CMTReorder: What is the Right Timeline of these Cross-Modal Fragments?

Anonymous ACL submission

Abstract

Timeline reordering is a crucial task in time series reasoning, where events need to be sorted along a temporal axis across various formats. While recent advancements in multimodal large language models (MLLMs) have shown promise in single-modal temporal reasoning, real-world data is often mixed and unstructured, with modalities existing independently without clear pairings. To address this gap, we introduce a novel task, Cross-Modal Timeline Reordering (**CMTReorder**), which evaluates the cross-modal temporal reasoning ability of MLLMs. The task consists of two tests: Cross-modal Direct Ordering, where models reorder the timeline directly, and Cross-modal Binary Decision, where models first make binary decisions on temporal relationships before reordering. We also present the MixStoryLine dataset, which includes text and image narratives from different time points. We evaluate CMTReorder using multiple MLLMs, including GPT-4o, LLaMA, and Deepseek. The results reveal significant challenges: GPT-4o achieves 24% consistent accuracy in direct ordering, 66.88% accuracy in binary judgment, and 9% consistent accuracy in the following reordering, with other models performing less effectively. These findings highlight the difficulty of cross-modal temporal inference and underscore the need for further improvements in model performance, while also offering insights for real-world applications.

1 Introduction

Timeline reorder is the task of correctly ordering events or items along a temporal axis, which can be represented in various media formats, such as text, images, or audio(Gangal et al., 2022). This task plays a critical role in applications like social media analysis(Wang et al., 2021; Chen et al., 2022), historical event reconstruction(Davis, 2011), medical diagnostics, and forensic analysis(Padilha et al., 2020), where the accurate sorting of mixed media is essential for understanding complex timelines.

Previous research on timeline reordering has primarily focused on single-modal studies. In the context of large language models, much of the work has concentrated on understanding long texts and reordering event descriptions in chronological order(Zeng et al., 2022; Gangal et al., 2022). Similarly, in multimodal settings, existing studies have mainly dealt with continuous storyline descriptions(Padilha et al., 2020; Li et al., 2019) that pair text with corresponding images.

These approaches, however, are limited to time series reasoning within a single modality, either text or image, and fail to address the fact that many real-world scenarios involve hybrid data. For example, in news reporting, event timelines may include multimedia content, such as articles, photographs and videos, which are often presented in a non-sequential order. In such cases, texts do not always have corresponding images, and vice versa. Without these explicit pairs, models often struggle to maintain temporal coherence, facing challenges in resolving ambiguous correlations between image and text data.

Given the shortcomings of current evaluation methods, we propose an extended **C**ross-**M**odal **T**imeline **R**eorder task (**CMTReorder**), which consist of two parts: *Cross-modal Direct Ordering* and *Cross-modal Binary Decision*. In the first part, the model is required to reorder the given options described by different modalities directly. In the second part, the model must make a binary decision on whether the temporal relationship between two options is correct. These two options are selected using the quick sort approach that humans use when reasoning, and the model should do the reorder task based on the above decisions.

To facilitate better experiments, we create the MixStoryLine cross-modal dataset based on the VIST dataset(Huang et al., 2016). It contains

200 carefully selected stories, each comprising sequences with mixed modalities.

We conducted our experiments using the latest MLLMs, such as GPT-4, ChatGLM, and LLaMA. The results highlight both the challenges and the potential of MLLMs in the CMTReorder task. Additionally, our comparative analysis of the two tests suggests that the logical reasoning employed by these models may not always be the most effective strategy, particularly for complex cross-modal tasks.

2 Related work

Time series inference has long been a critical area in exploring the capabilities of multimodal large language models (MLLMs) for temporal understanding(Liang et al., 2024; Jin et al., 2023b). As an essential task within time series inference, timeline reordering effectively reflects a model's ability to understand temporal and causal relationships in the progression of events(Rajani et al., 2019; Gangal et al., 2022), thus influencing model explainability, especially for explanation generation in MLLMs(Wiegreffe and Marasovic, 2021). This capability is valuable in various real-world applications, including historical event reconstruction, forensic analysis, and evidence investigations, among others(Jin et al., 2023a; Panaitescu, 2022).

Traditional research in timeline reordering has primarily focused on text-based data, aiming to reorder event descriptions, stories, or text snippets into chronological order. With the rise of large language models (LLMs), there has been a growing interest in leveraging their ability to process long contexts. This shift has spurred the development of various benchmarks focused on timeline reordering, such as LooGLE (Li et al., 2023) and Marathon (Zhang et al., 2023). For instance, Zhang et al. (2023) introduced the timeline reorder task in which a large language model is asked to rank three events, mentioned within a long context, in chronological order.

As LLMs continue to evolve, researchers are expanding the scope beyond text and incorporating multimodal approaches. This trend has led to the transfer of timeline reordering tasks into the realm of multimodal LLMs (MLLMs)(Ge et al., 2024), combining models from different modalities such as images and video. Early works on MLLMs mainly focused on generating single-sentence descriptions for visual content, but recent studies have explored more complex tasks, such as video storytelling(Gella et al., 2018), which involves generating coherent paragraph-length narratives. For example, Li et al. (2019) proposed a context-aware framework for multimodal embedding learning and developed a Narrator model to select video clips that best represent the underlying storyline. Yang et al. (2024)further advanced this area by refining methods to ensure that video descriptions align with a logical, coherent story, improving the narrative consistency across multimodal content.

However, a common limitation pervades all these studies: they predominantly focus on the model's comprehension of temporal relationships within the same modality, neglecting the fact that in the real world, temporal fragments often manifest in an interwoven, cross-modal fashion.

3 CMTReorder

To address the limitations of cross-modal time understanding tasks and to evaluate the performance of MLLMs in this domain, we propose a novel and challenging task, Cross-Modal Timeline Reordering (CMTReorder). As shown in Figure 1, this task consists of two main components: Cross-modal Direct Ordering and Cross-modal Binary Decision.

3.1 Dataset Construct

The Visual Storytelling Dataset (VIST) is a dataset designed for sequential vision-to-language tasks (Huang et al., 2016). It consists of distinct photos organized into different sequences or stories, each paired with descriptive and narrative text.

We use this dataset as the source corpus and reconstruct it through the following steps. First, we group all descriptions of a complete story together. Each description within the story is treated as an option, and the initial timeline order is disrupted. Next, we randomly select descriptions from the options to perform alignment replacement, which involves hiding the text and replacing it with the corresponding images. Finally, to ensure the accuracy and consistency of these scattered options, we manually verify them and use a large model to generate images for any sections where the timeline is unclear or images are missing. This process results in our MixStoryLine dataset, which consists of 200 carefully selected cross-modal stories.

3.2 Cross-modal Direct Ordering

In this test of the task, we ask the MLLMs to directly reorder the cross-modal options presented

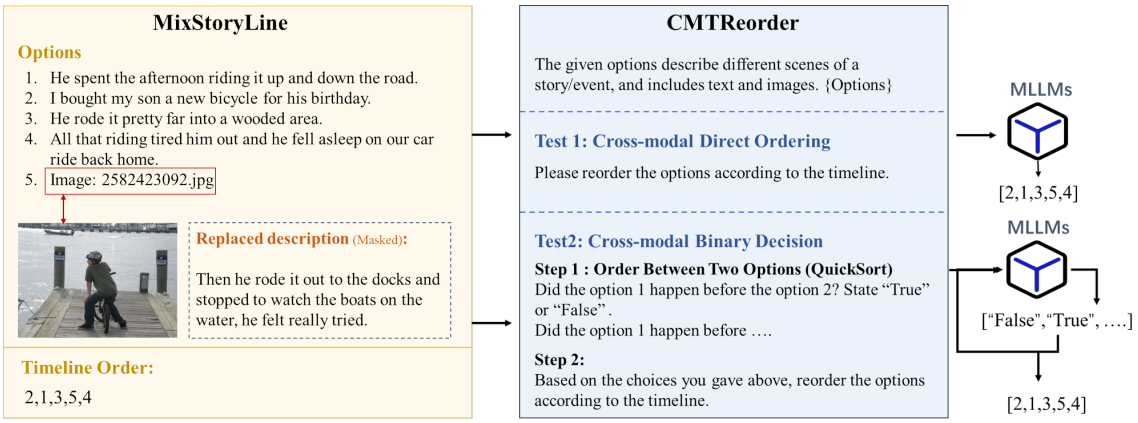

Figure 1: The description of task CMTReorder. Task CMTReorder requires MLLMs to reorder a given set of options according to a timeline. The options may consist of descriptive text or images. To better assess the temporal understanding capabilities of MLLMs, we design two tests. Test 1 follows the traditional question-and-answer method, where the model is tasked with directly completing the timeline reordering. In Test 2, the model is asked to make binary judgments about the options based on the principles of quick sorting, and then output the final reordered sequence based on these judgments.

in the timeline. Following a traditional question-and-answer format, we first prompt the model to describe the images in the options in detail. Then, the model is tasked with ranking all the options in chronological order. Throughout this process, the model must integrate both textual and visual information to restore the complete story. The task requires the model to effectively synthesize mixed-mode data (text and images) to generate a coherent sequence of events.

3.3 Cross-modal Binary Decision

In the second test, we want to find how well the model thinks in terms of human logic when sorting.Therefore, rather than testing timeline reordering through sequential ordering, we reframe the task as a binary classificate judgement problem, that is, through comparing each of the two options based on quick sort method to arrive at an interpretable sorting sequence. Specifically, The model is asked to determine whether the first event occurred before the second one for any given pair of options, answering "True" or "False." Based on these responses, the model was required to give the right order. Since options can appear in multiple modes, this part of the test specifically evaluates the MLLMs' ability to understand time-series relationships within the same modality, as well as its capacity to infer temporal connections across different modalities.

4 Experiment

4.1 Evaluate MLLMs

In our experiment, we incorporated a diverse array of MLLMs, including the open source models like Llama-3.2-11B-Vision-Instruct (Meta, 2024), VisualGLM-6B (Ding et al., 2021)[1], InternVL2_5-8B (Chen et al., 2024)[2] and deepseek-vl2-tiny (Wu et al., 2024)[3] and powerful closed-source model such as ChatGPT-4o[4] (GPT-4O-2024-08-06).

4.2 Environment and Setting

Our experiments are conducted on Linux with 10 A100 80GB GPUs. All the weight of open source models is download from hugging-face. For ChatGPT-4o, we just call it's API. The temperature of all models is 0.1 and the max_new_tokens is 1024.

4.3 Evaluation Metrics

To evaluate the model's performance on CMTReorder, we employed four metrics: **Accuracy** measures how closely the model's predicted order matches the actual labels, while **TF_Accuracy** evaluates the model's performance in the binary decision test. **Kendall's Tau** and **Spearman's Rank Correlation Coefficient** assess the correlation and consistency between two ranked sequences, offering insights into the alignment of the model's out-

[1]https://github.com/THUDM/VisualGLM-6B/tree/main
[2]https://huggingface.co/OpenGVLab/InternVL2_5-8B
[3]https://huggingface.co/deepseek-ai/deepseek-vl2
[4]https://platform.openai.com/docs/models/gpt-4o

Table 1: Experiment Results of MLLMs in CMTReorder

MLLMS	Cross-modal Direct Ordering			Cross-modal Binary Decision			
	Acc	Kendall's	Spearman	TF_Acc	Acc	Kendall's	Spearman
ChatGPT-4o	0.2400	0.7171	0.7822	0.6688	0.0600	0.3717	0.4091
Llama-3.2-VL	0.0500	0.3006	0.3477	0.5575	0.0450	0.1232	0.1333
VisualGLM	0.0450	0.1411	0.1600	0.1650	0.0250	0.0440	0.0280
InternVL	0.0900	0.3930	0.4598	0.5700	0.0550	0.3847	0.4260
DeepSeek-VL	0.0400	0.0220	0.0260	0.3588	0.0230	0.0230	0.0335

put with the true temporal order. The detailed calculation formulas for these metrics are provided in the appendix. Meanwhile, due to the randomness of large model generation, we use regular expressions to extract the model's responses for outputs that do not meet the required format.

4.4 Result and Analysis

The overall experiment results for various MLLMs on the CMTReorder task are presented in Table 1. The leading model, GPT-4o, achieves an accuracy of 24% in the direct ordering test, significantly outperforming Inter-VL, which achieved only 9%. Additionally, GPT-4o achieved high scores in the Kendall's Tau (0.717) and Spearman's Rank Correlation Coefficient (0.782), far surpassing other open-source models.

In the second test, the performance gap between models was smaller. GPT-4o achieved 66.88% accuracy in the timing judgment task between any two modes, but its accuracy in the final ranking dropped by 18% (from 24% to 6%) compared to the direct ordering test. The other two evaluation metrics also showed consistent declines. Similar trends were observed in the other models.

The performance drop can be attributed to several factors. While all options were provided upfront in both tests, the task of determining temporal relationships between pairs adds complexity. Analyzing each pair individually may limit the model's understanding of the broader context, causing misinterpretations. Meanwhile, inconsistencies between pairs may not be apparent in isolation, leading to errors that affect the final ranking.

Overall, the results suggest that even advanced models face challenges with the proposed task. For the models tested, direct guidance for timeline reorder led to better results than having the model follow a step-by-step logical process of determining temporal relationships between pairs. This highlights that providing more individual clues does not necessarily improve the model's temporal understanding, in fact, it may hinder the model's reasoning ability by forcing it to work with fragmented information. This finding suggests that MLLMs may struggle to mimic human-like reasoning, which often integrates multiple cues simultaneously rather than processing isolated judgments.

5 Conclusion and Future Work

In this paper, we introduce the Cross-Modal Timeline Reordering (CMTReorder) task and present the MixStoryLine dataset. Our evaluation of several multimodal large language models (MLLMs), including GPT-4o, ChatGLM, and Inter-VL, reveals that even state-of-the-art models struggle with the task, with GPT-4o achieving only 24% consistent accuracy in direct ordering. These findings suggest that cross-modal temporal comprehension remains a challenging problem, highlighting the need for further model improvement.

Our experiments also show that increasing temporal cues may actually hinder performance, particularly when models are asked to reason step-by-step across multiple modalities. Rather than providing isolated cues, a more holistic approach where the model has access to all information at once may lead to better results. This suggests that mimicking human-like logical reasoning in models might not always be the optimal strategy, especially for complex cross-modal tasks.

In our future work, there is a clear need to further explore the ability of models to understand and reason about cross-modal timing, a crucial aspect for real-world applications. Additionally, the interpretability of temporal reasoning in models remains a significant challenge. Enhancing models' understanding of timelines, while also improving transparency in their reasoning processes, will be key areas for future research.

Limitations

While CMTReorder provides a valuable method for evaluating the temporal inference ability of MLLMs, it has several limitations: (1) The MixStoryLine dataset is relatively small compared to other large-scale datasets, which may affect the accuracy and generalizability of the evaluation. Expanding the dataset would provide a more comprehensive assessment. (2) In the second test, we used the quicksort approach to simulate human reasoning. However, human thinking is more flexible, and this method may not fully capture how humans perform temporal reasoning tasks. (3) The MLLMs tested are only a subset of available models. Testing additional models is necessary to assess the task's generality and the robustness of the results.

Ethics Statement

All work in this paper adheres to the ACL Code of Ethics. However, some ethics problems may arise in the process of using MLLMs generation. We strictly adhere to the licenses and policies governing the use of released MLLMs. In the task, we try to limit the generation of the model to the scope of the given data. However, we do not guarantee that the content generated by these models is safe or harmless on the experiement.

References

Xiuying Chen, Mingzhe Li, Shen Gao, Zhangming Chan, Dongyan Zhao, Xin Gao, Xiangliang Zhang, and Rui Yan. 2022. Follow the timeline! generating an abstractive and extractive timeline summary in chronological order. *ACM Transactions on Information Systems*, 41:1 – 30.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Stephen Boyd Davis. 2011. Book review: Cartographies of time: A history of the timeline. *Journal of Visual Culture*, 10:269 – 271.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835.

Varun Gangal, Steven Y Feng, Malihe Alikhani, Teruko Mitamura, and Eduard Hovy. 2022. Nareor: The narrative reordering problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10645–10653.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Spandana Gella, Mike Lewis, and Marcus Rohrbach. 2018. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 968–974.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023a. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. 2023b. Large models for time series and spatio-temporal data: A survey and outlook. *arXiv preprint arXiv:2310.10196*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2019. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2):554–565.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. 2024. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6555–6565.

AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024.

Rafael Padilha, Fernanda Alcantara Andalo, and Anderson Rocha. 2020. Improving the chronological sorting of images through occlusion: A study on the notre-dame cathedral fire. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2972–2976.

Diana Maria Panaitescu. 2022. The use of time reorder as a literary plot device. *P'Arts' Hum*, 2(2):39–50.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Shang Wang, Zhiwei Yang, and Yi Chang. 2021. Bringing order to episodes: Mining timeline in social media. *Neurocomputing*, 450:80–90.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Dingyi Yang, Chunru Zhan, Ziheng Wang, Biao Wang, Tiezheng Ge, Bo Zheng, and Qin Jin. 2024. Synchronized video storytelling: Generating video narrations with structured storyline. *Preprint*, arXiv:2405.14040.

Ailing Zeng, Mu-Hwa Chen, L. Zhang, and Qiang Xu. 2022. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*.

Lei Zhang, Yunshui Li, Ziqiang Liu, Junhao Liu, Longze Chen, Run Luo, Min Yang, et al. 2023. Marathon: A race through the realm of long context with large language models. *arXiv preprint arXiv:2312.09542*.

A Appendix

A.1 Prompt Template

We designed prompt templates for two test sections of the CMTReorder task, each following the steps outlined in section 3.2 and 3.3. Figure 2 detailed describe the template of Cross-modal Direct Ordering.

In the Cross-modal Binary Decision test, the model is required to determine the temporal order of two options. Due to the variability in input types, the options may consist of text-text, image-image, or text-image pairs. To accommodate these possibilities, we adjusted the prompts accordingly. Examples of the samples and corresponding instructions are shown in Figure 3. Note that in Prompt 2.1, the two options are chosen by quicksort method and then reordered based on the results of multiple rounds of conversations with the model. Prompt 2.2 refers to the last round of task prompts after the above dialogue rounds.

A.2 Metrics

The formulas for calculating the evaluation metrics used in the task are provided in detail below.

- **Accuracy & TF_Accuracy:** Accuracy measures the degree to which the ranking predicted by the model aligns with the actual labels, while TF_Accuracy is used to evaluate the model's performance in the judgment test.

$$acc/tf\_acc = \frac{No.\ of\ correct\ items}{No.\ of\ items} \quad (1)$$

- **Kendall's Tau:** A non-parametric statistical measure used to evaluate the correlation between two ranked sequences. It assesses the consistency between the sequences by counting the number of concordant and discordant pairs, formulated as:

$$\tau = \frac{C - D}{\sqrt{(C + D + T_1)(C + D + T_2)}} \quad (2)$$

  where C and D are the numbers of concordant and discordant pairs, and T_1, T_2 are the numbers of ties in the predicted/ground truth variables.

- **Spearman's Rank Correlation Coefficient:** This metric evaluates the consistency between two sequences by calculating the Pearson correlation coefficient of their ranks. It is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (3)$$

  where d_i is the difference between the ranks of each pair of values, n is the number of data points. Similar to Kendall's Tau, it ranges from -1 to 1, indicating varying degrees of consistency between sequences.

#Prompt 1:

You are an expert at conducting temporal inference of a story/event. You will be given several options, each describing a scene of the story. Each option may be a text or an image. Your task is to carefully read the given options, understand the content and details of each, and answer the question accurately. Carefully consider these options and provide your answers along with reasoning.

Question:
Please inference the right order of the following options based on the timeline, use sequence number to indicate the option.

Options:
{The giving options}

The answer sequence should start by[BIO] and end with [EOF]. For example, {the right order is [BIO]2,3,1,4,5[EOF]}.

Figure 2: The prompt template of test 1 Cross-modal Direct Ordering.

#Prompt 2.1:
You are an expert at conducting temporal inference of a story/event. You will be given several options, each describing a scene of the story. Each option may be a text or an image. Your task is to carefully read the given options, understand the content and details of each, and answer the questions accurately. Carefully consider these options and provide your answers along with reasoning.

Question1:
Is there a picture format in Option {A} and Option {B}? If there is a picture, please describe it in details.

Question2: (Judgement)
Did the first scene {A} happen before {B} ? State "True" or "False".

Options:
{The giving options}

#Prompt 2.2:

Question3:
Based on the above **Judgement,** Please inference the right order of the following options based on the timeline, use sequence number to indicate the option.

Options:
{The giving options}

The answer sequence should start by[BIO] and end with [EOF]. For example, {the right order is [BIO]2,3,1,4,5[EOF]}.

Figure 3: The prompt template of test 2 Cross-modal Binary Decision. Questions 1 and 2 will involve multiple rounds of dialogue, while Question 3 will be part of the last round.