CinePile: A Long Video Question Answering Dataset and Benchmark

Ruchit Rawal ♦ **Khalid Saifullah** ♦ **Ronen Basri** ♣
David Jacobs ♦ **Gowthami Somepalli***♦ **Tom Goldstein***♦

♦ University of Maryland, College Park ♣ Weizmann Institute of Science

https://hf.co/datasets/tomg-group-umd/cinepile

Abstract

Current long-form video understanding datasets often fail to provide genuine comprehension challenges, as many tasks can be solved by analyzing only a few random frames. To address this, we introduce CinePile, a novel dataset and benchmark designed specifically for long-form video understanding. This paper details our approach to creating a question-answer dataset using advanced LLMs with human-in-the-loop, based on human-generated raw data. Our dataset includes 305,000 multiple-choice questions covering visual and multimodal aspects such as temporal comprehension, human-object interactions, reasoning about events, etc. We evaluate recent video-centric LLMs on our test split, revealing that even state-of-the-art models significantly lag behind human performance, underscoring the complexity of video understanding.

1 Introduction

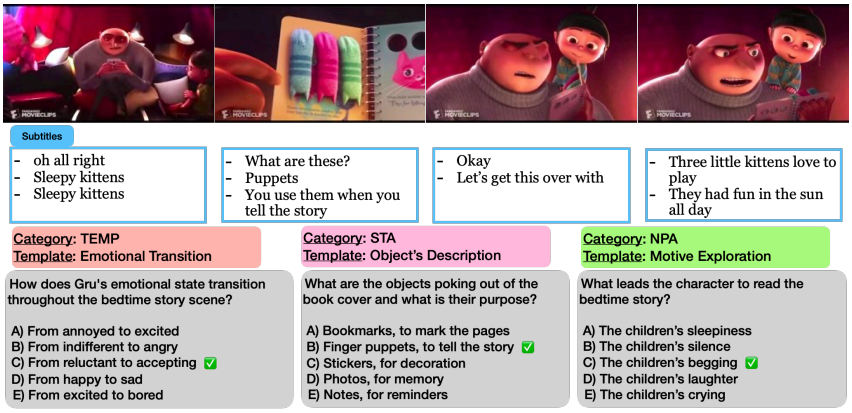

Figure 1: A sample clip (from here) and corresponding MCQs from CinePile.

Large multi-modal models offer the potential for analyzing long, complex videos. However, training and evaluating models on video data offers difficult challenges. Most videos contain dialogue and pixel data, requiring complete scene understanding. Furthermore, most existing vision-language models are pre-trained primarily on still frames, while understanding long videos requires the ability to identify interactions and plot progressions in the temporal dimension. We introduce CinePile, a large-scale dataset consisting of $\sim$ 305k question-answer pairs from 9396 videos, split into train and test

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

sets. Our dataset emphasizes question diversity, covering temporal understanding, perceptual analysis, complex reasoning, and more. It also emphasizes question difficulty, with humans exceeding the best commercial vision/omni models by approximately 25%. We present a scene and a few question-answer pairs from our dataset in fig. 1. Consider the first question, `How does Gru's emotional state transition throughout the scene?` For a model to answer this correctly, it needs to understand both the visual and temporal aspects, and even reason about the plot progression of the scene. To answer the second question, `What are the objects poking out of the book cover and what is their purpose`, the model must localize an object in time and space, and use its world knowledge to reason about their purpose.

CinePile addresses several weaknesses of existing video understanding datasets. First, the large size of CinePile enables it to serve as both an instruction-tuning dataset and an evaluation benchmark. We believe the ability to do instruction tuning for video at a large scale can bridge the gap between the open-source and commercial video understanding models. Also, the question diversity in CinePile makes it a more comprehensive measure of model performance than existing benchmarks. Unlike existing datasets such as LVU Wu and Krähenbühl [2021] and MAD Soldan et al. [2022], which focus primarily on genre classification, like ratio prediction, or scene captioning, CinePile emphasizes deeper video understanding that requires temporal understanding. CinePile's large size is made possible by our novel pipeline for automated question generation and verification using large language models. While previous datasets depend on fixed templates or manual curation (e.g., TGIF-QA Jang et al. [2017] and MoVQA Zhang et al. [2023b]), CinePile leverages detailed human descriptions of scenes to generate complex questions at scale. This automated yet scalable approach allows us to capture many more question categories than prior works. At test time, models must answer these questions from only the dialogue and raw video, without the hand-written descriptions used to build the questions.

Next, we discuss our method for dataset construction, verification, and model evaluation. Due to space constraints, we keep the discussion brief in the main draft and provide extensive additional details in the supplementary material.

2 Creating a long video reasoning benchmark

Data collection and consolidation: We obtain clips from the MovieClips YouTube channel, which hosts self-contained scenes highlighting major plot points, facilitating the creation of a dataset focused on understanding and reasoning. Next, we collected Audio Descriptions from AudioVault.

Getting visual descriptions of video for free. Audio descriptions (ADs) feature a narrator explaining crucial visual elements during pauses in dialogue, primarily for the vision impaired. Unlike conventional video caption datasets, ADs focus on contextual elements rather than being overly descriptive. We use ADs as a proxy for visual annotation in our dataset. Since our video clips are typically 2-3 minutes long and ADs cover entire movies, we use a rolling-window algorithm to align and extract the relevant parts from the ADs.

Automated Question Templates: While we use a template-based approach for question generation, rather than confining to a few predefined themes, we propose an automated method to create question templates from existing human-generated questions. We first cluster 30,000 human-generated questions across multiple existing datasets, then use GPT-4 to discern their underlying themes and generate prototypical questions for each template. In total, we generate 86 unique templates that we categorize into four high-level categories: Character and Relationship Dynamics (CRD), Narrative and Plot Analysis (NPA), Thematic Exploration (TE) and Setting, and Technical Analysis (STA).

Automated QA generation with LLMs: To generate scene-related questions, we used Gemini to select relevant templates from scene-text annotations, choosing 5-6 at random from the top 20. A commercial language model then generated questions based on the scene's audio description, selected templates, prototypical questions, and a system prompt aimed at producing deep, long-term questions. Prototypical examples helped minimize hallucinations and improved MCQ distractors' plausibility. Timestamps for dialogues and visual descriptions enhanced temporal questions, allowing us to generate about 32 questions per video on average. We use both GPT-4 and Gemini to generate questions.

Testing the quality of the dataset: While the process consistently produces well-formed and answerable questions, we found some questions to be trivial or related to basic world concepts that

don't require viewing the clip. To address this, we evaluated our dataset using a few LLMs on the following axes, and either removed such questions or computed metrics for users to utilize. **1. Degeneracy.** A question is considered degenerate if the answer is implicit in the question, e.g., `What is the color of the pink house?`. Manually reviewing all questions is impractical. Hence, we used three distinct LMs to automate this process: Gemini Anil et al. [2023], GPT-3.5 Achiam et al. [2023], and Phi-1.5 Li et al. [2023b]. Models received only questions and choices, and if all models answered correctly, the question was considered degenerate, and removed from the evaluation split. **2. Vision Reliance.** Some questions in the dataset might be answerable solely based on dialogue, without needing the video component. For this analysis, We used the Gemini model, providing it only with dialogue to assess performance. A correct answer scores 0 for visual dependence, while an incorrect one scores 1. **3. Hardness.** We developed a metric to gauge the difficulty of questions for the models, even when provided with full context. For this purpose, we selected the Gemini model, given its status as one of the larger and more capable models. This metric differs from accuracy; during evaluation, the models are only supplied with videos and dialogue information, excluding visual descriptions. However, in calculating the hardness metric, we include visual descriptions as part of the context given to the model.

3 Dataset Details & Model evaluation

Our dataset consists of 9396 video clips with average length of ≈160 seconds, split into train and test splits of 9248 and 148 videos each. Following the pipeline outlined in the methods section, we ended up with over $298,887$ training points and $4,941$ test-set points, with around 32 questions per video scene. Each MCQ contains a question, answer, and four distractors.

Given that our dataset consists of multiple-choice questions (MCQs), we evaluate a model's performance on our benchmark questions by measuring its ability to accurately select the correct answer from a set of options, which includes one correct answer and four distractors. One key challenge is reliably parsing the model's poorly formatted response to extract its chosen answer and mapping it to one of the predefined answer choices. Our two-stage evaluation addresses this by normalizing responses to extract the option letter and text, and then comparing this normalized response to the correct answer key. Scoring is based on the presence of both elements or either one if only one is present.

Table 1: **Model Evaluations.** We present the accuracy of various video LLMs on the CinePile's test split. We also present Human performance for comparison. We also ablate the accuracies across the question categories we discussed earlier.

Model	Average	CRD	NPA	STA	TEMP	TH
Human	73.21	82.92	75.00	**73.00**	75.52	64.93
Human (authors)	**86.00**	**92.00**	**87.5**	71.20	**100**	**75.00**
GPT-4o OpenAI [2024]	59.72	64.36	**74.08**	54.77	44.91	**67.89**
GPT-4 Vision Achiam et al. [2023]	58.81	63.73	73.43	52.55	46.22	65.79
Gemini 1.5 Pro Reid et al. [2024]	**61.36**	**65.17**	71.01	**59.57**	**46.75**	63.27
Gemini 1.5 Flash Reid et al. [2024]	57.52	61.91	69.15	54.86	41.34	61.22
Claude 3 (Opus) Anthropic [2024]	45.60	48.89	57.88	40.73	37.65	47.89
Video LLaVa Lin et al. [2023]	22.51	23.11	25.92	20.69	22.38	22.63
mPLUG-Owl Ye et al. [2023a]	10.57	10.65	11.04	9.18	11.89	15.05

We evaluate various commercial and open-source LLM models and we present their performance in table 1, along with general crowdsourced human results and author attempted results for comparison. Among the various commercial VLMs analyzed, Gemini 1.5 Pro (Video) emerge as top performer with 61.36% average accuracy, followed by GPT-4o and GPT-4 Vision. Gemini 1.5 Pro particularly outperforms the GPT-4 models in the "Setting and Technical Analysis" category that is dominated by visually reliant questions focusing on the environmental and surroundings of a movie scene, and its impact on the characters. On the contrary, we note that GPT-4 models performs substantially better on other question categories such as "Narrative and Plot Analysis" that revolve around the core storylines, and interaction between the key characters. Meanwhile, Claude 3 (Opus) ranks as the least effective commercial VLM model, trailing Gemini Pro Vision by approximately 4%. Gemini 1.5 Flash, a newly released lighter version of Gemini 1.5 Pro, performs quite competitively, achieving 57.52% accuracy. We also provide a "hard split" in the test set with particularly challenging questions. Most models experience a 15%-20% performance decline on this hard split, but the relative rankings

remain unchanged. Notably, Gemini 1.5 Flash suffers a 21% decline compared to 13% for Gemini 1.5 Pro, highlighting the trade-offs in optimizing for lightweight performance. A significant gap persists between human performance and state-of-the-art VLMs, with OSS models trailing both. To understand these trends, we conducted further qualitative and quantitative analyses, and found that the subpar performance of OSS models is only partially due to their inability to follow instructions.

Acknowledgements

The authors would like to thank Neel Jain, Yuxin Wen, Abhimanyu Hans, Monte Hoover, Sachin Shah, Sean McLeish, Kevin Zhang, Pedro Sandoval, John Kirchenbauer, Kaiyu Yue, Allen Tu, Steffen Jung, Arpit Bansal, Alex Hanson, Sravani Somepalli, Kamal Gupta, Alex Stein and many others who helped us with the human study.

Financial support was provided by the ONR MURI program and the AFOSR MURI program. Private support was provided by Capital One Bank, the Amazon Research Award program, and Open Philanthropy. Further support was provided by the National Science Foundation (IIS-2212182), and by the NSF TRAILS Institute (2229885). Computational resources were furnished by the Department of Energy INCITE Allocation Program.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. URL https://www.anthropic.com/claude.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.

Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*, 2023.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023b.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

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

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*, 2024.

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.

OpenAI. Gpt-4o release, 2024. URL https://openai.com/index/hello-gpt-4o/.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

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

Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. *arXiv preprint arXiv:2312.17742*, 2023.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021.

Chao-Yuan Wu and Philipp Krähenbühl. Towards Long-Form Video Understanding. In *CVPR*, 2021.

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*, pages 9777–9786, 2021.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023a.

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*, 2023b.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

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*, 2023a.

Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *arXiv preprint arXiv:2312.04817*, 2023b.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.