Language Repository for Long Video Understanding

Kumara Kahatapitiya   Kanchana Ranasinghe   Jongwoo Park   Michael S. Ryoo
Stony Brook University
`kkahatapitiy@cs.stonybrook.edu`

Abstract

Language has become a prominent modality in computer vision with the rise of multi-modal LLMs. Despite supporting long context-lengths, their effectiveness in handling long-term information gradually declines with input length. This becomes critical, especially in applications such as long-form video understanding. In this paper, we introduce a Language Repository (`LangRepo`) for LLMs, that maintains concise and structured information as an interpretable (*i.e.*, all-textual) representation. It consists of write and read operations that focus on pruning redundancies in text, and extracting information at various temporal scales. The proposed framework is evaluated on zero-shot video VQA benchmarks, showing state-of-the-art performance at its scale. Our code is available at `github.com/kkahatapitiya/LangRepo`.

1 Introduction

Video data is central to learning systems that can interact and reason about the world. Yet, they also associate with significant challenges such as increased compute requirements and redundant information. This is especially critical in long-form videos. Nevertheless, recent large-language-models (LLMs) [45, 57, 41, 59] have made significant strides in long-video reasoning, thanks to the scale of model/data and large context-lengths that capture long-term dependencies. However, recent studies show that the effectiveness of models declines with longer input sequences [22]. This promotes the search for techniques that effectively utilize the context of LLMs. Moreover, language as a modality has enabled benefits such as rich semantics [48, 27, 15], information sharing among specialized models [58, 25, 10] and interpretability [61, 38], to name a few. Among such, interpretability has a huge societal impact in the age of LLMs, in the context of managing adversities such as bias [24, 8] and hallucinations [60, 6]. Hence, interpretable representations have also been of interest to the community.

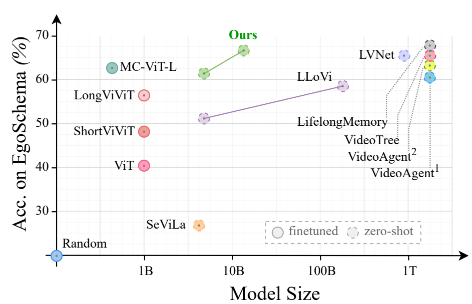
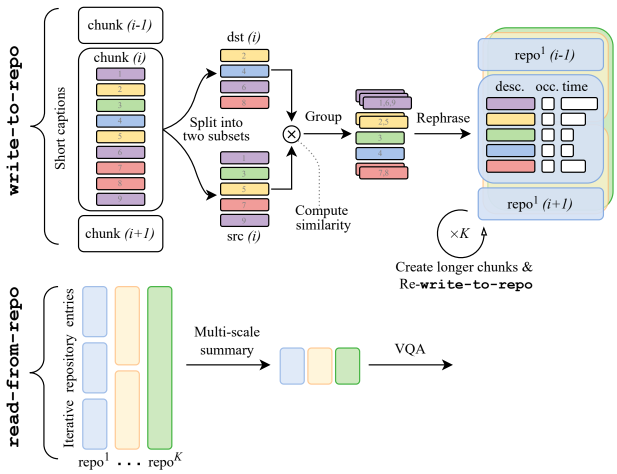


Figure 1: Comparison with prior-art on EgoSchema [26] subset: `LangRepo` outperforms finetuned and zero-shot pipelines of similar scale, while being competitive with pipelines based on much-larger proprietary models. Note that the x-axis is in log-scale.

Motivated by the above, we introduce Language Repository (`LangRepo`), an interpretable representation for LLMs that consists of *all-textual* write and read operations. The write operation (`write-to-repo`) consumes chunks of short-video captions, pruning redundant text and creating concise descriptions that keep the context-utilization of LLMs in-check. Its iterative application with increasingly-longer chunks enables it to learn high-level semantics (*e.g.* long temporal dependencies). The read operation (`read-from-repo`) extracts such stored language information at various temporal scales, together with other optional metadata within the repository entries (*e.g.* timestamps). Altogether, our proposed framework shows state-of-the-art performance on long-video reasoning benchmarks (*e.g.* EgoSchema [26], NExT-QA [52], IntentQA [23]) at its scale, while also being competitive with pipelines based on much-larger proprietary LLMs (see Fig. 1).

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).



Algorithm Long-video VQA pipeline in `LangRepo`

require captions of video chunks $C^0 = \{c_i^0 \mid i = 1, \cdots, n\}$, number of iterations K.

def `write-to-repo`(c):
  $c_{\text{dst}},\ c_{\text{src}} = \texttt{split}(c)$
  $\text{sim}_{\text{src-dst}} = \texttt{similarity}(\texttt{encode}(c_{\text{src}}),\ \texttt{encode}(c_{\text{dst}}))$
  $c_{\text{grp}} = \texttt{group}(c_{\text{dst}},\ c_{\text{src}},\ \text{sim}_{\text{src-dst}})$
  $c_{\text{reph}} = \texttt{rephrase}(\texttt{template}_{\texttt{reph}}(c_{\text{grp}}))$
  $r = (c_{\text{reph}},\ t,\ o)$ // t: timestamps, o: occurances
  return r

def `read-from-repo`(r):
  $d = \texttt{summarize}(\texttt{template}_{\texttt{sum}}(r))$
  return d

$r_i^0 = \texttt{write-to-repo}(c_i^0)$
$d_i^0 = \texttt{read-from-repo}(r_i^0)$

for k in range(K): // iterative write and read
  $C^{k+1} = \texttt{re-chunk}([\cdots,\ r_i^k,\ \cdots])$
  $r_i^{k+1} = \texttt{write-to-repo}(c_i^{k+1})$
  $d_i^{k+1} = \texttt{read-from-repo}(r_i^{k+1})$

ans $= \texttt{vqa}(q, [\ d_i^0, \cdots])$ // q: query

Figure 2: Overview of Language Repository: `LangRepo` consists of write and read operations. Given short-captions of video chunks, `write-to-repo` prunes redundant captions to generate repository entries by (1) grouping similar captions and (2) rephrasing them, while preserving additional metadata (*e.g*. #occurrences, timestamps). Next, `read-from-repo` generates concise descriptions at different semantic levels by summarizing multi-scale repository entries.

2 Related work

Video models have progressed over the years, going from primitive recognition tasks [39, 18] to complex and fine-grained reasoning tasks over long horizons [37, 56, 11]. More recently, long-video understanding has made a leap forward thanks to benchmark datasets [11, 26, 52] and model improvements [57, 59, 30], showing the importance of modeling long-term interactions. LLMs have further fueled this direction with breakthroughs in caching [9, 19, 16], model-sharding [62, 4, 21] and efficient attention [5, 20]. Even so, maintaining the effectiveness of reasoning over longer inputs is still challenging [22, 53, 36]. This motivates us to think about concise representations that can better-utilize LLM context. Such representations may come in the form of pruning [33, 1], latent memory [34, 12, 51], or external feature banks [50]. However, all these lack interpretability (*i.e*., being able to clearly identify which information gets preserved). Language being a dominant modality [32, 58, 25] that is also interpretable, can be a potential solution to address this limitation. Motivated by these, we introduce an interpretable language representation that can (1) prune redundant information, and (2) extract multi-scale (or, high-level) semantics, enabling better context-utilization in LLMs.

3 Language Repository

We present a Language Repository (`LangRepo`) that iteratively updates with multi-scale information from video chunks. It has multiple advantages including, requiring no training (*i.e*., zero-shot) and being compatible with both LLM-based processing and human interpretation (as it is fully-textual). `LangRepo` consists of two main operations: (1) information writing (`write-to-repo`), which prunes redundancies and iteratively updates language descriptions based on increasingly-longer video chunks, and (2) information reading (`read-from-repo`), which extracts preserved descriptions (together with any optional metadata) in multiple temporal scales. See Fig. 2 (left) for the overall pipeline.

Writing to repository: Our iterative write operation involves two stages: (1) Grouping and (2) Rephrasing redundant text. In the grouping stage, we identify most-similar captions. We first split the captions of each chunk into two sets, namely, source (*src*) and destination (*dst*) captions. Next, we embed all captions using a text-encoder (*e.g*. CLIP [32]), compute cosine similarities between *src-dst* sets, and group most-similar matches together. In the rephrasing stage, each group is rephrased to be a concise and coherent description via an LLM-call, dropping any redundant information. We further preserve additional metadata such as timestamps and #occurrences to avoid any loss of information due to rephrasing. Refer the supplementary for our rephrasing template and a qualitative example. In each subsequent iteration, we re-combine previous repository entries into increasingly-longer chunks with our `re-chunk`(·) operator, and re-write to repository, generating high-level information and forming a multi-scale language representation. See Fig. 2 (right) for the complete algorithm.

Table 1: (Left): Results on EgoSchema [26]. (Right) Results on NExT-QA [52] and IntentQA [23]. We focus on the zero-shot video VQA. `LangRepo` shows a strong performance at its scale. Open-source multi-modal LLMs with video-caption pretraining are de-emphasized for fair comparison.

Model	Params	Subset	Fullset
zero-shot (with proprietary LLMs)			
Vamos [43]	175B	-	41.2
LLoVi [59]	175B	57.6	50.3
ProViQ [3]	175B	-	57.1
MoReVQA [28]	340B	-	51.7
LVNet [31]	<1.8T	66.0	61.1
Vamos [43]	1.8T	-	48.3
VideoAgent[1] [44]	1.8T	60.2	54.1
VideoAgent[2] [7]	1.8T	62.8	-
IG-VLM [17]	1.8T	-	59.8
VideoTree [49]	1.8T	66.2	61.1
LifelongMemory [47]	1.8T	68.0	62.1
zero-shot (with open-source LLMs)			
InternVideo [45]	478M	-	32.1
FrozenBiLM [54]	890M	-	26.9
SeViLA [57]	4B	25.7	22.7
mPLUG-Owl [55]	7B	-	31.1
LLoVi [59]	7B	50.8	33.5
VideoLLaMA 2 [2]	12B	-	53.3
Vamos [43]	13B	-	36.7
InternVideo2 [46]	13B	-	60.2
Tarsier [42]	34B	68.6	61.7
`LangRepo` (ours)	7B	60.8	38.9
`LangRepo` (ours)	12B	66.2	41.2

Model	Params	NExT-QA	IntentQA
zero-shot (with proprietary LLMs)			
ViperGPT [40]	175B	60.0	-
ProViQ [3]	175B	64.6	-
MoReVQA [28]	340B	69.2	-
LVNet [31]	<1.8T	72.9	71.1
IG-VLM [17]	1.8T	68.6	64.2
LLoVi [59]	1.8T	67.7	64.0
TraveLER [35]	1.8T	68.2	-
VideoAgent[1] [44]	1.8T	71.3	-
VideoTree [49]	1.8T	73.5	66.9
zero-shot (with open-source LLMs)			
VFC [29]	164M	51.5	-
InternVideo [45]	478M	49.1	-
SeViLA [57]	4B	63.6	
Mistral [13]	7B	51.1	50.4
LLoVi [59]	7B	54.3	53.6
Tarsier [42]	7B	71.6	-
LLoVi [59]	12B	58.2	56.6
Tarsier [42]	34B	79.2	-
`LangRepo` (ours)	7B	54.6	53.8
`LangRepo` (ours)	12B	60.9	59.1

Reading from repository: As we make a single VQA prediction for a given long-video, our read operation (`read-from-repo`) is applied only after fully-forming the multi-scale repository (*i.e.*, after all write iterations). When reading, we generate summaries for each repo-entry separately via LLM-calls, allowing us to focus on varying temporal spans. Optionally, we can make use of additional metadata by prompting the read operator as "`[timestamps] description (`$\times$`#occurrences)`". Refer the supplementary for our summarizing template. Finally, we concatenate all output descriptions and prompt the LLM again to generate answer predictions.

4 Experiments

In our experiments, we rely on frame or short-clip captions pre-extracted using VLLMs [63, 25], and use open-source LLMs for modeling (*e.g.* Mistral [13] w/ 7B parameters or Mixtral [14] w/ 12B active parameters). We use CLIP-L/14 [32] as the text encoder in similarity-based pruning. In Table 1 (left), we present zero-shot performance of `LangRepo` on standard EgoSchema [26] splits. `LangRepo` shows significantly-better performance at its scale, among models with similar pretraining. On the fullset, we achieve $+10.1\%$ over mPLUG-Owl [55], $+7.7\%$ over LLoVi [59], and $+4.5\%$ over Vamos [43]. On the subset, ours is even competitive with much-larger models based on proprietary LLMs (*e.g.* GPT-4), showing $+6.0\%$ over VideoAgent[1] [44], $+3.4\%$ over VideoAgent[2] [7] and $+0.2\%$ over LVNet [31]. In Table 1 (right), we report zero-shot performance of `LangRepo` on standard NExT-QA [52] validation split and IntentQA [23] test split. On both benchmarks, we outperform models at its scale with similar pretraining (for fair comparison, as our captions have not seen any video pretraining). On NExT-QA, we gain $+9.8\%$ over Mistral [13] and $+2.7\%$ over LLoVi [59]. On IntentQA, `LangRepo` achieves $+8.7\%$ over Mistral [13] and $+2.5\%$ over LLoVi [59]. These results validate the effectiveness of our long-video VQA pipeline. Moreover, in Table 2, we evaluate the founding motivation of this work, showing that `LangRepo` can effectively utilize the LLM context and retain a stable performance over longer input lengths compared to other baselines.

Table 2: While other models drop performance with increasing #captions, `LangRepo` stays more-stable ($1\times$ is 180 captions in [26]).

Model	$0.5\times$	$1\times$	$2\times$
Mistral [13]	49.8	48.8	46.8
LLoVi [59]	57.2	55.4	53.6
`LangRepo` (ours)	56.4	57.8	56.4

5 Conclusion

In this paper, we introduced a Language Repository (`LangRepo`), which reads and writes textual information of long-video chunks, as a concise, multi-scale and interpretable language representation. Both our `write-to-repo` and `read-from-repo` operations are text-based and implemented as calls to a backbone LLM. Our empirical results show a strong performance on multiple VQA benchmarks at comparable settings, while also being (1) less-prone to performance drops at longer input lengths, and (2) interpretable, enabling easier human intervention if and when needed.

Acknowledgements: This work was supported by Electronics and Telecommunications Research Institute (ETRI) grants funded by the Korean government [24ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling and evolving ways]. This work was also supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. RS-2024-00336738, Development of Complex Task Planning Technologies for Autonomous Agents, 100%]. The authors thank the Robotics Lab at SBU for helpful discussions, and especially Xiang Li, for managing the lab servers and promptly responding to all the infrastructure-related issues.

References

[1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

[2] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[3] Rohan Choudhury, Koichiro Niinuma, Kris M Kitani, and László A Jeni. Zero-shot video question answering with procedural programs. *arXiv preprint arXiv:2312.00937*, 2023.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

[5] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

[6] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

[7] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *arXiv preprint arXiv:2403.11481*, 2024.

[8] Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

[9] Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*, 2023.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

[11] 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.

[12] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[13] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[14] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[15] Kumara Kahatapitiya, Anurag Arnab, Arsha Nagrani, and Michael S Ryoo. Victr: Video-conditioned text representations for activity recognition. *arXiv preprint arXiv:2304.02560*, 2023.

[16] Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. *arXiv preprint arXiv:1805.04623*, 2018.

[17] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *arXiv preprint arXiv:2403.18406*, 2024.

[18] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pp. 2556–2563. IEEE, 2011.

[19] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

[20] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. `https://github.com/facebookresearch/xformers`, 2022.

[21] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[22] Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*, 2024.

[23] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11963–11974, 2023.

[24] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pp. 6565–6576. PMLR, 2021.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[26] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[27] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

[28] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13235–13245, 2024.

[29] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15579–15591, 2023.

[30] Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. *arXiv preprint arXiv:2312.07395*, 2023.

[31] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. *arXiv preprint arXiv:2406.09396*, 2024.

[32] 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.

[33] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? *arXiv preprint arXiv:2106.11297*, 2021.

[34] Michael S Ryoo, Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanishka Rao, Austin Stone, Yao Lu, Julian Ibarz, and Anurag Arnab. Token turing machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19070–19081, 2023.

[35] Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. Traveler: A multi-lmm agent framework for video question-answering. *arXiv preprint arXiv:2404.01476*, 2024.

[36] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.

[37] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 510–526. Springer, 2016.

[38] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024.

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[40] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.

[41] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[42] Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024.

[43] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. *arXiv preprint arXiv:2311.13627*, 2023.

[44] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024.

[45] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[46] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.

[47] Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos, 2024.

[48] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497, 2022.

[49] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *arXiv preprint arXiv:2405.19209*, 2024.

[50] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 284–293, 2019.

[51] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13587–13597, 2022.

[52] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

[53] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*, 2023.

[54] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022.

[55] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[56] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126:375–389, 2018.

[57] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36, 2024.

[58] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[59] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023.

[60] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[61] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[62] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

[63] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.