MEMORY HELPS, BUT CONFABULATION MISLEADS: UNDERSTANDING STREAMING EVENTS IN VIDEOS WITH MLLMS

Gengyuan Zhang[♠,♦,*], **Mingcong Ding**[♠,*], **Tong Liu**[♠,♦], **Yao Zhang**[♠,♦], **Volker Tresp**[♠,♦]
[♠]Ludwig-Maximilians-Universität München, Germany
[♦]Munich Center for Machine Learning, Germany
zhang@dbs.ifi.lmu.de

ABSTRACT

Multimodal large language models (MLLMs) have demonstrated strong performance in understanding holistic videos. However, their ability to process streaming events—represented as sequences of visual clips—remains underexplored. Intuitively, leveraging prior events as memory can enrich the contextual and temporal understanding of current events. Inspired by this, we show in this paper that using preceding events as context, *i.e.*, memory, helps MLLMs better comprehend video events. However, such memory relies on MLLMs' predictions of prior events and inevitably accumulates misinformation in a streaming setting, leading to confabulation in contexts and degraded performance. To address this, we propose a confabulation-aware memory modification method that mitigates the impact of confabulated memory for improved memory-enhanced event understanding.

1 INTRODUCTION

Leveraging Multimodal Large Language Models (MLLMs) to understand events in videos (Yu et al., 2024; Zhang et al., 2023; Maaz et al., 2024; **?**) has shown their effectiveness in a wide range of tasks like video question answering and video captioning due to the excellent reasoning capabilities of backbone LLMs.

Much of the existing research focuses on understanding holistic videos (Zhang et al., 2023; Li et al., 2024a), where videos are treated as a whole, and MLLMs take videos' global representation as inputs without performing more fine-grained understanding. In real-world scenarios, however, videos are often streaming and multi-event in nature, with events unfolding sequentially. Events in videos exhibit significant temporal, causal, and contextual dependencies over preceding events due to their inherent semantic correlation. Akin to human cognition, knowledge of past events can help in understanding the present effectively. This intrigues us as to how MLLMs understand events in videos in a sequential manner.

We treat preceding events in videos as memory and thus begin by exploring how incorporating memory in LLMs can help enhance event understanding in videos. A common practice for forming memory in LLMs is to prepend relevant contexts to the inputs of LLMs/MLLMs (Behrouz et al., 2024; Zhang et al., 2024b; Wang et al., 2024b) where memory and knowledge are retained and recalled in tokens. Specifically, we specify the memory in the streaming video setting as event-triggered memory (Fountas et al., 2025; Hatalis et al., 2023) since they are contextualized and dependent on the current scenes in videos, other than external event-agnostic knowledge. The **memory** consists of *long-term* memory (Hatalis et al., 2023), recollected from other episodes or videos with relevant events, and *short-term* memory of events that directly precede the current one.

As shown in Fig. 1 and later in Tab. 1, integrating memory as contexts, which consist of events and their narrations (detailed in Sec. 2), significantly outperforms the off-the-shelf zero-shot model, highlighting that knowing the past notably helps MLLMs.

[*]Equal contribution

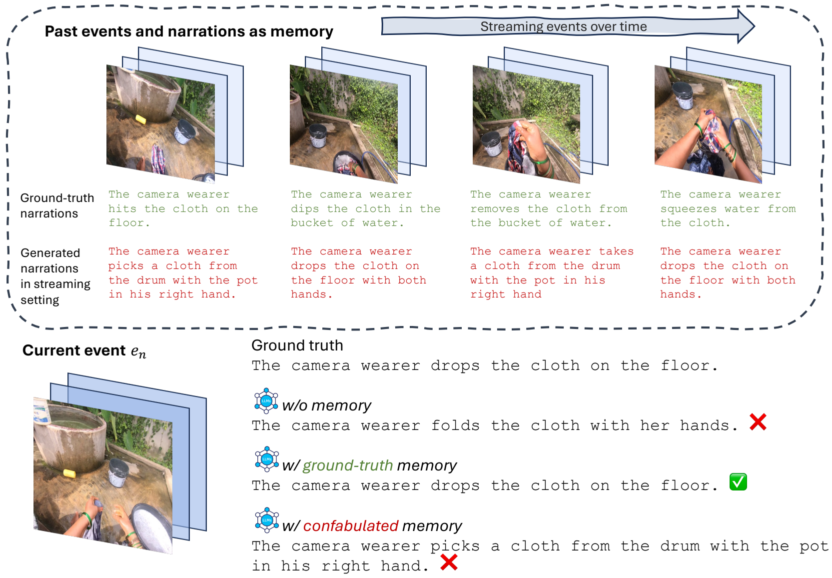


Figure 1: Possessing memory of past events can help understand the current event. However, for streaming events in videos, we cannot access ground-truth narrations for previous events and this leads to confabulation.

Nevertheless, this performance represents a gold-standard baseline—feasible only in ideal scenarios rather than in a practical streaming setting. In practice, MLLMs don't access the ground truth memory, rendering the leveraged short-term memory prone to mispredictions in the previous rounds due to hallucinations or incomplete observations. We reevaluate the model under streaming conditions and observe a substantial performance drop which can be attributed primarily to misinformation generated in the memory, a phenomenon referred to as **confabulation** (Sui et al., 2024).

To address the inevitable confabulated memory in the streaming setting, we introduce CAMEO, a confabulation-aware memory modification approach to mitigate the confabulation problem.

Our paper can be summarized as follows:

1. We demonstrate that leveraging memory as context in MLLMs improves understanding of video events by introducing contextual and temporal knowledge;
2. Yet ground-truth memory is not accessible in the real streaming setting since the memory suffers from misinformation in MLLMs' generation, namely confabulation;
3. We propose CAMEO, a memory modification approach, to mitigate confabulated memory in MLLMs.

2 Leverage Memory as Context in Event Understanding

2.1 Preliminary

Given a video episode $\mathcal{E}$ consisting of a sequence of N events $\{e_1, \ldots, e_N\}$ where an arbitrary event e_i in the episode is a short video clip consisting of multiple frames. LLMs need to predict the narration $\hat{t}_n$ of the current query event e_n with only access to the preceding events $e_{1:n-1}$.

Memory Memory (Fountas et al., 2025; Das et al., 2024) in LLMs, akin to human brains, refers to recollecting relevant events from past experiences. We denote the memory activated by the current event e_n as $\mathcal{M}$. We adopt two types of memory: long-term memory $\mathcal{M}^l$ from the set of other episodes as persistent memory $\mathcal{D}$ and short-term memory $\mathcal{M}^s$, which respectively means the memories that occurred earlier and memories in the current episode. Memory as contexts $\mathcal{M}$ is defined as a set of event and narration pairs recollected:

$$\mathcal{M} = \mathcal{M}^l \cup \mathcal{M}^s, \quad \mathcal{M}^l = \{(e_1^l, t_1^l), \ldots, (e_{N_l}^l, t_{N_l}^l)\}, \quad \mathcal{M}^s = \{(e_1^s, t_1^s), \ldots, (e_{N_s}^s, t_{N_s}^s)\}. \tag{1}$$

To collect events e_j^l in $\mathcal{M}^l$, we use a similarity-based retrieval method. In contrast, for events e_k^s in $\mathcal{M}^s$, we rely on recency, selecting the most recent events relative to the current event (detailed in Sec. C.1).

Memory as Context A prominent trend in memory-enhanced LLMs is to use memory as contexts in LLMs' inputs (Fountas et al., 2025; Behrouz et al., 2024), thereby leveraging the In-Context Learning (ICL) capability of LLMs (Brown et al., 2020; Gao et al., 2023). We formulate long-term and short-term memories as prepended contexts to the inputs.

Algorithm 1: Streaming Evaluation

Input: Current query event e_n ,
Current episode up to current event $\mathcal{E}_{:n-1}$,
Persistent memory $\mathcal{D}$
Output: Prediction of narration t_n
Initialize $\mathcal{M}^s \leftarrow \varnothing$;
for $n = 1$ *to* N **do**
  Collect long-term memory from $\mathcal{D}$
  for $k = 1$ *to* N_l **do**
    $\mathcal{M}^l \leftarrow (e_k^l, t_k) \in \mathcal{D}$;
  end
  $\hat{t}_n \leftarrow \mathrm{LLM}(\mathcal{M}, e_n)$;
  Update short-term memory with new prediction
  $\mathcal{M}^s \leftarrow (e_n, \hat{t}_n)$;
end
return $\hat{t}_n$;

2.2 Gain & Loss: Memory in Event Understanding

An Upper Bound: Ground-truth Memory helps We begin by introducing a golden baseline: we use $\mathcal{M}$ with ground-truth narrations in the short-term memory $\mathcal{M}^s$, denoted by $\mathcal{M}^s_{gt}$. In this setup, all memories are factual and plausible; therefore, providing an upper bound on the performance of memory-enhanced MLLMs.

Confabulated Memory Misleads Streaming Event Understanding We next evaluate LLMs' event understanding capabilities in a streaming scenario, as explained in Alg. 2. This setting is more practical and challenging since the model cannot access ground-truth short-term memory $\mathcal{M}^s_{gt}$ or rather the true narrations and can only utilize generated narrations up to the current event $\mathcal{M}^s_{gen}$.

As shown in Fig. 2, our streaming evaluation relies on the short-term memory that incorporates earlier events and their predicted narrations. Each incoming event's prediction is subsequently consolidated into the short-term memory of the current episode.

We will elucidate the experiments in the Sec. 3.3.

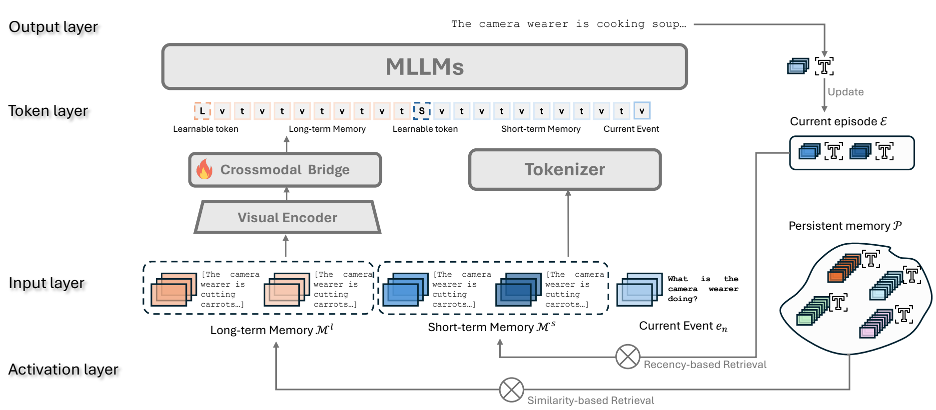


Figure 2: Model Pipeline for memory as contexts for streaming events reasoning. We interleave the events and narrations from the long-term and the short-term memory as contextual inputs.

2.3 CAMEO: Mitigating Confabulation

We propose a probing-and-surgical approach, CAMEO (**C**onfabulation-**A**ware **ME**mory m**O**dification), to mitigate confabulation in our setting. We treat confabulated memories as less factual and less credible contexts for LLMs. Accordingly, we propose to quantize the credibility of short-term memory and reduce the contribution of heavily confabulated memories.

Uncertainty Estimation We adopt the semantic entropy proposed by Kuhn et al. (2023); Farquhar et al. (2024) as a proxy for assessing the credibility of generated narrations. Specifically, we sample S generations of target narrations $\{\hat{t}_n^{(1)}, \dots, \hat{t}_n^{(S)}\}$ from the output distribution of the LLMs conditioned on the memory $\mathcal{M}$ and query event e_n from the posterior distribution $p(\hat{t}_n|\mathcal{M}, e_n))$. *Next*, we cluster the sampled narrations by their semantic equivalence, as described by Kuhn et al. (2023), to identify the semantic clusters C of generated narrations. The resulting semantic entropy is calculated as in Eq.2:

$$se(\hat{t}_n) = -\sum_c p(c|\mathcal{M}, e_n) \log p(c|\mathcal{M}, e_n)) \tag{2}$$

Probing Confabulation-Prone Heads Attention heads have been shown to represent different subspaces of contextual information in LLMs (Deng et al., 2024; Elhage et al., 2021; Meng et al., 2022) and (Deng et al., 2024) demonstrates that modifying attention heads can mitigate misinformation in textual contexts.

Inspired by (Deng et al., 2024), we aim to locate the confabulation-prone attention heads in the transformer architecture. To do so, we manipulate one ground-truth narration t_j with a random, irrelevant narration t_j^- in $\mathcal{D}$, ablate an arbitrary attention head h in the transformer, and then calculate the output logits to quantify the indirect effect (IE) (Meng et al., 2022; Deng et al., 2024), thereby identifying the influence of this attention head.

Confabulation-aware Memory Modification Once we identify the confabulation-prone heads, we can reweight the attention in the most influential attention heads to the confabulated memory $\hat{t}_n$ them according to their semantic entropy, as detailed in Appx. C.3. We convert semantic entropy into weights as follows:

$$w(\hat{t}_n) = 1/exp(-\tau \times se(\hat{t}_n)) \tag{3}$$

where τ is a temperature coefficient, with which we can tweak the modification scale.

3 EXPERIMENTAL STUDY

We elaborate the model design (Sec. 3.1), implementation details(Sec. 3.2), and results (Sec. 3.3).

3.1 MODEL DESIGN

Model Architecture Our models follow common MLLM paradigm (Liu et al., 2023; Dai et al., 2023; Liu et al., 2024; Li et al., 2023): a visual encoder, a cross-modal bridge, and an LLM backbone. Events, treated as short video clips, can vary in length, so we downsample them to a fixed length $l_e = 8$. We then encode each frame individually with the CLIP (Radford et al., 2021) visual encoder. Subsequently, we use an event-aware bridge to encode the events and align them to the language space of LLMs, as shown in Fig. 2. More details can be found in Appx. C.1.

Table 1: Model Performance. Memory improves event understanding, while confabulated memory may mislead MLLMs.

Model	STS	Rouge-L	BLEU
OPT-2.7B			
w/o memory	0.495	0.520	0.179
Δ	+0.160	+0.103	+0.167
w/ ground-truth memory (16 shots)	0.655	0.623	0.346
Δ	-0.152	-0.082	-0.163
w/ confabulated memory (16 shots)	0.503	0.541	0.183
Vicuna-7B			
w/o memory	0.615	0.588	0.240
Δ	+0.178	+0.145	+0.293
w/ ground-truth memory (16 shots)	0.793	0.733	0.534
Δ	-0.137	-0.142	-0.220
w/ confabulated memory (16 shots)	0.656	0.591	0.315

By encoding all the events in the memory $\mathcal{M}$, we interleave these events with their narrations and provide them as context to the model, alongside the new query event and the question.

Model Selection We employ two widely used Large Language Models in our experiments: OPT-2.7B (Zhang et al., 2022) and Vicuna-7B (Zheng et al., 2023). We also use the Q-Former proposed by Li et al. (2023) as an effective bridge to compress visual features from clips. Although MLP-based projection (Liu et al., 2024; 2023) is a popular method for adapting vision to the language space, it increases computational overhead when processing multiple video inputs, making compression essential. We initialize the Q-Former for OPT-2.7B from Dai et al. (2023) and for Vicuna-7B from Zhu et al., then unfreeze each with training data.

3.2 IMPLEMENTATION

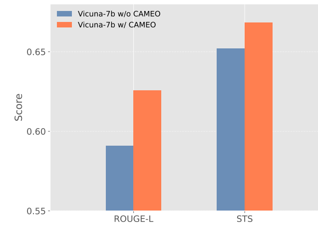
Training Details

To elicit the multimodal in-context learning capability of MLLMs, we retrain them with memory $\mathcal{M}$ as interleaved video-text data (Yu et al., 2024; Alayrac et al., 2022; Li et al., 2024b; Wang et al., 2024a), following Alg.2 in Appx. C.2.

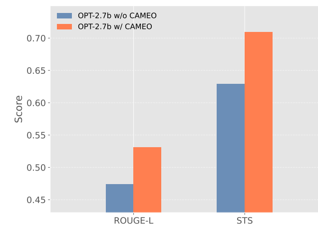
We pad 8 events sampled from $\mathcal{D}$ and 8 most recent events in the current episode $\mathcal{E}$ and adopt two new learnable tokens to distinguish different memories. The total shot number is 16. We unfreeze the Q-Former to adapt to visual events.

Dataset We use the Ego4D dataset (Grauman et al., 2022) for our task. To construct $\mathcal{D}$ from past episodes and for training, we utilize the training split. To probe the confabulation-prone heads, we rely on the validation set. Finally, for all evaluations, we employ the test set.

Evaluation We evaluate the narrations with Semantic Textual Similarity (STS) by Reimers & Gurevych (2019), ROUGE-L by Lin (2004), BLEU by Papineni et al. (2002).



(a) Vicuna-7B



(b) OPT-2.7B

Figure 3: Performance improvements with CAMEO.

3.3 RESULTS

Memory helps, but confabulation misleads We first compare 3 baselines: MLLMs without memory, MLLMs with ground-truth memory (w/ ground-truth narrations), and MLLMs with confabulated memory (w/ generated narrations) in Tab. 1.

First, we observe a significant performance gain when MLLMs leverage memory as contexts, yielding a large margin of improvement across all mentioned metrics for both models.

However, the performance drop in streaming evaluation is concerning. Its primary cause is that in the streaming scenario, we replace ground-truth narrations with previously generated narrations of recent events. This inevitably leads to a performance drop, stemming from the confabulation introduced by earlier predictions.

Mitigating Confabulation As shown in Fig. 3, with CAMEO, the reasoning capabilities of MLLMs for streaming event understanding improve significantly. This result underscores the effectiveness of CAMEO in combating confabulated memory.

4 CONCLUSION

In this work, we demonstrated how memory, leveraged as context, can inject temporal and contextual knowledge into MLLMs and enhance their understanding of events in streaming videos. However, in practical streaming scenarios, accessing the ground-truth memory of prior events is often impossible, leading to confabulation when the current predictions depend on previously generated narrations. We addressed this issue by proposing a confabulation-aware attention modification mechanism, CAMEO, which uses the semantic uncertainty of predicted narrations as a proxy for the credibility of potentially confabulation-prone memory. We show that this approach effectively mitigates confabulated memory.

Limitations In this work, we focus on forming memory as context for MLLMs. While this approach is training-free and allows us to exploit MLLMs' capabilities fully, future efforts could investigate parameterized memory components, as discussed by Behrouz et al. (2024).

Our current setting relies on a predefined semantic boundary between events in the original dataset to simplify the task. Thus, videos are streamed at the event level rather than the frame level. A more realistic streaming scenario would be more challenging because the semantic boundary is unknown.

Due to computational constraints, our model selection is limited. Implementing interleaved video-text inputs requires a powerful architecture to compress visual features and reduce token length (*e.g.*, Q-Former). This hinders our ability to explore other models that do not compress visual features, as their computational complexity grows quadratically with the amount of memory shots.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yixin Chen, Shuai Zhang, Boran Han, Tong He, and Bo Li. CaMML: Context-aware multimodal learner for large models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4056–4071, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.223. URL `https://aclanthology.org/2024.acl-long.223/`.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarath Swaminathan, Sihui Dai, Aurélie Lozano, Georgios Kollias, Vijil Chenthamarakshan, Soham Dan, et al. Larimar: Large language models with episodic memory control. *arXiv preprint arXiv:2403.11901*, 2024.

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. Cram: Credibility-aware attention modification in llms for combating misinformation in rag. *arXiv preprint arXiv:2406.11497*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Zafeirios Fountas, Martin Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou Ammar, and Jun Wang. Human-like episodic memory for infinite context LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=BI2int5SAC`.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. Memory matters: The need to improve long-term memory in llm-agents. In *Proceedings of the AAAI Symposium Series*, volume 2, pp. 277–280, 2023.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13504–13514, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. Textbind: Multi-turn interleaved multimodal instruction-following in the wild. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 9053–9076, 2024b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.679. URL https://aclanthology.org/2024.acl-long.679/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. Confabulation: The surprising value of large language model hallucinations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14274–14284, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.770. URL `https://aclanthology.org/2024.acl-long.770/`.

Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. Cosmo: Contrastive streamlined multimodal model with interleaved pre-training. *arXiv preprint arXiv:2401.00849*, 2024a.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36, 2024b.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Keunwoo Yu, Zheyuan Zhang, Fengyuan Hu, Shane Storks, and Joyce Chai. Eliciting in-context learning in vision-language models for videos through curated data distributional properties. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20416–20431, 2024.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13647–13657, 2024a.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL `https://aclanthology.org/2023.emnlp-demo.49/`.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

APPENDIX

A NOTATION TABLE

We summarize the notations used in the paper in Tab. 2.

Table 2: Notations used in the paper.

Notation	Description
$\mathcal{E}$	A video episode
$\mathcal{D}$	Persistent memory
$\mathcal{M}$	Memory (overall)
$\mathcal{M}^l$	Long-term memory
$\mathcal{M}^s$	Short-term memory
e_i	i-th event
e_n	the current query event
t	Narration of an event
$\hat{t}$	Generated narrations of an event
(e, t)	A pair of event and narration
N_l	The size of long-term memory
N_s	The size of short-term memory
c	Semantic cluster of sampled narrations
$\mathrm{LLM}(\cdot)$	generation of Large Language Model

B RELATED WORKS

B.1 TEMPORAL UNDERSTANDING IN VIDEOS

Understanding temporal ques in videos is challenging and MLLM-based methods draw great attention. A wide range of work (Maaz et al., 2024; Qian et al., 2025; Zhang et al., 2024a; He et al., 2024) leverages the reasoning ability of MLLMs and presents great performance in understanding holistic videos.

B.2 MEMORY AS CONTEXTS IN LLMS

Implementing memories has been a big topic. Due to the emergent abilities of LLMs (Wei et al., 2022) to understand and leverage context, using contexts as memory has been one mainstream approach for NLP tasks(Fountas et al., 2025) and also multimodal tasks(Fan et al., 2024; Chen et al., 2024).

C EXTENSION OF EXPERIMENTAL STUDY

C.1 MODEL IMPLEMENTATION

Similarity-based Retrieval We retrieve relevant events from $\mathcal{M}^l$ based on their semantic similarity with the current query event. We use the features in the output layer of the CLIP visual encoder and calculate their cosine similarity with the query event.

Recency-based Retrieval We collect the events from $\mathcal{M}^s$ based on recency, namely the most N_s recent events will be sampled from the current episode $\mathcal{E}$ from the current event e_n .

C.2 TRAINING

We elaborate our training algorithm as in Alg. 2. For training, we used 16 shots of memory events: 8 from long-term memory and 8 from short-term memory in the same episode. We use the ground-truth annotations in the training to adapt the MLLMs to learn the temporal and contextual knowledge from the memory.

We adopt the next token prediction loss. For training, we utilize 4 Nvidia A100 for 75 hours on the full training set of Ego4D with 5 epochs. The learning rate is set to 1e-5 and the batch size to 8. For streaming evaluation, we adopt 1 Nvidia A40 for 4 hours.

Algorithm 2: Training

Input: Current query event e_n ,
Current episode $\mathcal{E}$,
Persistent memory $\mathcal{D}$
Output: Prediction of narration t_n
Collect short-term memory from the current episode $\mathcal{E}$
for $j = 1$ *to* N_s **do**
  $\mathcal{M}^s \leftarrow (e_{n-j}, t_{n-j}) \in \mathcal{E}$;
end
Collect long-term memory from $\mathcal{D}$
for $k = 1$ *to* N_l **do**
  $\mathcal{M}^l \leftarrow (e_k^l, t_k) \in \mathcal{D}$;
end
$\mathcal{M} \leftarrow \mathcal{M}^l \cup \mathcal{M}^s$;
$t_n \leftarrow \mathrm{LLM}(\mathcal{M}, e_n)$;
return t_n ;

C.3 CAMEO Implementation

Probing Confabulation-prone Attention Heads We selected 9 episodes consisting of 189 events from the validation split of Ego4D. For each event, we compose the memory set $\mathcal{M}$ with 8 long-term memories and 8 ground-truth short-term memories and randomly replace one narration of a short-term event with a wrong narration. Then we ablate each attention head in each layer to calculate the influential effect. As shown in Fig. 8, we locate the top-k confabulation-prone attention heads with the largest IE.

In the end, we pick up top-96 confabulation-prone attention heads and use a temperature of 0.6 for Vicuna-7B; top-32 confabulation-prone attention heads and use a temperature of 0.8 for OPT-2.7B.

Attention Modification in CAMEO Considering the memory as contexts $\mathcal{C} = \{e_1^l, t_1^l, \ldots, e_1^l, t_{N_l}^l, e_{N_l}^s, \hat{t}_1^s, \ldots, e_{N_s}^s, \hat{t}_{N_s}^s\}$, each generated narration $\hat{t}$ has its semantic entropy score $se(\hat{t}^s)$ and its modification weight $w(\hat{t}^s)$. For other contextual inputs like e and t^l, the modification weight is set to 1. The resulting modification weight is $\mathbf{w}$ of which k-th token has the weight,

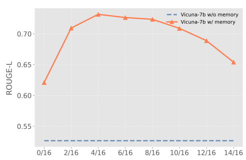
$$\mathbf{w}_k = \begin{cases} w(\hat{t}^s) & \text{if } token_k \text{ belongs to } \hat{t}^s, \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

For an arbitrary attention head h in transformers, we multiply the attention matrix $\mathbf{A}_h$ by the modification weight $\mathbf{w}$ elementwisely $\mathbf{A}_h \circ \mathbf{w}$.

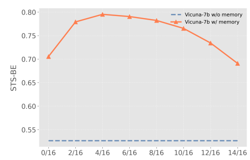
C.4 Extended Results

Different ratio of memory We explore different ratios of short-term memory $\mathcal{M}^s$ and long-term memory $\mathcal{M}^l$. We defined the total length of $\mathcal{M}$ as 16 and tried different ratios of $\mathcal{M}^s$ over a total length of $\mathcal{M}$. As shown in Fig. 4-5, we experiment with Vicuna-7B and OPT-2.7B under various ratios of memory in the training and evaluation stages (using ground-truth memory) respectively. With the memoryless baseline (0-shot) highlighted in the figures, we observe that a blend of short-term and long-term memory strikes a strong balance in event understanding performance. We also find that this trade-off is somewhat impacted by the ratio applied in the training phase. Furthermore, all memory-augmented settings consistently outperform the memoryless baseline, thus reinforcing our finding that memory indeed aids event understanding.

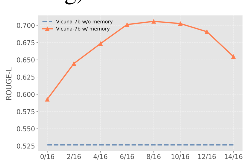
Different shots of memory We also study the performance impact of different shots. We use the same ratio $1/2$ where half of the memory is short-term memory as in Fig. 6-7. We find that an increasing number of events in memory can increase the performance. This aligns with our emphasis over the impact of memory and the in-context learning ability in LLMs (Yu et al., 2024; Brown et al., 2020).



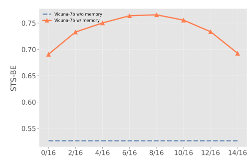
(a) ROUGE-L (25% short-term memory in training)



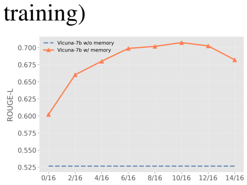
(b) STS (25% short-term memory in training)



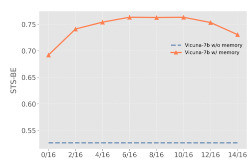
(c) ROUGE-L (50% short-term memory in training)



(d) STS (50% short-term memory in training)

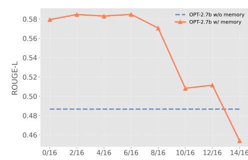


(e) ROUGE-L (75% short-term memory in training)

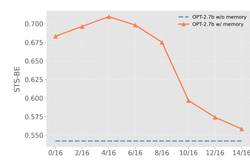


(f) STS (75% short-term memory in training)

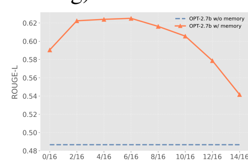
Figure 4: Evaluation of Vicuna-7b with different training ratios.



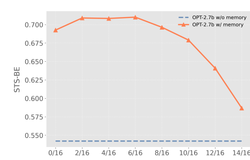
(a) ROUGE-L (25% short-term memory in training)



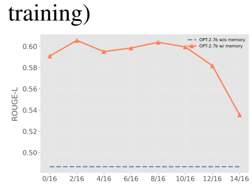
(b) STS (25% short-term memory in training)



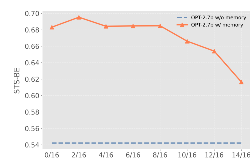
(c) ROUGE-L (50% short-term memory in training)



(d) STS (50% short-term memory in training)

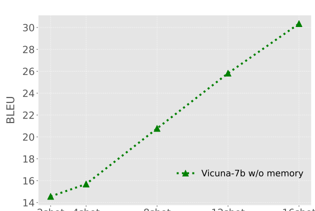


(e) ROUGE-L (75% short-term memory in training)

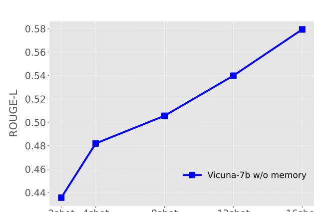


(f) STS (75% short-term memory in training)

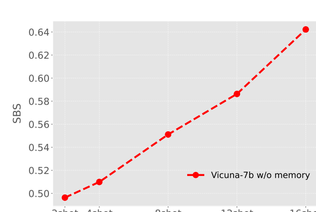
Figure 5: Evaluation of OPT-2.7B with different ratios.



(a) BLEU



(b) ROUGE-L

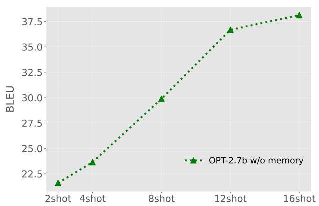


(c) STS

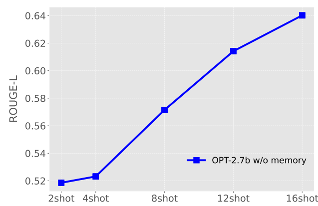
Figure 6: Results of Vicuna-7B with ground-truth (GT) memory of different shots. We employ 50% short-term memory in all the experiments.

C.5 Qualitative Studies

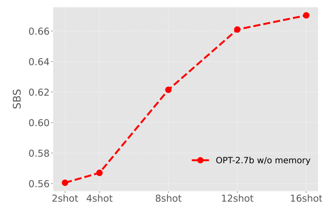
We showcase three examples as in Fig. 9. They demonstrate that leveraging short-term memory as contexts can significantly help to understand the current event.



(a) BLEU

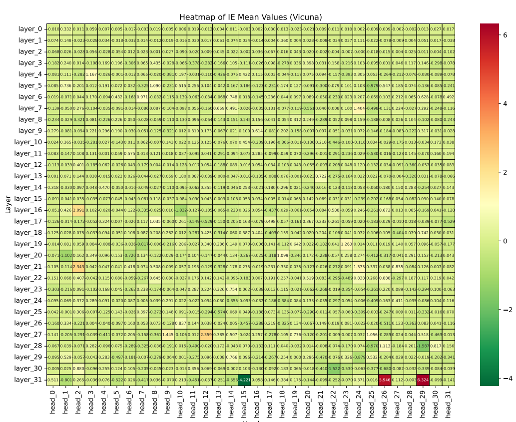


(b) ROUGE-L

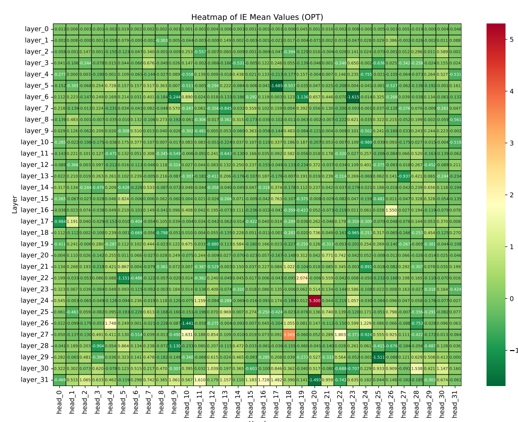


(c) STS

Figure 7: Results of Vicuna-7B with ground-truth (GT) memory of different shots.

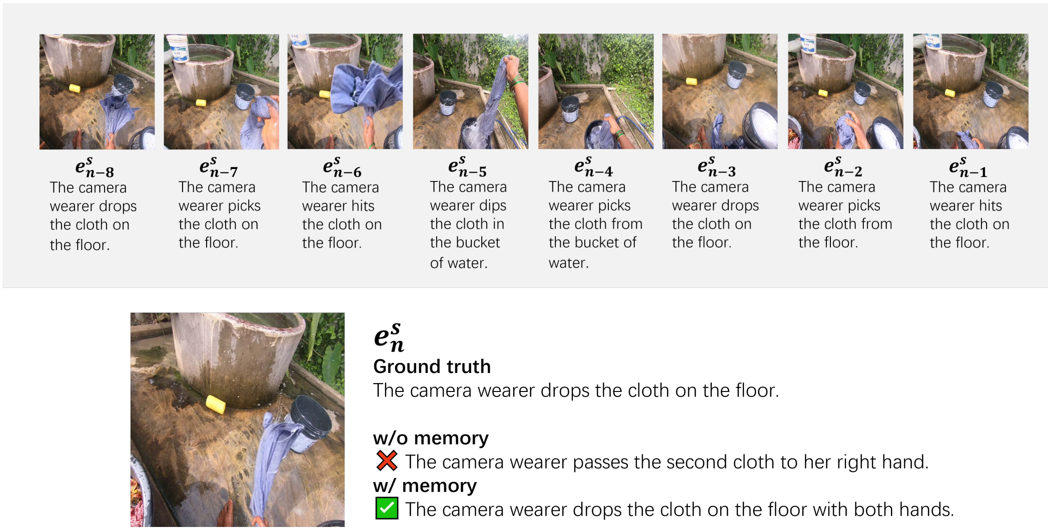


(a) Probing result of Vicuna-7B in different layers and heads

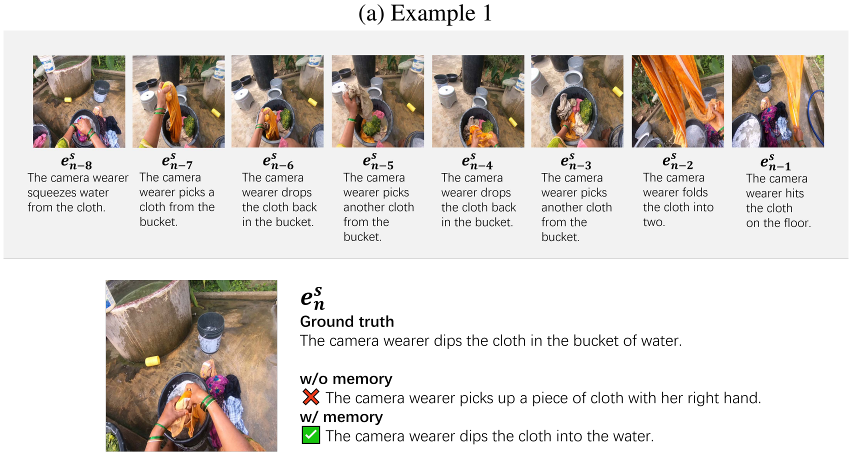


(b) Probing result of OPT-2.7B in different layers and heads

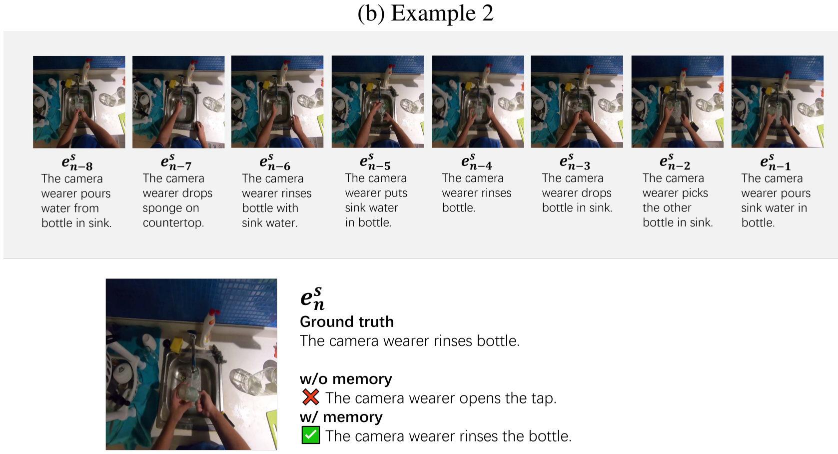
Figure 8: Comparison of probing results for Vicuna-7b and OPT-2.7b across different layers and heads.



(a) Example 1



(b) Example 2



(c) Example 3

Figure 9: Evaluation of OPT-2.7B with different ratios and metrics.