CINEPILE: A LONG VIDEO QUESTION ANSWERING DATASET AND BENCHMARK

Anonymous authors
Paper under double-blind review

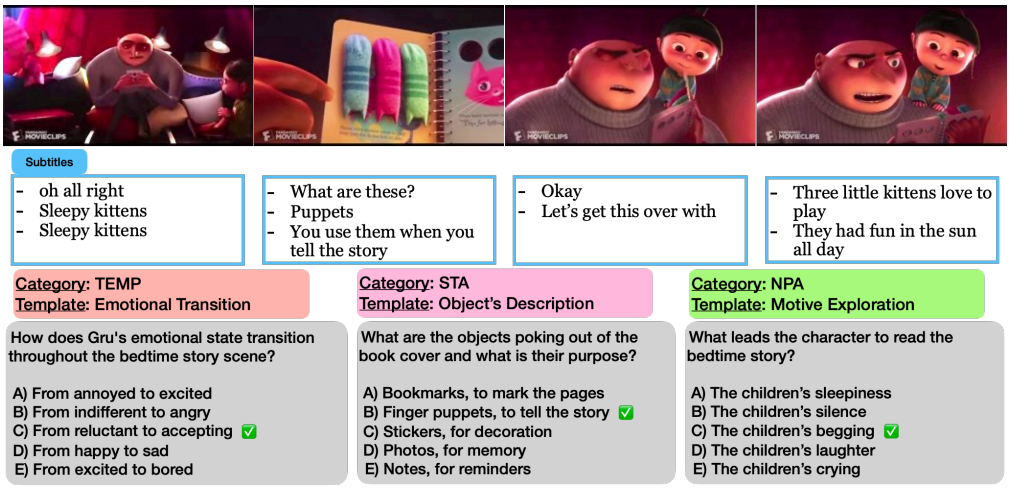


Figure 1: A sample clip (from here) and corresponding MCQs from CinePile.

ABSTRACT

Current datasets for long-form video understanding often fall short of providing genuine long-form comprehension challenges, as many tasks derived from these datasets can be successfully tackled by analyzing just one or a few random frames from a video. To address this issue, we present a novel dataset and benchmark, CinePile, specifically designed for authentic long-form video understanding. This paper details our innovative approach for creating a question-answer dataset, utilizing advanced LLMs with human-in-the-loop and building upon human-generated raw data. Our comprehensive dataset comprises 305,000 multiple-choice questions (MCQs), covering various visual and multimodal aspects, including temporal comprehension, understanding human-object interactions, and reasoning about events or actions within a scene. Additionally, we fine-tuned open-source Video-LLMs on the training split and evaluated both open-source and proprietary video-centric LLMs on the test split of our dataset. The findings indicate that although current models underperform compared to humans, fine-tuning these models can lead to significant improvements in their performance.

1 INTRODUCTION

Large multi-modal models offer the potential to analyze and understand long, complex videos. However, training and evaluating models on video data offers difficult challenges. Most videos contain dialogue and pixel data and complete scene understanding requires both. 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.

In this paper, we introduce CinePile, a large-scale dataset consisting of $\sim$ 305k question-answer pairs from 9396 videos, split into train and test sets. Our dataset emphasizes question diversity, and topics span 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%, and exceeding open source video understanding models by 37%.

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, CinePile does not over-emphasize on purely visual questions (e.g., `What color is the car?`), or on classification questions (e.g., `What genre is the video?`) that do not require temporal understanding. Rather, CinePile is comprehensive with diverse questions about vision, temporal, and narrative reasoning with a breakdown of question types to help developers identify blind spots in their models.

The large size of CinePile is made possible by our novel pipeline for automated question generation and verification using large language models. Our method leverages large existing sets of audio descriptions that have been created to assist the vision impaired. We transcribe these audio descriptions and align them with publicly available movie video clips from YouTube. Using this detailed human description of scenes, powerful LLMs are able to create complex and difficult questions about the whole video without using explicit video input. At test time, video-centric models must answer these questions from only the dialogue and raw video, and will not have access to the hand-written descriptions used to build the questions. We release the prompts for generating the question answers, the code for model evaluation, and the dataset splits in the Appendix.

2 Creating a long video understanding benchmark

Our dataset curation process has four primary components 1) Collection of raw video and related data. 2) Generation of question templates. 3) Automated construction of the Q&A dataset using video and templates, and 4) Application of a refinement pipeline to improve or discard malformed Q&A pairs.

2.1 Data collection and consolidation

We obtain clips from English-language films from the YouTube channel *MovieClips*[1]. This channel hosts self-contained clips, each encapsulating a major plot point, facilitating the creation of a dataset focused on understanding and reasoning. Next, we collected Audio Descriptions from AudioVault[2].
Getting visual descriptions of video for free. Audio descriptions (ADs) are audio tracks for movies that feature a narrator who explains the visual elements crucial to the story during pauses in dialogue. They have been created for many movies to assist the vision impaired. The key distinction between conventional video caption datasets and ADs lies in the contextual nature of the latter. In ADs, humans emphasize the important visual elements in their narrations, unlike other video caption datasets, which tend to be overly descriptive. We use the audio descriptions as a proxy for visual annotation in the videos for our dataset creation.

Scene localization in AD. The video clips we have gathered are typically 2-3 minutes long, while Audio Descriptions (ADs) cover entire movies. To align descriptions with video, we transcribe the audio from both the movie clip and the whole movie AD file using an Automatic Speech Recognition (ASR) system `WhisperX` (Bain et al., 2023), an enhanced version of `Whisper` (Radford et al., 2023) designed to offer quicker inference and more precise word-level timestamps. We then embed the first 3 and last 3 lines of the text transcription of a YouTube movie clip using a sentence embedding

[1] https://www.youtube.com/@MOVIECLIPS
[2] https://audiovault.net/movies

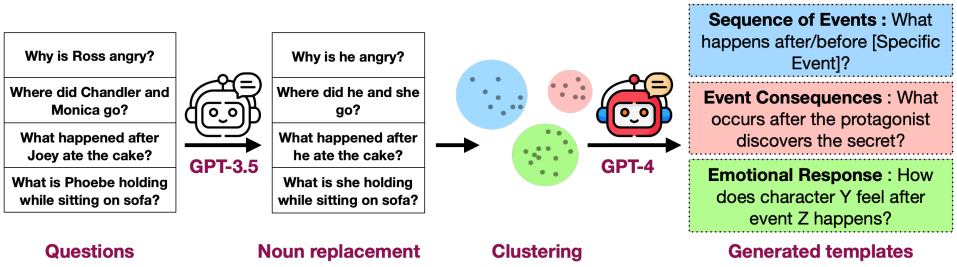


Figure 2: **Question template generation pipeline**: We begin by substituting the first names in human-written source questions and then cluster them. We then feed a selection of questions from each cluster into GPT-4, which outputs "question templates" used in the next stage of dataset creation. See Section 2.2 for more details.

model, `WhereIsAI/UAE-Large-V1`. We similarly embed all the sentences in the corresponding movie AD file. We then localize the YouTube clip within the AD file via the rolling window algorithm. We then extract all AD data that lies between the matched start and end of the movie clip embeddings. This localized text contains both the visual elements and the dialogue for the given YouTube clip. This serves as a base text for creating the QA dataset For the rest of the paper, we will refer to the human-written description of the scene as "visual description" and the speaking or dialogue part of the video as "dialogue". When combined, we will refer to both data sources as "**scene-text-annotation**".

Sentence classification. When we transcribe an AD file, the text contains a human's visual descriptions and the movie's dialogue. However, the transcription model does not label whether a given sentence belongs to a visual description or a dialogue. Since we planned to create a few questions solely on the visual components of the video, the distinction is important to us. To categorize each sentence as either visual or dialogue, we fine-tuned a BERT-Base model (Devlin et al., 2018) using annotations from the MAD dataset (Soldan et al., 2022), which contains labels indicating whether a sentence is a dialogue or a visual description. We applied a binary classification head for this task. For training the classification model, we split the MAD dataset annotations into an 80-20 training-evaluation split. The model achieves 96% accuracy on eval split after 3 epoch training. Qualitatively, we observed that the model accurately classifies sentences in the data we curated, distinguishing effectively between dialogue and visual description content.

Question Template Automation

Generate a few templates based on the following questions

`[Question 1]` Where did the couple meet?
`[Question 2]` What holiday is coming up when they meet?
`[Question 3]` What made him break with her?

Response Based on the nature of these questions, I can identify a few underlying templates that encompass most of them:

`[Template 1]` Catalytic Actions
`[Proto Question 1]` What event acts as a catalyst for the character's next major decision?

`[Template 2]` Setting and Context
`[Proto Question 2]` Where does this interaction take place, and how does the location impact the conversation?
...

Figure 3: **Extracting templates from human-generated questions.** We share 10 questions from each cluster, and prompt an LLM to create a few templates and a prototypical question. See Fig. 2 and Section 2.2 for details.

2.2 Automated Question Templates

Many prominent video question-answering benchmarks were written by human annotators. The question-answer pairs are typically curated in one of two ways: 1) Human annotators are given complete freedom to ask questions about a given scene (Tapaswi et al., 2016) 2) They are asked to focus on specific aspects and are trained or provided with examples of questions, encouraging them to write more questions in a similar style (Xiao et al., 2021; Li et al., 2020; Lei et al., 2018; Patraucean

[2]Icons in the figures are sourced from Flaticon.

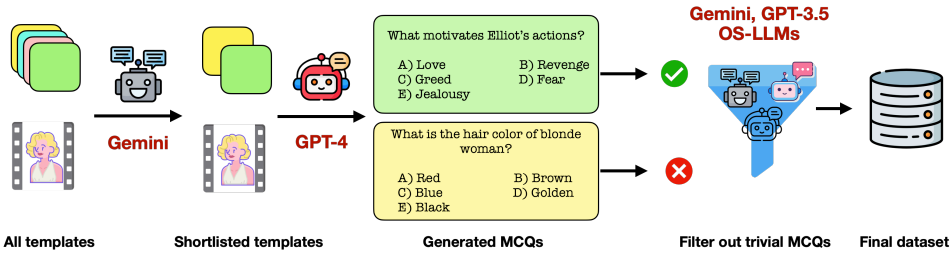


Figure 4: **Automated QA Generation and Filtering.** Begins with a set of automated templates and scenes. Filter out the templates relevant to each scene. Next, pass these templates along with the annotated-scene-text to GPT-4, which is then used to create multiple-choice questions (MCQs). The generated MCQs are then subjected to numerous filters to curate the final dataset. For more detailed information, refer to Section 2.3 and Section 2.4

et al., 2024). For instance, in the Perception Test Benchmark (Patraucean et al., 2024), annotators are directed to concentrate on temporal or spatial aspects, while for the Next-QA dataset (Xiao et al., 2021), annotators mainly focused on temporal and causal action reasoning questions.

During early experiments, we found that giving a range of templates and scene-text-annotation to an LLM helped create more detailed, diverse, and well-formed questions. Thus, we adopted a template-based approach for question generation. Instead of limiting questions to a few hand-curated themes, we propose a pipeline to create templates from human-generated questions (shown in Fig. 2).

Our starting point is approximately 30,000 human-curated questions from the MovieQA (Tapaswi et al., 2016), TVQA (Lei et al., 2018), and Perception Test (Patraucean et al., 2024) datasets. We cluster these questions, select a few representatives per cluster, and then use GPT-4 to discern the underlying themes and write a prompt. First, we preprocess the questions by replacing first names and entities with pronouns, as BERT (Reimers & Gurevych, 2019) embeddings over-index on proper nouns, hence the resultant clusters end up with shared names rather than themes. For instance, 'Why is Rachel hiding in the bedroom?' is altered to 'Why is she hiding in the bedroom?'. We used GPT-3.5 to do this replacement, as it handled noun replacement better than many open-source and commercial alternatives. The modified questions are then embedded using `WhereIsAI/UAE-Large-V1`, a semantic textual similarity model which is a top performer on the MTEB leaderboard[3]. When the first names were replaced, we observed significant repetition among questions, which prompted us to duplicate them, ultimately resulting in 17,575 unique questions. We then perform k-means clustering to categorize the questions into distinct clusters. We experimented with different values of $k = 10, 50, 100$. Qualitatively, we found $k = 50$ to be an optimal number of clusters where the clusters are diverse and at the same time clusters are not too specific. For example, we see a 'high-school dance' cluster when $k = 100$, and these questions are merged into an 'event' cluster when we reduce k to 50. The Perception Test questions are less diverse as human annotators were restricted to creating questions based on a small number of themes, so we used $k = 20$ for this set. The number of questions in each cluster ranges from 60 to 450. We selected 10 random questions from each, and used them to prompt GPT-4 to create relevant question templates, as illustrated in Fig. 3. We did ablations by selecting the closest 10 questions to the cluster center, however qualitatively observed that random questions produced more general/higher quality templates.

We generate four templates for each question cluster, resulting in around 300 templates across three datasets. We then manually reviewed all 300 templates, eliminating those that were overly specific and merging similar ones. Overly specific templates and their proto-questions looked like "**Pre-wedding Dilemmas:** `What complicates character Z's plans to propose marriage to their partner?`" and "**Crime and Consequence:** `What is the consequence of the character's criminal actions?`". The authors also added a many templates that were complimentary to the auto-generated ones. This process resulted in 86 unique templates. Following that, we manually binned these into five high-level categories: Character and Relationship Dynamics, Narrative and Plot Analysis, Thematic Exploration, Temporal, and Setting and Technical Analysis. For a detailed discussion on the category definitions, examples of templates, and prototypical questions from each category, please refer to the Appendix C & D.

[3] https://huggingface.co/spaces/mteb/leaderboard

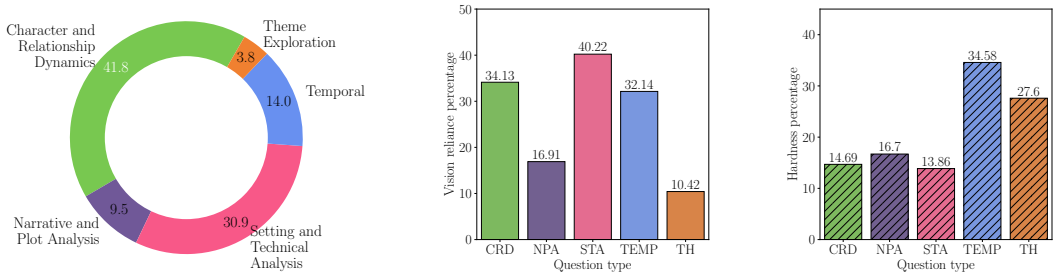


Figure 5: Test split statistics. **Left:** Question category composition in the dataset. **Middle:** Percentage of vision-reliant questions across categories. **Right:** Percentage of hard questions per question category type. TEMP - Temporal, CRD - Character and Relationship Dynamics, NPA - Narrative and Plot Analysis, STA - Setting and Technical Analysis, TH - Thematic Exploration. The colors correspond to the same categories across the plots. Refer to the Appendix for corresponding plots of train split.

2.3 Automated QA generation with LLMs

The pipeline for generating questions is shown in Fig. 4. While the question templates are general, they might not be relevant to all the movie clips. Hence for a given scene, we choose a few relevant question templates by providing Gemini with the scene-text-annotation of the scene, and asking to shortlist the 20 most relevant templates to that scene, out of which we randomly select 5-6 templates. We then provide a commercial language model with (i) the scene-text-annotation, which includes both visual descriptions and dialogue, (ii) the selected question template names (e.g. 'Physical Possession'), (iii) the prototypical questions for the templates (e.g. "What is [Character Name] holding"), and (iv) a system prompt asking it to write questions about the scene. Through rigorous experimentation, we devised a system prompt that makes the model attentive to the entire scene and is capable of generating deeper, longer-term questions as opposed to mere surface-level perceptual queries. We observed that providing the prototypical example prevents GPT-4 from hallucination, and also leads to more plausible multiple-choice question (MCQ) distractors. We also found that asking the model to provide rationale for its answer enhances the quality of the questions. Additionally, we found that including timestamps for the scene-text-annotation augments the quality of generated temporal questions. Through this method, we were able to generate ≈ 32 questions per video.

After experimenting with this pipeline, we analyzed the generated QA pairs and noticed a consistent trend: most questions are focused on reasoning or understanding. For diversity, we also wanted to include purely perceptual questions. To achieve this, we introduced additional hand-crafted prompt templates for perceptual questions and also templates for temporal questions. While GPT-4 performs well across all question templates, we found that Gemini excels particularly with perceptual templates. Therefore, we utilized Gemini to generate a segment of perceptual questions in the dataset, while using GPT-4 for reasoning templates. Our experiments with open-source models indicated subpar question quality, despite extensive prompt tuning. We present example questions and a quantitative investigation into the quality of the generations produced by GPT-4 and Gemini in Appendix E. Moreover, we provide the prompt we use question-answer generation in Appendix L.

2.4 Dataset quality evaluation and adversarial refinement

While the process above consistently produces well-formed and answerable questions, we observed that some questions are either trivial, with answers embedded within the question itself, or pertaining to basic world concepts that do not require viewing the clip. To identify these, we evaluated our dataset with the help of a few LLMs on the following axes and we improved the quality of those whenever possible. In the few instances where this was not possible, we removed the questions from the dataset or computed a metric that the users can use in the downstream tasks.

Degeneracy and educated guessing. A question is considered degenerate if the answer is implicit in the question itself, e.g., `What is the color of the pink house?`. Similarly, an educated guessing is the most probable answer to the question based on general knowledge, context, or common sense, e.g. `What is the bartender using the shaker for? a)` **`prepare a cocktail`** `b) do groceries c) collect tips .` Based on an investigation of a subset of the dataset, we found that such questions constituted only a small fraction.

However, since manually reviewing all the questions was impractical, we employed three distinct language models (LMs) to identify weak Q&As: Gemini (Anil et al., 2023), GPT-3.5 (Achiam et al., 2023), and Phi-1.5 (Li et al., 2023c). In order to do this, we presented only the questions and answer choices to the models, omitting any context, and calculated the accuracy for each question across multiple models. If multiple models with different pre-training or post-training setups all correctly answer a question, it is likely that the answer was implicit, rather than due to biases of any one.

Adversarial Refinement. After identifying weak Q&A pairs, wed an *adversarial refinement* process to repair these Q&A pairs. The goal was to modify the questions and/or answer choices so that a language model could no longer answer them correctly using only implicit clues within the question and answer choices themselves. To achieve this, we used a large language model (LLM), referred to as "deaf-blind LLM", to identify and explain why a question could be answered without extra context. Specifically, when the LLM answered a question correctly, we asked it to provide a rationale for its choice. This rationale helped us detect hidden hints or biases in the question. We then fed this rationale into our question-generation model, instructing it to modify the question and/or answer choices to eliminate these implicit clues. This process continued in a loop until the LLM could no longer answer the question correctly (after adjusting for chance performance), with a maximum of five attempts per question. Given the repetitive and computationally intensive nature of this process, we required a powerful yet accessible LLM that could run locally, avoiding issues with API limits, delays, and costs associated with cloud-based services. As a result, we selected LLaMA 3.1 70B (Dubey et al., 2024), an open-source model that met these desiderata. Through this adversarial refinement process, we successfully corrected approximately 90.94% of the weak Q&A pairs in the training set and 90.24% of the weak Q&A pairs in the test set. Finally, we excluded the unfixable Q&A pairs from the evaluation split ($\sim$ 80 Q&A) of our dataset but retained them in the training set ($\sim$ 4500 Q&A). We share more details about adversarial refinement in Appendix Sec. N

Vision Reliance. When generating the multiple-choice questions (MCQs), we considered the entire scene without differentiating between visual text and dialogue. Consequently, some questions in the dataset might be answerable solely based on dialogue, without the necessity of the video component. For this analysis, we utilized the Gemini model. The model was provided with only the dialogue, excluding any visual descriptions, to assess its performance. If the model correctly answers a question, it is assigned a score of 0 for the visual dependence metric; if it fails, the score is set at 1. In later sections, we present the distribution of the visual dependence scores across different MCQ categories.

Hardness. Hardness refers to the inability to answer questions, even when provided with full context used to create the questions in the first place (i.e., the subtitles & visual descriptions). For this purpose, we selected the Gemini model, given its status as one of the larger and more capable models. Unlike accuracy evaluation, which uses only video frames and dialogues (subtitles), the hardness metric includes visual descriptions as part of the context given to the model. After this, the authors reviewed all the questions flagged as "hard" for verification and fixed any minor issues, if present.

In addition, the authors went through the question in the evaluation split across multiple iterations, and fixed any systemic errors that arose in the pipeline. Furthermore, we conducted a human study to identify potential weaknesses, and we discuss our findings in Appendix I.

3 A LOOK AT THE DATASET

In the initial phase of our dataset collection, we collected $\sim$15,000 movie clips from channels like MovieClips on YouTube. We filtered out clips that did not have corresponding recordings from Audiovault, as our question generation methodology relies on the integration of visual and auditory cues—interleaved dialogues and descriptive audio—to construct meaningful questions. We also excluded clips with low alignment scores when comparing the YouTube clip's transcription with the localized scene's transcription in the Audio Description (AD) file as discussed in Section 2.1. This process resulted in a refined dataset of 9396 movie clips. The **average video length in our dataset is $\sim$160 sec**, significantly longer than many other VideoQA datasets and benchmarks. We split 9396 videos into train and test splits of 9248 and 148 videos each. We made sure both the splits and the sampling preserved the dataset's diversity in terms of movie genres and release years. We follow the question-answer generation and filtering pipeline which was thoroughly outlined in Section 2. We ended up with **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. As a post hoc step, we randomized the position of the correct answer among the distractors for every question, thus

Table 1: We compare our dataset, CinePile against the pre-existing video-QA datasets. Our dataset is both large and diverse. Multimodal refers to whether both the video and audio data is used for question creation and answering. For understanding different QA types, refer to Section 2.3

Dataset	Annotation	Domain	Num QA	Avg sec	Multimodal	QA Type			
						Temporal	Attribute	Narrative	Theme
TGIF-QA (Jang et al., 2017)	Auto	Tumblr GIFs	165,165	3	✗	✓	✗	✗	✗
MSRVTT-QA (Xu et al., 2017)	Auto	Multiple	243,690	15	✗	✗	✓	✗	✗
How2QA (Li et al., 2020)	Human	Instructional Videos	44,007	60	✗	✓	✓	✗	✗
NExT-QA (Xiao et al., 2021)	Human	Daily Life Videos	52,044	44	✗	✓	✓	✗	✗
EgoSchema (Mangalam et al., 2024)	Auto	Egocentric	5,000	180	✗	✓	✓	✓	✗
MovieQA (Tapaswi et al., 2016)	Human	Movies	6,462	203	✓	✓	✓	✓	✗
TVQA (Lei et al., 2018)	Human	TV Shows	152,545	76	✓	✓	✓	✓	✗
Perception Test (Patraucean et al., 2024)	Human	Scripted Videos	44,000	23	✓	✓	✓	✗	✗
MoVQA (Zhang et al., 2023b)	Human	Movies	21,953	992	✓	✓	✓	✓	✗
IntentQA (Li et al., 2023b)	Human	Daily Life Videos	16,297	Unknown	✓	✓	✗	✗	✗
Video-MME (Fu et al., 2024)	Human	Multiple	2,700	1017.9	✓	✓	✓	✓	✗
MVBench (Li et al., 2024)	Auto	Multiple	4,000	16	✓	✓	✓	✗	✗
Video-Bench (Ning et al., 2023)	Human + Auto	Multiple	17,036	56	✓	✓	✓	✗	✗
LVBench (Wang et al., 2024)	Human	Multiple	1,549	4,101	✓	✓	✓	✓	✗
CinePile (Ours)	Human + Auto	Movies	303,828	160	✓	✓	✓	✓	✓

eliminating any positional bias. We filtered out the degenerate questions from the test split, however, we left them in the train set, since those questions are harmless and might even teach smaller models some helpful biases the larger multimodal models like Gemini might inherently possess.

Our dataset's diversity stems from the wide variety of movie clips and different prompting strategies for generating diverse question types. Each strategy zeroes in on particular aspects of the movie content. We present a scene and example MCQs from different question templates in Fig. 1, and many more in the Appendix. In Fig. 5 (Left), we provide a visual breakdown of the various question categories in our dataset. A significant portion of the questions falls under "Character Relationship Dynamics". This is attributed to the fact that a large number of our automated question templates, which were derived from human-written questions belonged to this category. This is followed by "Setting and Technical Analysis" questions, which predominantly require visual interpretation. We display the metrics for vision reliance and question hardness, as discussed in Section 2.4, at the category level in Fig. 5 (Middle, Right). As anticipated, questions in the "Setting and Technical Analysis" category exhibit the highest dependency on visual elements, followed by those in "Character Relationship Dynamics", and "Temporal" categories. In terms of the hardness metric, the "Temporal" category contains the most challenging questions, with "Thematic Exploration" following closely behind. Finally, we compare our dataset with other existing datasets in this field in Table 1, showing its superiority in both the number of questions and average video length compared to its counterparts.

4 Model Evaluation

In this section, we discuss the evaluations of various closed and open-source video LLMs on our dataset, some challenges, and model performance trends. Given that our dataset consists of multiple-choice question answers (MCQs), we assess a model's performance by its ability to accurately select the correct answer from a set of options containing one correct answer and four distractors. A key challenge in this process is reliably parsing the model's response to extract its chosen answer and map it to one of the predefined choices. Model responses may vary in format, including additional markers or a combination of the option letter and corresponding text. Such variations necessitate a robust post-processing step to accurately extract and match the model's response to the correct option. To address these variations, we employ a two-stage evaluation method. First, a normalization function parses the model's response, extracting the option letter (A-E) and any accompanying text. This handles various formats, ensuring accurate identification. The second stage involves comparing the normalized response with the answer key, checking for both the option letter and text. If both match, a score of one is awarded; However, if only the option letter or text appears, the comparison is limited to the relevant part, and the score is assigned accordingly.

We evaluate 24 commercial and open-source LLM models and we present their performance in Table 2. We discuss additional details about the evaluation timelines, model checkpoints, and compute budget in Appendix G. We also present human numbers (author and non-author) for comparison. This distinction is important because the authors carefully watched the video (go back and rewatch the video if necessary) while answering the questions. This removes the carelessness errors from the human study. While commercial VLMs perform reasonably well, the very best of OSS models lag $\sim$10% behind the proprietary models. We present a few QA's which humans got wrong and GPT-4 got wrong and the plausible reason for errors in Appendix I.

Gemini 1.5 Pro leads overall; LLaVA-OV tops open-source models. Among the various commercial VLMs analyzed Gemini 1.5 Pro performs the best, and 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 offer competitive performance on question categories such as "Narrative and Plot Analysis" that revolve around the core storylines, and interaction between the key characters. It's important to note that Gemini 1.5 Pro is designed to handle long multimodal contexts natively, while GPT-4o and GPT-4V don't yet accept video as input via their APIs. Therefore, we sample 10 frames per video while evaluating them. Gemini 1.5 Flash, a newly released lighter version of Gemini 1.5 Pro, also performs competitively, achieving 58.75% overall accuracy and ranking second in performance. Its competitive edge over the GPT models is owing to the "Setting and Technical Analysis" category, where it performs significantly better. In open-source models, LLaVA-OV (One Vision) ranks as the best, achieving an overall accuracy of 49.34%. More broadly, while the accuracy of open-source models ranges from 49.34% to 13.93%, it's clear that recent models like LLaVA-OV (released August 2024), MiniCPM-V-2.6 (released August 2024), and VideoLLaMa2 (released June 2024) offer competitive performance compared to proprietary models.

Table 2: **Model Evaluations.** We present the accuracy of various video LLMs on the CinePile's test split. We also present Human performance for comparison. We ablate the accuracies across the question categories: TEMP - Temporal, CRD - Character and Relationship Dynamics, NPA - Narrative and Plot Analysis, STA - Setting and Technical Analysis, TH - Thematic Exploration.

Model	Params.	Avg	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**
Gemini 1.5 Pro–001	-	60.12	63.90	70.44	57.85	46.74	59.87
Gemini 1.5 Flash–001	-	58.75	62.82	69.76	55.99	44.04	62.67
GPT-4o	-	56.06	60.93	69.33	49.48	45.78	61.05
GPT-4 Vision	-	55.35	60.20	68.47	48.63	45.78	59.47
LLaVA-OV	7B	49.34	52.13	59.83	46.54	37.65	58.42
LLaVA-OV Chat	7B	49.28	52.47	58.32	46.28	37.79	58.42
MiniCPM-V 2.6	8B	46.91	50.10	54.21	44.52	35.61	54.74
Claude 3 Opus	-	45.60	48.89	57.88	40.73	37.65	47.89
VideoLLaMA2	7B	44.57	47.44	54.64	41.91	34.30	47.37
InternVL2	26B	43.86	47.10	56.16	39.03	34.16	52.63
LongVA DPO	7B	42.78	45.84	54.21	39.16	33.43	44.74
InternVL-V1.5	25.5B	41.69	45.07	51.19	38.97	30.09	45.79
LongVA	7B	41.04	43.28	51.84	38.45	33.58	38.42
InternVL2	4B	39.89	42.99	47.73	36.23	32.99	41.58
mPLUG-Owl3	8B	38.27	40.91	45.71	33.86	33.09	46.20
LLaVA-OV	0.5B	33.82	35.88	39.96	31.66	27.03	38.42
InternVL2	8B	32.28	35.25	40.39	28.46	24.71	38.42
InternVL2	2B	30.34	31.91	33.26	30.35	23.26	31.58
VideoChat2	7B	29.27	31.04	34.56	25.26	27.91	34.21
Video LLaVa	7B	25.72	26.64	32.61	23.63	23.26	24.74
CogVLM2	19B	17.16	18.33	17.06	17.23	13.08	18.95
InternVL2	1B	15.97	17.65	19.22	13.25	12.94	22.63
Video-ChatGPT	7B	15.08	17.06	16.34	15.17	7.26	18.58
mPLUG-Owl	7.2B	13.93	16.15	13.16	13.03	10.48	11.54

Performance significantly drops on the "hard-split". Additionally, as discussed in Section 2.4, we provide a "hard split" in the test set consisting of particularly challenging questions. In Fig. 6, we compare the performance of the top 6 models (in terms of average accuracy) on both the average and the hard splits of our dataset. We note that while most models suffer a performance decline of 15%-20% on the hard split; however, the relative ranking among the models remains unchanged. Interestingly, Gemini 1.5 Flash suffers a decline of $\approx 21\%$ compared to 13% for Gemini 1.5 Pro, underscoring the particularly severe trade-offs involved in optimizing the models for lightweight performance on more challenging samples.

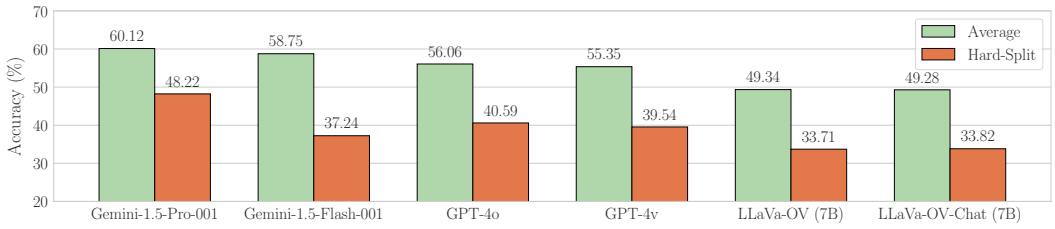
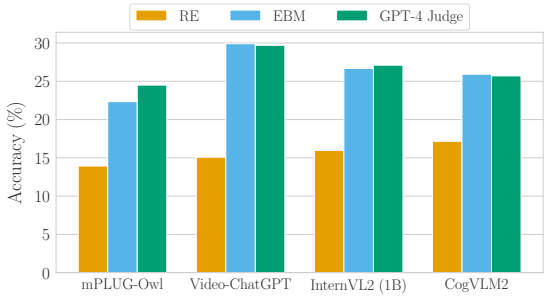
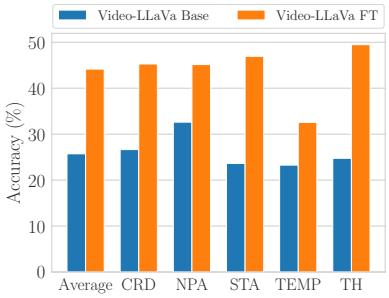


Figure 6: Models' performance on CinePile test split, all questions vs hard questions.



(a) Different strategies for evaluating performance on CinePile include: RE (Response Extraction), EBM (Embedding-Based Matching), and GPT-4 Judge (using GPT-4 to assess the raw response).



(b) Comparing the performance of Video-LLaVa after fine-tuning on CinePile's training set. 'Average' refers to the aggregate performance, while the remaining labels represent specific question types.

Figure 7

Why are (some) OSS models so far behind? We conducted further analyses to understand the poor performance of some open-source models, focusing on qualitative evaluations of their raw responses (Appendix H). Our findings indicate that a primary issue is their inability to follow instructions, often generating irrelevant or repetitive content, which hinders accurate extraction of the intended answer. To quantify these deviations, we introduced two alternative strategies for computing accuracy: a) Embedding Similarity Matching: We compute the similarity between the model's raw response and the various answer options within the embedding space of a sentence transformer (Zhang et al., 2019). The most similar option is selected as the predicted answer. b) GPT-4 as a judge: We use GPT-4 (Zheng et al., 2023) as an evaluator to extract the predicted answer key from the model's raw response. The results from these strategies are illustrated in Figure 7a. We observe that although these alternative evaluation strategies yield an improvement in the models' performance, their accuracy still falls significantly short compared to the best-performing open-source models. This suggests that the underperformance cannot be solely attributed to an inability to follow instructions. Rather, these models also exhibit fundamental limitations in video understanding capabilities. Notably, the two alternative evaluation strategies—embedding similarity matching and the use of GPT-4 as a judge—are highly consistent with each other, as well as largely aligning with the rankings obtained from the original response extraction strategy. We provide further details and additional results based on traditional video-caption evaluation metrics, such as BertScore (Zhang et al., 2019), CIDEr (Vedantam et al., 2015), and ROUGE-L (Lin, 2004), in Appendix H.

CinePile's train-split helps improve performance In this section, we investigate the impact of CinePile 's training split in enhancing the performance of open-source video LLMs. We selected Video-LLaVa as the baseline and fine-tuned it using CinePile 's training data. For efficient training, we load the model using 4-bit quantization. During fine-tuning, we freeze the base model, and conduct training using Low-Rank Adaptation (LoRA) (Hu et al., 2021). We fine-tuned the model for 5 epochs using the AdamW optimizer (Loshchilov & Hutter, 2017). We compare the performance of the fine-tuned Video-LLaVa against the base model, as shown in 7b. Our results indicate that fine-tuning led to an approximate 71% improvement in performance (increasing accuracy from 25.72% to 44.16%), with gains observed consistently across all question subcategories. These results demonstrate the significant utility of CinePile's training split in enhancing model performance.

Additional Ablations. We report additional results on the effect of removing video frames on model performance in Appendix K.1, performance on hard-split (for all models) in Appendix K.2.

5 Related Work

LVU (Wu & Krähenbühl, 2021), despite being one of the early datasets proposed for long video understanding, barely addresses the problem of video understanding as the main tasks addressed in this dataset are year, genre classification or predicting the like ratio for the video. A single frame might suffice to answer the questions and these tasks cannot be considered quite as "understanding" tasks. MovieQA (Tapaswi et al., 2016) is one of the first attempts to create a truly understanding QA dataset, where the questions are based on entire plot the movie but not localized to a single scene. On closer examination, very few questions are vision focused and most of them can be answered just based on dialogue. EgoSchema (Mangalam et al., 2024) is one of the recent benchmarks, focused on video understanding which requires processing long enough segments in the video to be able to answer the questions. However, the videos are based on egocentric videos and hence the questions mostly require perceptual knowledge, rather than multimodal reasoning. Another recent benchmark, Perception Test (Patraucean et al., 2024), focuses on core perception skills, such as memory and abstraction, across various reasoning abilities (e.g., descriptive, predictive, etc) for short-form videos that they collected by first preparing explicit video scripts. The MAD dataset introduced in (Soldan et al., 2022) and expanded in (Han et al., 2023) contains dialogue and visual descriptions for full-length movies and is typically used in scene captioning tasks rather than understanding. Another issue is this dataset does not provide raw visual data, they share only `[CLS]` token embeddings, which makes it hard to use. TVQA (Lei et al., 2018) is QA dataset based on short 1-min clips from famous TV shows. The annotators are instructed to ask What/How/Why sort of questions combining two or more events in the video. MoVQA (Zhang et al., 2023b) manually curates questions across levels multiple levels—single scene, multiple scenes, full movie— by guiding annotators to develop queries in predefined categories like Information Processing, Temporal Perception, etc. CMD (Bain et al., 2020) proposes a text-to-video retrieval benchmark while VCR (Zellers et al., 2019) introduces a commonsense reasoning benchmark on images taken from movies. Long video understanding datasets, such as EpicKitchens (Damen et al., 2018), tend to concentrate heavily on tasks related to the memory of visual representations, rather than on reasoning skills. More recently, multiple benchmarks focusing on long video understanding have been released, such as Video-MME (Fu et al., 2024), MVBench (Li et al., 2024), and LVBench (Wang et al., 2024), all having videos from multiple domains such as movies, sports, etc. Most of these datasets require significant human effort to generate questions, with costs increasing as you move toward longer video regimes. Hence, most of them range on a scale of a few thousand question-answer pairs (while CinePile ranges 70-75 $\times$ more). We discuss works utilizing synthetic data for dataset creation in Appendix B.

CinePile differs from all the above datasets, having longer videos and many questions to capture the perceptual, temporal, and reasoning aspects of a video. And it is truly multimodal where the person has to watch the video as well as dialogues to answer many questions. Unlike the previous datasets with fixed templates, we automated this process on previously human-generated questions, this let us capture many more question categories compared to previous works. Lastly, our approach to dataset generation is scalable, allowing us to fine-tune video models to improve performance. Moreover, CinePile can easily be extended in the future with additional videos, question categories, and more.

6 Discussion and Conclusion

In this paper, we introduced CinePile, a unique long video understanding dataset and benchmark, featuring $\sim$ 300k questions in the training set and $\sim$ 5000 in the test split. We detailed a novel method for curating and filtering this dataset, which is both scalable and cost-effective. Additionally, we benchmarked various recent commercial video-centric LLMs and conducted a human study to gauge the achievable performance on this dataset. To our knowledge, CinePile is the only large-scale dataset that focuses on multi-modal understanding, as opposed to the purely visual reasoning addressed in previous datasets. Our fine-tuning experiments demonstrate the quality of our training split. Additionally, we plan to set up a leaderboard for the test set, providing a platform for new video LLMs to assess and benchmark their performance on CinePile.

Despite its strengths, there are still a few areas for improvement in our dataset, such as the incorporation of character grounding in time. While we believe our dataset's quality is comparable to or even better than that of a Mechanical Turk annotator, we acknowledge that a motivated human, given sufficient time, can create more challenging questions than those currently generated by an LLM. Our goal is to narrow this gap in future iterations of CinePile.

Reproducibility Statement

To ensure the reproducibility of our work, we have taken several steps to provide all necessary details and materials. Our key contributions include: (a) a robust synthetic data generation pipeline for constructing a video question-answering dataset, (b) the final training and test splits derived from this pipeline, and (c) the fine-tuning and evaluation of video language models (LLMs) on these splits. To facilitate replication, we have included the exact prompt used for question-answer generation, the constructed train and test splits, and the fine-tuning and evaluation code in the supplementary materials and appendix. Specifically, the prompt can be found in Appendix L, while the train and test splits are available as Hugging Face objects (`dataset/cinepile/train` and `dataset/cinepile/test`) in the provided zip folder. The fine-tuning and evaluation code is also included in the zip folder under the `code/` directory. We believe these materials, along with the detailed explanations in the appendix and supplementary files, offer a comprehensive source for reproducing our dataset and experiments.

Ethics Statement

In accordance with the ICLR Code of Ethics, we acknowledge the potential for biases inherent in large language models, particularly regarding gender, race, and other demographic factors. Given our use of such models to generate question-answer pairs, there is a risk that these biases may be reflected in the generated content, potentially impacting downstream models trained on this data. While we manually reviewed and filtered problematic questions in the evaluation set, the scale of the training set made it infeasible to apply the same level of scrutiny. Additionally, as most of our movie clips originate from the "global west," there is a possibility that certain stereotypes may be perpetuated. Regarding our human study, we obtained an exemption from our Institute's Review Board (IRB) for the involvement of graduate students. For the dataset release, similar to many existing works (Lei et al., 2018; Tapaswi et al., 2016; Wang et al., 2024; Fu et al., 2024), we plan to release the dataset under the CC-BY-NC-4.0 license, limiting its use to non-commercial, academic purposes. We will host the dataset on Hugging Face, requiring users to agree to the license terms before access. Additionally, We do not distribute any raw video content directly; rather, we provide URLs redirecting to YouTube, ensuring compliance with YouTube's Terms of Service (YouTube, 2024).

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.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18930–18940, 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.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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*, pp. 2758–2766, 2017.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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.

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, 2023b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

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*, 2023c.

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

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2023.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

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.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023.

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*, pp. 28492–28518. PMLR, 2023.

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*, pp. 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*, pp. 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.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

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*, pp. 4566–4575, 2015.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.

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*, pp. 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*, pp. 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*, pp. 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.

YouTube. Terms of service, 2024. URL https://www.youtube.com/static?template=terms.

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

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

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.

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. *arXiv preprint arXiv:2304.10592*, 2023.