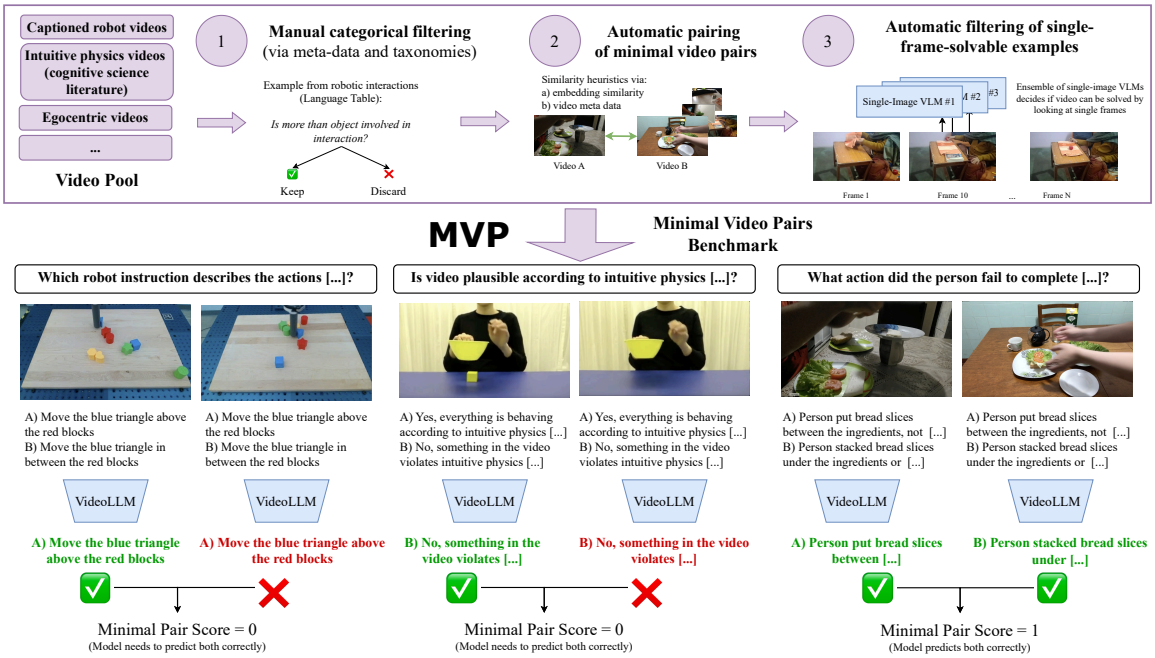
A Shortcut-aware Video-QA Benchmark for Physical Understanding via Minimal Video Pairs

Anonymous authors
Paper under double-blind review

Abstract

Existing benchmarks for assessing the spatio-temporal understanding and reasoning abilities of video language models are susceptible to score inflation due to the presence of shortcut solutions based on superficial visual or textual cues. This paper mitigates the challenges in accurately assessing model performance by introducing the Minimal Video Pairs (**MVP**) benchmark, a simple shortcut-aware video QA benchmark for assessing the physical understanding of video language models. The benchmark is comprised of 55K high-quality multiple-choice video QA examples focusing on physical world understanding. Examples are curated from nine video data sources, spanning first-person egocentric and exocentric videos, robotic interaction data, and cognitive science intuitive physics benchmarks. To mitigate shortcut solutions that rely on superficial visual or textual cues and biases, each sample in MVP has a minimal-change pair — a visually similar video accompanied by an identical question but an opposing answer. To answer a question correctly, a model must provide correct answers for both examples in the minimal-change pair; as such, models that solely rely on visual or textual biases would achieve below random performance. Human performance on MVP is 92.9%, while the best open-source state-of-the-art video-language model achieves 40.2% compared to random performance at 25%.



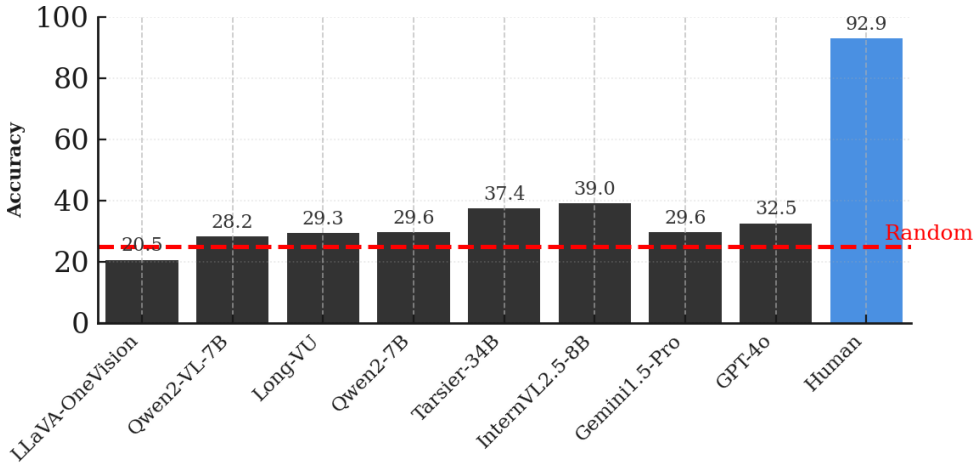


Figure 1: Performance of the strongest evaluated VideoLLMs on MVP (mini-version), compared to human performance.

1 Introduction

Moravec's paradox highlights a counterintuitive phenomenon: high-level reasoning tasks, often perceived as complex, are typically easier for AI agents to solve than sensorimotor and perception tasks, which are seemingly effortless for humans (Moravec, 1988).

Recently, large vision-language models have emerged as a promising paradigm for enabling perception capabilities in AI agents, demonstrating impressive progress on question-answering tasks across various domains including movies, documents, charts, and sciences (Alayrac et al., 2022; Team et al., 2024; Dubey et al., 2024; Wang et al., 2024a). This progress raises a natural question: do these models possess the spatiotemporal understanding and reasoning abilities essential for an agent to interact within the physical world, or do they buttress Moravec's paradox?

Various visual QA datasets have been proposed by the community to assess the spatiotemporal understanding of video-language models (Tapaswi et al., 2016; Maharaj et al., 2017; Li et al., 2024b; Patraucean et al., 2023; Zhang et al., 2023c; Xie et al., 2025; Wang et al., 2023c; Yi* et al., 2020); one of the most popular, MVBench (Li et al., 2024b), combines 11 video datasets into a single video QA benchmark.

While recent state-of-the-art video-language models obtain performance far superior to a random baseline on these benchmarks (Wang et al., 2024a; Shen et al., 2024; Li et al., 2024a), our investigation reveals that existing models can achieve strong performance on these tasks by relying on superficial visual or textual cues or biases. This is validated using simple baselines that discard the visual input or temporal aspect, yet achieve non-trivial performance. Similarly, recent work (Cores et al., 2024) shows that some of these tasks (Li et al., 2024b) fail to accurately measure the temporal understanding of a model.

In this work, we take inspiration from works in natural language processing (Levesque et al., 2012; Sakaguchi et al., 2021) and image processing (Thrush et al., 2022; Yuksekgonul et al., 2022) addressing visual and textual biases in evaluation, and introduce MVP, a video-QA benchmark containing minimal-change video pairs (MVP). Specifically, each video-question-answer sample in the benchmark is accompanied by a visually similar video possessing an identical question but an opposing answer (Figure 1). To answer a question correctly, a model must also provide the correct answer for its minimal-change pair while *processing them independently*. Many types of shortcut solutions are penalized under the minimal-pair scoring framework as a model relying on superficial visual or textual cues or biases would incorrectly output the same answer for both the samples in the pair.

While recent work created small sets of minimal-change video pairs for course-grained temporal reasoning (Zhang et al., 2024a; Liu et al., 2024), our key insight is that these pairs can be efficiently mined from existing video sources to test for several model capabilities through an automated process relying on visual

embeddings and video meta-data. We propose an automatic process to find minimally different video pairs with limited human intervention, and then build these into a video-question-answer tuple with identical questions and opposing answers, enabling the scaling of the benchmark to a broad set of videos spanning diverse situations. We further process the mined samples using a model ensemble to filter out single-frame solvable examples — questions that can be answered using any single randomly sampled frame from the video — to encourage a stronger focus on video understanding. We build MVP by running our process on nine video sources spanning intuitive physics understanding, spatiotemporal reasoning, action anticipation, and robotic manipulation, leading to a total of $54,828$ multiple-choice video QA examples with minimal-change pairs; i.e. $27,414$ minimal-change pairs.

Next, we assess recent proprietary and open-source state-of-the-art video-language models using MVP. Specifically, we evaluate 2 closed-source models (GPT4-o (Achiam et al., 2023) and Gemini-1.5 Pro (Team et al., 2024)), and 7 open-source video-language models: LLaVA-OneVision (Li et al., 2024a), VideoChat2 (Li et al., 2024b), Mini-CPM (Yao et al., 2024), Qwen2-VL (Bai et al., 2023), Tarsier (Wang et al., 2024a), LongVu (Shen et al., 2024), InternVL-2.5 (Chen et al., 2024b). We find that even proprietary models are only slightly above random and that the best accuracy achieved across models is only 40.2% , in stark contrast to human baseline performance at 92.9% accuracy. These findings suggest that video-language models may still struggle with seemingly simple physical reasoning tasks, despite achieving relatively high accuracy on standard spatio-temporal reasoning benchmarks.

In short, we make the following contributions:

1. Analyze potential shortcut solutions on all 11 datasets in the popular MVBench (Li et al., 2024b) benchmark suite, using simple baselines consisting of language-only models, single-frame/image models, and Socratic LLMs.

2. Introduce MVP, a video QA benchmark for physical world understanding comprising minimally different videos — the largest of its kind by an order of magnitude with ∼55K examples.

3. Benchmark closed-source and open-source state-of-the-art models and identify a gap in physical world understanding; human performance on MVP is around 92.9%, while even GPT4-o and Gemini achieve around 30% compared to random performance at 25%.

We are publicly releasing MVP as well as a smaller balanced version of the benchmark for faster inference (MVP-mini), together with a leaderboard and easy download scripts.

2 Robustness Analysis of Existing Video-QA Tasks

We begin by examining robustness of existing video QA benchmarks to shortcut solutions based on visual or textual cues or biases. Specifically, our analysis focuses on CLEVRER (Johnson et al., 2017), Perception Test (Patraucean et al., 2023), STAR (Wu et al., 2021), PAXION (Wang et al., 2023c), Moments in Time V1 (Monfort et al., 2020), FunQA (Xie et al., 2025), Charades-STA (Gao et al., 2017), MoVQA (Zhang et al., 2023b), NTU RGB+D (Liu et al., 2020), VLN-CE (Krantz et al., 2020) and TVQA (Lei et al., 2018), which are all included in the widely adopted MVBench (Li et al., 2024b) benchmark suite.

Empirical Setup. MVBench is comprised of 20 tasks from 11 datasets, collected in a multiple-choice video QA format, where a model is required to choose an answer a_i from a tuple of question, video, and answer candidates $(q, v, [\mathrm{a}_1, \mathrm{a}_2, ..])$. Following standard practice (Goyal et al., 2017), we study robustness to shortcuts by perturbing the task inputs, e.g., requiring the model to select an answer candidate without seeing the video or perhaps without reading the question, and compare to the accuracy achieved by a video LLM without perturbing the task inputs. We study 4 types of shortcut solutions by evaluating language-only models, video-only models, single-frame models, and simple Socratic LLMs. Results are reported in Table 1 using the original skill taxonomy outlined in MVBench.

Task	Avg	AA	AC	AL	AP	AS	CI	CO	EN	ER	FA	FP	MA	MC	MD	OE	OI	OS	ST	SC	UA
Random Chance	0.30	0.50	0.33	0.25	0.25	0.25	0.31	0.33	0.25	0.20	0.25	0.25	0.33	0.25	0.25	0.50	0.25	0.33	0.25	0.33	0.25
GPT-4V †	0.44	0.72	0.39	0.41	**0.64**	0.56	0.11	0.52	0.31	**0.59**	0.47	0.48	0.23	0.12	0.12	0.19	0.59	0.30	0.84	0.45	0.74
VideoChat2 (Mistral)	**0.61**	0.86	0.37	**0.44**	0.55	**0.76**	0.72	0.49	0.36	0.40	0.50	0.64	**0.88**	**0.69**	**0.49**	**0.87**	0.75	**0.41**	**0.85**	**0.50**	0.62
Language only: *Model considers question and answer choices, without access to the video.*																					
Llama 3-8B	0.36	0.63	0.38	0.27	0.25	0.28	0.35	0.43	0.29	0.43	0.29	0.29	0.38	0.27	0.21	0.46	0.29	0.36	0.52	0.40	0.52
Llama 3-70B	0.38	0.78	0.39	0.32	0.26	0.26	0.43	0.47	0.28	0.46	0.26	0.27	0.41	0.29	0.20	0.48	0.29	0.32	0.48	0.45	0.58
Video only: *Model considers video and answer choices only, without access to the question.*																					
VideoChat2 (Mistral)	0.50	**0.88**	**0.42**	0.25	0.49	0.68	**0.74**	0.44	0.28	0.39	**0.53**	**0.65**	0.47	0.29	0.26	0.53	**0.75**	0.34	0.81	0.32	0.55
Single-Frame only: *Model considers question, answer choices and a single key frame, without access to the full video.*																					
Idefics3	0.47	0.72	0.37	0.31	0.52	0.48	0.42	0.54	0.31	0.48	0.40	0.44	0.55	0.42	0.34	0.49	0.50	0.37	0.73	0.48	0.60
Qwen2-VL	0.51	0.87	0.37	0.31	0.55	0.54	0.57	**0.59**	**0.40**	0.45	0.46	0.53	0.6	0.43	0.37	0.53	0.54	0.39	0.74	0.42	**0.68**
Simple Socractic LLM: *Model considers the question, answer choices and a short generic description of the video.*																					
Llama 3-8B	0.44	0.56	0.38	0.28	0.49	0.57	0.35	0.53	0.29	0.42	0.30	0.35	0.56	0.42	0.32	0.50	0.56	0.35	0.68	0.44	0.56
Llama 3-70B	0.46	0.67	0.32	0.35	0.40	0.55	0.38	0.55	0.24	0.45	0.36	0.41	0.56	0.46	0.32	0.57	0.62	0.35	0.70	0.39	0.54

Table 1: **Shortcut Analysis on the 20 MVBench tasks from 11 datasets**: Optimal performance on these spatio-temporal reasoning benchmarks is frequently achieved by models relying on visual or textual biases (Single-Frame only, Video only, Simple Socratic LLM). †: GPT-4V accuracy from (Li et al., 2024b). Tasks: AA (Action Antonym), AC (Action Count), AL (Action Localization), AP (Action Prediction), AS (Action Sequence), CI (Counterfactual Inference), CO (Character Order), EN (Egocentric Navigation), ER (Episodic Reasoning), FA (Fine-grained Action), FP (Fine-grained Pose), MA (Moving Attribute), MC (Moving Count), MD (Moving Direction), OE (Object Existence), OI (Object Interaction), OS (Object Shuffle), ST (Scene Transition), SC (State Change), UA (Unexpected Action).

Language only. Language-only models do not observe the video, and therefore select an answer candidate by only considering the textual inputs q and the answer candidates $[\mathrm{a}_1, \mathrm{a}_2, ..]$. We leverage the Llama3-8B and Llama3-70B models due to their competitive performances (Dubey et al., 2024). In Table 1, we find that a Llama3-8B outperforms a random baseline by 6%, and a larger Llama3-70B outperforms a random baseline by 8%, suggesting that only a small subset of examples can be solved without considering the video input. However, digging into the individual datasets and sub-tasks in Table 1 reveals strong language-only performance on Action Antonym, where LLaMA3-70 achieves 78% compared to a random baseline at 50%. Upon closer inspection of the original dataset, we observe that many questions can be correctly selected by choosing the answer candidate with the highest marginal likelihood. For instance, given an example with answer candidates "book falling like a rock" versus "book rising like a rock," an LLM, just like a human, can rely on its language bias to infer that the former is probably the correct description without observing the video.

Video only. Video-only models do not observe the question, and therefore select an answer candidate by only considering the video input v and the answer candidates $[\mathrm{a}_1, \mathrm{a}_2, ..]$. Table 1 shows that a video LLM (VideoChat2-Mistral) can solve most sub-tasks without access to the question, reaching 50% overall accuracy; by comparison the same model achieves an accuracy of 61% when given the question in addition to the video, while a random baseline is at 30%. These findings indicate that the answer candidates for each question $[\mathrm{a}_1, \mathrm{a}_2, ..]$ are not sufficiently task-specific, as the model is able to discard the incorrect answers without knowing question. Recent work has found similar trends in language understanding QA benchmarks such MMLU-Pro (Wang et al., 2024c), where models are found to reach high scores without access to the question (Chandak et al., 2025).

This trend is particularly interesting on the counterfactual inference sub-task, where the counterfactual scenario such as "What happens if the cube is removed?" can only be known from the question. Manual inspection reveals that the correct answer in this task, e.g. "the red sphere will collide with the purple object", (based on CLEVRER (Yi* et al., 2020)) often occurs in the video regardless of the counterfactual scenario, e.g., the two objects in question will collide regardless of the causal intervention.

Single-frame only. Single-frame models do not observe the entire video, but rather are provided only a single frame $f_i \in v$ form the video. These models must therefore select an answer candidate by considering the frame f_i, the textual inputs q, and the answer candidates $[\mathrm{a}_1, \mathrm{a}_2, ..]$. We take f_i to be the center

frame from the video and leverage Idefics3-8B (Laurençon et al., 2024) and Qwen2-VL-7B (Wang et al., 2024b) for the single-frame baselines. In Table 1, Idefics3-8B achieves an overall accuracy of 47% and Qwen2-VL-7B achieves an overall 51% accuracy, which is comparable to the performance of full-fledged VideoLLMs. Moreover, on Action Antonym, Action Prediction, Character Order, Egocentric Navigation, Episodic Reasoning, Fine-grained Action, State Transition, and Unexpected Action, the single-frame models are on par with (or even exceed) the performance of the VideoLLMs. Concurrent work (Cores et al., 2024) also studies the related bag-of-frame bias by shuffling the video frames.

Simple Socratic LLM. A Simple Socratic LLM (Zhang et al., 2023a; Zeng et al., 2023) replaces the video input v with a short caption c_v that can only convey a low-bandwidth description of the video. In practice, c_v is 1 or 2 sentence-long caption generated by a separate VideoLLM (Zhang et al., 2024b) in a task-independent manner. The Socratic LLMs therefore select an answer candidate by only considering the low-bandwidth caption c_v, the question q, and the answer candidates $[\mathrm{a}_1, \mathrm{a}_2, ..]$. Following the text-only baselines, we use Llama3-8B and 70B. The performance of the Simple Socratic LLMs in Table 1 is significantly above random, with 44% for the LLaMA3-8B and 47% for the LLaMA-70B, suggesting that many sub-tasks (e.g. Character Order, Episodic Reasoning, Scene Transition) do not require fine-grained scene understanding.

Summary. The shortcut analysis reveals that existing models can often achieve strong performance on spatio-temporal reasoning benchmarks by relying on language cues (*Language only* shortcut) or visual cues, (*Video only* shortcut), and may not need to perform temporal reasoning (*Single-Frame only* shortcut), or possess fine-grained visual features (*Simplified Socratic LLM*).

	Domains of VideoQA-Examples							
Benchmark	Total	Natural Videos	Intuitive Physics	Robotics	Synthetic Videos	Minimally Diff. Videos	Procedural Single-Frame Bias Filtering	Format
CLEVRER	76.3K	0K	21.4K	0K	76.3K	×	×	MC-QA
Perception Test	11.5K	11.5K	0-0.2K	0K	0K	×	×	MC-QA
MVBench	4K	2.8K	0K	0.2K	1.2K	×	×	MC-QA
TVBench	2.5K	1.9K	0K	0.2K	0.6K	×	×	MC-QA
Vinoground	1K	1K	0K	0K	0K	✓	×	Group-Score
TempCompass	0.5K	0.5K	0K	0K	0K	✓	×	Group-Score
MVP	54.8K	22.3K	9.9K	25.8K	32.6K	✓	✓	Pair MC-QA

Table 2: **We compare with recent benchmarks that focus on similar skills**. Note that some videos may fall within several categories (e.g., synthetic intuitive physics videos). MVP contains minimally different videos at a much larger scale and across more diverse domains. From these benchmarks, MVP is the first to procedurally filter out examples due to single-frame bias. Group-Score = Present one video + two captions, and two videos + one caption. CLEVRER's intuitive physics entry is grayed as it only covers a narrow subset of intuitive physics concepts, largely based on collisions.

3 Testing Physical World Understanding via Minimal Change Pairs

In this section we discuss the construction of MVP to mitigate shortcut solutions based on visual and textual biases. MVP is comprised of 54,828 video QA examples covering various aspects of physical world understanding, including spatial reasoning, temporal understanding, human-object interaction, memory, counterfactuals, anticipation, and intuitive physics.

Task formulation. To improve robustness to the various shortcut solutions described in the previous section, we adopt a minimal-change pair approach (Levesque et al., 2012; Sakaguchi et al., 2021). An example in MVP consists of two video QA pairs $(q_1, v_1, [\mathrm{a}_1, \mathrm{a}_2])$ and $(q_2, v_2, [\mathrm{a}_1, \mathrm{a}_2])$ containing identical questions $q_1 = q_2$, visually similar videos $v_1 \sim v_2$, and two mutually exclusive (i.e., contradicting) answer candidates a_1 and a_2.

Benchmark Category	Sources (# paired video-QA examples)	Example
Fine-grained human-object interactions	Perception Test (3.5K), Something Something v2 (3.6K)	**Q:** *What stops the motion of the object placed on the slanted plane after being released [...]?* **A)** Person or collision with another object **B)** High friction with surface
Fine-grained robot-object interactions	Language Table (12.9K)	**Q:** *Which robot instruction best describes the actions in the video?* **A)** Move the green blocks in a vertical line below blue cube **B)** Move the green blocks and blue cube in a vertical line
Intuitive physics and collisions	IntPhys (0.2K), InfLevel (2.6K), GRASP (2.0K), CLEVRER (1.2K)	**Q:** *Is this video physically plausible/possible according to your understanding of e.g. object permanence, gravity, [...]* **A)** Yes, everything is behaving according to human intuitive physics understanding **B)** No, something in the video is off/strange or violates [...]
Coarse-grained temporal reasoning	STAR (1.0K), Vinoground (0.5K)	**Q:** *What is the best caption for this video?* **A)** The kayak flips over from facing upwards towards facing downwards **B)** The kayak flips over from facing downwards towards facing upwards

Table 3: **Overview of MVP**. Each answer option A/B is correct for only one video in the minimal-change pair, while acting as a hard negative for the other video (by curation design). Note that we show the number of *paired* video-QA examples, thus the number of videos in our data is twice that amount.

Minimal-change Pair Scoring. A model relying on superficial visual or textual cues or biases to solve a task will tend to produce the same output for each sample in the minimal-change pair. Thus, to penalize models for latching onto shortcuts, we only provide a positive score if the correct answer is produced for both minimal-change samples; the model receives each example $(q, v_1, [\mathrm{a}_1, \mathrm{a}_2])$ and $(q, v_2, [\mathrm{a}_1, \mathrm{a}_2])$ in isolation. Following a multiple choice QA framework, the model has to output a single answer letter (A or B) via task-specific prompts. In this setup, a random baseline achieves an accuracy of 25%.

Question Taxonomy. We wish to understand whether video LLMs possess the spatio-temporal understanding and reasoning abilities essential for an agent to interact within the physical world. As such we consider a coarse-grained taxonomy of question categories encompassing:

- Fine-grained human-object interactions,
- Fine-grained robot-object interactions,
- Intuitive Physics understanding,
- Coarse-grained temporal understanding.

We intentionally construct samples that are not overly reliant on cultural knowledge (Rawal et al., 2024; He et al., 2024; Li et al., 2024c) (e.g., movies) or specific domain knowledge (Tang et al., 2019) (e.g., detailed recipes) — tasks where language bias could contribute to the general performance.

We first manually filter videos from the sources described in Table 3 based on manual inspection (cf. Appendix B.1), then convert them into a question-answer format based on the associated meta-data (the textual captions for Language Table, the class labels for Something-Something-v2, QA annotations for PT, Vinoground, STAR, and CLEVERER, and the concept labels for IntPhys, InfLevel, and GRASP), yielding a starting set of 548K video QA examples.

Minimal-change Pair Mining. Next we procedurally identify minimal-change pairs from the 548K video QA examples produced from the previous stage. We note that 16% of the videos in our final benchmark ($\sim$ 8.8K examples) already possess explicit minimal visual pairs (even though most of these videos are not in

a suitable format for video QA benchmarking, they can be converted into a minimal-change pair question-answer format based on the associated meta-data). For the remaining 84% of the videos, we leverage the following procedure to construct visual minimal-change pairs. In this process, we search for samples that have visually similar videos, identical questions (based on semantic matching), and contradictory answers. To then determine whether two videos with the same question are suitable minimal pairs, we use a) **symbolic and neural rules to determine video similarity** and b) **entailment detection** (Bowman et al., 2015; Dagan et al., 2013) between the correct answers of each video. Whether we rely more on symbolic or or neural rules of similarity depends on the data source: If a dataset has rich annotations (positions or attributes of objects) or structured captions (such as CLEVRER or Something Something-v2), we use hand-crafted rules and the NLP toolkit spacy (Honnibal & Montani, 2017) to narrow down the candidate pool of minimal pairs. This step would match videos with a large intersection of objects or attributes mentioned in the annotation/caption, leading to highly similar videos (e.g., the same objects appearing in both videos). Once we have narrowed down the pool of candidate pairs, in the final step we rank video pairs by their cosine similarity in the ViCLIP (Wang et al., 2023b) video embedding space. We then select the top-ranked minimal video pairs such that each question or skill-type is sufficiently represented. At the same time, we ensure that the correct answers for samples in a minimal-change pair are sufficiently different, as the correct answer of one element in the pair must be a truly negative (*negative*) answer candidate for the other element, and vice versa: To avoid cases where both answers could be true at the same time (e.g., synonyms or more subtle cases) we define a set of textual rules to detect entailment for a subset of datasets. To illustrate this, in the *Fine-grained Robot-object interactions* category, our entailment-detection would discard the following pair of answers: A) "Move the blue cube towards the red heart" and B) "Move the blue cube to the left of the red heart", since A entails B. After this minimal-pair mining, we are down to 70K QA examples; cf. Appendix B.2 for technical details of the minimal pair mining process.

Single-frame Bias Filtering. Finally, to address single-frame bias, we remove examples that can be solved without the temporal information in the video; i.e., using only a single frame. We note that the input frame for this filtering stage should not be selected in a "smart way," since key-frame selection can be regarded as a basic form of temporal reasoning. In practice, five state-of-the-art multi-modal LLMs (LLama3.2-11B (Dubey et al., 2024), Molmo-7B (Deitke et al., 2024), Pixtral-12B (Agrawal et al., 2024), LLaVA-OneVision-7B (Li et al., 2024a), Idefics3-8B (Laurençon et al., 2024)) are prompted to answer the video-QA questions and "give their best commonsense guess given a single frame sampled from a video." If at least 4 out of 5 models in the ensemble predict the correct answer given the same frame, then we flag that frame as *solvable*. The minimal-change pair is then discarded if 30% of the frames in both videos are deemed solvable. This heuristic process removes around 20% of the samples from the previous stage.

MVP Statistics. We end up with $54,828$ examples in MVP, grouped into $27,414$ minimal-change video QA pairs. A breakdown of these examples is shown in Table 2 and Table 3 with a reasonably balanced split between natural videos, synthetic videos, robotics videos, and intuitive physics videos. An average video is 8.8 seconds long, the answer candidates contain an average of 8.1 words, and the datasets contains 2355 unique words in the questions and answers. Note that the word diversity is much less than MVBench (Li et al., 2024b), which has only 4K examples but twice the number of unique words (4338), reflecting our focus in testing for physical world understanding and not linguistically-diverse tasks with cultural or domain knowledge. Instead the task difficulty arises from the physical and perceptual aspects of MVP.

4 Empirical Results on MVP

We evaluate several state-of-the-art open-source VideoLLMs on MVP, summarized in Table 4: LLaVa-OneVision (Li et al., 2024a), VideoChat2 (Li et al., 2024b), Mini-CPM-v 2.6 (Yao et al., 2024), Qwen2-VL (Wang et al., 2024b), Tarsier (Wang et al., 2024a) 7B/34B, LongVU (Shen et al., 2024), InternVL2.5-8B (Chen et al., 2024b), Gemini-1.5 Pro (Team et al., 2024), and GPT4-o (Achiam et al., 2023). Most notably these models differ in their generality: The models we evaluate are either generalist models (GPT4-o, Gemini 1.5), specialized for any visual inputs (LLaVa-OneVision, Mini-CPM, Qwen2-VL, InternVL), or specialized primarily for videos (VideoChat2, LongVU). We also consider two baselines that are fed single-images, LLaVA-OneVision and Qwen2-VL, as they have been trained to process both single image and video. Note that we additionally evaluate on a smaller balanced version of MVP, dubbed MVP-mini, with

Model	MVP (macro-avg)	Fine-grained human-object interactions	Fine-grained robot-object interactions	Intuitive physics and collisions	Coarse-grained temporal reasoning
Random	25.0 (25.0)	25.0 (25.0)	25.0 (25.0)	25.0 (25.0)	25.0 (25.0)
Any text model†	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Single-Frame Baseline *(access to question, answer choices, and a single key frame from the video.)*					
LLaVA-OV (Qwen2-7B)	11.8 (11.7)	14.7 (12.2)	8.7 (10.5)	2.0 (2.3)	21.6 (21.9)
Qwen2-VL (7B)	16.7 (15.7)	16.9 (13.6)	20.1 (19.9)	3.7 (4.4)	26.3 (24.8)
VideoLLMs *(full access to the video, question, and answer choices.)*					
LLaVA-OV (Qwen2-7B)	20.7 (20.5)	24.3 (21.8)	5.2 (5.2)	5.8 (6.8)	47.5 (48.2)
VideoChat2 (Mistral-7B)	23.3 (22.0)	25.7 (21.0)	21.4 (20.1)	10.1 (11.5)	35.8 (35.3)
Mini-CPM-v 2.6	21.7 (22.3)	21.3 (20.2)	18.0 (17.9)	9.2 (11.9)	38.3 (39.2)
Qwen2-VL (7B)	30.0 (29.2)	27.1 (32.28)	27.6 (21.2)	20.0 (18.9)	45.2 (44.5)
LongVU (LLaMA3-3B)	20.6 (20.6)	15.8 (14.1)	14.8 (16.0)	16.2 (16.7)	35.4 (35.8)
LongVU (Qwen2-7B)	29.9 (29.3)	28.9 (26.3)	21.5 (21.8)	20.5 (22.3)	48.6 (46.7)
Tarsier-7B	26.0 (24.3)	31.3 (24.5)	18.7 (18.2)	15.0 (16.3)	38.9 (38.2)
Tarsier-34B	38.8 (37.4)	**45.2** (38.7)	36.3 (36.6)	21.0 (22.1)	52.7 (52.4)
InternVL2.5-8B	**40.2** (39.9)	43.7 (38.1)	**40.2** (38.7)	**22.8** (23.1)	**54.4** (59.8)
Gemini-1.5 Pro	– (29.6)	– (43.1)	– (15.5)	– (19.6)	– (40.2)
GPT4-o	– (32.5)	– (36.1)	– (32.8)	– (16.2)	– (45.0)
Human	92.9	91.3	91.7	97.6	90.9

Table 4: **Accuracy on MVP and MVP-mini in parentheses.** VideoLLM-performance is slightly greater than random chance, while humans achieve greater than 90% accuracy on all categories. Results for closed-source models are only shown on MVP-mini due to API costs. Performance is measured via Minimal Pair Score, wherein a model obtains a score iff the prediction for both QA examples of the pair is correct. †: If temperature of the LLM is zero, since from a text-side both examples in the minimal pair look the same.

around 1/3 of the original size.[1] For a fair comparison, we adopt the default parameters for all models, see Appendix E for details and for the full prompt to the models.

Overall performance of VideoLLMs. Despite their strong performances on other video QA benchmarks (Li et al., 2024b; Liu et al., 2024; Mangalam et al., 2024; Xiao et al., 2021), Table 4 shows that most models perform around random chance (25% accuracy) with the exception of the Tarsier-34B model and InternVL2.5, reaching an average accuracy of 38.1% and 40.2 respectively. This is in contrast to human performances which obtain an average accuracy of 92.9% on a representative subset of MVP (cf. Appendix D).

While average performance is close to random for most models, we do observe non-trivial performance on several sub-tasks and data sources. In particular, VideoLLMs achieve better than random performance on *Coarse-grained temporal reasoning*, meaning they possess some ability to distinguish the order of events in a video.

All models fall short on *Fine-grained robot-object interactions*, which involves understanding fine-grained object manipulation on a table with a robotic arm. This is particularly interesting given the proliferated usage of multi-modal LLMs for learning large-scale visuomotor control policies (Driess et al., 2023; Jiang et al., 2023). Most notably, the *Intuitive physics* category of MVP is by far the hardest with sub-random scores. As highlighted by previous works, intuitive physics reasoning is known to be a difficult task (Riochet et al., 2022; Jassim et al., 2024; Weihs et al., 2022; Du et al., 2023), as this involves reasoning about e.g. object permanence, gravity and trajectories.

VideoLLMs performance on dataset sub-tasks. Some sources in MVP are further divided into more fine-grained splits by their original designers, where each split tests for a specific ability (e.g., object permanence, shape consistency, motion consistency, etc.). In this section we summarize more detailed observations we gathered on them.

[1]We release MVP-mini for faster evaluation and lower costs of API models.

While performance on all intuitive physics tasks is close to 0%, we find that LongVU (Qwen2) obtains non-trivial performance on three splits: Gravity-Continuity (39.1%) and Unchangeableness (42.2%); with Tarsier-34B performing well on Gravity-Support (35.2%). Even some of the weaker models can achieve performance clearly above random on our Fine-grained human-object interactions category, when looking closer into Perception Test subsets such as Counterfactual (e.g., Qwen2-VL: 46.5% LongVU (Qwen2): 43.9%) and Memory (e.g., LongVU (Qwen2): 40.4%).

Importance of Data Curation. In Table 5, we explore the effects of the minimal-change pair mining and single-frame bias filtering on model performance. For this exploration we use the smaller MVP-mini (see Appendix A) and report the average performance of five VideoLLMs [2].

When pairing videos randomly instead of using minimal-change pairs, the average accuracy across tasks is at 45.4%, far superior to random chance. Using minimal-change pairs, the average VideoLLM performance significantly drops to 27.3%. This result shows the importance of the minimal-pair framework and suggests that VideoLLMs can frequently leverage shortcut solutions or spurious features to solve QA tasks. Additionally, the average VideoLLM performance drops again by another 2.2% to 25.1% by removing single-frame solvable videos, with much larger drops on certain subsets. Note that while *Fine-grained robot-object interactions* and the *Intuitive physics and collisions* categories contain almost no single-frame biases, we can see significant drops of 3.5% and 3.3% for the other two categories (*Fine-grained human-object interactions* and *Coarse-grained temporal reasoning*) with this additional filtering step. Overall, Table 5 confirms that the minimal-change pair mining and single-frame filtering pipeline is effective at mitigating potential shortcut solutions in MVP.

Model	Overall	Fine-grained human-object interactions	Fine-grained robot-object interactions	Intuitive physics and collisions	Coarse-grained Temporal reasoning
		Pairing of random videos (with same question)			
Avg. VideoLLM Acc.	45.4	36.8	40.9	19.7	84.3
		+ Pairing of minimally different videos			
Avg. VideoLLM Acc.	27.3 ↓18.1	28.7 ↓8.1	18.6 ↓22.3	16.7 ↓3.0	45.1 ↓39.2
		+ Remove single-frame-solvable examples = final version of MVP			
Avg. VideoLLM Acc.	25.1 ↓2.2	25.2 ↓3.5	18.3 ↓0.3	15.2 ↓1.5	41.8 ↓3.3

Table 5: **We ablate the effect of our main curation steps.** Both the automatic pairing of minimal pairs and the single-frame-bias filtering lead to lower average model performance, with an especially large drop once we introduce the minimal pair setup.

Influence of single-letter output format. When evaluating VideoLLMs on MVP, we ask models to produce a single-letter answer in the format *Answer: A/B* (full prompt in Appendix E). This design keeps inference simple and uniform across diverse APIs and architectures. However one critique might be that models would benefit from open-ended output formats that encourage reasoning before answering. To test this, we updated the prompt for Gemini 1.5 to explicitly request reasoning before committing to a final answer:

> Based on your observations, **reason about the following question [...] in 1–3 sentences**: {QUESTION + OPTIONS}. Then after the reasoning, select the best option [...]

Somewhat surprisingly, performance decreased slightly, in line with recent work on output formatting in LLMs (Long et al., 2025), see Table 6:

[2] LLaVA-OV, VideoChat, Qwen2-VL, LongVU (Qwen2), Tarsier-7B

Minimal Pair Accuracy	Gemini 1.5 (default)	+ reasoning prompt
MVP (macro-avg)	30.0	21.2
Human-Object	27.1	22.5
Robot-Object	20.0	23.9
Intuitive Physics	21.0	9.9
Temporal Reasoning	45.2	28.5

Table 6: **Allowing Gemini 1.5 to reason before answering slightly reduces performance.** While the robot-object category improves, other areas, especially intuitive physics, drop.

5 Related Work

Language biases in Vision-Language models. Vision-and-language benchmarks, such as Visual Question Answering (VQA) (Antol et al., 2015; Goyal et al., 2017; Marino et al., 2019) have been found to be vulnerable to language biases as evidenced by the performance of "blind" language-only models. Blind models are routinely shown to be efficient at solving many of the vision-and-language tasks (Goyal et al., 2017; Zeng et al., 2023; Chen et al., 2024a), and can also solve several image-text retrieval benchmarks (Yuksekgonul et al., 2022; Hsieh et al., 2024) using language biases (Lin et al., 2023). Visual Question Answering in the video-language domain (Video-QA) (Li et al., 2024b; He et al., 2024; Xiao et al., 2021; Lei et al., 2018; Majumdar et al., 2024; Tapaswi et al., 2016; Rawal et al., 2024) also exhibits language biases, as shown in the performance of strong language-only baselines (Zhang et al., 2023a; Cores et al., 2024).

Vision-centric biases in Vision-Language models. State of the art vision-language models are shown to be surprisingly unaware of the vision inputs, where they often struggle with simple questions due to incorrect visual grounding (Tong et al., 2024), despite leveraging sufficiently powerful visual embeddings. VLMs are shown to be imprecise at spatial information understanding and geometry (Rahmanzadehgervi et al., 2024; Kamath et al., 2023). Similar biases exists in video-and-language tasks, where VideoLLMs typically exhibit single-frame bias (Buch et al., 2022; Lei et al., 2023) or spatial bias (Cores et al., 2024), where either a single frame is enough to solve the task, or the ordering of the frames is not important. To overcome this bias, benchmarks propose computing temporal certificate sets (Mangalam et al., 2024), key-frame bias (Buch et al., 2022), or investigate temporal understanding through shuffled frame inputs (Cores et al., 2024). In MVP, we operationalize a looser definition of temporal understanding for our filtering pipeline (Section 3) in that we keep an example if it is only solvable given the right *key-frame*, but discard it if it can be solved with any randomly sampled frame — the intuition being that key-frame identification can already involve temporal reasoning.

Benchmarks addressing vision-and-language biases. Several approaches are proposed in the literature to reduce the aforementioned biases in Vision-Language systems. One promising approach is to use minimally different pairs of inputs (Thrush et al., 2022; Yuksekgonul et al., 2022; Hsieh et al., 2024; Krojer et al., 2022; Wang et al., 2023a), also known as Contrast Sets (Gardner et al., 2020), which stem from related work in natural language processing (Levesque et al., 2012; Sakaguchi et al., 2021; McCoy et al., 2019). Minimally different input pairs restrict the models' abilities to use these biases, as *both* samples in the pair must be answered correctly to achieve a non-zero score. Similar to MVP, some highly adopted examples of such image-language benchmarks build on top of existing image sources (ARO (Yuksekgonul et al., 2022)), or fix them explicitly (SugarCREPE (Hsieh et al., 2024)). Commonly, the focus is on *textual* minimal-change pairs, e.g., providing several answer candidates for a question with only slight variations in word order (Yuksekgonul et al., 2022; Cores et al., 2024; Park et al., 2022; Li et al., 2023; Cai et al., 2024). However, even **textual minimal-change pairs can be susceptible to the same language biases** (Hsieh et al., 2024; Wu et al., 2023), which why we focus on visual minimal pairs in MVP. Other works, such as in Video-QA, focuses on visual minimal-change pairs. TempCompass (Liu et al., 2024) creates a small set of less than 0.5K artificial minimally different videos by manipulating the original video, e.g., playing the video in reverse, at a faster speed, or playing one video above the other. Vinoground (Zhang et al., 2024a) scrapes 0.5K minimally different video pairs from YouTube with the majority following the same pattern: *event A before B* vs. *event B before A*. Our work differs in several aspects from these (summarized Table 2), notably as well in terms of

the scale of curation by showing that minimal video pairs can be procedurally extracted from existing video sources. While our Minimal Pair Score is inspired by Winoground (Thrush et al., 2022), unlike Vinoground, we intentionally do not adopt the Winoground metric directly since we want MVP to be agnostic to whether models can process several videos in one forward pass.

The language biases in existing vision-language benchmarks often stem from the over-reliance on world knowledge and plausible co-occurrences (Hsieh et al., 2024; Goyal et al., 2017). Thus, MVP focuses on short videos with "basic" perceptual skills (spatial, temporal, or intuitive physics), which requires understanding of physical world properties (Yi* et al., 2020; Chen et al., 2022; Jassim et al., 2024; Riochet et al., 2022; Bear et al., 2021; Margoni et al., 2024; Baillargeon et al., 1985), reducing the space for blind LLMs to rely on their cultural knowledge.

6 Discussion and Limitations

Going back to our initial question, our results suggest that VideoLLMs do not yet perceive and understand the world as reliably as humans. After evaluating various state-of-the-art VideoLLM models for physical world understanding on MVP, the best model obtains only 40.2% average accuracy, while human performance is 92.9%. Yet, VideoLLMs are not completely blind. On some sub-categories of spatio-temporal understanding and intuitive physics, VideoLLMs can perform significantly better than random chance. Overall, our empirical evaluation shows that current VideoLLMs are still far from matching human performances on all tested tasks, calling for more research in this direction to develop better training data for world modelling, as well as novel learning criteria and model architectures. We anticipate MVP to help the development of the next generation of visual systems to perceive the world as robustly as humans.

Limitations: No benchmark comes without limitations. First, it is possible that more elaborate prompting strategies for free-form reasoning (CoT) and higher frame rates could improve performance. Additionally, using an automated curation approach will not be able to fully remove noisy examples; through manual inspection, we found some of the examples to be too simple, and a few others to be ambiguous, although we note that these noisy samples only represent a small subset of the overall data.

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.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

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

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985.

Daniel Bear, Elias Wang, Damian Mrowca, Felix Binder, Hsiao-Yu Tung, Pramod RT, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Fei-Fei Li, Nancy Kanwisher, Josh Tenenbaum, Dan Yamins, and

Judith Fan. Physion: Evaluating physical prediction from vision in humans and machines. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/d09bf41544a3365a46c9077ebb5e35c3-Paper-round1.pdf.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2917–2927, 2022.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Answer matching outperforms multiple choice for language model evaluation. *arXiv preprint arXiv:2507.02856*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv [cs.CV]*, March 2024a. URL http://arxiv.org/abs/2403.20330.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. In *International Conference on Learning Representations*, 2022.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024.

Ido Dagan, Dan Roth, Fabio Zanzotto, and Mark Sammons. *Recognizing textual entailment: Models and applications*. Morgan & Claypool Publishers, 2013.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:257364842.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning, 2023. URL https://arxiv.org/abs/2310.10625.

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.

J. Gao, Chen Sun, Zhenheng Yang, and Ramakant Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36, 2024.

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 6297–6305. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/696. URL https://doi.org/10.24963/ijcai.2024/696. Main Track.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. In *Fortieth International Conference on Machine Learning*, 2023.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*, 2020.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3426–3440, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.241. URL https://aclanthology.org/2022.acl-long.241.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1167. URL https://aclanthology.org/D18-1167.

Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 487–507, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.29. URL https://aclanthology.org/2023.acl-long.29.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

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

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, 2024b.

Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. *arXiv preprint arXiv:2311.17404*, 2023.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *ArXiv*, abs/2406.11303, 2024c. URL https://api.semanticscholar.org/CorpusID:270559556.

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.

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. *TPAMI*, 2020.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517.

Do Xuan Long, Ngoc-Hai Nguyen, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen, and Min-Yen Kan. LLMs are biased towards output formats! systematically evaluating and mitigating output format bias of LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 299–330, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.15. URL https://aclanthology.org/2025.naacl-long.15/.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6884–6893, 2017.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16488–16498, 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.

Francesco Margoni, Luca Surian, and Renée Baillargeon. The violation-of-expectation paradigm: A conceptual overview. *Psychological Review*, 131(3):716–748, April 2024. ISSN 1939-1471, 0033-295X. doi: 10.1037/rev0000450. URL https://doi.apa.org/doi/10.1037/rev0000450.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: One million videos for event understanding. *TPAMI*, 2020.

Hans Moravec. Mind children: The future of robot and human intelligence. *Harvard UP*, 1988.

Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3574–3586, 2022.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens Continente, Larisa Markeeva, Dylan Sunil Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=HYEGXFnPoq.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 2024.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *ArXiv*, abs/2405.08813, 2024. URL https://api.semanticscholar.org/CorpusID:269761335.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, 2022. doi: 10.1109/TPAMI.2021.3083839.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

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.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11998–12008, 2023a.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024c.

Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. *Advances in Neural Information Processing Systems*, 36:20729–20749, 2023c.

Luca Weihs, Amanda Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. Benchmarking progress to infant-level physical reasoning in AI. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=9NjqD9i48M`.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Chenwei Wu, Li Erran Li, Stefano Ermon, Patrick Haffner, Rong Ge, and Zaiwei Zhang. The role of linguistic priors in measuring compositional generalization of vision-language models. In *Proceedings on*, pp. 118–126. PMLR, 2023.

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.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pp. 39–57. Springer, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

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 *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=G2Q2Mh3avow.

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*, abs/2312.04817, 2023b.

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*, 2023c.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024a.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024b.

Outline

Appendix A explains why and how we created a smaller version of MVP (MVP-mini).

Appendix B goes into all the nitty-gritty details of how we curated MVP.

Appendix C contains implementation details and further explorations for the shortcut analysis in Section 2 that motivated building MVP.

Appendix D explains how we arrived at the human baseline accuracy on MVP.

Appendix E provides details on how exactly we run inference, e.g. the exact prompt VideoLLMs (or single-image VLMs for the shortcut baselines) and how we extract answers.

Appendix F (Behind the Scenes) shows not just the final product (this paper) but also how we arrived here, what we discarded, and some personal reflections.

A MVP-mini

Next to the full MVP, we also release MVP-mini downsampled in a subset-aware manner to 18,290 video-QA examples (thus 9,145 pairs). MVP-mini will be faster to use, while MVP-full allows researcher to filter and curate derivatives at a large scale. For the most part our process simply involves selecting a random subset, except that we place additional conditions such that no dataset (or subset of a dataset) is underrepresented due to subsampling (i.e. never going below 500 examples). For example, IntPhys is already quite small with 360 videos so we don't remove any examples in this case.

B Details: The Curation of MVP

As illustrated in Section 3, we follow three main steps to curate MVP: 1) Manual categorical filtering 2) Automatic pairing of minimally different examples 3) Automatic filtering of single-frame-solvable examples. We also need to QAify 5 out of the 9 datasets.

While there are commonalities across datasets for how we implement Step 1 (categorical filtering) and Step 2 (pairing), there is also some differences that we describe here for reproducibility. Note that Step 3 is exactly the same across datasets, and is described in sufficient detail in Section 3. Hence, we focus on the first and second step.

B.1 Manual categorical filtering

For 6 out of 9 datasets we select subsets and categories of questions suitable for MVP.

Perception Test: We manually annotate the 132 question types in Perception. Specifically we filter out question that either do not require temporal understanding ("Where is the person?") or are ill-defined. Around 20% are discarded.

Language Table: We select the human-captioned and human-controlled split of Language Table which constitutes 440K. Thus we exclude other splits where the robot arm is automatically controlled and/or the robotic actions are synthetically captioned and not by a human. Additionally, we exclude any videos where the caption only mentions a single object such as "Move the arm to the left" to ensure complex enough interactions.

CLEVRER: We find an issue in the counterfactual split of CLEVRER and exclude these examples based on the meta-data associated with each video, e.g. object attributes and the exact position of each object at each frame. The issue is that most of the time, the object mentioned (*target object*) in the question ("What happens if the cube is not there?") is not actually involved in any collisions. As a result, the correct answer (e.g. "The red and yellow cube collide") is often depicted, whereas in a proper counterfactual example the correct answer should never be depicted but only happen in an "alternative world". Thus we

filter out an example if the target object is never near any other moving objects (i.e. a collision) based on their coordinates. Moreoever, we filter out an example if the two objects mentioned in the correct answer are in fact already colliding in the video, based on their coordinates.

InfLevel: We only use the split with real-life videos where humans conduct the experiments in front of a camera, similar to experimental designs in psychology (Weihs et al., 2022).

STAR: We exclude the *Feasibility* and *Interaction* splits since they are often ill-defined, lead to strong language biases or are too easy.

B.2 Automatic Pairing of Minimally Different Examples

We apply this step to only 5 out of the 9 datasets, since Vinoground (Zhang et al., 2024a) and the 3 intuitive physics datasets are already structured into minimal-change video pairs. Our pairing boils down to finding highly similar pairs, which we base on symbolic or visual similarity, and at the same time ensuring that both answers cannot be true at the same time, i.e. *mutually exclusive*. Especially the latter task, also known as *entailment* detection or *natural language inference*, has many nuanced edge cases.

This step is conducted on QA examples x consisting of a question, a video, and answer candidates: $(q, v, [\mathrm{a}_1, \mathrm{a}_2, \ldots])$.

Perception Test/STAR.

1. Group QA examples into sets with the same question:

$$P = \{X \mid \forall x_i, x_j \in X, q_i \equiv q_j\}.$$

2. For a given X, examples x_i and x_j are grouped into potential pairs if they have opposite (*mutually exclusive*) correct answers:

$$P' = \{(x_i, x_j) \mid a_i \neq a_j\}.$$

3. From this set of potential pairs P', we choose the top-k for a given question based on visual similarity, measured via cosine similarity of embeddings from the video encoder ViCLIP-ViT-L (Wang et al., 2023b):

$$P_k = \big\{(x_{i_1}, x_{j_1}), \ldots, (x_{i_k}, x_{j_k}) \;\big|\; \mathrm{sim}(v_{i_m}, v_{j_m}) \geq \mathrm{sim}(v_{i_{m+1}}, v_{j_{m+1}})\big\}.$$

   In practice, we choose $k = 50$ for each question.

Additionally, we use dataset-specific rules after manual inspection, e.g., for the Perception Test, we require that for two potential pairs x_i and x_j, neither correct answer a_i nor a_j is "Both the other options".

Language Table. Note that all examples in Language Table have the same question "Which robot instruction best describes the actions in the video?".

1. We group QA examples into sets such that a) both correct answers mention the same objects (e.g., both involve a "red triangle" and "green heart") and b) the set of tokens in a_i and a_j have a large enough overlap:

$$P = \big\{(a_i, a_j) \;\big|\; \mathrm{obj}(a_i) \equiv \mathrm{obj}(a_j) \wedge 0 < \mathrm{token\_diff}(a_i, a_j) < 4\big\}.$$

   Due to the finite number of attributes and objects in Language Table, $\mathrm{obj}(\cdot)$ checks for these attribute and object key-words.

2. We narrow this set of potential pairs P with a visual similarity threshold, measured via cosine similarity of embeddings from the video encoder ViCLIP-ViT-L (Wang et al., 2023b):

$$P' = \{(x_i, x_j) \mid \text{sim}(v_i, v_j) > 0.9\}\,.$$

3. Finally, we ensure that answers are mutually exclusive, i.e., $a_i \not\Longrightarrow a_j$ and $a_j \not\Longrightarrow a_i$. In practice, this involves several hand-crafted rules after inspecting failure cases: If the order of objects mentioned is different, there is no entailment (e.g., "Move yellow triangle to blue heart" and "Move blue heart to yellow triangle"); if otherwise one answer contains a *general direction* such as "towards", "to" or "into" but the other answer contains a *specific direction* such as "left" or "above", there is entailment (we discard the example). To illustrate: "move the X towards Y" entails "move X to the left of Y". We identify several of such situations.

Something Something v2. Note that Something Something v2 is a video caption dataset where each caption contains either one or two objects and a simple action, with in total 174 types of such actions. We QAify these examples with the question "Which action is being performed in the video¿' and use the caption with something-placeholders instead of objects as the answer a, e.g., "dropping something".

1. We group QA examples a_i and a_j into pairs such that the action in a_i is a well-defined antonym of the action in a_j:

$$P = \{(a_i, a_j) \mid \text{antonym}(a_i, a_j)\}.$$

   In practice, we identify a subset of 82 action types (47% of all actions) that have a well-defined opposite, e.g., "spinning something so it continues spinning" and "spinning something that quickly stops spinning".

2. We narrow down pairs further by selecting a pair x_i and x_j if the videos contain the same object(s) based on their captions:

$$P' = \{(x_i, x_j) \mid \text{obj}(v_i) \equiv \text{obj}(v_j)\}.$$

   If no pairs fulfill this strict criterion, we relax it such that only one object must overlap:

$$P' = \{(x_i, x_j) \mid \text{obj}(v_i) \cap \text{obj}(v_j) \neq \emptyset\}.$$

3. From this set of potential pairs P', we choose the top-k based on visual similarity, measured via cosine similarity of embeddings from the video encoder ViCLIP-ViT-L (Wang et al., 2023b):

$$P_k = \big\{(x_{i_1}, x_{j_1}), \ldots, (x_{i_k}, x_{j_k}) \;\big|\; \text{sim}(v_{i_m}, v_{j_m}) \geq \text{sim}(v_{i_{m+1}}, v_{j_{m+1}})\big\}.$$

   In practice, we choose $k = 4000$.

CLEVRER. Note that CLEVRER has detailed meta-data with a list of all objects throughout the video and their attributes (color, shape, material), with many videos featuring five or more objects.

1. Group QA examples into sets with the same question:

$$P = \{X \mid \forall x_i, x_j \in X, q_i \equiv q_j\}.$$

2. For a given X, examples x_i and x_j are grouped into potential pairs if they have opposite (*mutually exclusive*) correct answers:

$$P' = \{(x_i, x_j) \mid a_i \neq a_j\}.$$

   In the special case that the answers are both numerical, we require them both to be 1 apart, e.g., "How many objects are moving when the video ends? A) 2 B) 3".

3. We further filter the set P' by requiring a large overlap of objects with the exact same attributes in both videos. Specifically, we keep a pair if the set of objects in v_i is a "fuzzy subset" of the objects in v_j, or vice versa:

$$P'' = \{(x_i, x_j) \mid \text{fuzzy\_subset}(\text{obj}(v_i), \text{obj}(v_j)) \vee \text{fuzzy\_subset}(\text{obj}(v_j), \text{obj}(v_i))\}.$$

Here, fuzzy_subset$(\cdot, \cdot)$ allows one mismatch between the sets of objects and their attributes.

C Details: Shortcut Analysis

We provide additional experiments for the shortcut analysis on MVBench datasets from Section 2.

Language only shortcuts. Similar to (Li et al., 2024b) we also tested a VideoLLM as a language-only baseline by blacking out the video (replacing it with zeros). With VideoChat2 this gave slightly worse results with both LLM-versions: 33.0% (Mistral) and 34.6 (Vicuna).

Video only shortcuts. For the video only we remove the question and only provide the answer candidates to the model. In detail, we tested several ways of removing the question (empty string, replace with "what?", explain to answer without question, etc) and found that replacing the question with "[REDACTED]" yielded the best performance.

Single-frame only shortcuts. First we study how performance varies when selecting frames at different positions: first, middle, last, random and finally key-frame. We choose a key-frame based on the highest CLIP similarity among all frames and all answer candidates (with the question prefixed to the candidate). We find that the performance differs only by 1-2% among these selection strategies except for the first frame which performed more than 5% worse than the rest with Idefics3. Since middle has the highest MVBench accuracy, excluding key-frame, we show middle frame results in the paper. In the main paper we show results for Idefics3 and Qwen2-VL, models that mostly focus on non-video tasks. We also test VideoChat2 variants but found performance to be worse, with either showing a single frame once or copying it 16-times as a "video".

Simple Socratic LLM shortcuts. In this shortcut we test how well models can still perform when the video is replaced by a much lower bandwidth representation and presented to a text-only LLM: a short generic caption of the video. We generate these captions with InternLM-XComposer-2.5-7B (Zhang et al., 2024b). We investigate how model performance differs when increasing the bandwidth of this caption from short, medium to long caption: "[...] Briefly describe this video in one sentence.", "[...] Describe this video in 1-2 sentences." and "[...] Describe this video in as much detail and length as possible.". We also ask whether focus on objects or actions helps, i.e. by prompting the captioning model to list the objects or actions in the video. While the long caption variant achieves the highest performance when provided to LLaMA3 8B and 70B, followed by action caption, we choose to show the medium caption variant (i.e. asking the model to caption the video in 1-2 sentences) in the main paper since this is most in the spirit of a short (1-2 sentences) and generic (not asking for anything specific) caption as a simple baseline.

Additional robustness experiments beyond main paper. As a sanity check, we also study how well a perplexity baseline and answer frequencies perform on MVBench. For the perplexity baseline we compute the perplexity of each answer candidate sequence based on LLM (LLaMA3-7B), i.e. how plausible this string is by itself. For example, are common objects or scenarios more often the correct answer? This would be reflected in such a baseline. However we find that this baseline performs around random overall. Next, we also compute statistics to determine if some subtasks of MVBench have a skewed distribution of answer frequencies, i.e. whether option A is more often correct than the other option B, C, etc. or if it is more often "yes" than "no", etc. Here we also also find very little evidence of any issues in terms of frequencies.

D Human Annotation

We assigned the videos to 6 researchers from our lab, recorded their answer responses, and then computed the benchmark metric using the pair-wise scoring. Each person was assigned one video from the pair at a time, thus avoiding any advantage over VideoLLMs that would come from seeing both videos in a minimal-change pair (i.e., avoiding any knowledge that the answer can only be AB or BA).

E Details on prompting multi-modal LLMs

For a fair comparison (HuggingFace, API), we use each model's default inference setup (e.g. greedy decoding for all models) which are quite similar anyways. For example, videos are similarly processed (resolution, number of frames) and then fed through CLIP or ViT-based video encoders. Most models use either 8 or 16 frames by default, e.g. Tarsier-34B uses 16 which we believe is enough for most MVP examples, yet the model is still far below humans.

We use the following prompt for all VideoLLMs (filled with an example question from our benchmark):

VideoLLM Prompt

You are an expert video understanding AI system. Carefully watch the video and pay attention to the cause and sequence of events, the details and movements of objects, and actions of people. Based on your observations, select the best option that accurately addresses the following question:
Q: `{Question}`
A) `{Correct answer for video1}`
B) `{Correct answer for video2}`
Even when unsure, always answer with a single letter from A or B, format exactly like: 'Answer: A/B'.

We extract the answer letter via a simple regex and find that this approach fails in only less than 1% of examples.

F Behind the Scenes

In this section we go beyond what usually goes into a paper and discuss how the paper came about, what did not work, or what motivated the authors - so in essence: all the things that are usually deemed too subjective or "unscientific", yet would help other researchers, especially those joining the field, often much more than the polished narrative of the main paper.

F.1 Motivation and timeline

Several of the authors who work on video modeling felt a growing frustration with existing benchmarks that often rewarded the wrong things. So the direction of the project was quickly set after a short period of brainstorming: Quantify in what ways existing benchmarks are broken, and then fix it. We then spent a few months staring at hundreds of examples from the MVBench datasets, scouting for glaring issues or shortcuts and manually annotating lots of data. First, we tested the simple baselines with respect to frequency, text-only or single-frame biases, and soon included the less often discussed video-only (remove question) and Simple Socratic LLM shortcuts. In between we had philosophical discussions about benchmark design (bottom-down vs. top-down) or what it means for a video task to be truly temporal: is it temporal if two-frames are needed, or if a single frame is needed but it has to be a key-frame (needle in a haystack), or ...? Regarding benchmark design, should we adopt other paper's taxonomies or design or own? Should one collect all kinds of examples and ad-hoc define a taxonomy (bottom-up), or should one define a taxonomy, then systematically collect examples to fit the taxonomy (top-down)? From the beginning the idea of minimal video pairs generated excitement among us: Minimal visual pairs have led to much progress in the field of

vision-and-language compositionality (e.g. Winoground), yet had not been explored much in the realm of moving images.

There were, and perhaps still are, plans to crowd-source human shortcut performance, i.e. how much better are humans at solving video tasks when given single frames? At scale this could also be used to filter out examples with more precision than our five-model ensemble approach. Doing human crowd-sourcing well is not trivial, it is time-intensive and requires dedication but it can lead to much stronger insights than relying purely on automatic metrics and black-box models.

After the exploration phase, we executed on the benchmark building: From the start we had identified several promising datasets to mine minimal video pairs and continuously added a new source roughly every week. Perception Test was the first to go through our curation pipeline and hence took the longest as we were still refining the pipeline steps. Language Table was very hard to do well with many edge cases in the entailment detection, and also with its scale of 440K video-caption examples (imagine looking for potential pairs of videos, i.e. $440,000^2$ combinations).

F.2 Observations and lessons learned

1. There are too many edge cases to catch every single one at this scale of data curation.
2. Paper writing is smooth when the story and contribution is clear from the beginning of the project (this was not the case in the first author's last paper so it was nice to observe the contrast).
3. The intuitive physics datasets are (to the subjective taste of the first author) the cleanest and most fascinating sources in MVP, even though they are not directly suitable for video QA, and only make up a small fraction of the benchmark.
4. At the same time, Something Something v2 truly stood the test of time as a great video understanding dataset due to its low noise-ratio at a scale of 200K examples, and coverage of interesting yet simple phenomena.
5. Frame rate plays a big role for solving many examples in MVP. In order to push the field further we are now asking the models more and more nuanced questions, and the answer may lie only in a short span of a less than second. However many models may not have access to this short span in principle as they represent a video as 16 uniformly sampled frames.
6. Parsing outputs from LLMs into a structured format such as answer options can feel like the wild west sometimes: Could models perform better if we prompt them better, or have more flexible ways of extracting the answer?

F.3 Advice for others working on a similar direction

Video-QA is becoming an increasingly popular topic; it is a very exciting direction with enough dimensions for everyone to innovate on: long video benchmarking, intuitive physics, social common-sense/Theory-of-Mind/narratives, novel simulation engines, and so on.

Despite our best efforts studying shortcuts, we probably missed some shortcuts or issues in MVP. It is good to think two steps ahead what kind of shortcuts future more capable models could take. It is easier said than done, but in retrospect older video benchmarks from 2015-2020 might have at least been able to address single-frame biases, a priori, during benchmark design. Instead now the field took years to identify and clean up benchmarks.