Vinoground: Today's LMMs Don't Understand Short Counterfactual Videos

Jianrui Zhang[1*] **Mu Cai**[1,2*†] **Yong Jae Lee**[1]
[1]University of Wisconsin-Madison [2]Google DeepMind

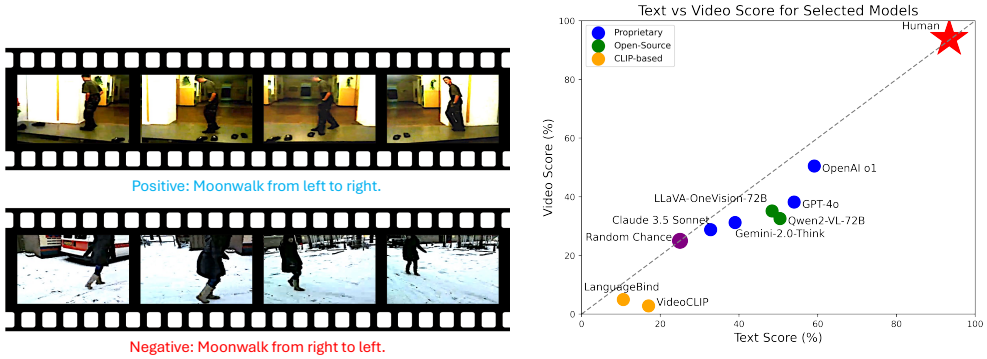

Figure 1: (Left) An example data point from the *spatial* category. Both videos last less than 10 seconds. Each data point contains two pairs of video-caption pairs counterfactual to each other. (Right) Models perform significantly poorer than humans, performing better on the text score metric than on the video score metric, as defined in Section 4.1.

Abstract

There has been growing sentiment recently that modern large multimodal models (LMMs) have addressed most of the key challenges related to short video comprehension. As a result, both academia and industry are gradually shifting their attention towards the more complex challenges posed by understanding long-form videos. However, is this really the case? Our studies indicate that LMMs still lack many fundamental reasoning capabilities even when dealing with short videos. We introduce Vinoground, a temporal counterfactual LMM evaluation benchmark encompassing 1000 short and natural video-caption pairs. We demonstrate that existing LMMs severely struggle to distinguish temporal differences between different actions and object transformations. For example, the best model OpenAI o1 only obtains ∼50% on our text and video scores, showing a large gap compared to the human baseline of ∼90%. All open-source multimodal models and CLIP-based models perform much worse, producing mostly random chance performance. Through this work, we shed light onto the fact that temporal reasoning in short videos is a problem yet to be fully solved. We will publicly share our benchmark.

1 Introduction

Large multimodal models (LMMs) have become very competitive in not only image comprehension but also short video comprehension. Proprietary models such as GPT-4o [1] and Gemini-1.5-Pro [2] as well as open-source models like LLaVA-OneVision [3] and Qwen2-VL [4] demonstrate strong performance in summarizing a short video's contents and answering questions regarding its details.

*Equal Contribution † Work done at UW-Madison

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Recent SoTA reasoning models such as OpenAI o1 [5] and Gemini-2.0-Flash-Thinking-Mode [6] show powerful multimodal reasoning capabilities over images and videos alike. This has led many researchers to believe that short video comprehension has mostly been solved, and consequently, the community's focus has been increasingly trending toward creating models that understand longer-form videos that are 10s of seconds or even minutes long. Our study, however, indicates that existing models are far from being capable of fully understanding short videos that are just a few seconds long, especially when there is dense temporal information.

As demonstrated in [7] and [8], for many existing video benchmarks like EgoSchema [8], ActivityNet-QA [9], MSVD and MSRVTT [10], the performance of most modern LMMs does not vary significantly with number of sampled frames. In fact, it is often the case that an LMM only needs to see a single frame to produce a correct response. This 'single-frame bias' [11] reduces the video comprehension problem into the much easier image comprehension problem, essentially discarding the temporal aspect of a video. Researchers have also proposed harder temporal counterfactual benchmarks [12, 13, 14] in order to better evaluate an LMM's temporal understanding capabilities. Existing counterfactual datasets test a model's ability to distinguish slight changes from a video's original (positive) caption to the new (negative) caption by asking the model to match the video with the correct caption. However, they either do not contain any negative videos corresponding to the negative caption, or simply swap the order of two unrelated videos to form the positive and negative videos, making it easy to distinguish the negative pair from the original positive pair due to the videos' unnaturalness. Hence, these benchmarks may be inflating the performances of modern LMMs in understanding short videos.

In this paper, we introduce Vinoground, a temporal counterfactual LMM evaluation benchmark composed of 1000 short and natural video-caption pairs. Vinoground is a challenging benchmark aimed to expose the incapabilities of state-of-the-art models in understanding temporal differences between different actions (e.g., "the man eats then watches TV" vs. "the man watches TV then eats") and object transformations (e.g., "water turning into ice" vs. "ice turning into water"). In each pair of captions, the positive and negative are the same in word composition but different in order. Our work is inspired by Winoground [15], a challenging counterfactual benchmark for visio-linguistic compositional reasoning in images. In Winoground, a model must correctly match two images with their corresponding captions, where both captions use the same set of words, but are rearranged to describe each image (e.g., "some plants surrounding a lightbulb" vs. "a lightbulb surrounding some plants"). This evaluates whether a model effectively encodes the text and images, paying attention to their compositional structures, and whether it can integrate and synthesize information across both modalities. Our benchmark's name changes the 'W' to a 'V' for "video", and further employs temporal counterfactuals to emphasize this unique element in video data. We use text score, video score, and group score to evaluate a model's ability to choose the right caption for a video, to choose the right video for a caption, and to match both positive and negative video-caption pairs correctly, respectively. These measure a model's textual, visual, and temporal reasoning capabilities in a balanced manner. Most of our videos are less than 10 seconds long, yet we find a very large performance gap between an average human and today's best models. An example can be found in Figure 1. We purposely focus on short videos as they efficiently expose deficiencies in temporal reasoning without the cost of long video curation and evaluation. Additionally, they prevent failures from being misattributed to limited context windows to process long videos rather than poor temporal understanding. If Video LLMs cannot handle short videos, tackling long ones is futile.

In sum, our main findings and contributions are:

- Existing temporal counterfactual benchmarks fail to fully expose the incapability of LMMs in temporal reasoning.
- We introduce Vinoground, the first temporal and natural counterfactual evaluation benchmark for evaluating video understanding models using only short videos.
- Modern SoTA LMM performance is subpar when it comes to temporal reasoning in short video comprehension tasks; most models perform at random-chance level on video score and even worse on group score, both being significantly lower than text score.
- We categorize our data into 3 major categories, 'object', 'action', and 'viewpoint', as well as 4 minor categories, 'interaction', 'cyclical', 'spatial', and 'contextual', in order to dissect each model's capabilities for each of these categories. We find that existing models are decent at analyzing video frames at coarse-level but tend to miss fine-grained details.
- Short video comprehension is a problem that is far from being solved.

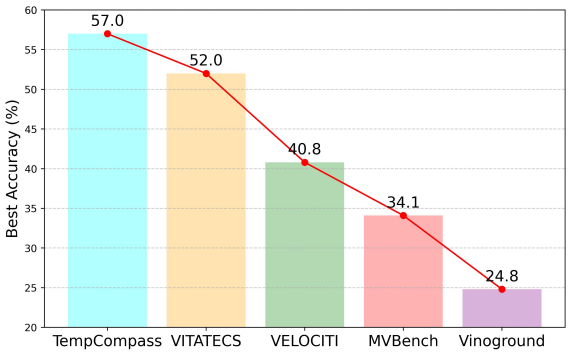

Figure 2: The best performances of Video-LLaVA/LLaMA-7B (the only models reported by all 5 benchmarks) on Vinoground and other temporal datasets. We use text score from Vinoground (as defined in Section 4.1), which matches TempCompass's Event Order category with Caption Matching format and VELOCITI's Mt2v metric. We report the average performances for VITATECS and MVBench. We can see that Vinoground is the most challenging benchmark.

2 Related Work

Counterfactual Reasoning. Counterfactual reasoning [16] in the context of computer vision typically involves curating negative images and captions by manipulating the original data and observing how the outcome changes [17, 18, 19, 20, 21, 22, 15, 23, 24]. The idea is that a model should understand cause and effect and be able to make predictions in unseen situations. For evaluation, curating meaningful and hard negatives is important. Winoground [15] is a pioneering benchmark for counterfactual reasoning where each data point contains two images and two corresponding captions. Given an image, a vision-language model is asked to find the matching caption from the provided two options, and vice versa. COCO-Counterfactual [23] explores simple linguistic rules to generate negative captions and uses an image editing model to produce negative images. We introduce a novel benchmark with counterfactuals that are temporal, an attribute specific to the video modality.

Single-Frame Bias and Temporal Reasoning. An important aspect of video data is its temporality, i.e., how events change as time progresses. Modern LMMs sample frames and treat the video as a set of images, both during training and evaluation. Benchmarks such as EgoSchema [8], MSVD and MSRVTT [10] exhibit a 'single-frame bias' [11] where only one video frame is needed for a model to predict correctly, as a model's performance does not vary significantly as the number of frames sampled increases [7, 8]. To better evaluate a model's temporal understanding capabilities, researchers have developed datasets such as YouCook2 [25], ActivityNet-QA [9] and COIN [26], which mainly involve procedural activities that often have a specific temporal dependency (e.g., if a video shows a person washing and slicing apples, and then baking an apple pie, a model would easily predict that "bake it to make a pie before washing the apple" is a wrong caption even without looking at the video). In contrast, Vinoground also includes actions that are entirely unrelated, such as "people are talking before drinking" vs "people are drinking before talking", making it more challenging for models to infer answers based solely on textual cues. MVBench [27] also includes temporal data that involves 20 different subcategories of temporal reasoning. However, even with this coverage, this benchmark does not contain any negatives like ours, reducing their difficulty since they do not contain any counterfactual examples. On top of not having any negative videos, NExT-QA [28] includes temporally rich questions but often mixes event inference with temporal reasoning. In contrast, Vinoground isolates pure temporal reasoning by presenting events e.g., A and B explicitly and asking about their order—removing confounding factors like causality or inference.

Temporal Counterfactuals. Recent benchmarks combine counterfactuals with temporal reasoning. EgoSchema [8] introduces long-form videos where each video has 1 positive caption and 4 negative captions to choose from, while VITATECS [12] introduces temporal counterfactual data where a word or phrase is swapped/replaced from the positive caption to form the negative caption. However, neither has any negative videos and thus do not fully evaluate an LMM's dense temporal reasoning capabilities like we do. VELOCITI [13] introduces positive/negative videos as a part of their intra-video association benchmark by clipping random portions in the same video, and asking the model to distinguish between the events. These videos, however, are not truly counterfactual pairs as different clips within the same movie are not guaranteed to have a positive-negative relation. TempCompass [14] includes videos that tests a model's ability to differentiate the order of events, but the videos are either concatenations of two completely unrelated videos with drastic frame changes in between the events, or reversed in time and thus impossible to happen in real life, and do not belong to the true data distribution. LMMs tend to do much better when it comes to such videos when compared to our benchmark's more natural negative videos, as shown in Fig. 2.

Similar to TempCompass, Paxion [29] uses reversed videos and caption edits (e.g., word swaps), which are often synthetically unnatural and detectable by models. Also, Paxion's perturbations are limited to captions, whereas Vinoground includes true negative videos, further increasing task difficulty. We summarize the comparisons between other temporal benchmarks and Vinoground in Table 1, demonstrating how Vinoground is the only benchmark unifying the four qualities, making it the most novel temporal reasoning benchmark.

Dataset	Negative Videos	Counter-factual	Short (Avg $\leq$10s)	Natural Videos
Paxion	✓	✓	✗	✗
NExT-QA	✗	✓	✗	✓
MVBench	✗	✗	✗	✓
EgoSchema	✗	✓	✗	✓
VITATECS	✗	✓	✗	✓
VELOCITI	✓	✗	✓	✓
TempCompass	✓	✓	✗	✗
Vinoground (Ours)	✓	✓	✓	✓

Table 1: Comparison between Vinoground and other temporal datasets. Ours is the only one possessing natural negative videos that are counterfactual and mostly less than 10 seconds long.

3 Vinoground

In this section, we introduce our data curation and categorization process. In order to curate Vinoground's video-caption pairs, we first explain how we generate the required captions in Section 3.1, how we find the corresponding videos in Section 3.2, and finally the details of categorizing the videos in Section 3.3. An illustration of the overall process can be found in Appendix A.

3.1 Generating Counterfactual Captions

The first step in curating our data is to find counterfactual caption pairs. We want to ensure that the captions we curate are of high-quality and temporal in nature. While human annotation is a possible solution, it is costly and difficult to scale up. Instead, we leverage a SoTA LLM, specifically the GPT-4 [30] model, as it is much cheaper, follows the multiple requirements we impose, and guarantees that there are no duplicate candidates. We require our caption pairs to be composed of the exact same words, only permuted into different orders. We also want to avoid candidates that could easily be solved by looking at a single frame of the video such as "a man is waving at a woman" vs. "a woman is waving at a man". Hence, we ask GPT-4 to create *temporal* counterfactuals that require one to process and understand the entire video, and in particular, understand the order of events in which they happen, such as "a man waves at a woman before he talks to her" vs. "a man talks to a woman before he waves at her". We will later showcase in Section 4.3 that we can already expose LMMs greatly with such videos (i.e., by swapping the order of two events), making more complicated scenarios unnecessary. We include the detailed prompt fed to GPT-4 in Appendix E.

3.2 Video Curation

After curating counterfactual caption candidates, we next find corresponding videos for those captions. We make use of the VATEX [31] dataset, which contains 5 distinct captions for each maximum 10-second long video. We only use the validation and test subsets of VATEX to make sure none of Vinoground is ever used as training data. This results in a pool of 9000 videos and 45000 captions.

We retrieve potential matches in VATEX according to the generated caption candidates. We leverage sentence transformers [32], which are good at summarizing sentence-level information into feature vectors, to extract the features of both our GPT-generated captions and VATEX's captions. We subsequently use the Faiss library [33] to efficiently index and retrieve the top 20 most similar VATEX captions for each GPT-4 generated caption. We manually examine if any retrieved caption is a good match, and if its corresponding video reflects the caption as well. The primary criterion during manual review is straightforward: Does the caption accurately and unambiguously describe the video content? While this process does involve some degree of semantic judgment—as is inevitable in aligning language and vision—we mitigate subjectivity by (1) cross-validating questionable cases, and (2) filtering out ambiguous matches. We also ensure that only caption/video pairs where multiple authors independently agree are retained. The quality of the dataset yielded under this process can be justified by our human performance (Table 3). For some cases where none of the retrieved captions are a good match, we search YouTube with the caption candidate to find a matching video.

In the end, we curate 500 counterfactual pairs of video-caption pairs (1000 video-caption pairs in total) for evaluation. Each video-caption pair is provided in the form of the original YouTube ID,

the clip's starting and ending timestamps, and the corresponding caption. We also put Vinoground through 3 rounds of human evaluation by the authors, making sure that the pair of captions truly contain the same word composition and that the video clips indeed reflect their respective captions.

3.3 Categorization

Finally, we want to be able to evaluate LMMs in a fine-grained manner on multiple aspects represented by our dataset. Hence, we categorize Vinoground according to the unique characteristics discovered through the data curation process. We report the number of counterfactual data pairs assigned under each category in Table 2. We define each category as follows.

Major	Object	Action	Viewpoint	Total
Count	160	257	83	500
Minor	Interaction	Cyclical	Spatial	Contextual
Count	73	111	103	63

Table 2: The number of data points in Vinoground assigned under each category, separated by major and minor groups. All 500 pairs have one and only one major category assigned to them, while minor category assignments are content-based.

We divide Vinoground into 3 major categories: *object*, *action*, and *viewpoint*. Each counterfactual pair must be in one and only one of the three major categories.

- **Object** requires LMMs to detect changes in the status of one specific object, such as "water turning into ice" vs. "ice turning into water." This category is similar to the "Reversing" category in TempCompass [14] that evaluates a model's ability to detect attribute and directional changes. While TempCompass reverses positive videos in time to create negatives and thus can be unnatural, we curate real, natural videos that correspond to the negative captions.
- **Action**, on the other hand, simply asks models to distinguish the order in which two or more different actions happened, e.g. "the man eats and then watches TV" vs. "the man watches TV and then eats." The two actions need not be correlated at all, and thus less logical comprehension is necessary for a correct prediction.
- **Viewpoint** specifically describes changes in the camera angle, perspective, or focus within the video, such as "a person films the car in front of him before he films himself" vs. "a person films himself before he films the car in front of him." The change in viewpoint is usually accompanied by a drastic difference in between the frames, whereas other events most likely happen within the same context or background.

We also introduce 4 minor categories: *interaction*, *cyclical*, *spatial*, and *contextual*. Some pairs belong to a multitude of these minor categories, while some do not belong to any.

- **Interaction** involves videos where a human changes their way of interacting with an object in the course of the video, e.g. "the calligrapher writes with his pen before he dips it into the ink" vs. "the calligrapher dips his pen into the ink before he writes with it."
- **Cyclical** tests a model's ability to identify either procedural temporal activities or two actions that are dependent on each other. The calligrapher example earlier is also cyclical as the person repeats the procedure "write, dip, write, dip...", and the action "dip" happens as a result of "write" in the positive, while "write" is enabled after "dip" in the negative. In contrast, the "action" category can involve completely unrelated actions.
- **Spatial** It has been shown that LMMs struggle to distinguish physical locations between objects in image-caption pairs [24]. We want to further evaluate this deficiency when it comes to temporal understanding as well. Thus, this category involves object movements and requires positional understanding, such as "the man ran from left to right" vs. "the man ran from right to left." Note that this does not include movement of the background; e.g., when the camera is moving along with the object in question, which belongs to the next category.
- **Contextual** requires LMMs to understand changes in the background or general information of entire video frames. An example is the pair "the biker rides down the street before he goes down the stairs" vs. "the biker goes down the stairs before he rides down the street" where the camera that records the videos is strapped on the biker's forehead, making the background the only changing aspect. One cannot infer positional changes only by observing object movements like the "spatial" category, but instead must focus on the background as the object in question can appear motionless due to the camera moving along with the object.

We provide in-depth analysis of models' performances on our benchmark based on the above categories in Section 4.4.2. A detailed teaser can be found in Appendix M.

4 Experiments

In this section, we evaluate state-of-the-art vision-language models on our benchmark. We first describe the models and evaluation metrics in Section 4.1; then we explain our experimental setup, including prompting methods and human studies, in Section 4.2; we analyze the performances of the models in Section 4.3, and provide further ablation studies in Section 4.4.

4.1 Models and Evaluation Metrics

We evaluate both CLIP-based [34] and large generative models, both proprietary and open-source. The exact list of models we evaluate can be found in Table 3. CLIP-based models use contrastive learning between videos and captions, while text-generation LMM models use next-token prediction to generate a response. Due to the different nature of the CLIP-based vs. LMM methods, we introduce our metrics in different fashions accordingly.

We use C to denote captions and V to denote videos. For each positive and negative set of counterfactual video-caption pairs, (C_i, V_i) and (C_i', V_i'), $\forall i \in \{1, 2, ..., 500\}$, we ask CLIP-based models to compute a similarity score e between not only the correct pairs but also the incorrect pairs (C_i, V_i') and (C_i', V_i) (identical to Winoground [15]). For generative LMMs, we can only provide inputs (e.g., 2 captions and 1 video) to the model and ask it to choose between the captions/videos.

We first evaluate the text score s_t where the model is presented with both positive and negative captions but only one of the videos, forming the triplets (C_i, C_i', V_i) and (C_i, C_i', V_i'). For each triplet, the model is then asked to choose the caption that describes the contained video. We denote the score function of a model response given any triplet as s; for instance,

$$s(C_i, C_i', V_i) = \begin{cases} 1 & \text{if LMM chooses } C_i \text{ or} \\ & e_{(C_i,V_i)} > e_{(C_i',V_i)} \text{ for CLIP-based} \\ 0 & \text{otherwise} \end{cases} \qquad s(C_i, C_i', V_i') = \begin{cases} 1 & \text{if LMM chooses } C_i' \text{ or} \\ & e_{(C_i',V_i')} > e_{(C_i,V_i')} \text{ for CLIP-based} \\ 0 & \text{otherwise} \end{cases}$$

Then the text score for the given counterfactual pair (C_i, V_i) and (C_i', V_i') is:

$$s_t(C_i, C_i', V_i, V_i') = s(C_i, C_i', V_i) \wedge s(C_i, C_i', V_i')$$

where $\wedge$ is the logical and operator; i.e., s_t is 1 only if both triplets are correct. This exposes the models when they guess randomly.

Similarly, for video score s_v, the model is presented with one caption and both positive and negative videos, forming triplets (C_i, V_i, V_i') and (C_i', V_i, V_i'). For each triplet, the model is asked to choose the video that is described by the caption. In this case, the response scoring becomes:

$$s(C_i, V_i, V_i') = \begin{cases} 1 & \text{if LMM chooses } V_i \text{ or} \\ & e_{(C_i,V_i)} > e_{(C_i,V_i')} \text{ for CLIP-based} \\ 0 & \text{otherwise} \end{cases} \qquad s(C_i', V_i, V_i') = \begin{cases} 1 & \text{if LMM chooses } V_i' \text{ or} \\ & e_{(C_i',V_i')} > e_{(C_i,V_i')} \text{ for CLIP-based} \\ 0 & \text{otherwise} \end{cases}$$

Then the video score is:

$$s_v(C_i, C_i', V_i, V_i') = s(C_i, V_i, V_i') \wedge s(C_i', V_i, V_i')$$

We also include a group score metric s_g:

$$s_g(C_i, C_i', V_i, V_i') = s_t(C_i, C_i', V_i, V_i') \wedge s_v(C_i, C_i', V_i, V_i')$$

s_g serves as the ultimate test for a model to demonstrate its temporal reasoning capabilities in both the textual and visual domains, as both s_t and s_v must be 1. For all three metrics, we report the mean over all test instances. We include an illustration of the metrics in Appendix B.

4.2 Experimental Setup

Since for each pair of counterfactuals, we have 2 text-score questions and 2 video-score questions, we have 2000 questions in total. To evaluate CLIP-based models, we use the evaluation code provided by the authors to calculate video-caption embeddings and similarity scores. Evaluating text-generative models is slightly more complicated. We first introduce the different prompts we use. For text score, we provide the model with the video and the two corresponding captions, and prompt "⟨video⟩ Which caption best describes this video? A. {Caption 1}, B. {Caption 2}". For video score, however, since some LMMs only support 1 video input, we concatenate the positive and negative videos into a single video with a 2 second black screen in between.

When sampling N frames for the model's input, we make sure we sample $(N-1)/2$ frames from the positive and negative video fragments and at least 1 frame of black screen in between. More details can be seen in Appendix L. For the sake of consistency, we provide all models with the single concatenated video, regardless of how many videos they can actually take as input. We then prompt the model with "⟨video⟩ Which video segment matches this caption? Note: The video contains two segments separated by a 2-second black frame. Caption: {Caption}. A. First segment (before black frame), B. Second segment (after black frame)" to choose between the two video segments. For text score, we shuffle the caption orders so that both answer choices "A" and "B" have 50% probability as ground truths. For video score, we concatenate videos in random orders while also making sure both answer choices appear evenly. We also report the results with respect to the number of frames sampled by the model from the video, if supported, to evaluate the effect of temporality in Section 4.4.1. All experiments are done with 4xA100-80GB GPUs.

In addition, we also use Prolific (`https://www.prolific.com`) to evaluate human performance, and find that our dataset is fairly easy for an average human to complete with high accuracy. Prolific is a platform similar to Amazon MTurk which recruits workers to complete tasks such as data annotation. The interface we present to the workers is in Appendix H. To filter out unfaithful workers, we employ a qualification process prior to evaluating on Vinoground. We sample 10 video-question pairs from TempCompass [14] that are of the event order category, which contains concatenated videos with no correlation, such as "a man lifts weights in a gym, then a cat plays on the grass". Such examples are easy enough for an average human to obtain 100% accuracy. We ask the workers the 10 beginner-level questions first, and they are qualified only if they answer every question correctly. This process results in 170 qualified workers, whose demographics are also included in Appendix H.

We conduct human evaluation under two settings. First, the Prolific workers are provided the full videos with audio. We want to create another environment where the workers see the same input as the models. Hence, we uniformly sample 32 frames from each video and concatenate them into a new 10-second video with no audio. The results for the two settings are also compared in Section 4.4.1. Each question is answered by 10 unique workers. For the 10 answers from a single question, we calculate the *average* human response by taking the mode of the 10 answers. We then report the mean over all the questions as the final result.

Model	# Fr	s_t	s_v	s_g
Random Chance	N/A	25.0	25.0	16.7
Prolific Human	All	***93.4***	***94.0***	***90.0***
	32	***91.4***	***90.8***	***85.2***
Proprietary Large Multimodal Models				
OpenAI o1 [5]	32	**59.1**	**50.5**	**36.0**
GPT-4o (CoT) [35]	32	**59.2**	**51.0**	**35.0**
GPT-4o [1]	32	**54.0**	**38.2**	**24.6**
GPT-4o	0	10.0	24.6	2.0
Gemini-2.0-Thinking [6]	32	**39.0**	**31.2**	14.6
Gemini-1.5-Pro (CoT)	1fps	**37.0**	27.6	12.4
Gemini-1.5-Pro [2]	1fps	**35.8**	22.6	10.2
Claude 3.5 Sonnet (CoT)	4	**39.4**	27.0	13.6
Claude 3.5 Sonnet [36]	4	**32.8**	28.8	10.6
Open-Source Large Multimodal Models				
Qwen2-VL-72B (CoT)	32	**53.0**	26.6	15.2
Qwen2-VL-72B [4]	32	**50.4**	**32.6**	17.4
Qwen2-VL-7B [4]	4fps	**40.2**	**32.4**	15.2
LLaVA-Video-72B [37]	64	**49.2**	**34.0**	**20.2**
LLaVA-Video-7B [37]	64	**42.4**	**30.0**	17.0
LLaVA-OneVision-72B [3]	32	**48.4**	**35.2**	**21.8**
LLaVA-OneVision-7B [3]	16	**41.6**	**29.4**	14.6
VideoLLaMA3 [38]	16	**47.4**	**30.4**	15.6
Apollo-7B [39]	4	**43.8**	**30.2**	17.2
VideoLLaMA2-72B [40]	8	**36.2**	21.6	8.4
InternVideo2.5-8B [41]	32	**35.0**	**29.0**	11.4
MiniCPM-2.6 [42]	16	**32.6**	**29.2**	11.2
Aria [43]	32	**34.8**	28.8	12.0
InternLM-XC-2.5 (CoT)	1fps	**30.8**	28.4	9.0
InternLM-XC-2.5 [44]	1fps	28.8	27.8	9.6
Video-LLaVA-7B [45]	8	24.8	25.8	6.6
Phi-3.5-Vision [46]	16	24.0	22.4	6.2
MA-LMM-Vicuna-7B [47]	4	23.8	25.6	6.8
LLaVA-NeXT-34B (CoT)	32	25.8	22.2	5.2
LLaVA-NeXT-34B [48]	32	23.0	21.2	3.8
LLaVA-NeXT-7B (CoT)	32	21.8	26.2	6.8
LLaVA-NeXT-7B [48]	32	21.8	25.6	6.2
M^3 [49]	6	21.2	25.8	6.8
VTimeLLM [50]	100	19.4	27.0	5.2
CLIP-based Models				
VideoCLIP [51]	60	17.0	2.8	1.2
LanguageBind [52]	8	10.6	5.0	1.2
ImageBind [53]	20	9.4	3.4	0.6

Table 3: Vinoground results for different models and number of sampled frames. Performances significantly better than random chance are bolded. There are four groups separated by double lines: random chance and human performance, proprietary text-generative models, open-source text-generative models, and CLIP-based models from top to bottom. The best performances of proprietary and open-source models are highlighted in red.

4.3 Main Results

Table 3 presents the results. (Appendix K presents more detailed results, as we only include each model's best performances here.) First, all CLIP-based models (VideoCLIP, LanguageBind, ImageBind) perform much worse than random chance, suggesting that contrastive learning does not provide models with enough knowledge of temporality. Among text-generative models, OpenAI o1

performs best, achieving $\sim 59\%$ on the text score metric. Chain-of-Thought (CoT) prompting [35] further improves the performance of models including GPT-4o, especially on the video score metric where it improves by 12.8% and 10.4% on group score, matching that of OpenAI o1. We include the full CoT prompt and parsing process in Appendix F. We do not use CoT on reasoning models because there always is a hidden reasoning process even without explicitly asking them to do so. Amongst the open-source models, LLaVA-Video, LLaVA-OneVision and Qwen2-VL demonstrate competitive performance compared to proprietary models, especially with Qwen2-VL-72B's 50.4% performance on text score. Using CoT on open-source models, however, helps much less, especially if they perform at near chance level. Current reasoning models also show limited improvements compared to base models in terms of temporal reasoning. All other models perform at or worse than random chance, showing that dense temporal reasoning is still very challenging for LMMs. We further provide failure case analysis for some of GPT-4o's responses in Appendix I.

Similar to Winoground [15], we find that for models that perform better than chance level, their text score is significantly higher than video score, while group score is the lowest amongst all three. A clear pattern can also be seen in Fig. 1. This shows that they are better at identifying textual differences compared to visual/temporal differences. For example, GPT-4o's video score (38.2%) is significantly lower compared to its text score (54.0%). Many open-source models only have non-random outcomes on the text score but equal or lower than random chance on video and group scores. Notably, LLaVA-OneVision-72B and LLaVA-Video-72B are the only open-source models that perform better than chance group score.

Human evaluators perform significantly better than any model, with scores around 90%. This indicates that Vinoground can be tackled relatively easily within human capacity. When the human evaluators are provided with 32-frame videos, the scores decrease by a few points, but are still much higher than those of any model.

Finally, we also report performance for GPT-4o with 0 frames sampled as a control to test for text bias. For text score, we hypothesize that the model will choose the more likely caption since it cannot see the video, and for the video score, we hypothesize it will choose an answer at random, which is indeed what happens. The lower than chance performance for text score of 10.0% indicates that there is some language bias in GPT4o, where it prefers to select one caption over the other (if it consistently did that for all questions, the text score would be 0). Thus, our balanced way of computing the scores (i.e., both $s(C_i, C_i', V_i)$ and $s(C_i, C_i', V_i')$) prevents a model from doing well only via its language bias. This is in contrast to existing benchmarks like VITATECS [12] and EgoSchema [8] which lack negative videos, and hence enable models to potentially answer a question correctly only based on which caption is more likely.

All in all, even the very best models exhibit subpar performance when it comes to dense temporal reasoning, and this is only using short videos (less than 10 seconds) as well. This strongly indicates that short video comprehension in LMMs is still far from human-level intelligence. We provide further insights on model design and data utilization strategies in Appendix N.

4.4 In-Depth Analysis of Performance Variations

4.4.1 Frames Sampled

We demonstrate Vinoground's temporal understanding requirements by varying the number of frames sampled, either from the video entirely, or as measured by frames-per-second (fps). If a dataset suffers from 'single-frame bias', a model would not perform very differently when only 1 or more frames are sampled. The results of the strongest proprietary and open-source models in Table 4 (additional results in Appendix K) show that the more frames a model takes, the better its performance. This indicates that a model does need the entirety of each video to fully comprehend the task at hand. Interestingly, too many sampled frames can hurt a model's performance; for GPT-4o, its 64-frame variant performs 5% worse on all three metrics compared to its 32-frame variant. We suspect that current models are not good at discarding redundant information and isolating signal from noise when given too many visual tokens. Appendix G further explains the novelty of this finding.

Note that for our video score metric to function as intended, a model must sample at least one frame from each video, and at least one black frame in between. This means that the number of frames sampled must be no fewer than 3. We hence gray out the video score and group score performances of models sampled at 1 or 2 frames and only focus on their text scores.

Finally, for human evaluators, the 'All' group performs better than the 32 frame group, which indicates that humans can answer Vinoground questions better when the full videos are shown. In contrast, modern LMMs generally lack the ability to process inputs of an entire video without coarse sampling of frames. This suggests that further research into creating models that can handle more frames will be an important research direction for temporal reasoning.

Model	# Fr	s_t	s_v	s_g
Prolific Human	All	***93.4***	***94.0***	***90.0***
	32	***91.4***	***90.8***	***85.2***
GPT-4o	64	**49.0**	**34.8**	19.0
	32	**54.0**	**38.2**	**24.6**
	8	**53.6**	**31.4**	**20.6**
	1	28.2	28.0	10.0
LLaVA-OneVision-72B	64	**46.2**	**31.8**	18.6
	32	**48.4**	**35.2**	**21.8**
	16	**47.2**	**33.8**	**20.4**
	8	**46.8**	**29.8**	19.0
	4	**40.4**	24.8	13.0
	2	**33.4**	25.2	10.2

Table 4: Results of the strongest closed- and open-source models with different frames sampled. Performances significantly higher than random chance are highlighted, while the best overall performance of each model are highlighted in red. More frames do lead to better performance, but too many frames can worsen the results.

4.4.2 Category

Fig. 3 shows results per category as defined in Section 3.3. Interestingly, many models perform significantly better on the *viewpoint* and *contextual* categories, while being significantly worse on other categories. Here, we only report the group score for a selected set of models due to space. See Appendix J for the full results.

Both *viewpoint* and *contextual* bring forth drastic changes in between the video frames whenever the events change, as *contextual* involves background changes that occupy most of the frame while in *viewpoint*, as the camera angle changes, the entirety of the video frame changes as well. On the other hand, *interaction* and *cyclical* not only require a model to have strong logical understanding of the connection between events, but also the ability to focus on small temporal changes for the different actions involved. *Spatial*, as previously hypothesized, also poses a difficult challenge for models in understanding changes in object location. Overall, today's models are much better at understanding coarse-level information over a set of frames in their entirety than understanding fine-grained details from a part of each video frame. This also demonstrates how fine-grained comprehension is also crucial for dense temporal reasoning.

5 Conclusion

We introduced Vinoground, a novel temporal counterfactual benchmark encompassing 1000 short and natural video-caption pairs. We demonstrated that existing video LMMs are quite incapable in terms of temporal reasoning, even for short (<10s) videos. While an average human can easily and accurately complete our benchmark, the best model, OpenAI o1, performs much worse, and most models barely perform better than random chance. Our work demonstrates that there is much more to do still in the area of short video comprehension. We believe Vinoground can serve as an important checkpoint in evaluating a model's true performance for temporal understanding of short videos.

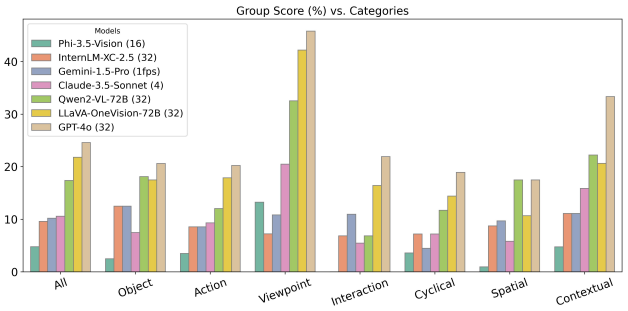

Figure 3: Group score of selected models per category. Models do better on contextual & viewpoint, worse on others.

Limitations. One cannot fully analyze the behavior of proprietary models included in this paper due to the lack of access, namely OpenAI o1, GPT-4o, the Gemini series and Claude 3.5 Sonnet.

Broader Impacts. Vinoground thoroughly stress tests the temporal reasoning capabilities of large multimodal models. Having shown that modern SoTA LMMs lack such elementary skills only with using short videos, we demonstrate great impact to the research community by prompting further work for researchers to use our benchmark as a true standard for evaluating temporal understanding and short video comprehension, and lastly to improve future LMMs based on our metrics.

References

[1] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

[2] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[3] 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: 2408.03326*, 2024.

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

[5] OpenAI. Openai o1 system card. https://cdn.openai.com/o1-system-card-20241205.pdf, 2024.

[6] Google. Gemini 2.0 flash thinking experimental. https://deepmind.google/technologies/gemini/flash-thinking/, 2025.

[7] Wenhao Wu. Freeva: Offline mllm as training-free video assistant. *arXiv preprint arXiv:2405.07798*, 2024.

[8] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in neural information processing systems (NeurIPS)*, 2023.

[9] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, pages 9127–9134, 2019.

[10] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.

[11] 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, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] 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. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2024.

[13] Darshana Saravanan, Darshan Singh, Varun Gupta, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. Velociti: Can video-language models bind semantic concepts through time? *arXiv preprint: 2406.10889*, 2024.

[14] 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, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 8731–8772, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[15] 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 (CVPR)*, pages 5238–5248, June 2022.

[16] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[17] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[18] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity for explanations. *Advances in neural information processing systems (NeurIPS)*, 2019.

[19] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR, 09–15 Jun 2019.

[20] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *Advances in neural information processing systems (NeurIPS)*, 2020.

[21] Hangzhi Guo, Thanh Hong Nguyen, and Amulya Yadav. Counternet: End-to-end training of prediction aware counterfactual explanations. In *Proceedings of the 29th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2023.

[22] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

[23] Tiep Le, Vasudev Lal, and Phillip Howard. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *Advances in neural information processing systems (NeurIPS)*, 2023.

[24] Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. CounterCurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 15481–15495, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[25] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[26] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13853–13863, June 2022.

[27] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, June 2024.

[28] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021.

[29] 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, 2023.

[30] OpenAI. Gpt-4 technical report. *arXiv preprint: 2303.08774*, 2024.

[31] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[32] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems (NeurIPS)*, 2020.

[33] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint: 2401.08281*, 2024.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems (NeurIPS)*, 35:24824–24837, 2022.

[36] Anthropic. Introducing claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024.

[37] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024.

[38] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025.

[39] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. Apollo: An exploration of video understanding in large multimodal models, 2024.

[40] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[41] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025.

[42] 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.

[43] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, Chongyan Zhu, Xiaoyi Ren, Chao Li, Yifan Ye, Peng Liu, Lihuan Zhang, Hanshu Yan, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model, 2025.

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

[45] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[46] Microsoft. Discover the new multi-lingual, high-quality phi-3.5 slms, 2024.

[47] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[48] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[49] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*, 2024.

[50] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14271–14280, June 2024.

[51] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2021. Association for Computational Linguistics.

[52] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[53] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, June 2023.

[54] Raymond R. Panko. Thinking is bad: Implications of human error research for spreadsheet research and practice, 2008.

[55] Yifan Du, Yuqi Huo, Kun Zhou, Zijia Zhao, Haoyu Lu, Han Huang, Wayne Xin Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. Exploring the design space of visual context representation in video mllms, 2024.

Appendix

A Data Curation Process

We include an overall illustration of the data curation process in Figure 4.

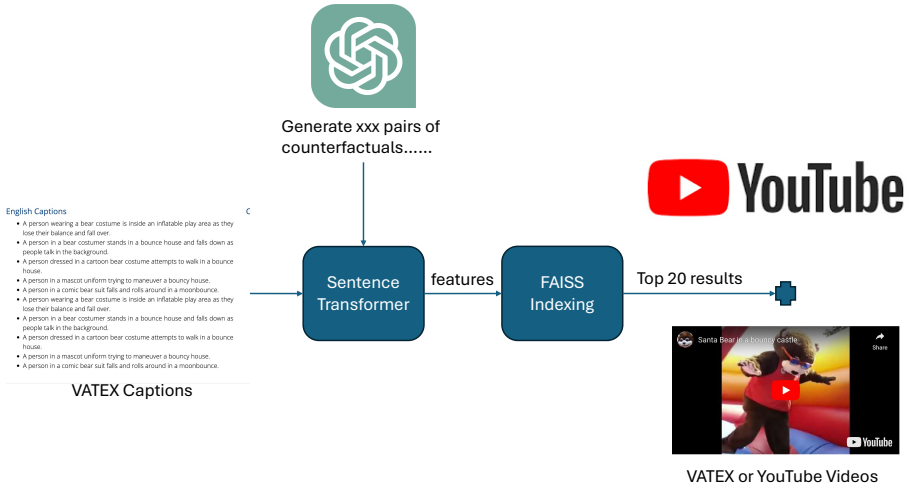

Figure 4: The data curation process.

B Metrics Illustration

We visualize our text and video score metrics in Figure 5. This shows the 4 possible questions that can be derived from one counterfactual data point in the dataset.

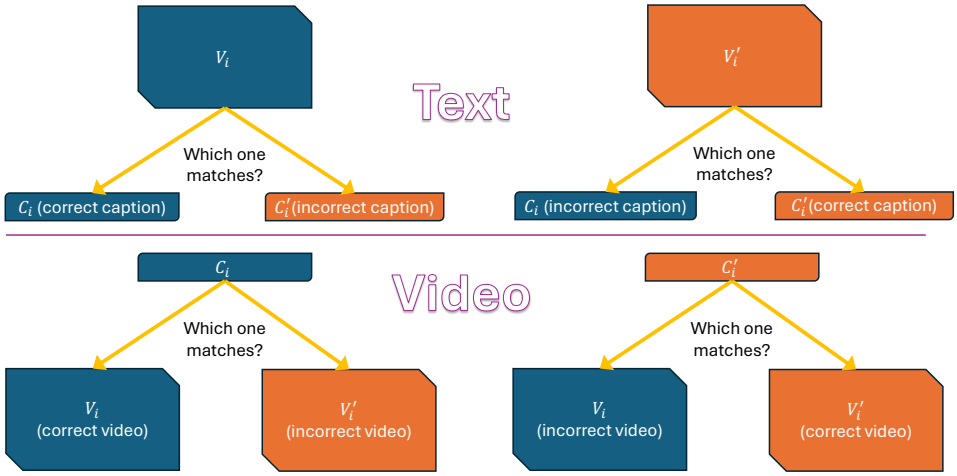

Figure 5: Visualization of the text and video score metrics.

C Random Chance Performance

We set the random chance performance for text, video, and group score as 25%, 25%, and 16.67%. It is intuitive to understand the setup for both text and video score since there are two questions in the same counterfactual pair for each metric, and the probability of guessing correctly is 50% each. For the counterfactual pair (C_i, C_i', V_i, V_i'), a model can only produce six possible permutations of video-caption matchings: $\{(C_i, V_i), (C_i', V_i')\}$, $\{(C_i, V_i), (C_i, V_i')\}$, $\{(C_i, V_i), (C_i', V_i)\}$, $\{(C_i, V_i'), (C_i', V_i')\}$, $\{(C_i', V_i), (C_i', V_i')\}$, and $\{(C_i', V_i), (C_i, V_i')\}$. This is why the random chance performance for group score is $1/6 \approx 16.7\%$.

D Human Performance

The 90% group score is because of human error, not because of data quality issues. Upon carefully examining the error cases, we find no particular pattern or poor quality examples. On the other hand, [54] shows how humans have a 5% error rate even for the simple task of entering spreadsheets, which closely models the human text and video scores 93% and 94%. Since group score is a composite metric of both, the combined correctness is $0.95 \cdot 0.95 = 0.9025$, which matches our human group score performance. This confirms the high-quality of Vinoground.

E Caption Curation Prompt

The prompt we give GPT-4 to generate potential caption candidates is: "I am trying to find videos that have appropriate temporal counterfactuals; e.g., I want to find video pairs that can be described with the following captions: "a man eats then watches TV" vs "a man watches TV then eats"; "the old man is working hard before the young man is playing" vs "the young man is working hard before the old man is playing". Note that for both elements of the same pair, they use the exact same words. Give me 10 examples." Then in the same conversation, we prompt the model "give me 10 different ones" until we have 500 pairs of candidates.

F CoT Prompt and Parsing

For chain-of-thought prompting, we simply add "please think step by step" at the end of our questions (as mentioned in Section 4.2). We then use GPT-4 as the judge with the prompt: "Please parse the following model response into either A or B. If the model response is just A or B, then it denotes the model answer, just output it. The model response starts after ====, and end before ====):\n==== ⟨MODEL RESPONSE⟩ ====\nProvide output your answer as a single character (A or B): "

G Comparison with Prior and Concurrent Analyses on Impact of Number of Frames Sampled

In [11], the authors only demonstrated how they can use a model trained with one frame of the video and perform better than all SoTA methods trained on many frames. They did not conduct a comprehensive analysis like we do. In Section 4.4.1, we demonstrate how each model has an optimal number of frames sampled based on their model structure and size that affects their performance. It is not only too many frames but also too few frames that can cause performance issues.

In [55]'s Tables 3 and 4, which show the performance of a model with different number of frames sampled, the results only illustrate how an increasing number of up to 128 frames (with a fixed number of visual embeddings per frame) mostly improves performance, while in our case, using 64 frames and sometimes 32 frames can show significantly degraded performance. Since [55] had to change both the # of frames sampled and # of visual embeddings per frame to demonstrate the performance tradeoff/degradation, our findings are more direct since our Section 4.4.1 only changes 1 factor (# of frames) instead of both.

In conclusion, our work provides more significant analysis beyond the single-frame bias shown in [11] and the performance comparison in [55].

H Prolific Survey Interface and Worker Demographics

We first upload all the videos to Google Drive and embed them into our surveys using Qualtrics. The 2000 questions from Vinoground are split into 50 surveys, with each survey having 40 random questions. We then distribute our surveys on Prolific where we pay everyone who completed a survey \$2, or \$0.05 per question. The interface is illustrated in Figure 7.

Out of the 170 qualified workers, 91 were male and 79 were female; 13 were students, 119 were non-students, and 38 have no data/did not consent to provide information; 69 were full-time working, 22 were part-time working, 20 were not in paid work (e.g. homemaker, retired or disabled), 17 were unemployed, and 42 have no data/did not consent to provide information; all of the 170 workers were fluent in English; we report the histogram of workers' ages in Figure 8.

I Analysis of Selected Failure Cases of GPT-4o

Here we analyze 3 example failure cases of GPT-4o on our benchmark. One can find the corresponding video clips from the supplementary materials or the HuggingFace link. Extracted frames for video 8_neg are shown in Figure 6.

Figure 6: For this video, GPT-4o summarizes the video's event in the exact opposite order of the video itself, indicating its lack of understanding temporal dynamics between the frames.

For video 8_neg, the model response is: "a box with a lid placed on top of it => The hand interacts with the lid => The hand lifts the lid off the box => The hand moves the lid to the side of the box => the lid placed to the side of the box. Based on this sequence, the man moved the lid from the top of the box to the side". In this response, the model summarized the video's event in the exact opposite order of the video itself, indicating its lack of understanding temporal dynamics between the frames.

For video 248_pos, the model's response is: "the man is seen using the machine from left to right => the man continues moving the machine towards the right => the man is seen moving the machine back towards the left. Based on this analysis, the man first uses the machine from left to right and then goes from right to left." In this analysis, the model also reported the order of events in their exact opposite, but note the fact that here the model specifically failed to understand spatial movements as well.

For video 361_neg, the model response is: "two scallops on a grill, with the one on the left being more prominently visible => the camera appears to shift focus slightly, but the scallop on the left remains the primary focus => the camera seems to zoom in or shift slightly, but it does not distinctly pan from one scallop to the other. Since the video does not show a clear pan from one scallop to the other, neither caption A nor B describes the video." In this response, the model ignores huge frame shifts that is easily identifiable by the human eye. This shows how models can fail even with coarse-grain large frame changes.

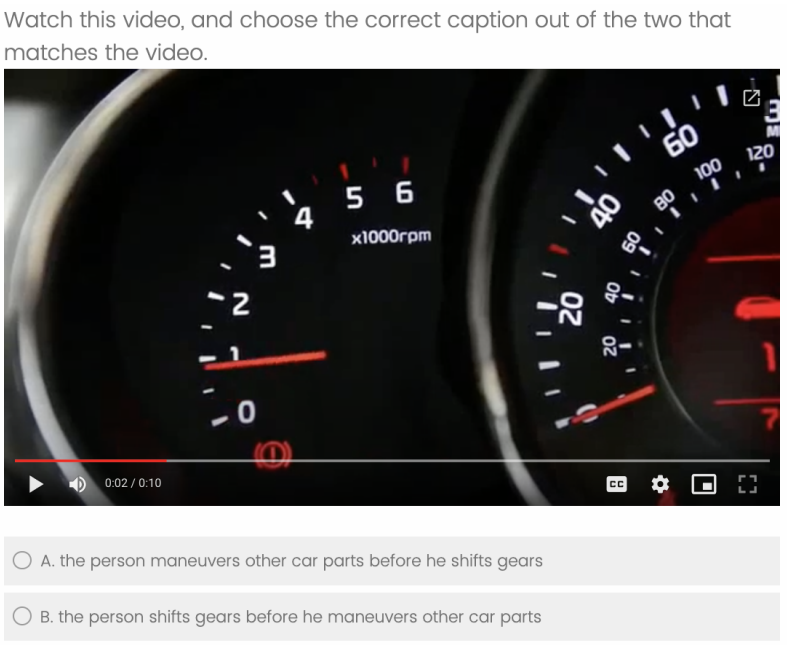

Figure 7: The Qualtrics survey that Prolific workers see.

Figure 8: Distribution of Prolific workers' ages.

J Full Categorical Results

Here we include the selected top-6 strongest models we evaluated and report their results by category in Tables 5 and 6. We also include the text and video score bar plots in Figures 9 and 10. We can see that the general trend is the same as reported in Section 4.4.2, where models perform much better on contextual and viewpoint, and worse on other categories.

	GPT-4o			Gemini-1.5-Pro			Claude 3.5 Sonnet		
category	text	video	group	text	video	group	text	video	group
all	54.00	38.20	24.60	35.80	22.60	10.20	32.80	28.80	10.60
object	52.50	35.62	20.62	36.25	25.62	12.50	30.00	25.00	7.50
action	47.47	35.41	20.23	30.74	22.18	8.56	27.63	28.79	9.34
viewpoint	77.11	51.81	45.78	50.60	18.07	10.84	54.22	36.14	20.48
interaction	50.68	42.47	21.92	30.14	27.40	10.96	20.55	21.92	5.48
cyclical	39.64	41.44	18.92	22.52	19.82	4.50	27.03	25.23	7.21
spatial	47.57	30.10	17.48	37.86	24.27	9.71	31.07	20.39	5.83
contextual	53.97	49.21	33.33	38.10	31.75	11.11	52.38	28.57	15.87

Table 5: The best performances of proprietary models grouped by category. Significantly high performances are highlighted in blue, while significantly low performances are highlighted in red.

	LLaVA-OneVision-72B			Qwen2-VL-72B			InternLM-XC-2.5		
category	text	video	group	text	video	group	text	video	group
all	48.40	35.20	21.80	50.40	32.60	17.40	28.80	27.80	9.60
object	42.50	33.75	17.50	46.88	33.75	18.12	28.75	28.12	12.50
action	42.80	31.91	17.90	44.75	28.79	12.06	25.68	29.96	8.56
viewpoint	77.11	48.19	42.17	74.70	42.17	32.53	38.55	20.48	7.23
interaction	36.99	36.99	16.44	34.25	31.51	6.85	23.29	36.99	6.85
cyclical	36.04	29.73	14.41	36.94	32.43	11.71	18.92	36.04	7.21
spatial	37.86	25.24	10.68	53.40	31.07	17.48	23.30	29.13	8.74
contextual	57.14	31.75	20.63	49.21	39.68	22.22	26.98	26.98	11.11

Table 6: The best performances of selected open-source models grouped by category. Significantly high performances are highlighted in blue, while significantly low performances are highlighted in red.

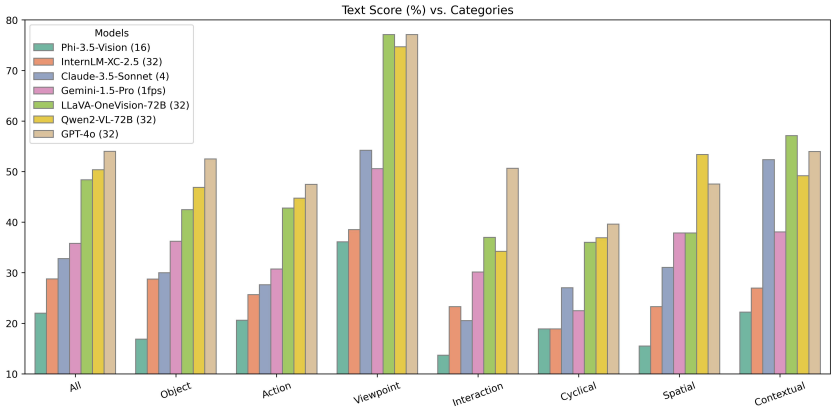

Figure 9: Text score bar plot based on category grouped by model.

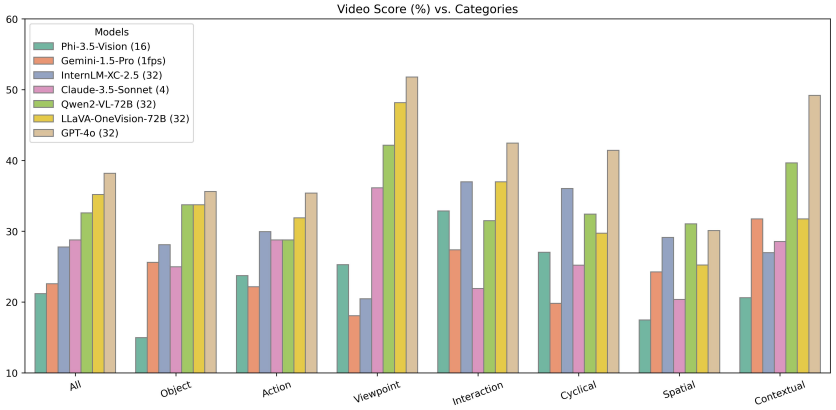

Figure 10: Video score bar plot based on category grouped by model.

K Full Results on Evaluated Models

Due to the extensive number of models evaluated and different number of frames sampled as hyperparameters, we include the full results of our evaluation that are not mentioned in the main paper in Tables 7 and 8.

Model	Frames	Text	Video	Group
Claude-3.5-Sonnet	16	**30.0**	22.6	8.4
	8	**32.2**	25.4	9.4
	4	**32.8**	28.8	10.6
	2	29.4	24.0	8.4
	1	26.2	30.0	10.8
Qwen2-VL-72B	32	**50.4**	**32.6**	17.4
	8	**37.4**	23.0	7.8
	4	26.2	23.8	6.2
	2	15.6	24.4	4.0
Qwen2-VL-7B	32	**40.0**	26.4	11.8
	16	**36.8**	25.8	10.2
	8	27.6	23.4	7.8
	4	22.2	22.8	5.6
	2	21.4	25.8	5.2
	4fps	**40.2**	**32.4**	15.2
	2fps	**34.8**	27.4	10.6
	1fps	26.8	26.6	7.6
	0.5fps	23.2	19.6	4.8
MiniCPM-2.6	32	28.4	27.0	9.4
	16	**32.6**	**29.2**	11.2
	8	**33.4**	25.6	9.0
	4	25.8	27.4	8.6
	2	22.8	23.2	4.6
	1	27.0	27.0	8.0
LLaVA-NeXT-Video-34B	32	23.0	21.2	3.8
	16	21.0	21.8	4.4
	8	21.2	22.0	5.2
	4	16.6	21.6	3.4
	2	15.4	21.6	2.2
	1	13.2	21.8	2.0
LLaVA-NeXT-Video-7B	32	21.8	25.6	6.2
	16	22.2	25.6	6.4
	8	21.8	25.6	6.4
	4	21.8	25.6	6.4
	2	21.2	25.4	6.0
	1	22.4	25.6	6.4
Phi-3.5-Vision	32	22.0	21.2	4.8
	16	24.0	22.4	6.2
	8	21.8	21.2	5.0
	4	21.2	22.8	5.6
	2	20.4	21.6	3.8
	1	22.6	22.8	3.8
MA-LMM-Vicuna-7B	32	22.4	25.6	6.8
	16	22.0	26.0	6.0
	8	23.0	26.0	6.4
	4	23.8	25.6	6.8
	2	23.8	25.6	6.8

Table 7: The full evaluation results based on model type, frames sampled, and the metrics aforementioned. Only the model settings that are not mentioned in the main paper are listed here. Performances significantly better than random chance are bolded.

Model	Frames	Text	Video	Group
VideoLLaMA3	64	**46.2**	**29.8**	17.0
	32	**47.4**	**29.2**	15.0
	16	**47.4**	**30.4**	15.6
	8	**43.4**	28.6	12.0
	4	**38.8**	24.6	8.8
	2	**35.6**	22.8	7.4
	1	22.8	22.4	6.2
LLaVA-Video-72B-Qwen2	128	**45.2**	28.8	16.4
	64	**49.2**	**34.0**	**20.2**
	32	**48.4**	**33.2**	**20.0**
	8	**44.0**	27.6	16.0
	4	**37.2**	23.0	10.2
	2	**31.4**	23.6	9.4
	1	25.2	26.8	8.0
LLaVA-Video-7B-Qwen2	128	**41.4**	27.6	14.0
	64	**42.4**	**30.0**	17.0
	32	**40.8**	**30.4**	15.4
	16	**36.8**	28.0	13.0
	8	**33.6**	25.6	11.4
	4	29.0	24.6	10.0
	2	27.0	23.6	6.2
	1	27.8	22.4	6.4
Aria	32	**34.8**	28.8	12.0
	16	**32.4**	27.6	9.4
InternVideo2.5-8B	64	**36.0**	28.2	11.0
	32	**35.0**	**29.0**	11.4
	16	**30.6**	25.6	8.6
	8	23.4	25.0	6.0
	4	17.4	25.2	3.6
Apollo-7B	64	**41.5**	**31.5**	17.5
	32	**44.3**	28.5	15.8
	16	**43.6**	28.8	16.6
	8	**42.6**	28.4	14.2
	4	**43.8**	**30.2**	17.2

Table 8: Continuation of Table 7.

L Video Lengths and The Use of Black Frames

We report the video length distribution of our benchmark in Figure 11. We also report that out of the 1000 videos in Vinoground, there are a total of 992 videos with length ≤ 20 seconds, and 930 of them are ≤ 10 seconds.

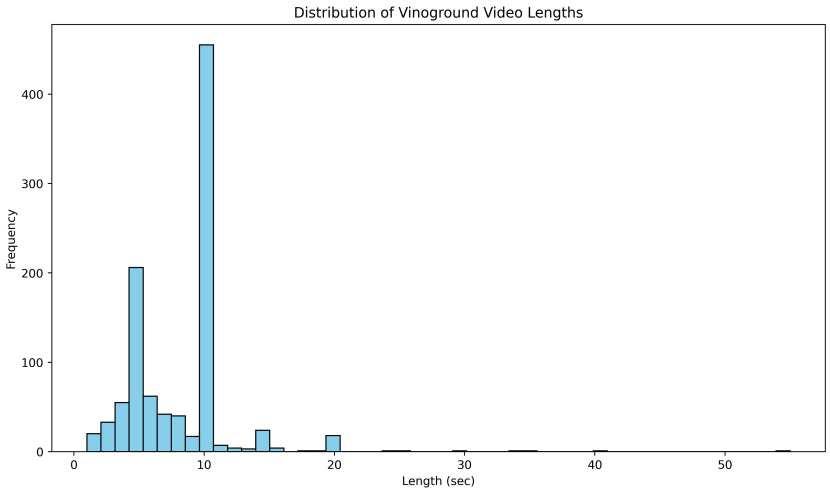

Figure 11: Video length distribution of Vinoground.

We show another histogram regarding—in all 500 concatenated videos for the video score metric—how much of each video is composed of black frames in Figure 12. We can see that for the majority, black frames only consist of less than two-tenths of the videos. This ensures that data loss due to sampling black frames is kept at a minimum.

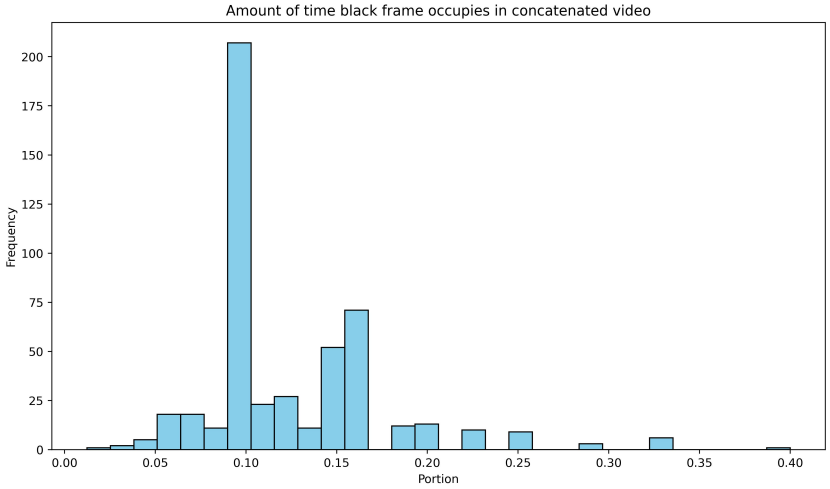

Figure 12: The portion of black frames in each concatenated video for video score questions.

M Detailed Categorical Teaser

Figure 13: Examples of Vinoground video-caption pairs under each category.

N Insights into Model Design and Data Utilization Strategies

The goal of our paper is to introduce a highly challenging benchmark to expose existing video models' weakness in temporal understanding, while future work can use our benchmark to showcase true improvements. Hence, improving models is beyond the scope of our work. Nevertheless, we believe that the following findings can be valuable for future researchers:

In Table 3, we include CLIP-based models and observe that contrastive learning models perform much worse compared to SoTA text-generative LLMs. We hypothesize that two key factors contribute to the significant performance gap:

Feature Representation: CLIP-style models encode each modality with a single vector (e.g., 768–2048 tokens). This limited feature representation makes it difficult to capture the fine-grained temporal details essential for video understanding. In contrast, video LLMs like Video-LLaVA use thousands of visual tokens (e.g., 2048 tokens in our benchmark) that can represent more details containing temporal dynamics.

Model Scale: CLIP-based text encoders are smaller in scale (in terms of the number of parameters and the amount of pretraining data) compared to the large language models used in video LLMs. Video LLMs' richer pretraining datasets and larger capacities might have also made them better at understanding temporality.

Case Study: For example, both Video-LLaVA and LanguageBind use the same video encoder, yet Video-LLaVA outperforms LanguageBind greatly as demonstrated in Table 2. This suggests that the difference between encoder architectures and learning objectives/training paradigms significantly influence performance on tasks requiring temporal reasoning.

Regarding potential improvement suggestions:

Architectural Improvements: Methods such as incorporating hierarchical temporal modeling or cross-modal attention mechanisms tailored for temporal reasoning could improve performance. For example, attention layers that focus on sequential dependencies across frames may help capture causality more effectively.

Data Utilization Strategies: Current datasets often suffer from single-frame bias and fail to emphasize temporal consistency. Including more counterfactual training data designed to model temporality (as in Vinoground) could mitigate these biases.

Learning Objectives: Transitioning from contrastive learning to visually conditioned next-word prediction, as seen in Video-LLaVA, could enhance temporal understanding. Fine-tuning LLMs on datasets emphasizing temporality is another promising direction.

O Temporal Localization

Vinoground inherently requires models to exhibit temporal localization abilities to answer questions correctly as our dataset highly focuses on the understanding of temporal ordering. For example, for the data pair "the cat moves before the person touches it" vs. "the person touches the cat before it moves", the model is implicitly required to localize the temporal events "the cat moves" and "the person touches", or simply to understand in which part of the video did the cat move; otherwise it cannot determine if an event happened before the other. Thus, even though we evaluate the models in the format of multiple-choice, without the ability to localize temporal events, these models cannot perform well on our benchmark.

Yet again, we emphasize how easy to humans our benchmark is as reflected by the human baseline. No matter what form of task Vinoground can take, be it multiple choice or localization, models still perform much worse than humans, demonstrating the critical lack of temporal reasoning in modern video LLMs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our results match claims made in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 contains our limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include detailed methodologies and experimental setups in Sections 3 and 4 and in Appendices A, B, E, F, and H.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our dataset in the supplementary materials and through the HuggingFace link.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The entire Section 4 explains this in great detail.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We only report model performance on benchmarks which conventionally does not necessitate the use of error bars or statistical significance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.2 reports the details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We uphold the Code of Ethics in every way we can.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5 contains our broader impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our data is obtained either from the VATEX dataset or from YouTube directly. We carefully inspected each piece of data before including them in our dataset, as described in Section 3.2.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The videos in Vinoground either come from VATEX or YouTube, whose original YouTube IDs we also provide along with the dataset files we submit. We properly cited these works as well as all the models we evaluated upon.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Section 3 well-documents the dataset we introduce.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Section 4.2 and Appendix H discusses about our human experiments using Prolific, how our surveys looks like, the demographics, etc.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve any risks involved with study participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Section 3.1 and Appendix E explains how we use GPT-4 to curate caption candidates, while Appendix F describes how we use GPT-4 as judge during evaluation.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.