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

Supplementary Material

The supplementary material is structured as follows:

1. We first present schematic diagram of the memory consolidation algorithm of MovieChat in Section A.
2. We provide detailed supplementary statistical information for MovieChat-1K in Section B.
3. The prompt template we use for LLM-Assisted Evaluation is shown in Section C.
4. We also list the hyperparameter settings of MovieChat in Section D.
5. We mention the specifical LLM-Assisted Evaluation method employed for the assessment of short video generative performance in Section E.
6. To avoid the impact of misjudgments by LLM assistants on the results, we introduce the manual filtering strategy in Section F.
7. To demonstrate the outstanding performance of MovieChat across a wide range of categories, we calculate the Pearson correlation coefficient of different score methods in Section G.
8. We then list the evaluaotion results with GPT-3.5 [5], Claude [1] and human blind rating in Section H.

A. Memory consolidation algorithm of MovieChat.

As shown in Fig. A1, for each sampled frame x_i, we calculate its similarity with adjacent frames. After that, we select the pair with the greatest similarity, merge and replace these two frames, resulting in a new sequence. We conduct the merge operation repeatedly until the count of existing frames in short-term memory reaches the predefined value.

B. MovieChat-1K Statistics Information

Distribution of video categories. MovieChat-1K contains videos from 15 popular categories with varying distribution. As shown in Tab. B1, every video comprises multiple alternating scenes.

Category	Percentage
Documentary Film	21.80%
Animation Film	17.00%
Detective Film	15.10%
Epic Film	11.40%
Action Film	6.70%
Family Film	4.90%
Crime Film	3.80%
Science Fiction Film	3.70%
War Film	3.70%
Adventure Film	3.50%
Romance Film	3.30%
History Film	2.10%
Suspense Film	1.30%
Fantasy	0.90%
School Film	0.80%

Table B1. Distribution of video categories in MovieChat-1K.

Video information and visual question-answer data format. To the best of our knowledge, a long video understanding dataset has not yet been established. Our work represents the initial step in creating and making it publicly available.We create MovieChat1K, containing 1k long videos and corresponding 1k dense captions, and 13k visual question-answer pairs.One visual example of these arrangements is provided in Figure B2.

Sentence length distribution of question-answer pairs. MovieChat1K exhibits diverse lengths of question-answer pairs in the segmented clip level. Fig. B3 and Fig. B4 demonstrate the length distribution of question-answer pairs in different modes. Despite the distribution of question-answer pairs varies between the global mode and breakpoint mode, the majority of questions tends to concentrate between 5-15 words in length, while the length of answers generally have fewer than 10 words.

Comparison between MovieChat-1K and other benchmarks. MovieChat-1K provides a large-scale benchmark for long video understanding, which contains 1K movies, 1K dense captions and 13k question-answer pairs. The comparison between different datasets are shown in Tab. B2. It is evident that MovieChat-1K provides the

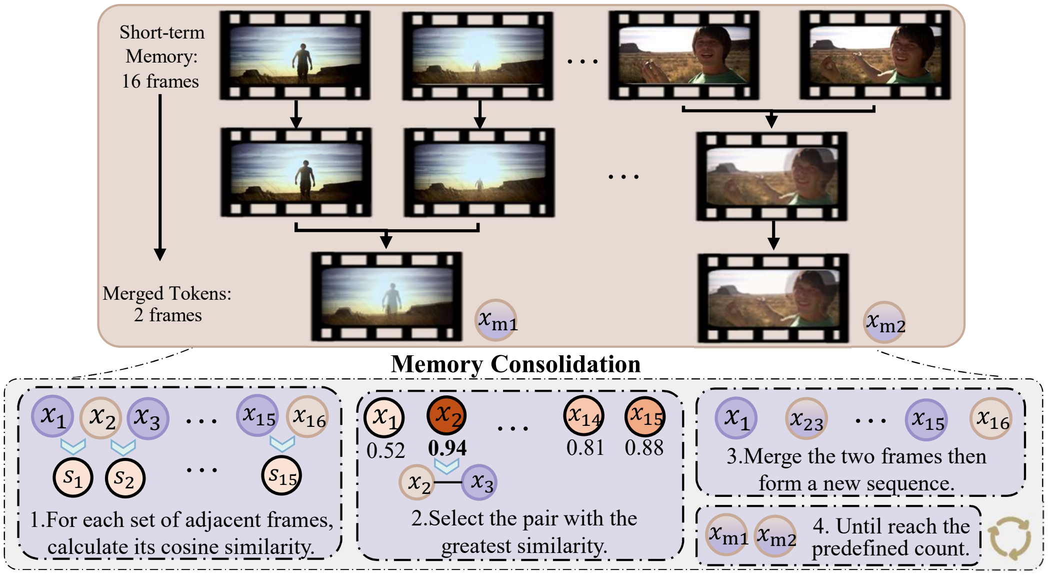

Short-term Memory: 16 frames
Merged Tokens: 2 frames
x_{m1}
x_{m2}
Memory Consolidation
x_1 x_2 x_3 ... x_{15} x_{16}
s_1 s_2 ... s_{15}
1.For each set of adjacent frames, calculate its cosine similarity.
x_1 0.52
x_2 0.94
x_{14} 0.81
x_{15} 0.88
x_2 x_3
2.Select the pair with the greatest similarity.
x_1 x_{23} ... x_{15} x_{16}
3.Merge the two frames then form a new sequence.
x_{m1} x_{m2}
4. Until reach the predefined count.

Figure A1. Question and answer about clips from *YouTube*, which is a tutorial on how to cook steak. The entire instructional process begins with marinating the steak, followed by pan-searing it, preparing side dishes, and ultimately plating the meal.

longest average duration for movie clips. MovieQA [6] exclusively offers question-answer pairs related to movies, while MovieGraphs [7] supplies captions associated with movies. Unlike other datasets, MovieNet [2] encompasses three main types of texts: subtitle, synopsis, and script, excluding question-answer pairs. Additionally, the synopsis category is designed for the entire movie rather than video clips. Consequently, MovieChat-1K is more suitable for studying long video comprehension compared to other datasets.

C. LLM-Assisted Evaluation for the short video question-answering task.

Following [4], we use LLM-Assisted Evaluation for the short video question-answering task. Given the question, correct answer, and predicted answer by the model, the LLM assistants should return the *True* or *False* judgement and relative score (0 to 5). The whole prompt is shown in Fig. C1. It takes about 250 tokens per question. We report the baseline results of short video question-answering from https://github.com/mbzuai-oryx/Video-ChatGPT.

D. Hyperparameter Setting

We report the detailed hyperparameter settings of MovieChat in Tab. D3. The sliding window size of MovieChat is set to 16, which means that every slide involves the extraction of 16 frames. We configure the short-term memory to consist of 18 frames, with each frame containing 32 tokens. When the short-term memory reaches its capacity, it is directed to the memory consolidation module to be merged into 2 representative frames. The 2 frames are simultaneously input into the long-term memory with a total length of 256 and used to reinitialize the short-term memory.

E. LLM-Assisted Evaluation for short video generative performance.

We use LLM-Assisted Evaluation proposed by [4] for short video generative performance. The evaluation pipeline assesses various capabilities of the model and assigns a relative score (1 to 5) to the generated predictions, in the following five aspects: *Correctness of Information*, *Detail Orientation*, *Contextual Understanding*, *Temporal Understanding* and *Consistency*. We follow the corresponding prompts provided in https://github.com/mbzuai-oryx/Video-ChatGPT and report the baseline results of short video generative performance.

F. Manual filtering strategy for LLM-Assisted Evaluation.

For each test data, [4] utilized GPT-3.5 [5] to provide an evaluation result in terms of a ’yes/no’ response and a corresponding score, as demonstrated in Fig. C1. The score is an integer value ranging from 0 to 5, where a score of 5 indicates the highest degree of meaningful correspondence.

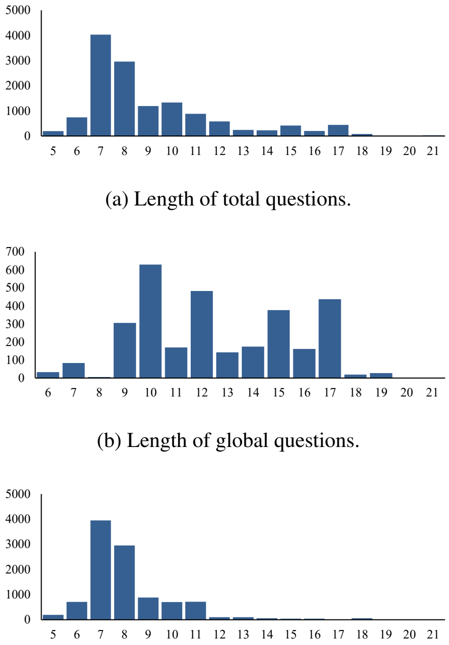
```
"info": {
  "video_path": "MI6-19.mp4",
  "url": "",
  "class": "action film",
  "w": 720,
  "h": 480,
  "num_frame": 10500,
  "fps": 25
},
"caption": "It is a part of a movie. The theme of the movie is spy and agent. The helicopter is crashed on the snow land on the cliff. At the same time, in a room, a man is tied on the rope, and another man is trying to kill a woman. The fight against each other. The man tied on the rope helps the woman, but then he is nearly killed by the rope. Luckily, the woman finally kills the man and saves the man who is nearly killed by the rope. At the cliff, a man is kicked down the cliff by another man. ",
"global": [
  {
   "question": "When does the things in the video happens, ancient age, modern age or future?",
   "answer": "Modern age."
  },
  ...
  {
   "question": "Does it happen during day or night?",
   "answer": "Day."
  }
],
"breakpoint": [
  {
   "time": 750,
   "question": "What are the people doing?",
   "answer": "Fighting."
  },
  ...
  {
   "time": 9750,
   "question": "Are there any plants?",
   "answer": "Yes."
   }
]
```

Figure B2. Video information and visual question-answer data format in MovieChat1K.

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

Table B2. Comparison between MovieChat-1K and other benchmarks.

However, we observe instances where GPT-3.5 [5] offered judgments and scores that do not align, such as providing a 'yes' response with a score of 0 or a 'no' response with a score of 5. This discrepancy has the potential to impact the accuracy of results and introduce fluctuations. We adapt the prompts used for GPT-3.5 [5] with the aim of addressing



(a) Length of total questions.

(b) Length of global questions.

(c) Length of breakpoint questions.

Figure B3. Length distribution of questions.

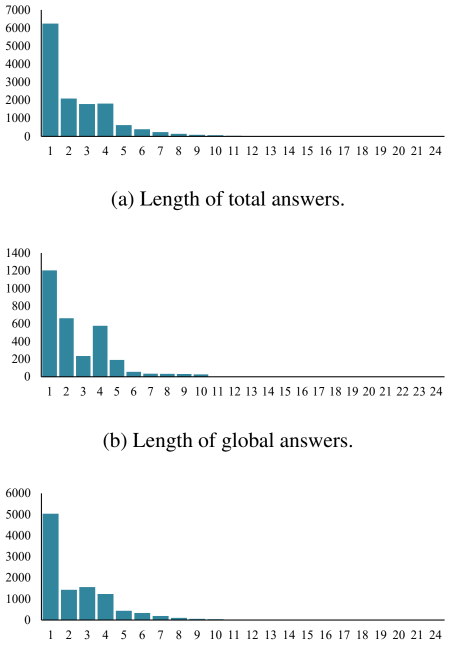
Description	Default Value
size of sliding window	16 frames
size of short-term memory	18 frames × 32 tokens per frames
size of long-term memory	256 frames
consolidation length	2

Table D3. Hyper-parameter settings of MovieChat.

this concern and did not yield the desired mitigation. Hence, we introduce an artificial filtering strategy. For each evaluation result generated by GPT-3.5 [5], we conduct manual screening. We retain only those outcomes that exhibited consistency between the 'yes/no' judgments and the associated scores, thus enhancing the reliability of the evaluations. Similarly, we applied the same filtering strategy to the evaluation results generated by Claude [1].

G. Pearson correlation coefficient of different score methods.

The Pearson correlation coefficient is represented by the formula:



(a) Length of total answers.

(b) Length of global answers.

(c) Length of breakpoint answers.

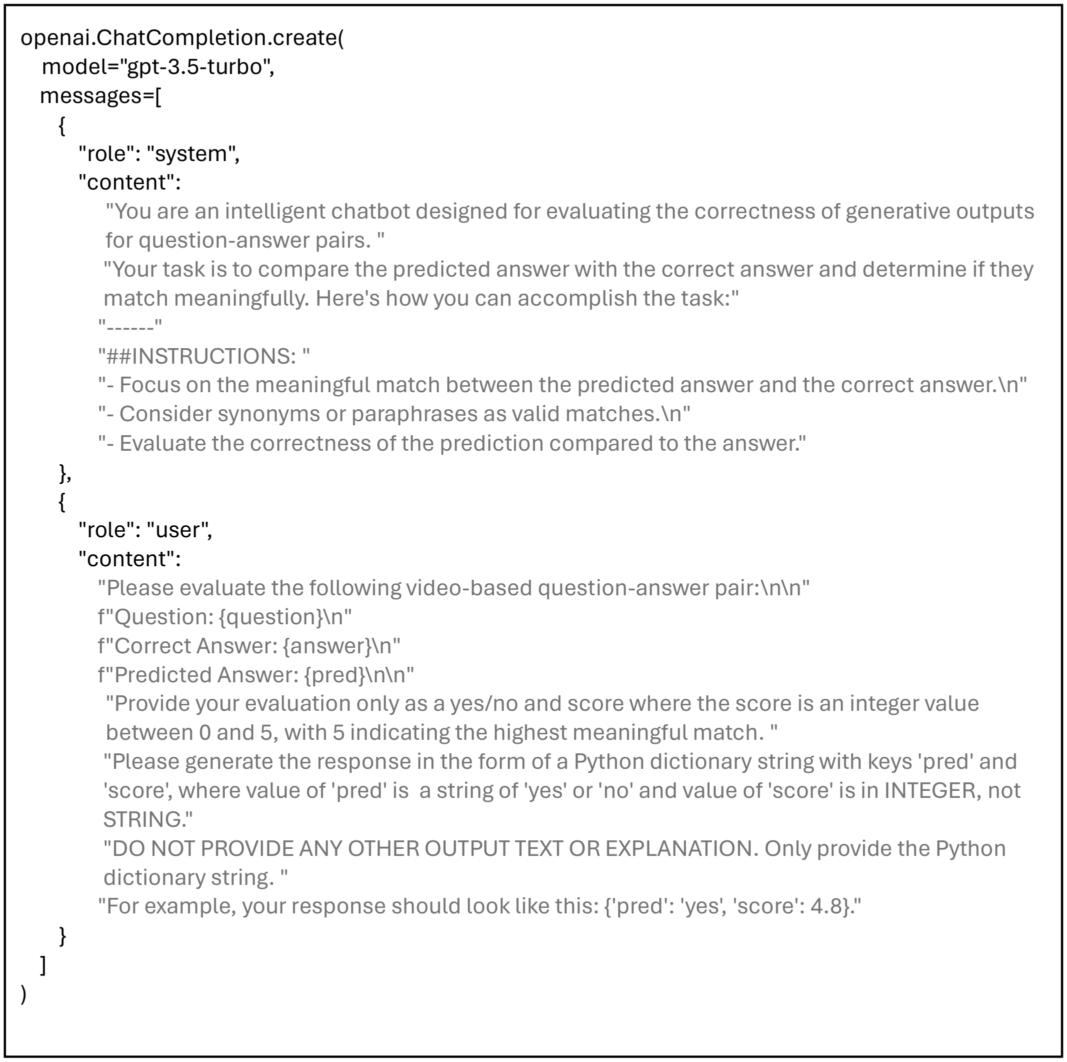
Figure B4. Length distribution of answers.

Evaluation Method	Pearson Correlation Coefficient
GPT3.5 VS. Claude	0.927
GPT3.5 VS. Human Blind Rating	0.955
Claude VS. Human Blind Rating	0.978

Table G4. Pearson correlation coefficient of GPT-3.5 [5], Claude [1], and human blind rating on score. We calculate the mean score across each score dimensions for MovieChat and previous methods [3, 4, 8, 9], and then computes the Pearson correlation between these means for each pair of evaluation methods. The Pearson correlation coefficient, which ranges from -1 to +1, indicates a stronger positive linear relationship between the two sets of data when the coefficient is higher (closer to +1).

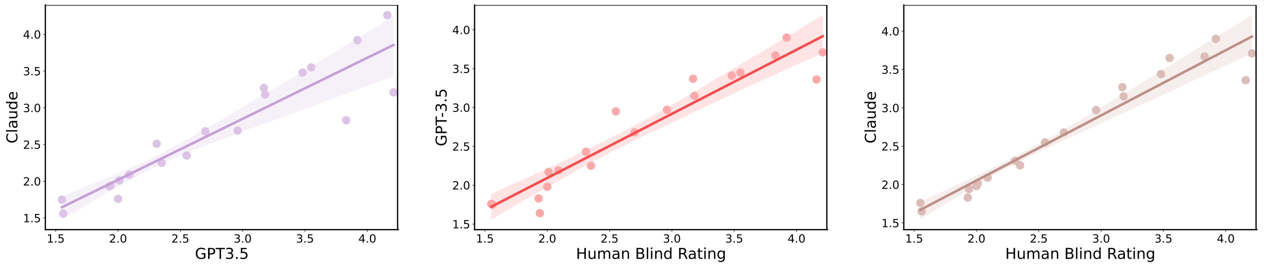
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where r_{xy} is the Pearson correlation coefficient between two variables x and y, x_i and y_i are the individual sample points for variables x and y, $\bar{x}$ and $\bar{y}$ are the averages of the x and y samples respectively, and n is the number of sample points. The formula essentially assesses the extent of linear correlation between two variables by evaluating the product of their deviations from their respective means. The numer-



```
openai.ChatCompletion.create(
  model="gpt-3.5-turbo",
  messages=[
    {
      "role": "system",
      "content":
         "You are an intelligent chatbot designed for evaluating the correctness of generative outputs
         for question-answer pairs. "
         "Your task is to compare the predicted answer with the correct answer and determine if they
         match meaningfully. Here's how you can accomplish the task:"
        "------"
        "##INSTRUCTIONS: "
        "- Focus on the meaningful match between the predicted answer and the correct answer.\n"
        "- Consider synonyms or paraphrases as valid matches.\n"
        "- Evaluate the correctness of the prediction compared to the answer."
    },
    {
      "role": "user",
      "content":
        "Please evaluate the following video-based question-answer pair:\n\n"
        f"Question: {question}\n"
        f"Correct Answer: {answer}\n"
        f"Predicted Answer: {pred}\n\n"
         "Provide your evaluation only as a yes/no and score where the score is an integer value
         between 0 and 5, with 5 indicating the highest meaningful match. "
         "Please generate the response in the form of a Python dictionary string with keys 'pred' and
         'score', where value of 'pred' is  a string of 'yes' or 'no' and value of 'score' is in INTEGER, not
         STRING."
         "DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python
         dictionary string. "
        "For example, your response should look like this: {'pred': 'yes', 'score': 4.8}."
    }
  ]
)
```

Figure C1. Prompt for ChatGPT in LLM-Assisted Evaluation for the short video question-answering task.



(a) **GPT-3.5 VS. Claude**
PCC = 0.927

(b) **GPT-3.5 VS. Human Blind Rating**
PCC = 0.955

(c) **Claude VS. Human Blind Rating**
PCC = 0.978

Figure G2. Results of the Pearson correlation analysis between three evaluation methods, including GPT-3.5 [5], Claude [1], and human blind rating. PCC stands for Pearson Correlation Coefficient.

ator represents the covariance between the two variables, and the denominator normalizes this value, ensuring that the coefficient remains between -1 and +1. The Pearson correlation coefficient quantifies the extent to which two vari-

Method	# Frames	Global Mode		Breakpoint Mode	
		Accuracy	Score	Accuracy	Score
Video Chat [3]	32	61.0	3.34	48.3	2.43
Video LLaMA [8]	32	51.4	3.10	38.2	2.31
Video-ChatGPT [4]	100	44.2	2.71	49.8	2.71
MovieChat *(ours)*	2048	**67.8**	**3.81**	**50.4**	**2.96**

Table H5. Quantitative evaluation for long video question answering on MovieChat-1K test set with GPT-3.5 [5]. The best result is highlighted in bold, and the second best is underlined.

Method	# Frames	Global Mode		Breakpoint Mode	
		Accuracy	Score	Accuracy	Score
Video Chat [3]	32	52.1	2.59	43.8	2.12
Video LLaMA [8]	32	47.3	2.19	33.2	1.69
Video-ChatGPT [4]	100	39.8	2.04	**46.4**	2.21
MovieChat *(ours)*	2048	**55.3**	**2.73**	**46.4**	**2.28**

Table H6. Quantitative evaluation for long video question answering on MovieChat-1K test set with Claude [1]. The best result is highlighted in bold, and the second best is underlined.

ables co-vary in comparison to their individual variations.

As shown in Tab. G4 and Fig. G2, we conduct pearson correlation analysis between GPT-3.5 [5], Claude [1], and human blind rating. The result indicates a substantial agreement among these evaluation methods. The alignment of scores across different score methods strengthens the reliability of our assessment. Crucially, our proposed method, MovieChat outperforms previous methods [3,4,8,9] in long video understanding tasks. The superior performance of MovieChat is evident across a broad spectrum of categories, suggesting that our model not only has a deeper understanding of long videos and respective questions but also exhibits a more accurate and consistent ability to generate relevant responses.

H. Evaluation results with GPT, Claude and human blind rating.

As shown in H5–H13, we provide detailed scoring results for GPT-3.5 [5], Claude [1], and human blind rating across various experiments.

Method	# Frames	Global Mode		Breakpoint Mode	
		Accuracy	Score	Accuracy	Score
Video Chat [3]	32	60.2	3.08	46.3	2.32
Video LLaMA [8]	32	56.3	2.72	45.8	2.11
Video-ChatGPT [4]	100	58.7	2.89	47.8	2.43
MovieChat *(ours)*	2048	**63.7**	**3.15**	**48.1**	**2.46**

Table H7. Quantitative evaluation for long video question answering on MovieChat-1K test set with human blind rating. The best result is highlighted in bold, and the second best is underlined.

Method	CI	DO	CU	TU	CO
Video Chat [3]	3.26	3.20	3.38	2.97	3.47
Video LLaMA [8]	3.30	2.53	3.28	2.77	3.42
Video-ChatGPT [4]	2.48	2.78	3.03	2.48	2.99
MovieChat *(Ours)*	**3.32**	**3.28**	**3.44**	**3.06**	**3.48**

Table H8. Quantitative evaluation for long video generation performance in global mode with GPT-3.5 [5]. CI stands for correctness of information, DO stands for detail orientation, CU stands for contextual understanding, TU stands for temporal understanding, and CO stands for consistency. The best result is highlighted in bold, and the second best is underlined.

Method	CI	DO	CU	TU	CO
Video Chat [3]	2.83	2.43	3.02	2.87	2.93
Video LLaMA [8]	2.04	1.66	2.46	2.07	2.36
Video-ChatGPT [4]	1.81	1.65	2.05	2.07	2.07
MovieChat *(Ours)*	**2.88**	**2.82**	**3.11**	**3.04**	**2.96**

Table H9. Quantitative evaluation for long video generation performance in global mode with Claude [1]. CI stands for correctness of information, DO stands for detail orientation, CU stands for contextual understanding, TU stands for temporal understanding, and CO stands for consistency. The best result is highlighted in bold, and the second best is underlined.

Method	CI	DO	CU	TU	CO
Video Chat [3]	3.03	2.61	2.87	3.15	3.23
Video LLaMA [8]	2.91	2.54	2.74	3.01	3.12
Video-ChatGPT [4]	2.83	2.47	2.66	2.92	3.01
MovieChat *(Ours)*	**3.12**	**2.68**	**3.17**	**3.41**	**3.31**

Table H10. Quantitative evaluation for long video generation performance in global mode with human blind rating. CI stands for correctness of information, DO stands for detail orientation, CU stands for contextual understanding, TU stands for temporal understanding, and CO stands for consistency. The best result is highlighted in bold, and the second best is underlined.

Method	**CI**	**DO**	**CU**	**TU**	**CO**
Video Chat [3]	2.96	3.09	3.24	2.46	3.22
Video LLaMA [8]	2.42	2.85	2.87	2.00	2.87
Video-ChatGPT [4]	**3.11**	**3.32**	3.29	2.62	3.29
MovieChat *(Ours)*	3.07	3.24	**3.31**	**2.70**	**3.45**

Table H11. Quantitative evaluation for long video generation performance in breakpoint mode with GPT-3.5 [5]. CI stands for correctness of information, DO stands for detail orientation, CU stands for contextual understanding, TU stands for temporal understanding, and CO stands for consistency. The best result is highlighted in bold, and the second best is underlined.

Method	**CI**	**DO**	**CU**	**TU**	**CO**
Video Chat [3]	2.12	2.20	2.30	1.97	2.37
Video LLaMA [8]	1.62	1.85	2.20	1.34	2.02
Video-ChatGPT [4]	2.36	**2.26**	2.34	2.23	**2.70**
MovieChat *(Ours)*	**2.38**	2.16	**2.35**	**2.43**	2.68

Table H12. Quantitative evaluation for long video generation performance in breakpoint mode with Claude [1]. CI stands for correctness of information, DO stands for detail orientation, CU stands for contextual understanding, TU stands for temporal understanding, and CO stands for consistency. The best result is highlighted in bold, and the second best is underlined.

Method	**CI**	**DO**	**CU**	**TU**	**CO**
Video Chat [3]	2.17	2.24	2.89	1.87	2.75
Video LLaMA [8]	2.09	2.18	2.82	1.74	2.68
Video-ChatGPT [4]	2.39	2.36	**2.96**	2.10	2.89
MovieChat *(Ours)*	**2.48**	**2.41**	2.94	**2.33**	**3.12**

Table H13. Quantitative evaluation for long video generation performance in breakpoint mode with human blind rating. CI stands for correctness of information, DO stands for detail orientation, CU stands for contextual understanding, TU stands for temporal understanding, and CO stands for consistency. The best result is highlighted in bold, and the second best is underlined.

References

[1] Anthropic. Meet claude, 2023. 1, 4, 5, 6, 7

[2] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 2, 3

[3] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 4, 6, 7

[4] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2, 4, 6, 7

[5] openai. Gpt3.5, 2021. 2021. 1, 2, 3, 4, 5, 6, 7

[6] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2, 3

[7] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos, 2018. 2, 3

[8] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 4, 6, 7

[9] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 4, 6