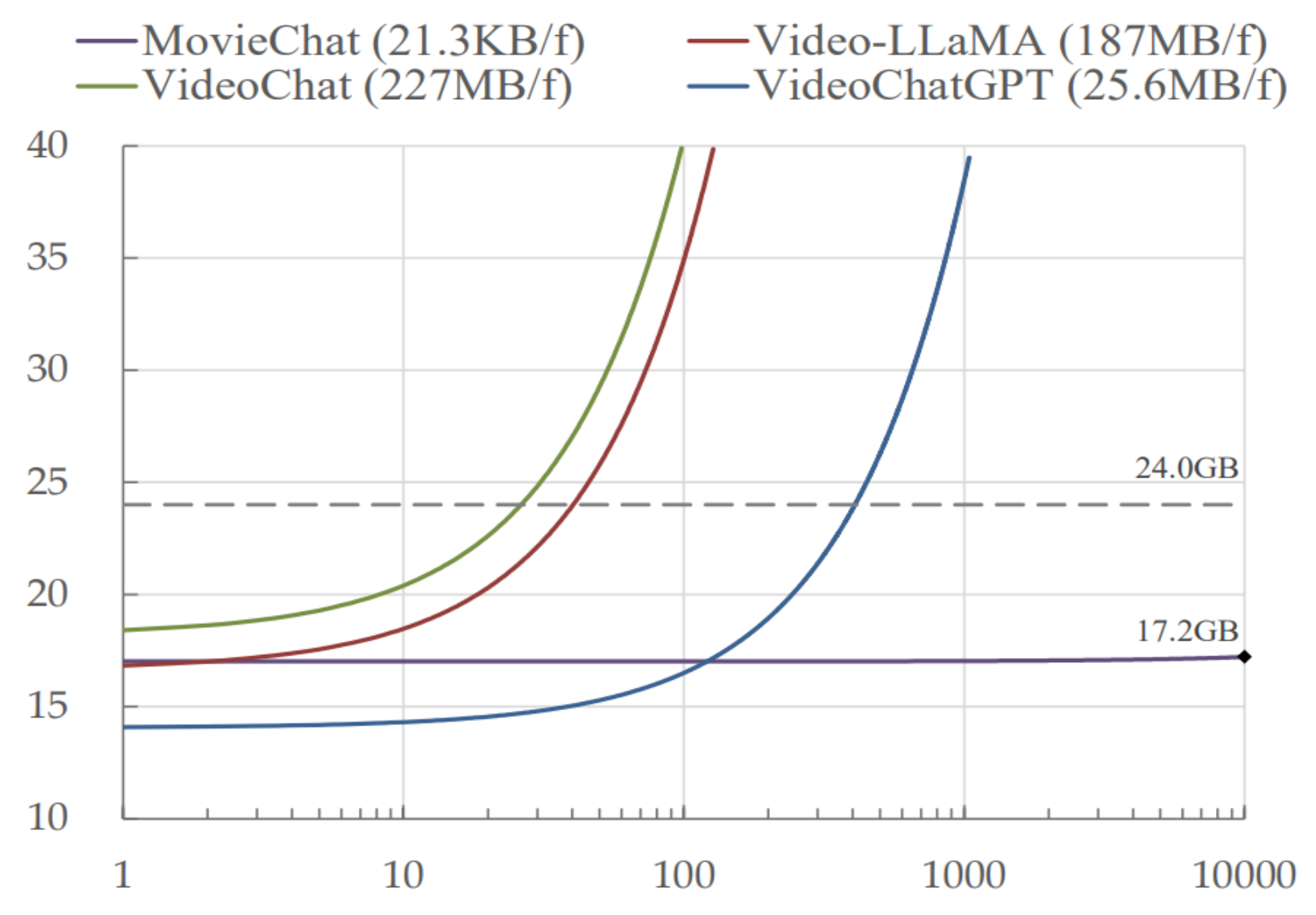
Demo

MovieChat: From Dense Token to Sparse Memory for Long Video Understanding

Enxin Song [1,*] Wenhao Chai [2,*,†] Guanhong Wang [1,*] Yucheng Zhang [1,‡] Haoyang Zhou [1,‡] Feiyang Wu [1,‡] Haozhe Chi[1] Xun Guo [3] Tian Ye[4] Yanting Zhang[5] Yan Lu[3] Jenq-Neng Hwang[2] Gaoang Wang [1,✉]
[1] Zhejiang University [2] University of Washington [3] Microsoft Research Asia [4] Hong Kong University of Science and Technology (GZ) [5] Donghua University
*Equal contribution, † Project lead, ‡ Data collection, ✉ Corresponding Author.

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Motivation & Contribution



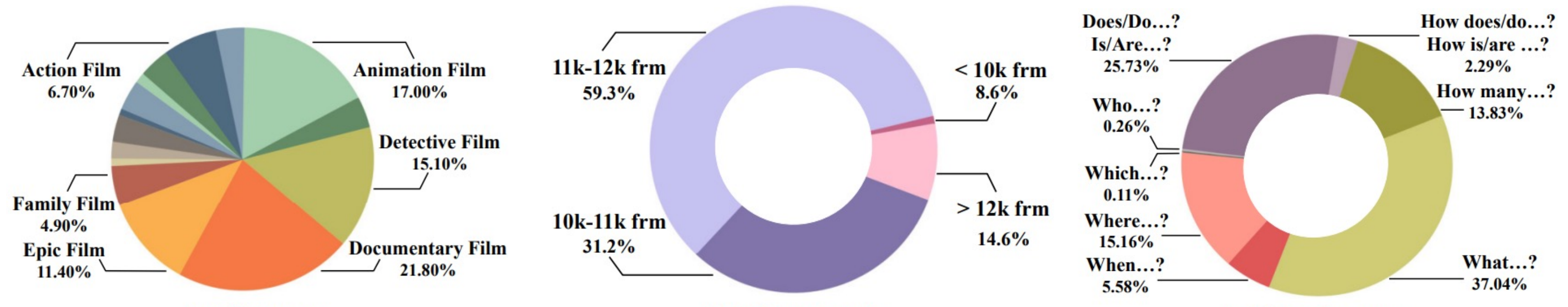
Motivation

Recently, integrating video foundation models and large language models to build a video understanding system can overcome the limitations of specific pre-defined vision tasks. Yet, existing systems can only handle videos with very few frames. **For long videos, the computation complexity, memory cost, and long-term temporal connection impose additional challenges.**

Contribution

- We present MovieChat, a novel framework that integrates vision models and LLMs, which is the first to support long video (>10K frames) understanding tasks.
- We propose an effective memory management mechanism to reduce the computation complexity and memory cost, while enhancing the long-term connection.
- We release the first long video understanding benchmark, MovieChat-1K, with manual annotations and conduct extensive quantitative evaluation and case studies to evaluate the comparable performance of both understanding capability and inference cost.

A New Benchmark: MovieChat-1K



(a) Category. (b) Video length. (c) Question type.

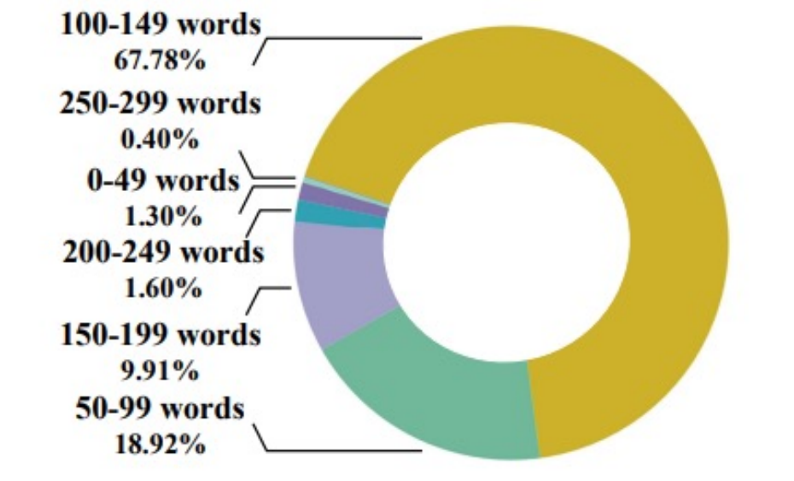


Figure B5. **Distribution of caption length.**

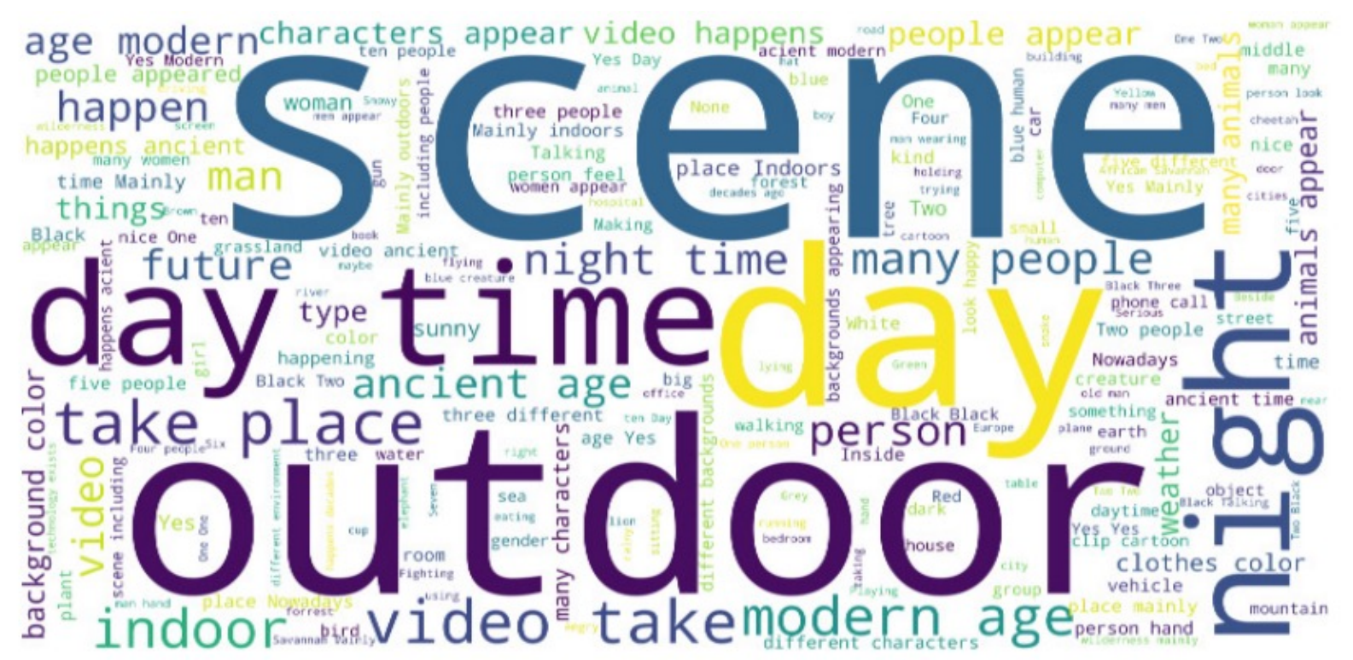
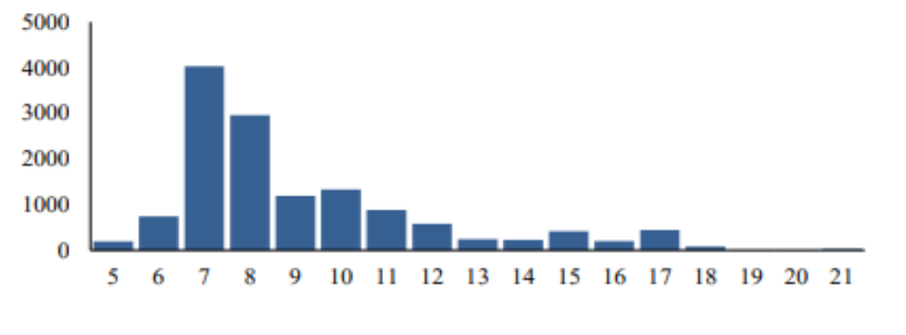
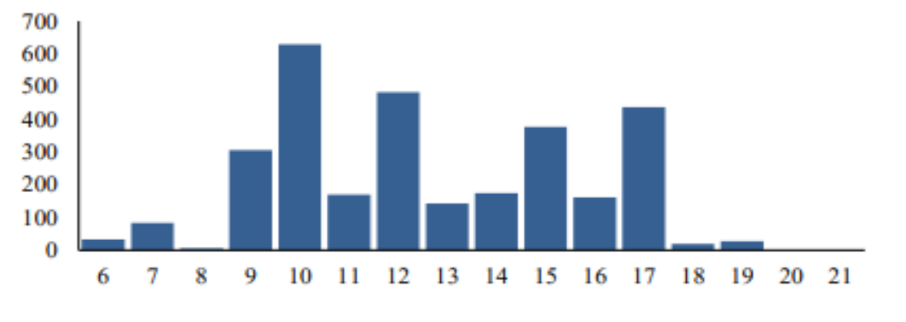


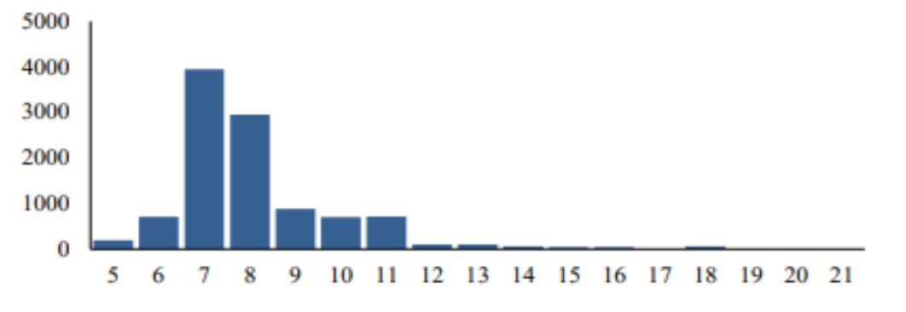
Figure 4. **Word Cloud** of the answer set in MovieChat-1K.



(a) Length of total questions.

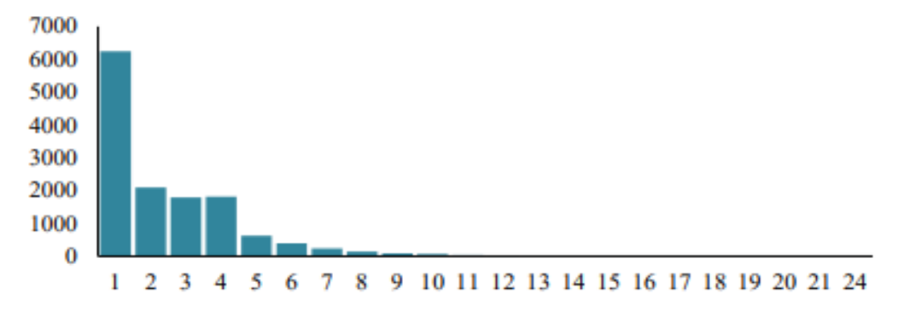


(b) Length of global questions.

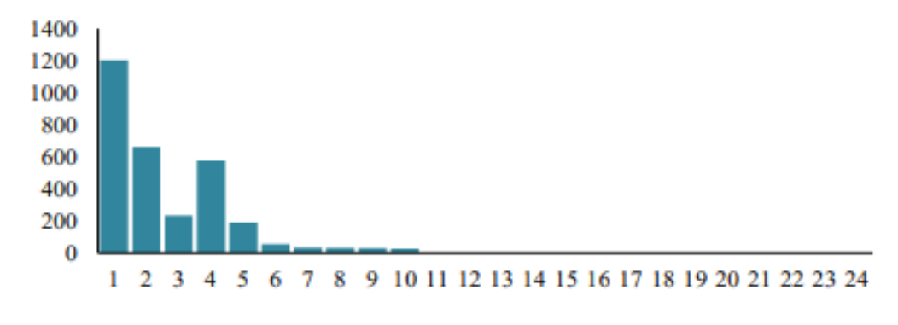


(c) Length of breakpoint questions.

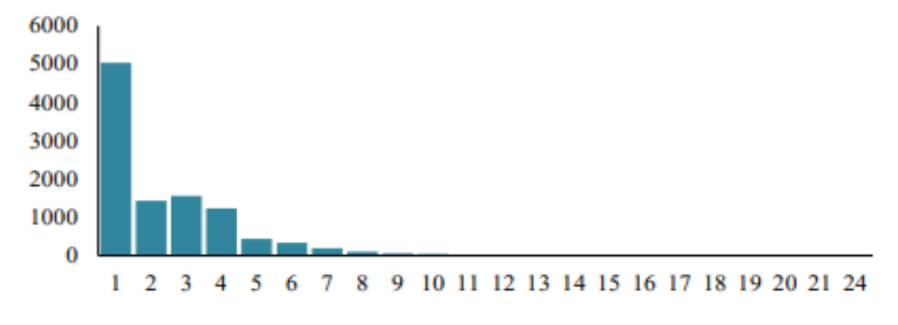
Figure B3. Length distribution of questions.



(a) Length of total answers.



(b) Length of global answers.



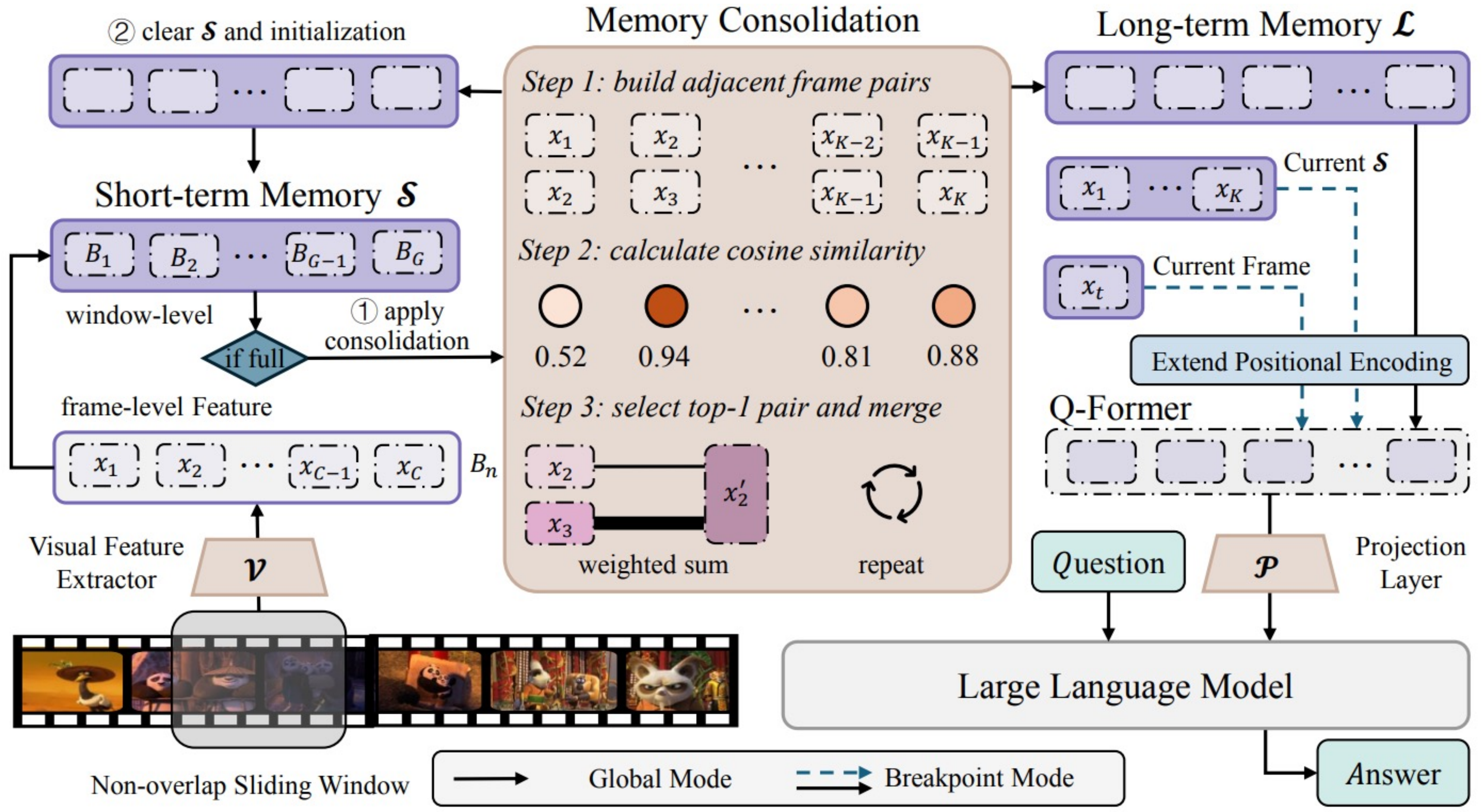
(c) Length of breakpoint answers.

Figure B4. Length distribution of answers.

Figure B6. **Word Cloud** of the caption set in MovieChat-1K.

Dataset	Avg. Duration (min)	Number of Captions	Avg. Caption Length	Number of Question-Answer Pairs	Avg. Question Length	Avg. Answer Length
MovieQA [6]	3.5	-	-	14.9K	9.3	5.1
MovieGraphs [7]	0.73	15K	35	-	-	-
MovieNet [2]	2.1	2.5K	-	-	-	-
MovieChat-1K	9.4	1K	121	13K	7.8	2.3

Method

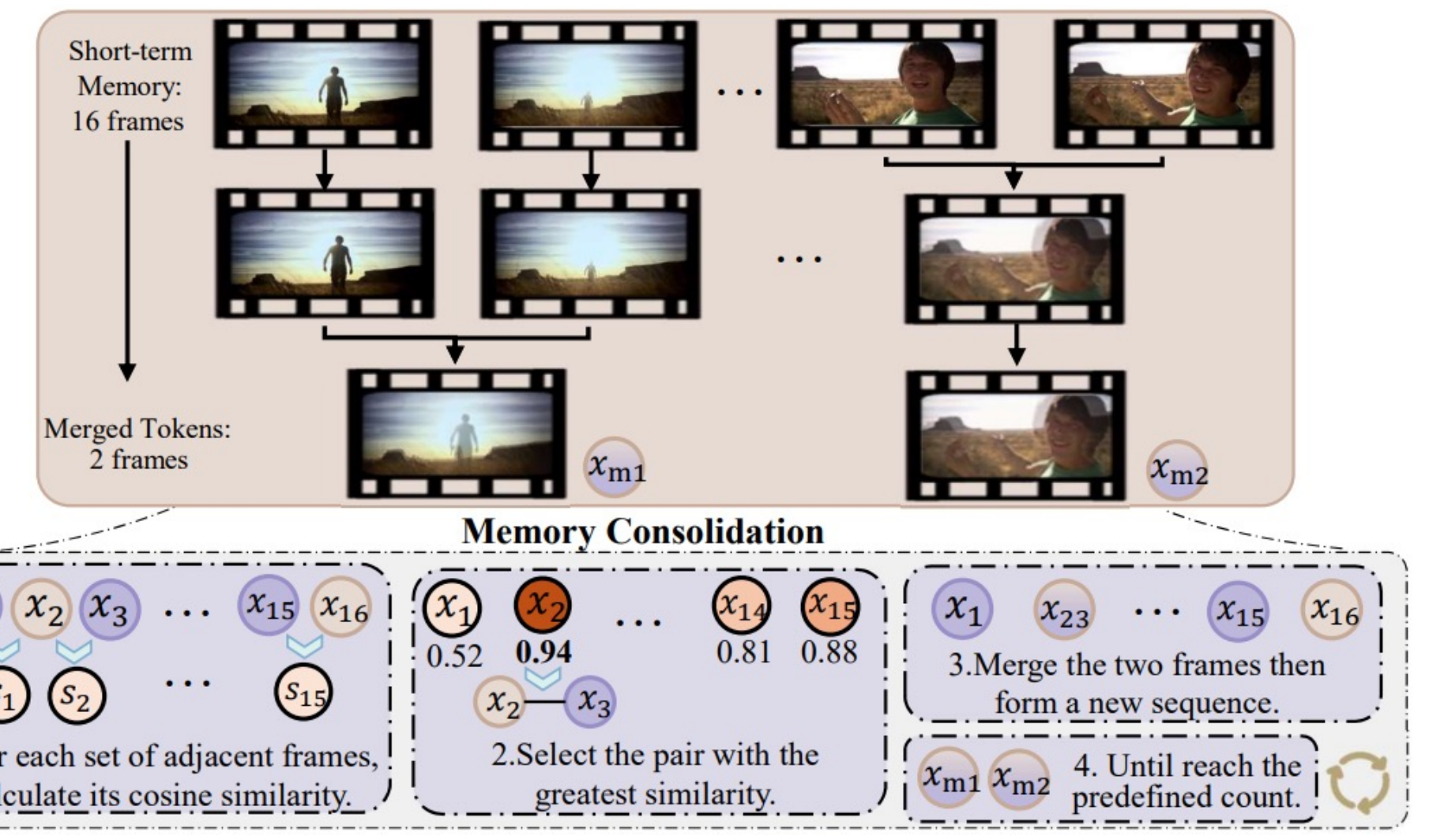


Global Mode is defined as the understanding and question-answering for the whole video.

Breakpoint Mode is distinctly defined as understanding specific moments in a video.

Algorithm 1 Memory consolidation

```
Require: $\mathcal{S}$                                  ▷ short-term memory
  while $len(\mathcal{S}) > R_L$ do                      ▷ iterative merge
      for $\mathbf{x}_i$ in $\mathcal{S}$ do
          $s \leftarrow sim(\mathbf{x}_i, \mathbf{x}_{i+1})$       ▷ tokens similarity
      end for
      $m \leftarrow max(s)$                              ▷ the maximum value index
      $\mathbf{x}_m \leftarrow merge(\mathbf{x}_m, \mathbf{x}_{m+1})$  ▷ merge
      del $\mathbf{x}_{m+1}$
  end while
```



1.For each set of adjacent frames, calculate its cosine similarity.

2.Select the pair with the greatest similarity.

3.Merge the two frames then form a new sequence.

4. Until reach the predefined count.

Experiments

Quantitative Evaluation

Short video question-answering.

Method	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM [72]	32.2	–	16.8	–	24.7	–
Video Chat [34]	56.3	2.8	45.0	2.5	26.5	2.2
LLaMA Adapter [79]	54.9	3.1	43.8	2.7	34.2	2.7
Video LLaMA [78]	51.6	2.5	29.6	1.8	12.4	1.1
Video-ChatGPT [40]	64.9	3.3	49.3	**2.8**	35.2	2.7
MovieChat *(Ours)*	**75.2**	**3.8**	**52.7**	2.6	**45.7**	**3.4**

Short video generative performance.

Method	CI	DO	CU	TU	CO
Video Chat [37]	2.23	2.50	2.53	1.94	2.24
LLaMA Adapter [83]	2.03	2.32	2.30	1.98	2.15
Video LLaMA [82]	1.96	2.18	2.16	1.82	1.79
Video-ChatGPT [43]	2.40	2.52	2.62	1.98	2.37
MovieChat *(Ours)*	**2.76**	**2.93**	**3.01**	**2.24**	**2.42**

Long video question-answering.

Method	# Frames	Global Mode		Breakpoint Mode	
		Accuracy	Score	Accuracy	Score
Video Chat [34]	32	57.8	3.00	46.1	2.29
Video LLaMA [78]	32	51.7	2.67	39.1	2.04
Video-ChatGPT [40]	100	47.6	2.55	48.0	2.45
MovieChat *(ours)*	2048	**62.3**	**3.23**	**48.3**	**2.57**

Long video generative performance.

Method	CI	DO	CU	TU	CO
Video Chat [37]	3.04	2.75	3.09	3.00	3.21
Video LLaMA [82]	2.75	2.24	2.83	2.62	2.97
Video-ChatGPT [43]	2.37	2.30	2.58	2.49	2.69
MovieChat *(Ours)*	**3.11**	**2.93**	**3.24**	**3.17**	**3.25**

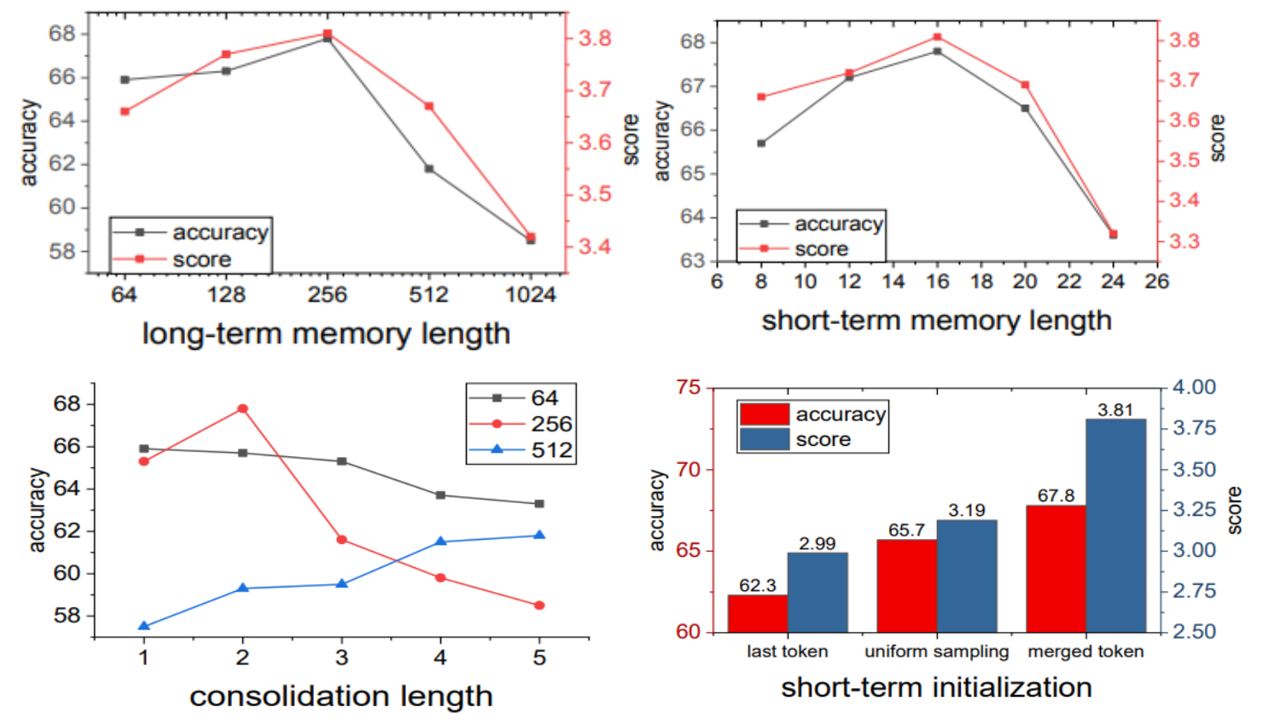
Ablation Study

Short-term and long-term memory buffers.

Method	Global Mode		Breakpoint Mode	
	Accuracy	Score	Accuracy	Score
w/o MM	51.4	3.10	38.2	2.31
base	**67.8**	**3.81**	**50.4**	**2.96**

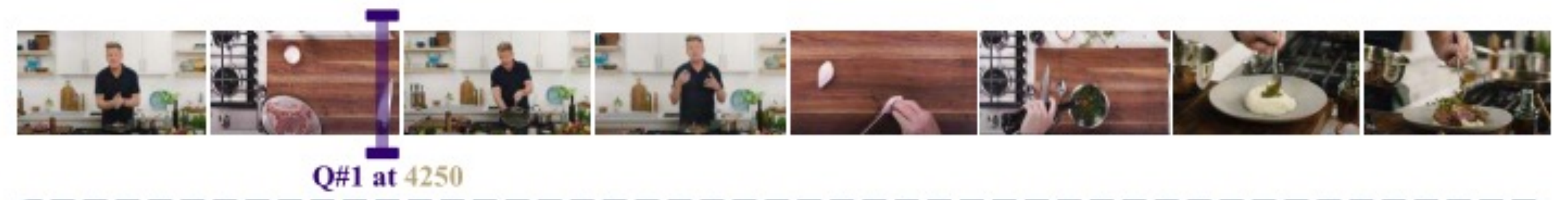
Ablation Study

Hyper-parameter ablations.



We assess the impact on performance by altering **the length of memory buffers**, **consolidation length**, and **short-term initialization**, and compare the average outcomes from GPT-3.5, Claude, and human blind ratings.

Case Study



Question #1 **What may happen next step?** **Breakpoint Mode**

Video Chat Based on the video, it appears that the person is cooking meat and vegetables on a stove top using different utensils and ingredients such as onions, garlic, mustard, and olive oil. It is likely that the person will continue cooking the meat until it is fully cooked, and then serve it to their family or friends. Additionally, they may add other ingredients such as mushrooms or potatoes to the dish. It is possible that they will use different types of cooking methods such as sautéing or grilling to ensure the meat is cooked evenly and to perfection.

Video LLaMA At frame 4250, the video shows the hand of a man cutting meat on a cutting board. It's not clear what happens next, but it is possible that the man continues cutting the meat.

Video-ChatGPT The next step in the video is to place the cooked meat on a plate and serve it. The person in the video is seen cutting the meat into smaller pieces and placing it on a plate. The person then takes a fork and begins to eat the meat.

MovieChat I think the man will prepare the food in the kitchen. It seems like he is going to cook it with spices and herbs. He has a large wooden board on the kitchen counter and a large white pot on the stove, and he will probably use some tools to prepare the food. He might also use his hands or a spatula to cook the meat.

Question #2 **Which step took the longest to complete this dish?** **Global Mode**

Video Chat Based on the video, it appears that the longest step in making the dish is preparing the pan for cooking the steak. Specifically, the man takes the pan out of the oven and starts cleaning it by wiping it with a rag, then rubbing it with olive oil, and finally heating the pan to high heat. It is also mentioned that the man has to wait for the pan to heat up before putting the steak in the pan.

Video LLaMA The dish that this video portrays took several minutes to complete. However, the most time-consuming task seems to have been preparing the ingredients on the white kitchen counter. The man in the video is seen chopping vegetables and using a sharp knife to prepare a meal. Finally, he will sit down to enjoy the meal, savoring each bite and reflecting on the process that led to this culinary creation.

Video-ChatGPT The longest step in preparing the dish was the time it took to cook the meat. The man in the video cooks the meat on a stove for a long time, which is a crucial step in the process. The video shows the man cooking the meat on a stove for a considerable amount of time, which indicates that it took a considerable amount of effort and patience to achieve the desired result.

MovieChat The longest step to complete this dish would be the preparation of the food on the cutting board. It requires chopping the meat and adding other ingredients like broccoli and carrots, and mixing them together before putting them in the pan. The chef had to carefully chop and arrange the ingredients before cooking the dish.

More Cases

Question #1 **What is she doing?**

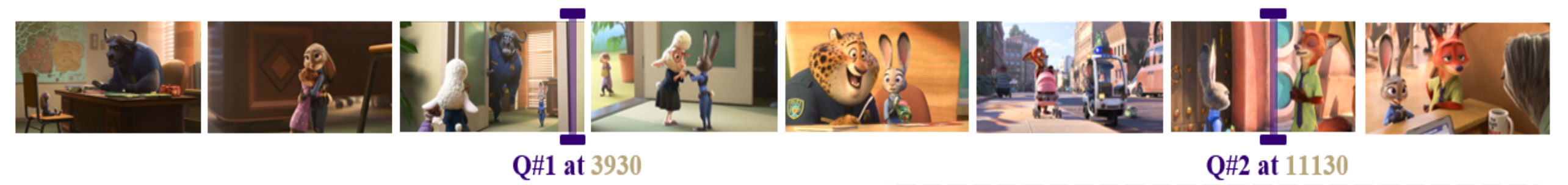
Answer #1 The young woman is seen strolling down a city street, followed by a young man walking in the same direction behind her. Later the woman is seen walking alone. The woman is dressed in an outfit that combines red and black.

Question #2 **Where is this couple?**

Answer #2 In the scene, a couple is positioned in the center of an urban street, with the man wearing a black coat. Then they are walking in a park, surrounding by several trees and bushes.

Question #3 **Can you tell me what happened in this movie?**

Answer #3 First, the video shows a vibrant scene set in a park with a green hill in the background. You can also see people walking on a street and in a place with trees. There is a woman in a red coat walking in the park. Later, this woman stands by a hill with a man behind her. In the next scene, she is depicted strolling down a street under the cloak of night, with a man trailing behind her. Next, we see a couple standing on a balcony overlooking a street in the evening, where a man is observed passing by before continuing down the street. At the end of the video, there is a scene of people walking in a town covered with snow.



Question #1 **What is Chief Bogo doing?**

Answer #1 Chief Bogo, dressed in a brown outfit, is talking to someone through a laptop. He is seated at a table in a dimly lit room. Then he is standing inside the office, talking to a sheep holding some documents.

Question #2 **Who is this rabbit named Judy talking to?**

Answer #2 Judy, who is a rabbit and a police officer, is talking to a fox. They are walking down the street and standing in front of a movie theater.

Question #3 **Can you describe this video?**

Answer #3 The video shows an animation in settings such as offices, restaurants and theaters. Initially, it featured a scene of a cartoon rabbit and a bull talking on a desk. The next scene shows a rabbit dealing with a sheep in an office. The narrative then shifts to a conversation between a rabbit and a cat sitting together at a restaurant table. The rabbit and the fox talk on the road and end up arguing in front of a building. Another notable scene is of a dog sitting in a room with a computer on a table. The video goes on to revisit the characters in different contexts, including a moment where a rabbit and a fox are working together on a table, and a moment when a rabbit and a dog are talking in the background. The collection includes different animal characters such as foxes, mice and cats, with each shot set against a different background.

Question #1 **In which scene does the movie mainly focus on what happened? Are there other scenarios?**

Answer #1 The video highlights a scene where people are gathered in front of some buildings. In the next scene, individuals are standing around a table, with a man positioned behind them.

Question #2 **Do they let the soldiers go first or negotiate with the general first?**

Answer #2 First, two soldiers is talking face to face. Then the general releases the soldiers. Following this, the general engages in negotiations with other soldiers.

Question #3 **What plot does this movie clip include?**

Answer #3 The clip begins on a black screen with white text. In the next scene, a man dressed in a black robe, with a vast white flag waving in the background. Following this, the scene shifts to a group of individuals in front of a grand building, all set against a black backdrop with more text. Next, the focus moves to two men, one with a beard and the other wearing a robe, standing on solid ground. The video then transitions to show two other men, one in a shirt and the other in a blue shirt, also standing on the ground. The clip concludes with a black screen featuring a large black bear and a big bird.