Lost in Time: A New Temporal Benchmark for VideoLLMs

Daniel Cores*†1
daniel.cores@usc.es
Michael Dorkenwald*2
m.l.dorkenwald@uva.nl
Manuel Mucientes1
manuel.mucientes@usc.es
Cess G. M. Snoek2
cgmsnoek@uva.nl
Yuki M. Asano3
yuki.asano@utn.de

1 CiTIUS
University of Santiago de Compostela
2 QUVA Lab
University of Amsterdam
3 Fundamental AI Lab
University of Technology Nuremberg

Abstract

Large language models have demonstrated impressive performance when integrated with vision models even enabling video understanding. However, evaluating video models presents its own unique challenges, for which several benchmarks have been proposed. In this paper, we show that the currently most used video-language benchmarks can be solved without requiring much temporal reasoning. We identified three main issues in existing datasets: (i) static information from single frames is often sufficient to solve the tasks (ii) the text of the questions and candidate answers is overly informative, allowing models to answer correctly without relying on any visual input (iii) world knowledge alone can answer many of the questions, making the benchmarks a test of knowledge replication rather than video reasoning. In addition, we found that open-ended question-answering benchmarks for video understanding suffer from similar issues while the automatic evaluation process with LLMs is unreliable, making it an unsuitable alternative. As a solution, we propose TVBench, a novel open-source video multiple-choice question-answering benchmark, and demonstrate through extensive evaluations that it requires a high level of temporal understanding. Surprisingly, we find that many recent video-language models perform similarly to random performance on TVBench, with only a few models such as Aria, Qwen2-VL, and Tarsier surpassing this baseline.

1 Introduction

Vision language models [1, 22, 28] have gained popularity, benefiting from both the progress made in natural language processing [4, 10, 33] and the surge of foundation models for vision [6, 31, 52] tasks with strong generalization capabilities. Recently, video-language models have been introduced [20, 45, 48], aiming to replicate the success achieved in the image domain. To evaluate their performance, visual question answering has emerged as a

* Equal contribution. † Research conducted while at VISLab, Unversity of Amsterdam.

key task requiring both textual and visual reasoning. With the rapid model development and release cycles, having a reliable and robust benchmark is crucial in measuring progress and guiding research efforts.

There are two main approaches to designing question-answering benchmarks for videos: multiple-choice question answering (MCQA) [19, 26, 29, 39, 42] and open-ended question answering (OEQA) [41, 44, 46, 51]. Given the critical role these benchmarks play in evaluating video understanding, their reliability is paramount. This raises an important question: To what extent do they truly capture and assess video understanding?

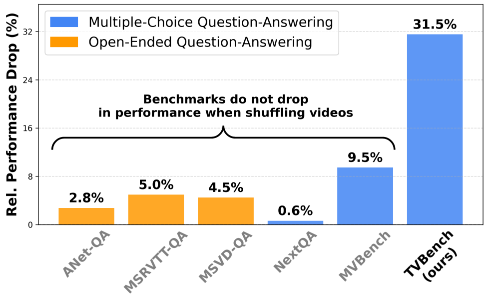

Figure 1: **Existing VideoLLM benchmarks are time-invariant.** The performance of a SOTA model [35] on commonly used benchmarks hardly drops when shuffling the input videos. This suggests that these benchmarks do not effectively measure temporal understanding. In contrast, in our proposed TVBench, shuffling input frames results in random accuracy, as it should be.

Previous analysis in image question answering benchmarks [14] has demonstrated that poorly formulated benchmarks could bias the development of new models towards learning strong text representations while ignoring visual information. This is especially relevant for the video-language community, where benchmarks must account not only for visual but also for temporal understanding.

In this work, we conduct a comprehensive analysis of widely used video question-answering benchmarks, revealing that temporal information is poorly evaluated (see Fig. 1). Furthermore, in MCQA tasks, prior world knowledge, combined with overly informative questions and answer choices, often allows questions to be answered solely through text without the need for visual input. Our results also indicate that automatic open-ended evaluation is unreliable, with significant evaluation discrepancies in results for different models.

We reveal the shortcomings of existing benchmarks such as MVBench [19], NextQA [42], MSVD-QA [41], MSRVTT-QA [46] and ActivityNet QA [51] and based on those insights propose a new benchmark, TVBench, that *requires* temporal understanding to be solved, providing an effective evaluation tool for current video-language models: i) We provide only temporal challenging candidate answers, requiring models to leverage temporal information to answer correctly. ii) We design task-specific templates to generate questions that are not overly informative such that they cannot be answered solely by text. iii) We design questions that can only be answered from the video content, without relying on prior world knowledge.

As a result, TVBench measures the temporal understanding of video-language models in contrast to previous benchmarks. In this setting, text-only and single-frame models, such as Gemini 1.5 Pro and GPT-4o, perform at random chance levels on TVBench despite achieving competitive results on other benchmarks. Surprisingly, even recent state-of-the-art video-language models perform close to random chance on TVBench, with only a few models, such as Qwen2-VL [36] and Tarsier [35], outperforming the random baseline. Shuffling the videos for these models lead to significant performance drops, unlike prior benchmarks, further verifying TVBench as a temporal video benchmark. Moreover, TVBench has already been used in several recent SOTA methods such as Seed1.5-VL [15] or PerceptionLM [8].

2 Related Work

Traditional video evaluation benchmarks focused on specific tasks such as action recognition [13, 17] or video description [9, 38, 46]. With the emergence of Vision Language Models (VLMs), there is a growing need for more comprehensive evaluation protocols to effectively evaluate models with increasingly advanced generalization capabilities. There are two major trends in the QA format: open-ended QA and multiple-choice QA (MCQA).

Open-ended question answering. Evaluating open-ended QA introduces new challenges, as traditional evaluation metrics such as ROUGE [21], METEOR [3], and CIDEr [34] fail to analyze discrepancies of more complex and elaborated answers. Alternatively, Maaz et al. [24] introduces a novel quantitative evaluation pipeline for open-ended QA datasets. The proposed method relies on GPT-3.5 to determine the correctness of the predicted answer and provides a matching score with the ground truth. Commonly used datasets for evaluating models in this context include MSRVTT-QA [44], MSVD-QA [44], TGIF-QA [16] and ActivityNet-QA [51]. In general, any open-ended QA benchmarks can be evaluated following this protocol. Our analysis shows that Large Language Model (LLM) based evaluations are prone to hallucinations, leading to unreliable conclusions. In contrast, MCQA benefits from a more straightforward evaluation process based on the accuracy score.

Multiple-choice question answering. CLEVRER [50] assesses reasoning about object interaction in synthetic videos. Perception Test [29] was introduced to evaluate visual perception in multimodal settings, mainly in indoor scenes. EgoSchema [26] focuses on long egocentric videos. NextQA [42] aims to evaluate temporal explanation of actions. VideoHallucer [39] was introduced as a first attempt to define a video-language benchmark specifically designed for hallucination detection. Lately, several approaches [2, 6, 32, 37] emphasize long video understanding. MVBench [19] defines 20 dynamic tasks designed to require temporal reasoning throughout the entire video. However, our experiments demonstrate that many of these tasks are highly spatial and textual biased, failing to evaluate temporal understanding effectively. We propose a new benchmark that requires a high level of spatiotemporal understanding across different tasks to be solved.

3 Problems in Video MCQA Benchmarks

In this section, we identify two key shortcomings in current video multiple-choice question-answering (MCQA) benchmarks, as demonstrated on MVBench [19] and NextQA [42]. First, we show that these benchmarks contain strong spatial bias, meaning that questions can be answered without requiring temporal understanding. Secondly, we find also a strong textual bias, as many questions can be answered without even looking at the visual input.

3.1 Does Time Matter?

Video benchmarks must define tasks that cannot be solved using solely spatial information to evaluate the temporal understanding of a model effectively. Questions should not be answerable using spatial details from single random or multiple frames, e.g., after shuffling them. However, if no understanding of the sequence of events and temporal localization is needed, the benchmark fails to assess temporal understanding, focusing only on spatial information, which we define as spatial bias. To analyze this bias, state-of-the-art image

	Input	Action Count	Unexp. Action	Action Antonym	Episodic Reasoning	Avg.
Random	–	33.3	25.0	33.3	20.0	27.9
Llama3 70B	text-only	44.5	63.5	74.5	50.5	58.2
Gemini 1.5	text-only	49.0	68.0	85.5	49.0	62.9
GPT-4o	text-only	44.0	69.5	57.5	51.5	55.6
Tarsier-34B	text-only	37.0	39.5	66.0	44.0	46.6
Gemini 1.5	video	41.2	82.4	64.5	66.8	63.7
GPT-4o	video	43.5	75.5	72.5	63.0	63.6
Tarsier-34B	video	46.5	72.0	97.0	54.5	67.5

Table 2: **Textual bias of MVBench.** Text-only LLMs perform nearly as well as video models, indicating vision is not essential. Average is over the four tasks.

	Input	FG Action	Scene Transition	FG Pose	Avg.
Random	–	25.0	25.0	25.0	25.0
Gemini 1.5	image	47.0	78.0	46.5	57.2
GPT-4o	image	49.0	84.0	53.0	62.0
Tarsier-34B	image	48.5	67.0	22.5	46.0
Gemini 1.5	shuffle	49.5	90.0	54.5	64.7
GPT-4o	shuffle	52.0	84.5	69.0	68.5
Tarsier-34B	shuffle	51.0	89.0	56.5	65.5
Gemini 1.5	video	50.0	93.3	58.5	67.3
GPT-4o	video	51.0	83.5	65.5	66.7
Tarsier-34B	video	48.5	89.5	64.5	67.5

Table 3: **Spatial bias of MVBench.** Near random for image and shuffled videos.

	Input	NextQA
Random	–	20.0
Tarsier-34B	text-only	47.6
Tarsier-34B	image	71.3
Tarsier-34B	video shuffle	78.5
Tarsier-34B	video	79.0

Table 1: **Spatial and textual bias on NextQA.**

and video-language models like GPT-4o [28], Gemini 1.5 Pro [12], and Tarsier-34B [35] are tested on MVBench (Table 3) and the NextQA (Table 1) dataset by comparing their performance using single frames, shuffled videos, and original videos.

The models receiving only a random frame as input show strong performance across all four tasks in Table 3, surpassing the random baseline. GPT-4o achieves the highest average performance of 62.8% across the four tasks, nearly matching its video performance with 65.8% and other state-of-the-art video-language models. The lower image performance of Tarsier-34B might stem from its training data composition, which contains five times more video data than image data. These findings are unexpected, as task names like *Fine-grained Action* suggest a need for temporal understanding. For this fine-grained task, the image-model GPT-4o achieves 49%, which is even slightly better than the state-of-the-art Tarsier model, which scores 48.5%. Similarly, for the other three tasks. Overall, GPT-4o achieves an average accuracy across all 20 tasks of 47.8%, which is 20.5% higher than the random performance of 27.3% on MVBench. Also, on NextQA in Table 1, the Tarsier model significantly outperforms the random baseline of 20.0% with 71.3%, processing a single random frame. In Fig. 2 we show examples of such. In the Appendix, in Fig. 15 -22 we show 34 more examples of spatial bias in MVBench.

Additionally, shuffling the videos has minimal impact on the MVBench performance of all video-language models, with an average difference of 2.3%, indicating that temporal information is not necessary to solve these tasks. Similarly, for NextQA, the Tarsier model achieves the same performance for shuffling or non-shuffling. Note, as confirmed in Sec. 5.2, the Tarsier model shows a significant drop in performance when videos are shuffled for tasks that require temporal understanding. This problem goes beyond these tasks as shown in Table 5, Gemini 1.5 Pro and Tarsier achieve an average accuracy across all 20 MVBench tasks of 60.5% and 67.6%, respectively. Shuffling video frames causes a performance drop of only 3.8% and 6.4%, respectively, indicating that the spatial bias affects not only the tasks analyzed in Table 3 but the entire dataset. The agreement between the correct responses of Tarsier-34B across modalities is 91.0% between image and video inputs, and 93.9% between video and shuffled video. This confirms that current models heavily rely on spatial biases to solve MVBench.

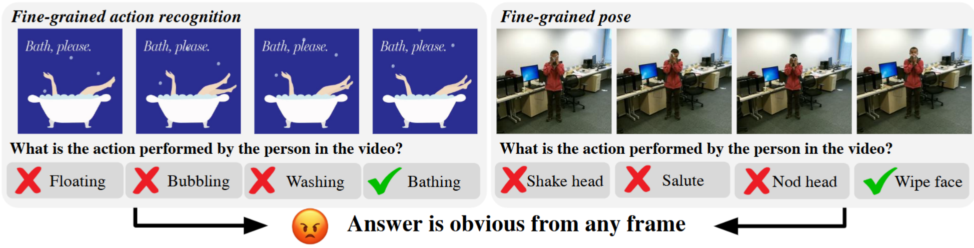

Figure 2: **Spatial bias of MVBench.** We show different tasks of the MVBench benchmark and observe that the question can be answered without requiring temporal understanding.

Problem 1

MVBench and NextQA have a strong spatial bias, meaning questions can be answered without requiring temporal understanding.

3.2 Does Vision Matter?

Video benchmarks must be designed to prevent questions from being answered solely through common sense reasoning. Modern LLMs possess strong reasoning skills, which can exploit the information within the question and candidate sets in MCQA video language evaluation benchmarks. This creates textual bias, enabling models to answer questions without leveraging the video content.

We analyze the impact of textual bias on MVBench in Table 2. We evaluate the performance of state-of-the-art text-only LLMs, Llama 3 [27], and multi-modal LLMs such as Gemini 1.5 Pro [12], GPT-4o [28] and Tarsier [35]. Our findings reveal that LLMs can eliminate incompatible candidates easily, greatly outperforming the random baseline. Models using only text achieve competitive results compared to video-language models across these four tasks. For instance, Gemini 1.5 Pro achieves an average performance of 62.3% using text-only, compared to Tarsier-34B's 67.4% using videos. Additionally, we verify an 85.3% agreement between Tarsier-34B's correct text and video responses, confirming its strong reliance on textual biases in MVBench.

This goes beyond the four tasks, as Gemini Pro 1.5 achieves an average performance across all 20 tasks of 38.2% with text-only, which is 10.9% higher than the random chance baseline of 27.3%. Similarly, for NextQA in Table 1, Tarsier achieves 47.6% performance across the whole dataset, an increase of 27.6% over the random baseline. We have identified three key sources of this textual bias on MVBench:

Bias from LLM-based QA generation. Collecting and manually annotating large datasets for training and evaluation is very costly. Automatic and semi-automatic collection and annotation processes are commonly used [19, 26]. This includes techniques such as automatic QA pair generation with LLMs. ChatGPT plays a fundamental role in QA generation for 11 of the 20 tasks in the MVBench dataset. However, this introduces unrealistic candidates and QA pairs with excessive information. Fig. 3 presents examples of QA pairs that can be resolved merely with text information. Questions 1 belong to the *Action Antonym* task, where an LLM is prompted to generate the antonym of the actual action shown in the video. The answers generated are either unrealistic, as one cannot "remove something into something," or consistently incorrect, such as "not sure".

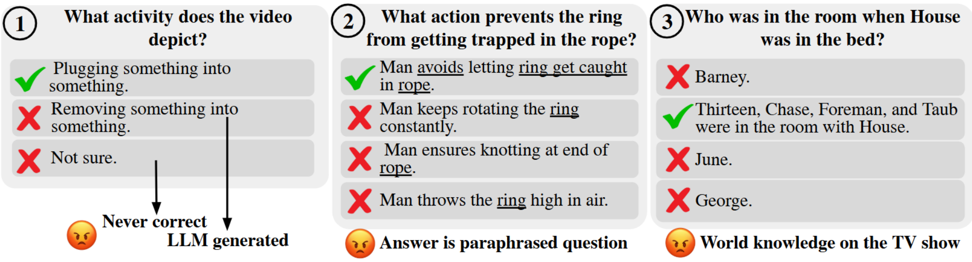

Figure 3: **Textual bias of MVBench.** We show the shortcomings of the QA generated by MVBench and find that questions can be answered without considering the visual part.

Bias from unbalanced sets. Unbalanced QA sets also hinder a robust evaluation process. For instance, the correct answer for the Action Count task on MVBench is '3' for 90 out of 200 questions, while '9' is only the correct answer for one question. A model with a similar bias might get higher results than random by chance. We have observed in our experiments that some text-only models such as GPT-4o have this bias, predicting '3' for 88 out of 200 samples. This makes GPT-4o perform on par with the best video model with an accuracy of 44.0% and 46.5% respectively.

Overreliance on world knowledge in questions. Video benchmarks should ensure models cannot rely solely on memorized world knowledge from an LLM to guess answers without using visual input. Even with well-designed questions, models might bypass visual reasoning and rely on prior knowledge to answer correctly. An example of this can be seen in question 3 of Fig. 3. The question does not exhibit an obvious bias in the QA generation. Still, it can be correctly answered if the model has world knowledge of the TV show from which the question was derived as answer 2 are character names from the House TV show.

In the Appendix, in Fig. 23 -29, we show 26 more examples of textual bias in MVBench.

Problem 2

MVBench and NextQA can be partially solved without visual information due to the bias from LLM QA generation, unbalanced dataset, and world knowledge.

4 Open-ended QA to the rescue?

Contrary to multiple-choice question answering (MCQA), open-ended question answering can be seen as an alternative to solving the aforementioned issues. Without a predefined candidate answer set, the model cannot rely on textual information to eliminate implausible candidates. However, open-ended evaluation presents new challenges compared to MCQA. Following Maaz et al. [24], LLMs have been widely used for the evaluation of open-ended question-answering in video datasets such as MSVD-QA [41], MSRVTT-QA [46] and ActivityNet QA [51]. Specifically, Maaz et al. [24] proposed GPT-3.5 as the evaluator model, which makes the entire evaluation process rely on a private API model. The evaluation model determines if the predicted answer is correct given the question and the ground-truth answer. In addition, the evaluator also computes a score to measure the answer quality.

We conducted a comparative analysis to assess the influence of the evaluation model on the results. Table 4 shows the accuracy and average score for different models on two open-ended datasets, using two evaluators: GPT-3.5 and Llama3-70B. The evaluators produced

Model	Input	Dataset	GPT-3.5 Acc.	GPT-3.5 Score	Llama3-70B Acc.	Llama3-70B Score	ΔAcc.	Dataset	GPT-3.5 Acc.	GPT-3.5 Score	Llama3-70B Acc.	Llama3-70B Score	ΔAcc.
Lama3 70B	text	MSRVTT	23.9	2.4	47.8	2.6	+23.9	ActivityNet	25.3	2.6	32.6	1.9	+7.3
GPT-4o	text	MSRVTT	23.3	2.3	42.4	2.3	+19.1	ActivityNet	27.1	2.5	33.9	1.8	+6.8
GPT-4o	image	MSRVTT	34.2	2.7	50.8	2.7	+16.6	ActivityNet	46.4	3.2	56.2	2.9	+9.8
Tarsier-34B	shuffle	MSRVTT	63.1	3.5	62.7	3.4	-0.4	ActivityNet	59.9	3.6	60.8	3.4	+0.9
Tarsier-34B	video	MSRVTT	66.4	3.7	63.0	3.4	-3.4	ActivityNet	61.6	3.7	61.3	3.4	-0.3

Table 4: **Unreliability and biases of open-ended video-language benchmark evaluation.** Different LLMs used for evaluation produce varying results, see Δ column. Additionally, open-ended benchmarks also exhibit spatial and textual bias, similar to MCQA.

significantly different results for the same method on the same dataset, with discrepancies of more than 20 points. Specifically, Llama3 highly increases the accuracy of text-only and single-image models, while providing similar or even lower results than GPT-3.5 for video models. Llama3 assigns better metrics to predictions made by the same model. If both the prediction model and the evaluator contain similar biases, the hallucinations in the predictions may be classified as correct responses by the evaluator. This includes cases where the model gives completely unrelated answers to the video —for example, responding to the question "Which plants can be seen in the desert?" with a generic list of desert plants— yet the evaluation model incorrectly assigns a high score classifying the response as correct. Additional qualitative examples are provided in Appendix A.3. These findings raise doubt about the reliability of these evaluations, as different models give completely different results.

Moreover, as shown in Table 4 open-ended QA does not solve the main issues of MCQA. The performance of text-only models is surprisingly strong; LLMs can guess the answer solely from the question text for a significant number of questions, even without a candidate list. This includes questions such as *Which hand of the person in black wears a watch?* or *What color is the pants of a person wearing black clothes?*, which correct answers are *Left hand* and *Black*. The first question can be answered just with prior knowledge as people commonly wear the watch on the left hand, while in the second one, the question contains the answer. Similar to the findings for MCQA on spatial bias in Sec. 3.1, when using a single random frame for image-text models such as GPT-4o, performance reaches 60.6% and 46.4%, approaching the video-language model's 80.3% and 61.6%, respectively. In addition, the performance of Tarsier-34B does not significantly drop—on average less than 3%—when the input videos are shuffled, indicating the low temporal understanding required for solving the benchmarks. This shows that open-ended benchmarks also exhibit strong spatial bias, not requiring temporal understanding to be solved.

In summary, current open-ended benchmarks are unreliable due to their use of LLMs as evaluators. This makes them unsuited for evaluating video-language models, especially as they also suffer from spatial and textual bias. In addition, they rely on closed-source LLMs for evaluation, which incurs costs to access, and becomes unreproducible when newer versions are released.

5 TVBench: A Temporal VQA Benchmark

We propose TVBench, a new benchmark for evaluating temporal understanding in video QA. We adopt a multiple-choice QA approach to prevent the problems of open-ended VQA described in Sec. 4. The main design principles of TVBench are derived from and address the problems listed in Sec. 3. Appendix A.2 provides an overview of the tasks, questions, and answers candidates used in our benchmark. We verify our choice of tasks and QA templates in Sec. 5.2 by the performance of multi-modal LLMs with a random frame or shuffled videos.

5.1 Designing TVBench

This section explains the key strategies implemented in TVBench to address the issues identified in Sec. 3 of current video MCQA evaluation benchmarks.

Strategy 1: Define Temporally Hard Answer Candidates. To address Problem 1, it is crucial that the temporal constraints in the question are essential for determining the correct answer. This involves designing time-sensitive questions and selecting temporal challenging answer candidates.

1. We select 10 temporally challenging tasks that require: repetition counting (Action Count), properties of moving objects (Object Shuffle, Object Count, Moving Direction), temporal localization (Action Localization, Unexpected Action), sequential ordering (Action Sequence, Scene Transition, Egocentric Sequence), and distinguishing between similar actions (Action Antonyms).

2. We define hard-answer candidates based on the original annotations to ensure realism and relevance, rather than relying on LLM-generated candidates that are often random and easily disregarded, as seen in MVBench. For example, in the Scene Transition task (Fig. 4), we design a QA template that provides candidates based on the two scenes occurring in the videos for this task, rather than implausible options like "From work to the gym." Similarly, for the Action Sequence task, we include only two answer candidates corresponding to the actions that occurred in the video. More details for the remaining tasks are in Appendix A.2.

Strategy 2: Define QA pairs that are not overly informative. Contrary to LLM-based generation, we apply templates to mitigate the effect of text-biased QA pairs (Problem 2).

1. We design QA pairs that are concise and not unnecessarily informative by applying task-specific templates. These templates ensure that the QA pairs lack sufficient information to determine the correct answer purely from text. An example of Unexpected Action is illustrated in Fig. 4. QA pairs require the same level of understanding for the model to identify what is amusing in the video, but without providing additional textual information. Unlike MVBench, the model cannot simply select the only plausible option containing a dog. We use the same candidate sets across tasks like Action Count, Object Count, Object Shuffle, Action Localization, Unexpected Action, and Moving Direction, to ensure balanced datasets with an equal distribution of correct answers, keeping visual complexity while reducing textual bias. Appendix Table A.2.2 provides an overview of all tasks, demonstrating that the QA templates are carefully crafted without unnecessary textual information.

2. Solving the overreliance on world knowledge requires providing questions and candidates that contain only the necessary information, specifically removing factual information that the LLM can exploit. We remove tasks such as Episodic Reasoning, that are based on QA pairs about TV shows or movies.

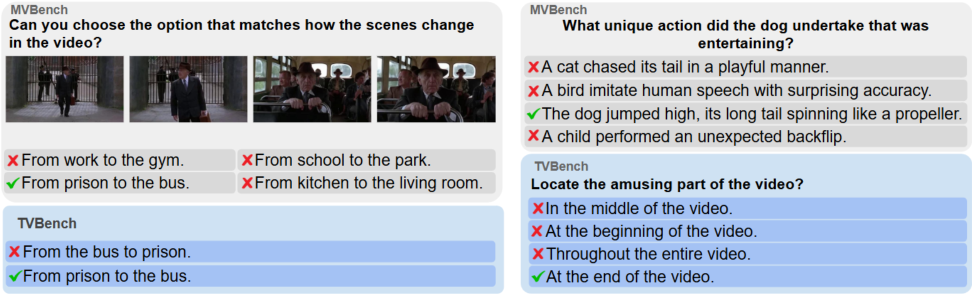

Figure 4: **TVBench Strategies.** Strategy 1 (left) mitigates spatial bias by defining temporally challenging answer candidates. Strategy 2 (right) reduces textual bias by minimally informative QA templates.

5.2 TVBench Evaluation

Table 5 provides a detailed performance breakdown of state-of-the-art text [27] and multi-modal LLMs [7, 12, 18, 19, 25, 28, 35, 36, 47, 49, 54, 55] across the 10 TVBench tasks. In addition, we also report a human baseline to verify the quality of the benchmark, more details see appendix A.2.3. We also include the average performance of these models on MVBench and TVBench, with the upward arrow $\uparrow$ indicating the improvement over random chance, and a human baseline to verify our benchmark's solveability.

Does time matter? For TVBench, multi-modal LLMs with a single image perform at random chance, verifying that a random frame is not sufficient for accurate question answering. Specifically, Gemini 1.5 Pro, the top image-language model on TVBench, outperforms random chance by only 3.0%, compared to a 21.2% improvement on MVBench. Shuffling videos has minimal impact on the performance of video-language models on MVBench, but significantly degrades their accuracy on TVBench, where it drops to near-random levels. For example, Tarsier-34B's accuracy is 33.9% higher than the random baseline on MVBench when videos are shuffled, while on TVBench, it is only 4.7% higher under the same conditions. This suggests that temporal understanding is crucial for TVBench, where visual data alone is insufficient to outperform random chance, unlike MVBench.

Does vision matter? For TVBench, state-of-the-art LLMs with text-only perform at random levels, highlighting the effectiveness of our Strategy 2 for Problem 2. Notably, Llama 3 achieves the best performance, just 1.4% above random chance on TVBench, whereas it performs 10.8% better on MVBench. This indicates that LLMs cannot determine the answer solely by analyzing the question and answer candidates or by relying on prior world knowledge. Thus, visual information becomes key for solving TVBench.

6 Discussion

A sobering view on current models. With our new TVBench, we can accurately assess the temporal understanding of existing video-language models. Surprisingly, we find that recent state-of-the-art and highly popular models, such as VideoChat2, ST-LLM, PLLava, VideGPT+, GPT-4o, mPLUG-Owl3 perform close to random chance on our temporal benchmark. Only five models, Qwen2-VL, LLaVA-Video, IXC-2.5, Aria, and Tarsier, achieve above 50% accuracy, significantly outperforming the random baseline. From these results,

Model	**Input**	**MVBench** Average	**TVBench** Average	**TVBench** AC	OC	AS	OS	ST	AL	AA	UA	ES	MD
Random	–	27.3	33.3	25.0	25.0	50.0	33.3	50.0	25.0	50.0	25.0	25.0	25.0
GPT-3.5 Turbo	text-only	$35.0_{\uparrow 7.7}$	$33.1_{\downarrow 0.2}$	27.2	18.2	44.9	32.0	53.5	26.9	45.9	29.2	26.5	26.7
Llama 3 70B	text-only	$38.1_{\uparrow 10.8}$	$34.7_{\uparrow 1.4}$	30.2	27.0	48.7	32.9	55.1	26.9	49.1	23.3	25.5	28.0
GPT-4o	text-only	$34.8_{\uparrow 7.5}$	$33.8_{\uparrow 0.5}$	28.0	21.0	48.7	33.3	53.5	25.6	50.9	25.8	28.5	22.4
Gemini 1.5 Pro	text-only	$38.2_{\uparrow 10.9}$	$33.6_{\uparrow 0.3}$	25.0	17.6	53.0	33.3	51.9	25.0	54.7	23.3	28.0	24.1
Tarsier-34B	text-only	$35.7_{\uparrow 8.4}$	$34.4_{\uparrow 1.1}$	28.7	25.0	50.9	33.3	49.7	26.2	54.7	22.5	23.5	29.7
Idefics3	image	$44.2_{\uparrow 16.9}$	$34.5_{\uparrow 1.2}$	27.1	23.0	56.8	36.9	49.2	25.6	52.2	29.2	25.5	19.8
GPT-4o	image	$47.8_{\uparrow 20.5}$	$35.8_{\uparrow 2.5}$	30.8	19.6	62.1	33.3	52.4	25.0	52.5	27.5	33.0	21.6
Gemini 1.5 Pro	image	$48.5_{\uparrow 21.2}$	$36.3_{\uparrow 3.0}$	22.8	21.0	55.9	36.4	51.4	36.9	54.7	35.8	30.0	23.7
Tarsier-34B	image	$45.1_{\uparrow 17.8}$	$35.0_{\uparrow 1.7}$	30.0	26.4	61.0	27.1	53.0	32.5	49.4	27.5	21.5	22.0
VideoChat2	video shuffle	$49.8_{\uparrow 22.5}$	$34.7_{\uparrow 1.4}$	25.6	27.0	54.0	32.9	56.2	23.1	48.1	33.3	24.5	22.4
PLLaVA-34B	video shuffle	$56.7_{\uparrow 29.4}$	$37.2_{\uparrow 3.9}$	25.9	29.1	58.5	35.1	56.2	34.4	55.9	28.3	25.0	23.7
Gemini 1.5 Pro	video shuffle	$56.8_{\uparrow 29.5}$	$36.1_{\uparrow 2.8}$	26.5	23.7	55.9	35.1	51.4	31.3	51.9	31.7	29	24.6
Tarsier-34B	video shuffle	$61.2_{\uparrow 33.9}$	$38.0_{\uparrow 4.7}$	30.2	35.8	59.8	34.7	52.4	35.6	55.6	35.0	23.0	18.1
VideoChat2	video	$51.0_{\uparrow 23.7}$	$35.0_{\downarrow 1.7}$	25.9	27.0	56.8	32.9	56.2	23.1	48.1	32.9	24.5	22.4
ST-LLM	video	$54.9_{\uparrow 27.6}$	$35.7_{\uparrow 2.4}$	25.0	35.1	51.7	36.0	54.4	31.0	45.6	34.2	24.0	20.3
GPT-4o	video	$49.1_{\uparrow 21.8}$	$39.9_{\uparrow 6.6}$	26.1	21.3	59.3	33.2	52.4	25.0	78.4	41.7	31.0	30.6
PLLaVA-7B	video	$46.6_{\uparrow 19.3}$	$34.9_{\uparrow 0.9}$	32.1	25.7	55.6	33.3	52.4	23.8	53.1	30.5	20.5	21.6
PLLaVA-13B	video	$50.1_{\uparrow 22.8}$	$36.4_{\uparrow 3.1}$	37.3	24.3	61.8	33.3	55.1	28.1	47.8	39.0	19.5	17.2
PLLaVA-34B	video	$58.1_{\uparrow 30.8}$	$42.3_{\uparrow 9.0}$	27.6	32.4	67.0	35.6	77.8	44.4	58.8	32.9	27.0	19.4
mPLUG-Owl3	video	$54.5_{\uparrow 27.2}$	$42.2_{\uparrow 8.9}$	27.4	32.4	69.8	37.3	76.8	43.1	56.9	34.2	18.0	26.3
VideoLLaMA2.1	video	$57.3_{\uparrow 30.0}$	$42.1_{\uparrow 8.8}$	25.4	30.4	71.2	29.8	76.6	46.9	58.1	36.8	23.5	22.4
VideoLLaMA2 7B	video	$54.6_{\uparrow 27.3}$	$42.9_{\uparrow 9.6}$	30.6	37.8	69.6	33.8	68.1	40.6	56.9	43.9	24.5	22.8
VideoLLaMA2 72B	video	$62.0_{\uparrow 34.7}$	$48.4_{\uparrow 15.1}$	26.7	46.6	73.5	40.4	81.6	53.1	76.6	36.6	19.5	29.7
VideoGPT+	video	$58.7_{\uparrow 31.4}$	$41.7_{\uparrow 8.4}$	30.6	52.0	61.3	36.9	69.2	40.0	53.1	29.7	16.0	28.5
Gemini 1.5 Pro	video	$60.5_{\uparrow 33.2}$	$47.6_{\uparrow 14.3}$	33.5	22.3	71.5	34.5	82.6	51.6	77.8	43.9	25.0	33.6
Qwen2-VL 7B	video	$67.0_{\uparrow 39.7}$	$43.8_{\uparrow 10.5}$	27.1	61.5	63.8	38.7	75.7	41.3	63.1	22.0	22.5	22.4
Qwen2-VL 72B	video	$73.6_{\uparrow 46.3}$	$52.7_{\uparrow 19.4}$	32.6	73.0	68.6	31.1	82.2	45.0	75.3	37.8	19.5	62.1
LLaVA-Video 7B	video	$58.6_{\uparrow 31.3}$	$45.6_{\uparrow 12.3}$	32.6	23.6	78.5	32.9	81.6	58.8	71.9	29.3	22.5	24.1
LLaVA-Video 72B	video	$64.1_{\uparrow 36.8}$	$50.0_{\uparrow 16.7}$	38.6	27.0	80.3	33.3	85.9	65.6	66.6	35.4	25.5	41.8
IXC-2.5-7B	video	$69.1_{\uparrow 41.8}$	$51.6_{\uparrow 18.3}$	32.1	50.7	77.1	37.3	78.4	46.9	60.0	41.7	21.0	70.3
Aria	video	$69.7_{\uparrow 42.4}$	$51.0_{\uparrow 17.7}$	58.4	60.1	75.1	42.2	81.6	43.1	70.3	32.9	21.0	25.0
Tarsier-7B	video	$62.6_{\uparrow 35.3}$	$46.9_{\uparrow 13.6}$	22.9	58.8	73.2	35.6	64.9	46.9	75.6	40.2	25.5	25.0
Tarsier-34B	video	$67.6_{\uparrow 40.3}$	$55.5_{\uparrow 22.2}$	31.7	64.2	81.7	32.1	77.8	58.1	84.4	52.4	24.5	48.3
Human Baseline	video	-	$94.8_{\uparrow 61.5}$	100.0	94.9	100.0	90.6	90.0	96.0	100.0	86.0	90.0	100.0

Table 5: **Results on TVBench** where text-only and image models perform near-random as well as several recent VideoLLMs With TVBench we can identify temporally strong models like Aria and Tarsier as these models drop significantly when the videos are shuffled.

we observe that TVBench amplifies the performance gaps between models with the strongest temporal understanding and those with weaker capabilities.

Conclusion. In this work, we highlight major limitations in existing language-video benchmarks, particularly in the widely used MVBench and open-ended benchmarks. Key issues include inadequate temporal evaluation and tasks that do not require visual information, making tracking progress in this domain ineffective. To address these problems, we introduce TVBench, a benchmark designed to assess the temporal understanding of video-language models explicitly. Our experiments reveal that on TVBench, text-only and visual models lacking temporal reasoning perform randomly, and only a handful of models achieve moderately high scores, showing the potential for progress supported by a human baseline. TVBench provides a reliable yardstick for evaluating future advancements in VideoLLMs and has already been adopted by the community as an evaluation standard for recent SOTA.

Acknowledgement

This work has received financial support from the Agencia Estatal de Investigación (Spain) (PID2023-149549NB-I00), the Xunta de Galicia - Conselleria de Educación, Ciencia, Universidades e Formación (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04 and the European Union (European Regional Development Fund - ERDF). It is also financially supported by Qualcomm Technologies Inc., the University of Amsterdam, and the Top Consortia for Knowledge and Innovation (TKIs) allowance from the Netherlands Ministry of Economic Affairs and Climate Policy.

References

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

[2] Kirolos Ataallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding. *arXiv preprint arXiv:2406.19875*, 2024.

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640. IEEE, 2021.

[6] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

[7] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs, 2024. URL https://arxiv.org/abs/2406.07476.

[8] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al.

Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025.

[9] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, pages 2634–2641, 2013.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[11] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017.

[12] Gemini. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

[13] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842–5850, 2017.

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017.

[15] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[16] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017.

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[18] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. ARIA: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024.

[19] 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 *CVPR*, pages 22195–22206, 2024.

[20] 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.

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

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[23] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE TPAMI*, 42(10):2684–2701, 2019.

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

[25] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. VideoGPT+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024.

[26] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 36, 2024.

[27] MetaAI. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[28] OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[29] 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. *NeurIPS*, 36, 2024.

[30] Yicheng Qian, Weixin Luo, Dongze Lian, Xu Tang, Peilin Zhao, and Shenghua Gao. SVIP: Sequence verification for procedures in videos. In *CVPR*, pages 19890–19902, 2022.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[32] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

[33] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor

Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

[34] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.

[35] Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for training and evaluating large video description models, 2024. URL https://arxiv.org/abs/2407.00634.

[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yasng Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

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

[38] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4581–4591, 2019.

[39] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Video-Hallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.

[40] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.

[41] Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. Deep learning for video classification and captioning. In *Frontiers of multimedia research*, pages 3–29. 2017.

[42] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021.

[43] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. FunQA: Towards surprising video comprehension. *arXiv preprint arXiv:2306.14899*, 2023.

[44] 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 *ACM MM*, pages 1645–1653, 2017.

[45] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021.

[46] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[47] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. PLLaVA: Parameter-free llava extension from images to videos for video dense captioning, 2024.

[48] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. CLIP-ViP: Adapting pre-trained image-text model to video-language alignment. In *ICLR*, 2023.

[49] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. URL https://arxiv.org/abs/2408.04840.

[50] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *ICLR*, 2020.

[51] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, volume 33, pages 9127–9134, 2019.

[52] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11941–11952. IEEE, 2023.

[53] 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*, 2023.

[54] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.

[55] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.