Can you SPLICE it together? A Human Curated Benchmark for Probing Visual Reasoning in VLMs

Anonymous ACL submission

Abstract

In this work, we introduce SPLICE, a human-curated benchmark derived from the COIN instructional video dataset, designed to probe event-based reasoning across multiple dimensions: temporal, causal, spatial, contextual, and general knowledge. SPLICE includes 3,381 human-filtered videos spanning 12 categories and 180 sub-categories, such as sports, engineering, and housework. These videos are segmented into a total of 11,423 event clips. We evaluate both human participants and state-of-the-art vision-language models (VLMs) on the task of rearranging these clips into coherent event sequences to assess visual reasoning capabilities. Results reveal a significant gap: VLMs struggle to match human performance. While human-annotated textual descriptions improve model accuracy, they do not affect human performance, suggesting that models rely more on language priors than on visual understanding. Even with annotations, VLMs fall short of human-level reasoning, underscoring persistent challenges in visual reasoning. A deeper analysis across sub-categories shows that VLMs perform relatively better on videos where temporal and causal reasoning are dominant, compared to those where contextual and spatial reasoning are dominant. They also perform better on everyday tasks than on specialized ones.

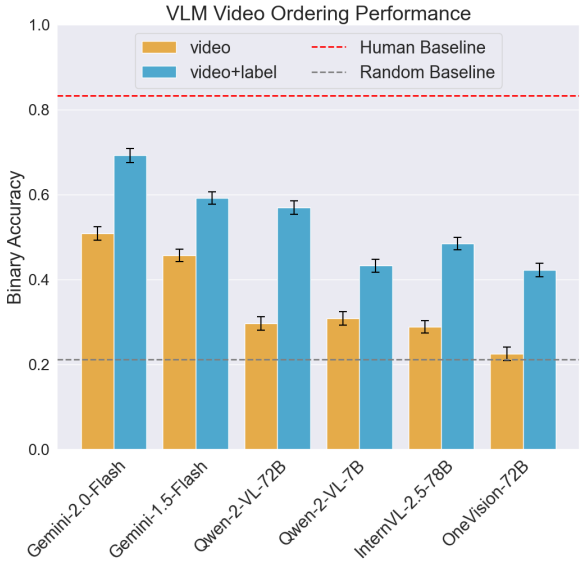


Figure 1: Video clip ordering accuracy of VLMs across 12 video categories (3,381 videos) in two modalities, compared to the human and random baseline.

1 Introduction

Transformer-based models (Vaswani, 2017) initially focused on pre-training with language data alone (Radford et al., 2018; Devlin et al., 2019; Raffel et al., 2020). They later evolved to multimodal pre-training with the introduction of patch-based training (Dosovitskiy, 2020). Since then, vision large language models (VLMs) have advanced rapidly, increasingly matching or even surpassing human performance across various domains, including coding, mathematics, scientific knowledge, and reasoning. For example, benchmarks like ARC-AGI (Chollet, 2019), where models scored 0% in 2019, now report state-of-the-art models achieving scores between 33% and 55.5% (Chollet et al., 2024). While this remarkable progress in reasoning capabilities is essential for enhancing the utility of current AI systems, our understanding of how well these models reason about purely visual sequences remains incomplete. Unlike textual reasoning tasks, where progress is well-documented, the field lacks benchmarks that rigorously evaluate visual reasoning without heavy reliance on language priors.

Alongside advancements in performance, VLMs have become significantly more efficient, enabling them to process long videos. For instance, open-source models like Qwen2-VL (Wang et al., 2024b) can understand videos exceeding 20 minutes in length and handle multiple input videos simultaneously. However, despite these capabilities, current benchmarks do not sufficiently probe models' ability to infer event sequences from purely visual cues.

Leveraging this new capability, we propose an intuitive yet challenging benchmark, SPLICE[1] (**S**equential **P**rocessing for **L**earning and **I**nference in **C**hronological **E**vents), where the task is to order shuffled clips, cut from original videos, based on the events depicted in them. SPLICE fills this gap by leveraging a dataset where the correct ordering of clips requires multiple types of reasoning, such as causal, temporal, spatial, contextual, and general knowledge reasoning.

Unlike previous datasets that rely on automatic or unsupervised segmentations, SPLICE is constructed through a rigorous human curation process. We adapt the COIN instructional video dataset (Tang et al., 2019), which was originally created for video understanding and event localization, by extracting 3,600 videos spanning 180 tasks across 12 domains (e.g., vehicles, gadgets, cooking). In the "vehicles" domain, for instance, tasks may include changing tires, lights, or fuses. COIN provides event-based annotations, which we use to segment each video into distinct clips before shuffling them. This repurposing necessitates careful filtering and validation to ensure that only meaningful and well-defined tasks remain. Consequently, SPLICE eliminates ambiguous or trivial sequences, making it a stronger probe of true visual reasoning.

This event-based structure aims to prevent models from relying solely on the first and last frames or other shortcuts, instead requiring deeper reasoning. For instance, a Karate practice video may be divided into three clips: an opening salutation, practicing movements, and a closing salutation. Since the salutation clips are visually identical but occur at different points, the model must rely on other cues such as breathing, sweat levels, spatial positioning, or background actions to determine their order.

In this work, we compare the performance of multiple state-of-the-art models that support multi-video input, including Qwen2-VL (Wang et al., 2024b), Gemini-Flash (Gemini Team, 2024), InternVL2.5 (Chen et al., 2024), and LlavaOnevision (Li et al., 2024a), across three different input settings: videos only, text only, and videos+text. Additionally, we provide human performance benchmarks and compare them with the performance of these VLMs in both the video-only and video+text settings. Our results indicate that VLMs fall significantly behind humans, particularly in the vision-only setting, where there is a substantial performance gap.

The main contributions of this paper can be summarized as follows:

- We introduce a simple and yet challenging, human-curated benchmark designed to test a model's ability to reconstruct event sequences from shuffled video clips. The dataset consists of 3,381 human-validated videos, each segmented into multiple clips that must be ordered correctly.
- We show that state-of-the-art models struggle with this task, achieving only 23% to 51% accuracy, while humans consistently score around 85%.
- We perform a reasoning-type analysis and show that VLMs perform better when the dominant type of reasoning is causal or temporal, as opposed to contextual or spatial.

2 Related Work

As this work evaluates VLMs on various aspects of reasoning, we provide a comprehensive review of reasoning evaluations for VLMs, along with tasks similar to our approach, where videos or images were shuffled and reordered.

2.1 Reasoning in VLMs

While reasoning in large language models has been extensively explored (Huang and Chang, 2023; Plaat et al., 2024; Xu et al., 2025), covering aspects such as temporal reasoning (Chu et al., 2023; Tang and Belle, 2024; Li et al., 2023; Maruthi et al., 2022), causal reasoning (Zhang et al., 2023; Hobbhahn et al., 2022), general knowledge reasoning (Zhang et al., 2024; Wu et al., 2024b), and spatial reasoning (Li et al., 2024b; Hu et al., 2024), vision language models (VLMs) are considered relatively underexplored (Wang et al., 2024c). This is partly due to their novelty, higher computational demands, and the complexity of evaluating their performance. Nevertheless, several studies have investigated different aspects of reasoning in VLMs (Wu et al., 2024a; Li et al., 2024c; Ko et al., 2023; Zhang et al.).

Several benchmarks have also been developed to evaluate VLMs on different reasoning aspects, such as intuitive physics (Jassim et al., 2023; Weihs et al.,

[1]Samples of the benchmark: https://drive.google.com/file/d/19lvuzTNgQLOkpg9pWLM4mwzK5ZzNHWNv/view?usp=sharing. The full dataset will be made available upon acceptance.

2022), mathematics (Gupta et al., 2024; Chen et al., 2021), spatial reasoning (Mayer et al., 2025), and science in general (Bubeck et al., 2023; Nori et al., 2023). With the increasing capabilities of VLMs and claims of achieving AGI, more challenging benchmarks—such as MMbench (Liu et al., 2025), MMMU (Yue et al., 2024), M4U (Wang et al., 2024a), and ARC-AGI (Chollet et al., 2024)—have been introduced to assess the general capabilities of these models and compare them to human performance. In this context, we present our benchmark, which is human-baselined, to challenge the visual reasoning abilities of VLMs using diverse real-world instructional videos that require multiple aspects of reasoning to solve.

2.2 Ordering Videos

With the new capability of models to take multiple videos as input, most earlier work has focused on reordering videos (Misra et al., 2016; Lee et al., 2017; Xu et al., 2019; Sharma et al., 2020; Hu et al., 2021; Wang et al., 2023) or images (Sevilla-Lara et al., 2021; Kim and Sabuncu, 2023) based on their extracted embeddings.

Additional works (Yang et al., 2025; Li et al., 2020; Fernando et al., 2015) in the literature use image or video ordering for pre-training or training models to enhance their temporal reasoning abilities. Other studies (Wei et al., 2018; Zhou and Berg, 2015) also employ video ordering, but with a primary focus on temporal reasoning. While our presented dataset could be used as a pre-training task, our main goal is to benchmark state-of-the-art vision-language models against human performance on event-based video ordering. This task serves as a proxy to evaluate not only temporal reasoning but also other types of reasoning, including causal, contextual, spatial, and general knowledge.

A closely related line of work (Xu et al., 2019; Hu et al., 2021) treats clip reordering as a self-supervised task by extracting embeddings from uniformly sampled clips. In contrast, we input raw, non-overlapping clips directly into models, leveraging their ability to process multiple clips simultaneously. We also adopt an event-driven clip extraction strategy, where clip lengths vary based on event duration. This approach introduces reasoning challenges beyond temporal understanding, as discussed in Section 7.5.

The approach in Sharma et al. (2020) is similar to Xu et al. (2019) but incorporates additional modalities such as audio, text, and visual features. In our work, we also leverage the multi-modal capabilities of the models, employing two different settings: one using vision only, and another combining vision and text

3 The SPLICE Benchmark

3.1 Step-Annotated Video Dataset

Our video ordering benchmark, SPLICE, is built on a subset of the COIN dataset (Tang et al., 2019), a hierarchically organized collection of 11,827 instructional videos covering 180 tasks across 12 everyday domains[2]. Originally designed for event localization and video understanding, COIN provides detailed, human-annotated step-by-step segments with precise timestamps.

Each video is divided into distinct action steps, with Figure 4 showing an example sequence. Although multiple videos may depict the same task, they vary in execution, step grouping, camera angles, lighting, and backgrounds. This diversity makes COIN a strong foundation for evaluating video ordering tasks in realistic instructional scenarios.

3.2 From COIN to SPLICE: Data Preparation

Due to the complexity and high cost of human cleaning and ordering, we randomly selected a 3,600-video subset from COIN before dividing them into smaller clips for ordering. This subset spans all domains and tasks, with each video originally containing up to 7 clips, segmented based on the original annotations.

The decision to limit the number of clips to 7 was motivated by several factors. First, computational memory constraints in the models make handling longer sequences impractical. Second, longer sequences are more challenging and time-intensive for humans to order accurately. Finally, many high-step videos in the COIN dataset involve repetitive loop actions (e.g., prolonged mixing), where it becomes infeasible to determine an order from shuffled clips, as there are no discernible changes after the action has been completed.

3.3 Human Ordering Protocol

Our dataset includes two modalities: videos-only and videos combined with text. The audio modality

[2]Domains include: Nursing & Caring, Vehicles, Leisure & Performance, Gadgets, Electric Appliances, Household Items, Science & Craft, Plants & Fruits, Snacks & Drinks, Dishes, Sports, and Housework.

was excluded because not all models support audio alongside videos. We have 3,600 uncut videos in total, which effectively becomes 7,200 ordering tasks when considering both modalities. The videos were divided into eight sets of 900: four sets for the video-only modality and four sets for the video + text modality. Four annotators, all authors in this paper, were grouped into two teams, each consisting of a PhD student and a Master's student in cognitive science. Each team handled 1,800 videos across both modalities (900 per modality) without access to the original video order, ensuring that they relied solely on the provided clips for sequencing. The annotators used the following cross-checking procedure:

- **Annotator A:** orders 900 videos in the video-only modality and another 900 in the video + text modality, completing 1,800 ordering tasks in total.
- **Annotator B:** orders the same 900 videos as Annotator A, but flips the modality for each set. Specifically, the 900 videos Annotator A ordered in video-only are ordered by Annotator B in video + text, and the 900 videos Annotator A ordered in video + text are ordered by Annotator B in video-only.

Each set of 900 videos is ordered by both annotators in both modalities to enable cross-checking and ensure consistency. Videos are either ordered or marked as *inconclusive*. Annotators provide their best guess regardless, and a video is labeled inconclusive only if both annotators independently agree. If one annotator provides an order and the other marks it inconclusive, the latter's best guess is used as the final sequence.

We opted against crowd-sourcing to ensure transparency and consistency, as the publicly available dataset and metadata could compromise the reliability of crowd-sourced annotations.

3.4 Criteria for Excluding Videos

Annotators were instructed to mark a video as *inconclusive* only in specific cases: when identical actions were performed separately, such as two people independently demonstrating fire extinguisher use; when actions spanned multiple clips without clear temporal cues, like continuous CPR compressions; or when unrelated actions appeared without contextual links, such as cutting different vegetables with no indication of sequence. This systematic filtering ensured a high-quality dataset suitable for evaluating model reasoning. The final dataset includes 3,381 videos, available in both video-only and video+text modalities. Table 1 presents the video distribution by clip count, and additional statistics are provided in Table 5 in Appendix A.4.

# of Clips	# of Videos	Average Duration (s)
2	1020	46.83
3	1026	53.31
4	734	62.41
5	333	72.67
6	172	67.86
7	96	73.50

Table 1: Distribution of videos by the number of clips, with a total of 3,381 videos segmented into 11,423 clips. The average duration per video is reported.

4 Types of Reasoning

Cutting, shuffling, and reordering clips may seem like a simple task requiring basic temporal reasoning. However, when applied to event-based instructional videos, it demands a richer set of reasoning skills. Below, we outline key reasoning types, with examples from our benchmark:

Contextual Reasoning: Understanding the context of actions to predict what follows. E.g., when changing a printer cartridge, opening the door is typically followed by replacing the cartridge and then closing the door.

Spatial Reasoning: Interpreting spatial relationships and orientations. E.g., aligning a car to park or tracking an athlete's movement across pole vault stages.

General knowledge Reasoning: Applying everyday knowledge to infer plausible actions. E.g., drying follows washing potatoes, or avoiding a hot plate after oven use.

Temporal Reasoning: Grasping the correct sequence of events. E.g., assembling a sofa by attaching legs before placing the mattress.

Causal Reasoning: Recognizing cause-effect links. E.g., mixing flour and water forms dough; cutting a bike chain results in it falling loose.

5 Evaluation

5.1 Metrics

As outlined earlier, each video in the dataset is segmented into clips based on the original COIN dataset's step localization. These clips are then randomly shuffled and renamed as $C = \{c_1, c_2, \ldots, c_n\}$. Models were tested on different modalities, receiving either video input, text

annotations input (short textual descriptions of events in the video) or a combination of both. The model's task is to predict the correct permutation of the clip order. Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ denote the ground-truth sequence of clip indices, and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n]$ represent the predicted sequence. For n clips, there exist $n!$ possible permutations. We evaluate performance using two main metrics: binary accuracy, where a prediction is correct only if the entire sequence matches the ground truth, and position-wise (Hamming) accuracy, which measures the proportion of correctly placed clips. Additional metrics, such as Longest Common Subsequence (LCS) and Edit Distance (Levenshtein Distance), are included in Appendix A.1. Due to the high cost of inference on this benchmark, which includes 22,846 clips for two modalities, we evaluated each model only once, as preliminary tests indicated minimal variation across multiple runs.

5.2 Models

Our benchmark requires models to process multiple videos and reference them accurately. Simply merging frames from different clips is not sufficient; the model must correctly identify and arrange them in sequence, a capability still uncommon in current state-of-the-art models. To assess this, we ran a sanity check by inputting multiple clips into candidate models to evaluate whether they could reference and describe them coherently. Based on these results, we selected Qwen2-VL-Instruct (Wang et al., 2024b), Gemini-Flash (Gemini Team, 2024), InternVL2.5 (Chen et al., 2024), and LlavaOnevision (Li et al., 2024a) as they were among the few able to perform this task reliably. Full test settings, prompts, and model details are provided in Appendices A.2 and A.3.

6 Results

The performance of the models compared to humans is shown in Table 2, with different input settings and metrics, along with random accuracy calculated based on the number of clips in the videos. In terms of binary accuracy, where a prediction is considered correct only if it exactly matches the ground truth, humans score 84.86%, while the highest-performing model, Gemini-2.0-Flash-Exp, scores 51.08%, and random accuracy is 21.14% when using video-only input. This demonstrates that although the model performs well above random accuracy, it still lags behind human performance. In contrast, while human performance did not benefit from the additional text modality, models showed substantial improvement.

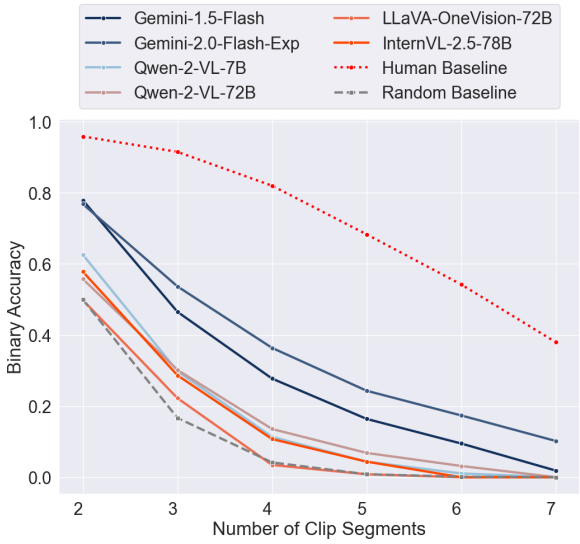


Figure 2: Binary accuracy versus the number of clips (2–7), comparing various state-of-the-art VLMs against human and random baselines

Even on videos that humans misordered, models rarely outdo them. For instance, Qwen2-VL-72B solved none of the 57 seven-clip videos humans got wrong, while Gemini-2.0-Flash-Exp solved only three. Likewise, out of 77 six-clip videos misordered by humans, Qwen2-VL fixed two, and Gemini-2.0-Flash-Exp five, highlighting the persistent gap in visual reasoning.

7 Discussion

7.1 Models Performance

The results indicate that open-source models still lag behind closed-source models like Gemini, particularly in visual reasoning tasks. However, this performance gap narrows when text input is introduced or in text-only evaluations, where Qwen2-VL-72B outperforms Gemini. Notably, Qwen2-VL-7B performs on par with Qwen2-VL-72B in visual reasoning, suggesting that increasing the language model size does not specifically enhance visual reasoning capabilities, given that both models utilize the same vision encoder. Furthermore, SPLICE proves to be a particularly challenging benchmark, as models like LlavaOneVision perform at random levels, while InternVL-2.5-78B scores just above random chance with video only settings.

	Vision Only		Vision+Text		Text	
	Binary	Hamming	Binary	Hamming	Binary	Hamming
Random	0.2114	0.3385	0.2114	0.3385	–	–
Human	0.8486	0.8855	0.8332	0.8904	–	–
Qwen2-VL-7B	0.3091	0.4432	0.4354	0.5683	0.3318	0.4924
Qwen2-VL-72B	0.2990	0.4170	0.5708	0.6820	**0.5402**	**0.6907**
Gemini-1.5-Flash	0.4599	0.5825	0.5936	0.7115	0.4642	0.6029
Gemini-2.0-Flash-Exp	**0.5108**	**0.6188**	**0.6939**	**0.7931**	0.5271	0.6652
InternVL2.5-78B	0.2899	0.4243	0.4856	0.6046	0.4768	0.6062
LLaVA-OneVision-72B	0.2260	0.3514	0.4256	0.5597	0.4210	0.5545

Table 2: Binary and Hamming accuracy scores for various VLMs across different input modalities: Vision Only, Text Only, and Vision+Text. Human and random baselines are included for comparison.

7.2 Human Performance

Human performance on the binary metric reaches around 84%, reflecting the challenge of tasks requiring commonsense and domain-specific knowledge. This ceiling highlights the difficulty of instructional videos, where distinguishing sub-steps in technical domains like automotive or medical tasks depends on expert knowledge. Notably, performance did not improve with text, likely because the short captions (averaging 4.84 words) rarely add information beyond the video. As a result, performance remains consistent across metrics.

7.3 Input Modality

In our study, we report results across three modalities: video-only, text-only, and video+text. However, the text-only results are not intended to assess the reasoning capabilities of language models, as the dataset was curated with a focus on video and video-text inputs. The text often lacks sufficient detail on its own and is primarily included to evaluate the impact of combining modalities. When paired with video, the descriptions become clearer through visual context. Thus, we use text-only performance solely as a baseline to measure gains from incorporating visual information.

The results show that models benefit significantly from the additional text modality, unlike humans. This could be due to two factors. First, the text is a human-generated summary of what is happening in the video, meaning there is a form of knowledge distillation from humans to the model. In contrast, with video-only input, the model receives no additional human-provided context. Additionally, this improvement suggests that language models are still more capable of reasoning compared to Vision-Language Models (VLMs), as they were able to benefit from information that was not necessary for humans to correctly order the videos.

7.4 Factors Influencing Performance

One clear factor influencing the performance of both models and humans is the number of clips. As shown in Figure 2, performance declines for both groups as the number of clips to be ordered increases. However, models are more adversely affected than humans by this increase, and the benefit of including human annotations diminishes with longer sequences as shown in Figure 7 in the appendix. This trend aligns with findings from prior work (Sharma et al., 2024). As a result, most models perform near random when ordering seven clips, where the random-chance baseline is much lower due to the 7! possible permutations.

Figure 3 (left) shows varying performance across categories that does not always concur with human outcomes. For instance, electrical appliances is hard for most models but easy for humans, while both struggle with sports. Models in the same family typically show similar trends, but cross-family comparisons can differ. Another interesting finding illustrated is that Gemini, similar to humans, is not affected by the duration of the videos for both video only input (Figure 3 right) and video+text (Appendix, Figure 6). It maintains stable performance even on longer videos, whereas Qwen2-VL-72B, and to an even greater extent Qwen2-VL-7B, exhibit performance degradation as video length increases. This suggests that Gemini has robust performance across long contexts.

7.5 In-depth Analysis of Reasoning in Models

To better understand model failures, we analyzed sub-domains within the 12 main COIN domains.

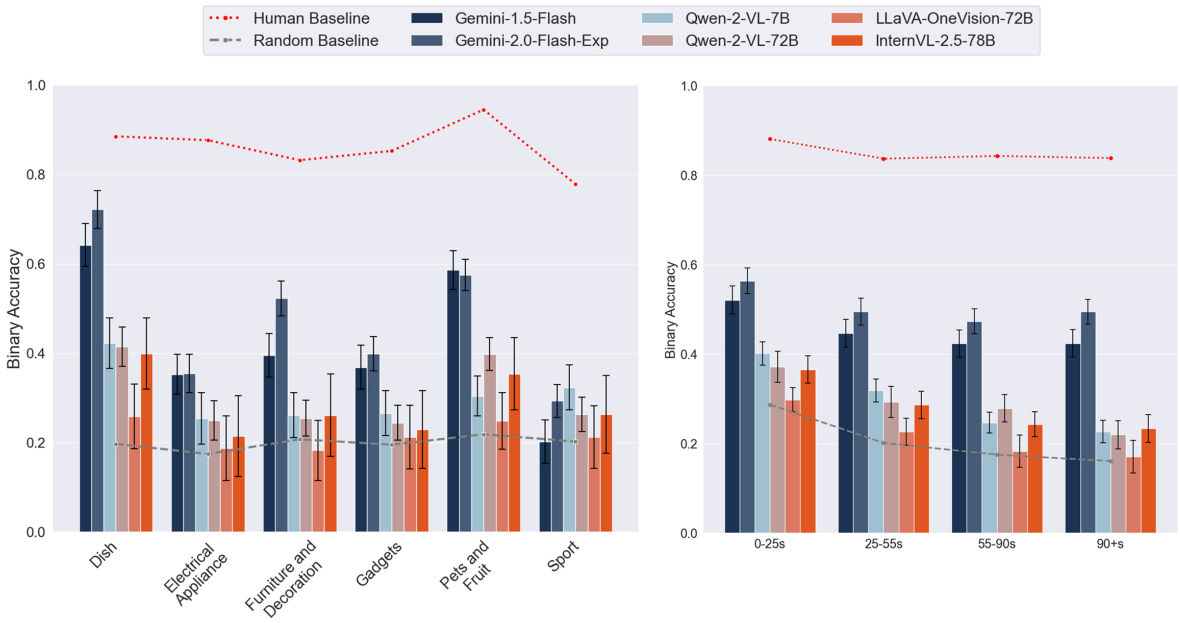


Figure 3: Binary accuracy performance of various state-of-the-art VLMs across different domains (left) and video durations (right), compared to a human baseline (red dashed line) and a weighted random baseline (gray dashed line). Error bars represent the 95% confidence interval (CI)

For example, the "sport" domain includes subclasses like "Practice Skiing Aerials" and "Practice Karate." Our analysis used only the video modality.

Contextual vs. Temporal Reasoning We first examined the "electrical appliance" domain, which exhibited the second-lowest model performance and contained 485 of the dataset's 3381 videos, to understand the reasons for this low performance. A closer look revealed that approximately 70% of the videos involve actions like "replacing" or "changing". In these "change/replace" videos, the initial and final states are often visually similar. For example, in the "replace toner cartridge" subdomain, a typical sequence involves opening the printer door, removing the old cartridge, inserting the new one, and closing the door. Consequently, the first and last clips (door closed) are visually similar to each other, as are the intermediate clips involving the cartridge exchange. Therefore, within the "electrical appliance" domain, we compared sub-domains involving "change/replace" actions against those without. We found a significant performance disparity: Gemini-2.0-Flash scored 26.83% on "change/replace" videos but 53.5% on other videos within the same domain. Results for all models are shown in Table 6 in the appendix. While models performance drops sharply for the "change/replace" group, human performance remains largely unaffected.

To investigate this performance gap further, we examined the models' predictions for "change/replace" videos. We found that models frequently predicted the first and last clips as being sequential. For instance, in videos with four or more clips, Gemini-2.0 predicted the first and last clips were adjacent 57.36% of the time. In contrast, human annotators made this prediction only 2.45% of the time, while random chance would be 27.29%. This indicates a model bias towards ordering visually similar clips adjacently. We attribute this to a potential lack of contextual understanding, where models take shortcuts by associating visually similar clips as sequential rather than performing deeper contextual reasoning (e.g., recognizing that opening a door precedes intermediate actions and closing the door comes last). However, in these same videos, models like Gemini-2.0-Flash demonstrated relatively better temporal reasoning concerning the first and last clips (despite their visual similarity), correctly identifying the first clip as preceding the last about 70% of the time.

Contextual vs. Causal Reasoning We further investigated this hypothesis across all domains using a larger set of videos, filtered into two groups: "make" and "change/replace". Our hypothesis was that "make" videos primarily require causal and temporal reasoning, whereas "change/replace" videos, as previously argued, depend more heav-

Models	Causal vs. Contextual		General Knowledge		Spatial
	Change/Replace	Make	Everyday	Technical	
Number of videos	844	765	342	300	138
Random	18.02	19.14	23.26	22.49	19.19
Human	84.95	86.01	85.38	80.00	76.09
Gemini-1.5-Flash	28.44	61.18	56.14	44.33	24.64
Gemini-2.0-Flash-Exp	32.11	68.37	59.65	47.00	36.96
Qwen2-VL-72B	20.38	36.86	36.55	32.00	21.01

Table 3: Models and human video ordering accuracy (%) on task subsets probing different reasoning types

ily on contextual reasoning. For example, in "make" tasks like preparing food, mixing ingredients causally changes the food's state; similarly, when making a bed, folding the covers has a causal relationship to the finished state. The dataset contains 844 "change/replace" videos (average 3.57 clips) and 765 "make" videos (e.g., make ice cream, make bed; average 3.53 clips), indicating similar complexity based on clip count and thus a comparable random baseline. The results in Table 3 reveal a significant performance difference between the "change/replace" and "make" groups, suggesting models possess stronger causal reasoning abilities compared to contextual reasoning with models like Gemini-2.0-Flash-Exp scoring 32.11% on "change/replace" videos, but 68.37% on "make videos".

General Knowledge Reasoning

Next, we defined two groups: "Everyday tasks" and "Specialized/Technical tasks". "Everyday tasks" include activities like washing dishes or ironing clothes, while "Specialized/Technical tasks" involve actions such as drawing blood or installing a ceiling fan. To avoid earlier issues with "change/replace" scenarios, we excluded such videos from both groups. This resulted in 342 videos labeled as "Everyday tasks" and 320 as "Specialized/Technical tasks". Results in Table 3 show that both models and humans perform better on "Everyday tasks". These tasks likely rely more on commonsense reasoning and are better represented in the models' training data. The performance drop on specialized tasks suggests limitations in the models' knowledge and reasoning in less familiar technical contexts.

Spatial Reasoning Finally, we identified a set of sub-domains requiring spatial reasoning. Examples include "Practice Skiing Aerials", where the model must recognize that the skier in the air precedes landing; pole vaulting, involving jumping, being airborne, and landing; and weightlifting, where the weight on the ground precedes being lifted. Results indicate that models struggle with these spatial reasoning sub-domains; for instance, Gemini-2.0-Flash achieved only 36.96% accuracy, compared to its overall dataset average of 51.08% as shown in Table 3. This likely explains the lower performance observed in the "sport" domain, as 8 out of its 10 sub-domains heavily involve spatial reasoning.

8 Conclusion

In this paper, we presented a novel, human-curated benchmark designed to assess multiple facets of visual reasoning, including temporal, causal, contextual, visual-spatial, and general knowledge. We evaluated various open-source and closed-source VLMs under different input modalities and compared their performance against human participants. Despite improvements from combining video and text (highlighting the value of cross-modal alignment), all models lag significantly behind human performance, especially without human-annotated descriptions. Moreover, open-source models lag further behind their closed-source counterpart, revealing a persistent gap in visual reasoning. The low performance of several models, with some scoring just above random chance, highlights the benchmark's effectiveness as a rigorous probe of visual reasoning. Our reasoning analysis further reveals that models perform better on videos dominated by causal and temporal reasoning than on those requiring contextual or spatial reasoning. In the future, we aim to incorporate voice to enhance cross-modal alignment and assess how models integrate audio with visual reasoning.

9 Limitations

Currently, only a few state-of-the-art VLMs support the ability to input multiple videos and reference them appropriately. Even single-video processing capabilities are limited in many models, restricting our evaluation to the handful that do offer this functionality. Nonetheless, the field is evolving rapidly, and we expect that most models will soon be able to handle multi-video inputs, enabling broader application of our benchmark.

References

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523.

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

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. 2024. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. 2015. Modeling video evolution for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5378–5387.

Gemini Team. 2024. Gemini 2.0 flash (experimental).

Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. 2024. Polymath: A challenging multi-modal mathematical reasoning benchmark. *arXiv preprint arXiv:2410.14702*.

Marius Hobbhahn, Tom Lieberum, and David Seiler. 2022. Investigating causal understanding in llms. In *NeurIPS ML Safety Workshop*.

Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. 2024. Chain-of-symbol prompting for spatial reasoning in large language models. In *First Conference on Language Modeling*.

Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. 2021. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. 2023. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*.

Heejong Kim and Mert R Sabuncu. 2023. Learning to compare longitudinal images. *arXiv preprint arXiv:2304.02531*.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316.

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2017. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 667–676.

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

Fangjun Li, David C Hogg, and Anthony G Cohn. 2024b. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18500–18507.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065.

Xingxuan Li, Liying Cheng, Qingyu Tan, Hwee Tou Ng, Shafiq Joty, and Lidong Bing. 2023. Unlocking temporal question answering for large language models using code execution. *arXiv e-prints*, pages arXiv–2305.

Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Tom Weidong Cai. 2024c. Multimodal causal reasoning benchmark: Challenging vision large language models to infer causal links between siamese images. *CoRR*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Srihari Maruthi, Sarath Babu Dodda, Ramswaroop Reddy Yellu, Praveen Thuniki, and Surendranadha Reddy Byrapu Reddy. 2022. Temporal reasoning in ai systems: Studying temporal reasoning techniques and their applications in ai systems for modeling dynamic environments. *Journal of AI-Assisted Scientific Discovery*, 2(2):22–28.

Julius Mayer, Mohamad Ballout, Serwan Jassim, Farbod Nosrat Nezami, and Elia Bruni. 2025. ivispar–an interactive visual-spatial reasoning benchmark for vlms. *arXiv preprint arXiv:2502.03214*.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 527–544. Springer.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. 2021. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 535–544.

Aditya Sharma, Michael Saxon, and William Yang Wang. 2024. Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5429–5451, Miami, Florida, USA. Association for Computational Linguistics.

Vivek Sharma, Makarand Tapaswi, and Rainer Stiefelhagen. 2020. Deep multimodal feature encoding for video ordering. *arXiv preprint arXiv:2004.02205*.

Weizhi Tang and Vaishak Belle. 2024. Ltlbench: Towards benchmarks for evaluating temporal logic reasoning in large language models. *arXiv preprint arXiv:2407.05434*.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024a. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*.

Jianghui Wang, Yuxuan Wang, Dongyan Zhao, and Zilong Zheng. 2023. Moviepuzzle: Visual narrative reasoning through multimodal order learning. *arXiv preprint arXiv:2306.02252*.

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

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024c. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. 2018. Learning and using the arrow of time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8052–8060.

Luca Weihs, Amanda Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Benchmarking progress to infant-level physical reasoning in ai. *Transactions on Machine Learning Research*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's transformers: State-of-the-art natural language processing. Publication Title: arXiv e-prints ADS Bibcode: 2019arXiv191003771W.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024a. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024b. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. 2019. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10334–10343.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. Towards large reasoning models: A survey on scaling llm reasoning capabilities. *Preprint*, arXiv:2501.09686.

Charig Yang, Weidi Xie, and Andrew Zisserman. 2025. Made to order: Discovering monotonic temporal changes via self-supervised video ordering. In *European Conference on Computer Vision*, pages 268–286. Springer.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, et al. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.

Yifei Zhang, Xintao Wang, Jiaqing Liang, Sirui Xia, Lida Chen, and Yanghua Xiao. 2024. Chain-of-knowledge: Integrating knowledge reasoning into large language models by learning from knowledge graphs. *arXiv preprint arXiv:2407.00653*.

Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Yipin Zhou and Tamara L Berg. 2015. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506.

A Appendix

A.1 Additional Metrics and full results

Each video in the dataset is segmented into clips based on the original COIN dataset's step localization. These clips are then randomly shuffled and renamed as $C = \{c_1, c_2, \ldots, c_n\}$. Depending on the modality being tested, the model receives either video-only input or a combination of video and annotations (short textual descriptions of events in the video). The model's task is to predict the correct permutation of the clip order. Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ denote the ground-truth sequence of clip indices, and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n]$ represent the predicted sequence. For n clips, there exist $n!$ possible permutations. The models are evaluated on four metrics:

Binary Accuracy. The prediction is correct only if the entire sequence matches the ground truth:

$$\text{Binary Accuracy} = \begin{cases} 1 & \text{if } \hat{\mathbf{y}} = \mathbf{y}, \\ 0 & \text{otherwise.} \end{cases}$$

Position-Wise (Hamming) Accuracy The proportion of correctly placed clips:

$$\text{Hamming Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\hat{y}_i = y_i)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Longest Common Subsequence (LCS). The LCS measures the longest sequence of elements appearing in the same relative order in both $\mathbf{y}$ and $\hat{\mathbf{y}}$. Let $c(i, j)$ denote the length of the LCS for substrings $\mathbf{y}_{1:i}$ and $\hat{\mathbf{y}}_{1:j}$:

$$c(i,j) = \begin{cases} 0 & i = 0 \text{ or } j = 0, \\ c(i-1, j-1) + 1 & y_i = \hat{y}_j, \\ \max\{c(i-1, j), c(i, j-1)\} & \text{otherwise.} \end{cases}$$

The LCS ratio normalizes this value:

$$\text{LCS Ratio} = \frac{\text{LCS Length}}{n}$$

Edit Distance (Levenshtein Distance). The minimum number of insertions, deletions, or substitutions required to transform $\hat{\mathbf{y}}$ into $\mathbf{y}$. Define a matrix D where $D(i, j)$ is the edit distance between $\mathbf{y}_{1:i}$ and $\hat{\mathbf{y}}_{1:j}$:

$$D(i, 0) = i, \quad D(0, j) = j \quad \text{(boundary conditions)},$$

$$D(i,j) = \begin{cases} D(i-1, j-1) & y_i = \hat{y}_j, \\ 1 + \min \begin{pmatrix} D(i-1, j) \\ D(i, j-1) \\ D(i-1, j-1) \end{pmatrix} & \text{otherwise.} \end{cases}$$

The final edit distance is $D(n, n)$.

	Vision Only				Vision+Text			
	Binary	Hamming	LCS	Edit	Binary	Hamming	LCS	Edit
Random	0.2114	0.3385	0.6554	2.0970	0.2114	0.3385	0.6554	2.0970
Human	0.8486	0.8855	0.9359	0.4105	0.8332	0.8904	0.9337	0.3875
Qwen2-VL-7B	0.3091	0.4432	0.7130	1.7891	0.4354	0.5683	0.7896	1.4377
Qwen2-VL-72B	0.2990	0.4170	0.7011	1.8465	0.5708	0.6820	0.8483	1.0677
Gemini-1.5-Flash	0.4599	0.5825	**0.7980**	1.3458	0.5936	0.7115	0.8633	0.9734
Gemini-2.0-Flash-Exp	**0.5108**	**0.6188**	0.7927	**1.2511**	**0.6939**	**0.7931**	**0.9030**	**0.7788**
InternVL2.5-78B	0.2899	0.4243	0.7050	1.8364	0.4856	0.6046	0.7694	1.3602
LlavaOnevision-72B	0.2260	0.3514	0.6615	2.0636	0.4256	0.5597	0.7866	1.4312

Table 4: Performance comparison of various VLMs across different input modalities (Vision Only and Vision+Text) using Binary Accuracy, Hamming Distance, Longest Common Subsequence (LCS), and Edit Distance metrics. Human and random baselines are included for reference. Models perform significantly better with textual input, highlighting the benefit of cross-modal information

A.2 Testing Settings

Below are the details about test settings of each model:

Qwen2-VL-Instruct Family. Qwen2-VL was tested with both 7B and 72B parameters. The number of frames was set to 1 fps, and the highest image resolution was set to 448 pixels, while the other dimension was automatically adjusted based on the aspect ratio of the input frames.

Gemini-Flash Family. We used Gemini Flash 1.5 and 2.0 (experimental) versions, with the fps set to 1. The model was loaded using the official Google API, and the image resolution was left at the default setting, allowing the model to handle it automatically.

InternVL2.5 Family. InternVL2.5 was tested with the 78B parameters model only. The 8B model did not pass the sanity check. We used the default settings of the uniformal distribution of frames input for each clip and we set it to 16 frames instead of fps.

LlavaOnevision Family LlavaOnevision was tested with 72B parameters. The 7B model did not pass the sanity check. We used the default settings of the uniformal distribution of frames input for each clip and we set it to 16 frames instead of fps.

All of the open source models were used from the Hugging Face library (Wolf et al., 2019) and adopted with the Flash Attention approach. All of these models are tested with three different modalities, vision only, text only, and vision + text. Samples of the prompts are shown in the Appendix A.3. All jobs were submitted to a cluster of A100 and H100 GPUs, which were used interchangeably based on availability.

A.3 Prompts

Three samples of prompts are shown below, for each model the prompts were slightly tuned for better performance:

Here is a sample prompt for video-only input: **prompt:** f"A video has been split into len(clips) clips, shuffled randomly." "Your task is to analyze each clip deeply to reorder them into the correct temporal sequence. Focus on:" "1. Visual content: Examine the actions, transitions, scene details, and context within each clip." "Provide the reordered sequence strictly within order tags in this format: " "'<order>Video X, Video Y, Video Z, ...</order>'."

Here is a sample prompt for video+text input: **prompt:** f"A video has been split into len(clips) clips, shuffled randomly." "Your task is to analyze each clip deeply to reorder them into the correct temporal sequence. Focus on:" "1. *Visual content*: Examine the actions, transitions, scene details, and context within each clip." "2. *Temporal logic*: Identify the logical progression of events based on what happens before or after." "3. *Annotations*: Leverage the annotations to infer their proper chronological sequence." "Provide the reordered sequence strictly within order tags in this format: " "'<order>Video X, Video Y, Video Z, ...</order>'."

Here is a sample prompt for text models: **prompt:** "Analyze the following video clips descriptions and order them chronologically as they are part of one continuous video. " "Focus on temporal clues, event progression, scene transitions and other cues " "Each video clip is labeled as 'Video X', where Video X corresponds to one shuffled clip. " "Maintain these labels in your response. "

"Return the ordered video strictly within <order> tags in this format: " "<order>Video X, Video Y ...</order> ")

A.4 Videos Stats

Table 5 provides statistics on the segmented video dataset, detailing how videos are divided into segments and their distribution across different segment counts.

Segments	Videos	Clips	Mean Time (s)	Std Dev	(2, 35]	(35, 68]	(68, 100]	(100, 133]	(133, 166]	(166, 198]	(198, 330]
2	1020	2040	46.84	40.75	531	245	128	65	35	12	4
3	1026	3078	53.32	42.68	434	309	142	79	36	19	7
4	734	2936	62.41	40.76	209	260	146	70	32	13	4
5	333	1665	72.67	43.20	62	123	69	46	21	10	2
6	172	1032	67.86	37.49	38	56	45	22	9	2	0
7	96	672	73.50	38.28	19	26	30	14	5	2	0

Table 5: This table summarizes the distribution of videos based on their segmentation. It includes the number of segments(2-7), total videos per segment number, total clips, mean duration (seconds), and standard deviation. The rightmost columns show the distribution of videos across predefined video duration intervals, providing insights into the dataset's temporal structure for event ordering analysis.

A.5 Instructions for Annotators

Instruction Brief Task: Reorder the video parts for each folder into their correct sequence. Steps: Download and Open the Folder assigned to you: You will receive a folder containing several subfolders, each labeled with a unique number (e.g., 1, 2, 3, etc.). Each subfolder corresponds to a video task with shuffled parts. View the Video Parts: Inside each subfolder, you will find video parts named random_part_1.mp4, random_part_2.mp4, etc. These parts contain embedded labels as secondary context for your understanding of the video context. Reorder the Parts: Watch each video part carefully. Determine the correct sequence of these parts based on the visual and textual cues. Write down the sequence in the format: Folder Number: Correct Order (e.g., 1: random_part_3, random_part_1, random_part_2). For simplicity use [2, 3, 4, 5, 1], where each number represents the Random number video. Use "unk" in these cases:

1- Repeated instructions: If the video contains two separate instances of the same instruction.

2- Continuous actions without sufficient context: An action extends across multiple clips with insufficient background information to establish a clear sequence.

3- Unrelated actions: The video includes unrelated actions with no contextual clues to establish order.

Submit Your Results: Compile the correct order for all folders in the attached spreadsheet Use "unk" for any task sample you believe makes no sense or as discussed during the meeting, Notes: Do not use any external sources Complete all tasks to the best of your ability.

A.6 Additional Results

Table 6: Video Ordering Accuracy of Electrical Appliance domain for sub-domains that include change/replace compared to others that dont

Models	Accuracy (%)
Group: change/replace, 328 videos	
Random	15.02
Human	88.11
Gemini-1.5-Flash	26.22
Gemini-2.0-Flash-Exp	26.83
Qwen2-VL-72B	21.04
Group: other, 157 videos	
Random	22.68
Human	86.62
Gemini-1.5-Flash	54.14
Gemini-2.0-Flash-Exp	53.50
Qwen2-VL-72B	33.12

A.7 Additional Figures

A.7.1 Example of Video Input

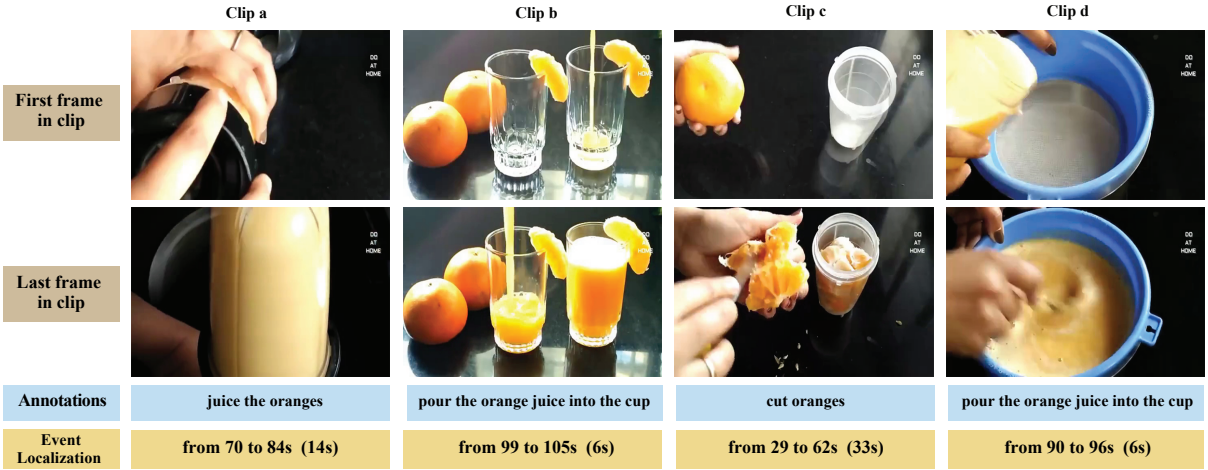
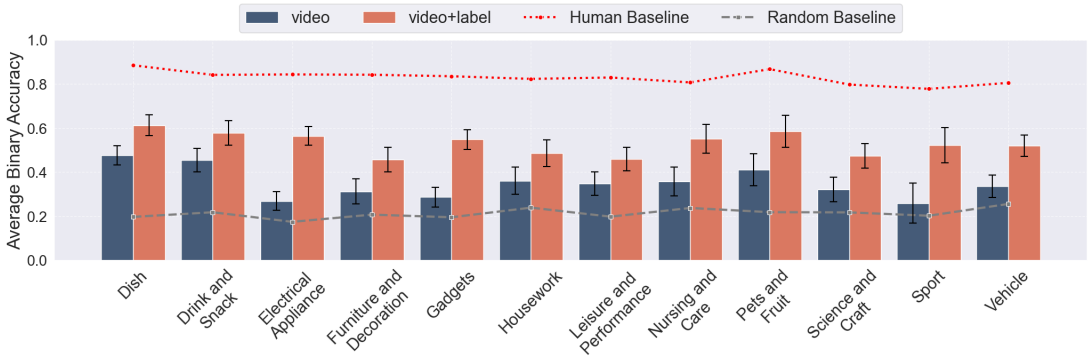
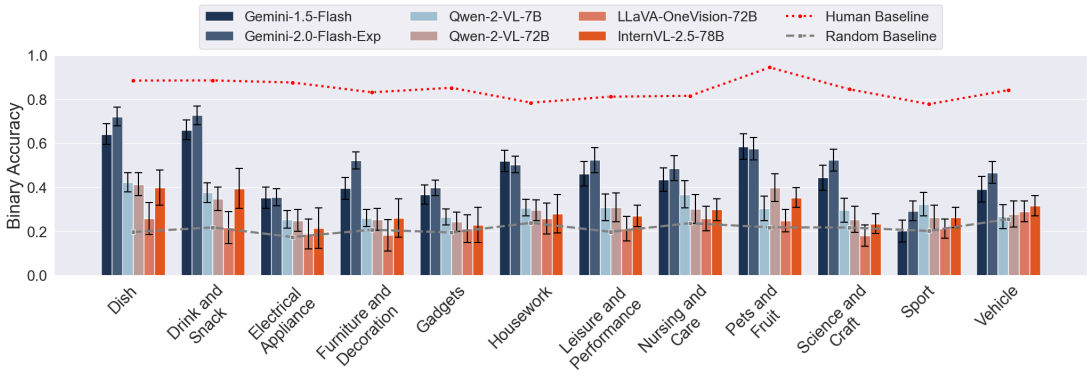


Figure 4: An example of a set of clips that the models need to order correctly. The figures show the first and last frames of each clip. The clips are segmented based on events, reducing the reliance on shortcuts. Clip durations vary based on the event, with gaps where moments not relevant to the main steps are omitted. In this video, models must infer that the oranges are cut, juiced, filtered, and then served. Clips order: c, a, d, b.

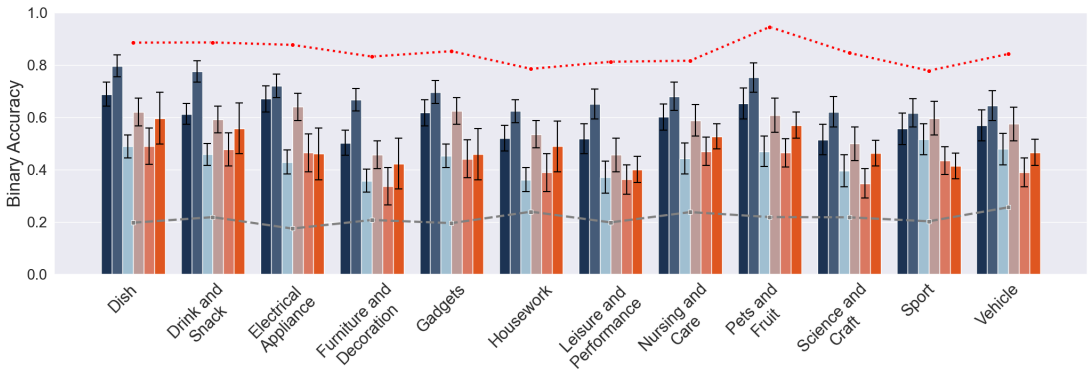
A.7.2 Modalities Across Domains



(a) Average VLM performance on both modalities across video domains.



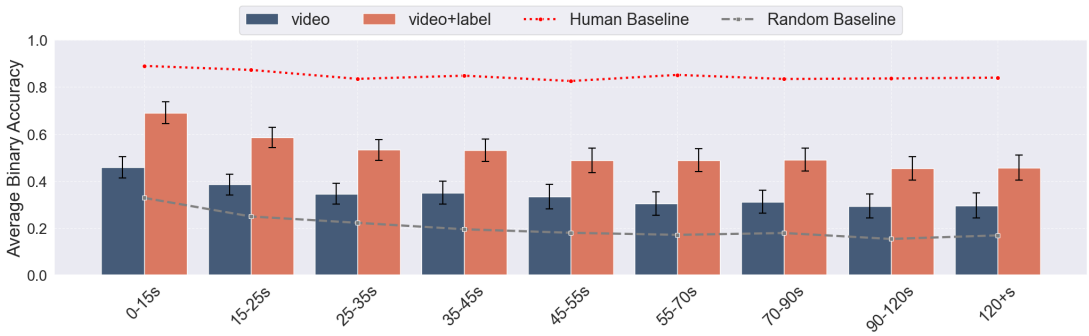
(b) VLM performance by video domains with video only input



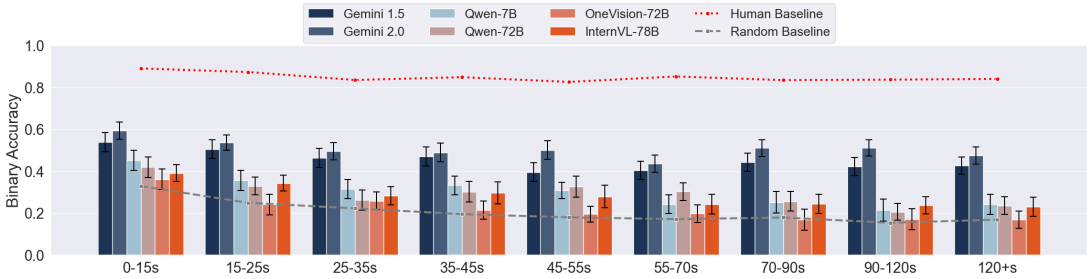
(c) VLM performance by video domains with video and text input

Figure 5: Binary accuracy performance of various state-of-the-art VLMs across different domains and modalities compared to a human baseline (red dashed line) and a weighted random baseline (gray dashed line). Error bars represent the 95% confidence interval (CI).

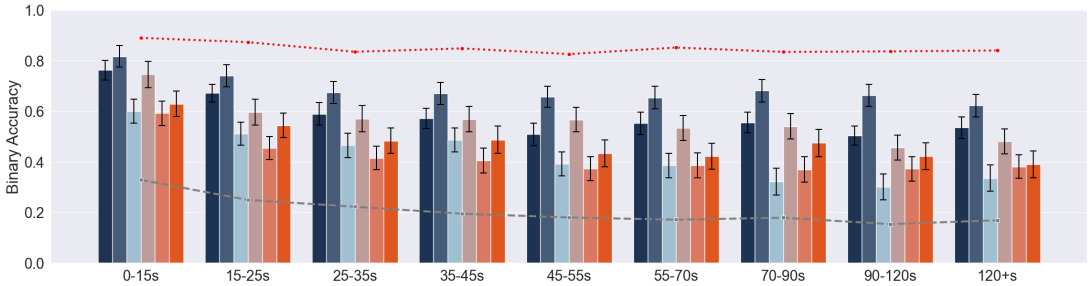
A.7.3 Modalities Across Video Length



(a) Average VLM performance on both modalities across video length.



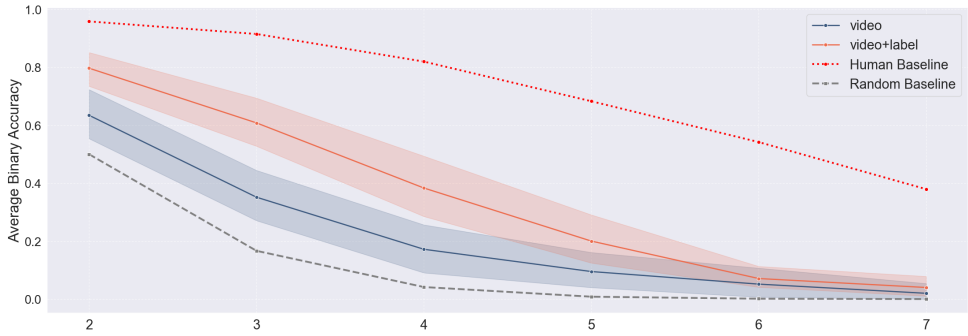
(b) VLM performance by video duration with video only input



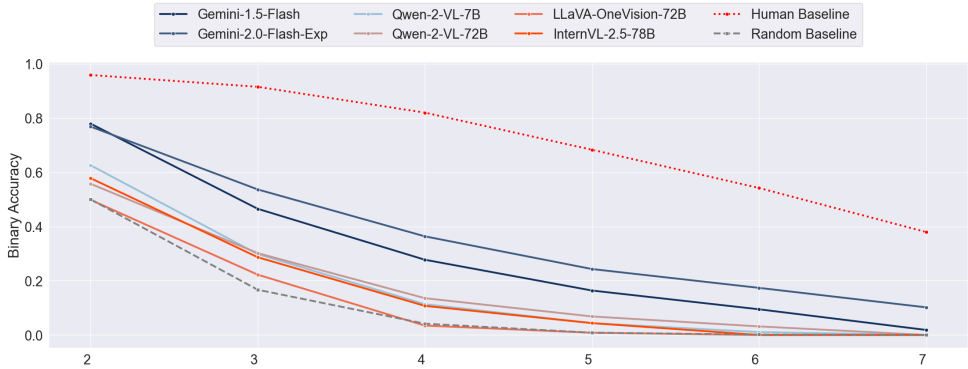
(c) VLM performance by video duration with video and text input

Figure 6: Binary accuracy performance of various state-of-the-art VLMs across video duration and modalities compared to a human baseline (red dashed line) and a weighted random baseline (gray dashed line). Error bars represent the 95% confidence interval (CI).

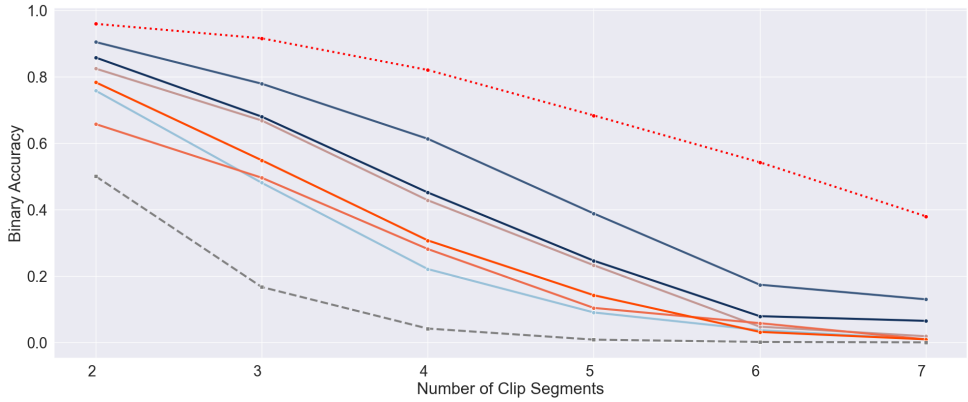
A.7.4 Modalities Across Number of Clips



(a) Average VLM performance on both modalities across number of Clips.



(b) VLM performance comparison by number of clips with video only input



(c) VLM performance by number of clips with both video and text as input

Figure 7: Binary accuracy performance of various state-of-the-art VLMs across different number of clips and modalities, compared to a human baseline (red dashed line) and a weighted random baseline (gray dashed line). Error bars represent the 95% confidence interval (CI).