

# SVLTA: Benchmarking Vision-Language Temporal Alignment via Synthetic Video Situation

Hao Du<sup>1\*</sup>

Bo Wu<sup>2\*</sup>

Yan Lu<sup>3</sup>

Zhendong Mao<sup>1†</sup>

<sup>1</sup> University of Science and Technology of China

<sup>2</sup> MIT-IBM Watson AI Lab

<sup>3</sup> The Chinese University of Hong Kong

<https://svlta-ai.github.io/SVLTA>

dh97@mail.ustc.edu.cn, bo.wu@ibm.com, yanlu@cuhk.edu.hk, zdmao@ustc.edu.cn

## Abstract

*Vision-language temporal alignment is a crucial capability for human dynamic recognition and cognition in real-world scenarios. While existing research focuses on capturing vision-language relevance, it faces limitations due to biased temporal distributions, imprecise annotations, and insufficient compositionally. To achieve fair evaluation and comprehensive exploration, our objective is to investigate and evaluate the ability of models to achieve alignment from a temporal perspective, specifically focusing on their capacity to synchronize visual scenarios with linguistic context in a temporally coherent manner. As a preliminary step, we present the statistical analysis of existing benchmarks and reveal the existing challenges from a decomposed perspective. To this end, we introduce SVLTA, the Synthetic Vision-Language Temporal Alignment derived via a well-designed and feasible control generation method within a simulation environment. The approach considers common-sense knowledge, manipulable action, and constrained filtering, which generates reasonable, diverse, and balanced data distributions for diagnostic evaluations. Our experiments reveal diagnostic insights through the evaluations in temporal question answering, distributional shift sensitivity, and temporal alignment adaptation.*

## 1. Introduction

Multimodal Large Language Models (MLLMs) [1, 35, 39] are pioneering a new direction beyond LLMs [10, 67, 68]. They have demonstrated remarkable advancements in mainstream evaluations including vision-language comprehension [23, 42, 76, 87], analysis [1, 33, 81], alignment [9, 58, 65], and even reasoning [69, 73]. However, current

assessments primarily focus on the models’ performances of semantic-aligned vision-language inferences, often neglecting if models perform well on the capacities in real-world situations that evolve over time. This temporal perspective is crucial for effective human cognition and adaptation to the surrounding environment [3, 4]. In this work, we investigate and assess the ability of models to achieve vision-language temporal alignment, specifically how they can synchronize visual events with linguistic context in a temporal manner. This remains a significant challenge for the comprehensive evaluation of multimodal models.

To address similar goals, one category of existing work builds datasets for traditional video grounding models on top of collected videos and temporal segment annotations, such as TACoS, DiDeMo, and Charades-STA [2, 14, 28, 59, 63]. But human-crafted annotations solely on semantic correlations inevitably result in unreliable labels [51, 86], due to the inherent subjectivity and ambiguity of linguistic and visual semantic descriptions. Additionally, while constructing data combinations with semantic correlation as the goal allows for consideration of semantic diversity, it has been found to suffer from significant temporal distribution imbalances [51, 80]. As multimodal large model capabilities advance, these shortcomings in evaluations become critical bottlenecks affecting the accuracy and objectivity of evaluations. We conduct a comprehensive examination of the current video benchmarks in depth. We identify three types of temporal alignments (processes, compositions, and entities) from a decomposed perspective, propose the metric Temporal Jenson-Shannon Divergence (TJSD), and visualize the temporal distributions for clearer understanding. The observed biased distributions at multiple levels reflect the “unbalanced influences” of the existing evaluations.

Distinct from previous work, we propose SVLTA, exploring the synthetic vision-language temporal alignment, enabling compositional, unbiased, and large-scale video

---

\*Both authors contributed equally.

†Corresponding author.

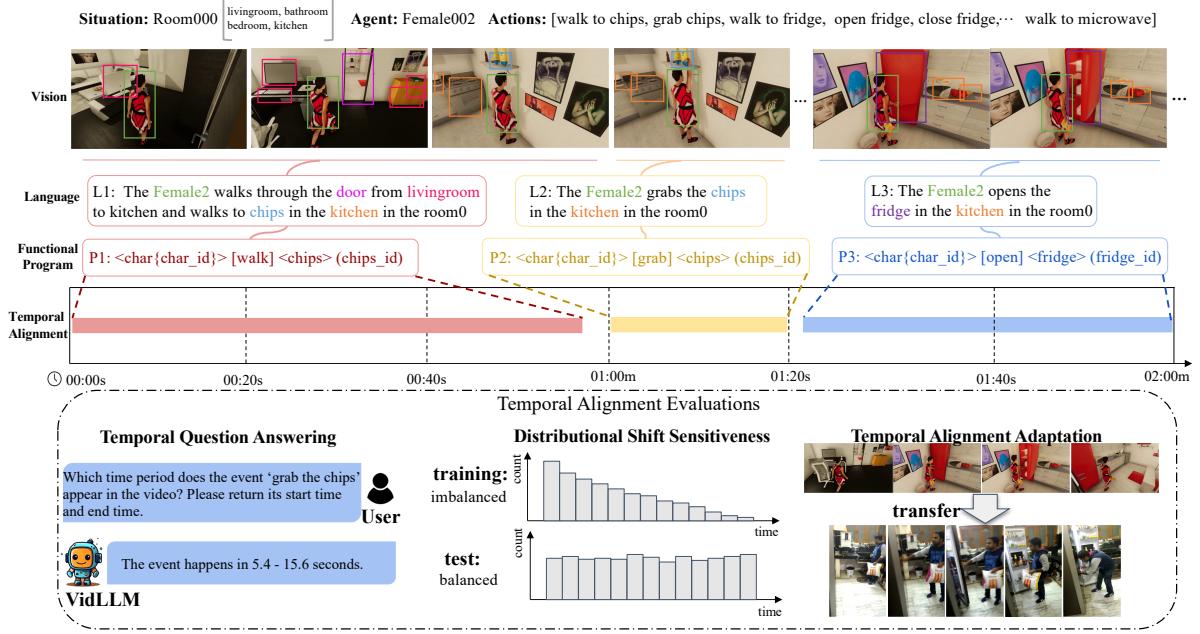


Figure 1. Overview of the SVLTA benchmark, which consists of synthetic videos, language, and high-quality temporal alignment.

evaluations with precise temporal alignments. SVLTA encompasses 96 different compositional actions, 25.3K synthetic video situations, and 77.1K high-quality temporal annotations with consistent visual-language semantics. Our benchmark addresses the limitations of existing datasets by generating videos through synthetic simulations, providing better control over temporal alignment, which is challenging in realistic videos. The benchmark overview is illustrated in Figure 1. Synthetic video situations are created by executing functional programs derived from a series of raw data (predefined agents, actions, and situations) in a human-centric 3D VirtualHome simulator [38, 55, 56], while sentences are generated using templates defined for different scenes. For timestamp annotations, the programs automatically record the time and duration of each action during execution. This method allows us to easily associate the annotated timestamps with the corresponding actions in the sentences, resulting in high-quality annotations.

Through SVLTA, we evaluate temporal alignment from three perspectives: 1) temporal question answering, 2) distributional shift sensitiveness, and 3) temporal alignment adaptation. Our experiments yield the following results: 1) by using simple temporal-related questions, current popular Video Large Language Models (VidLLMs) rarely give correct answers, even some time-aware VidLLMs or close-sourced models, which means current models lack temporal alignment capability, 2) most specific temporal alignment models are easily affected by temporal bias, even some debiased models, which indicates these models have poor generalization and cannot address temporal distribution shifts, and 3) several specific temporal alignment models may

have potential to transfer temporal knowledge, which means these models can adapt to new situations in some degree. Our main contributions are as follows:

- We conduct three levels of temporal distribution analysis and visualization from a decomposed perspective and propose an appropriate measurement.
- We introduce SVLTA, a synthetic benchmark for vision-language temporal alignment. Through our proposed approaches, we generate both human activity situations, language descriptions, and temporally aligned samples with fewer human resources based on commonsense inference, compositional optimization, and constrained sampling.
- We design multiple types of vision-language temporal alignment tasks enabling comprehensive evaluations for both pre-trained VidLLMs and specific temporal alignment models, providing detailed experimental analysis with insightful conclusions.

## 2. Related Works

### 2.1. Visual-Language Temporal Alignment

Visual-language temporal alignment aims to link video content and language in the temporal dimension, which also appears in action localization or temporal grounding tasks. Most of the benchmarks in temporal grounding or localization [2, 14, 16, 31, 59, 64] collected videos from web or recorders and utilized the crowd-sources to annotate temporal segments (e.g. action timestamps, event order, etc). Additionally, some of them [19, 63] used Automatic Speech Recognition (ASR) to generate text transcripts from speech and constructed their related timestamps which require less

costs and resources. These benchmarks played an important role in evaluating the development of temporal alignment systems [6, 22, 32, 37, 40, 78, 79, 82], which pushes the boundaries of temporal alignment research. According to our analysis (refer to § 3.2), current video benchmarks may be influenced by multiple levels of bias as they are primarily sourced from the real world and rely on human-provided annotations. Our conclusion also consists of the observations of several recent work [51, 72, 80], and they noticed serious temporal biases of sentence-level video segments in video datasets, which can cause the model to have poor generalization when training on these datasets [29, 41, 49]. Moreover, Otani et al. [51] found that the temporal annotations from multiple annotators are inconsistent since they have different perceptions that lead to unreliable annotations. These drawbacks limit the accuracy and effectiveness of the assessment, which demonstrates current benchmarks cannot provide a detailed and valid diagnosis environment for temporal alignment models.

## 2.2. Synthetic Situation Generation

Synthetic data generation is gaining significant interest within the research community because of its cost-effectiveness and ease of scale, with applicability across various research fields. Synthetic videos were widely adopted in data augmentation for video understanding or action recognition [5, 18, 24, 27, 57, 61, 70]. Meanwhile, some research [15, 21, 77] utilized the templates or physics engine to generate associated questions for diagnostic evaluations on model capacities in neighbor tasks, such as video question answering [77], action recognition [15, 27], and multi-object tracking [13]. However, the aforementioned methods are not designed to study vision-language temporal alignment and ignore the explicit control of temporal alignment as the primary generation objective. Unlike the others, our SVLTA generates human activity situations, language descriptions, and temporally aligned samples for multiple types of program-generated evaluation tasks based on commonsense knowledge and compositional optimization.

## 2.3. Temporal Understanding

With the emergence of Video Large Language Models, there has been a significant increase in the collection of video-language benchmarks, to evaluate them. Most previous work mainly considered semantic diversity and video length duration when collecting the data, while the video has a temporal dimension that depicts the objects' motion and the corresponding interaction state. To fill this blank and comprehensively evaluate the video-language models, some works aim to explore the temporal understanding ability of these multimodal video models. AGQA [17] utilized templates combined with the spatio-temporal scene graph to generate well-designed questions and answers to assess the

temporal reasoning ability. Furthermore, ViLMA [26] and Perception Test [52] developed a series of temporal-related tasks (for example, action location and action counting) to diagnose whether the model has strong temporal modeling capabilities. TempCompass [44] contributed a comprehensive benchmark to evaluate temporal perception using the single frame, language debiasing strategies, and various task formats. E.T.Bench [45] also designed multiple fine-grained temporal sensitive tasks to assess event-level video understanding from open-ended scenarios. However, none of the above benchmarks focus on the assessment of temporal alignment with fairness and comprehensiveness, which is also an important part of temporal understanding.

## 3. SVLTA

SVLTA is a synthetic and scalable benchmark with diverse, compositional, and controllable temporal distribution, to provide a fair diagnostic framework for evaluating the temporal alignment ability of models. However, constructing such a benchmark is challenging, especially in controlling the temporal distribution, as it requires detailed types of temporal distribution and carefully designed methods to maintain the balance of the data set.

In this work, we first formulate the visual-language temporal alignment problem (§ 3.1), then thoroughly analyze the temporal distribution in current mainstream datasets and design a metric to measure them (§ 3.2). Following this, common sense activity and action chain generation, controllable temporal distribution strategies, and synthetic generation are adopted to build a benchmark that contains three processes: 1) synthetic video generation (§ 3.3.1), 2) language sentence generation (§ 3.3.2), and 3) visual-language temporal alignment (§ 3.3.3), a post-processing filtering method is also developed to further adjust the temporal distribution (§ 3.3.4). Finally, we compare our SVLTA with other major benchmarks (§ 3.3.5).

As summarized in Table 1, SVLTA comprises 25.3K dynamic situations derived from human activity videos, featuring 77.1K language descriptions and temporal-aligned activity sequences, covering 96 distinct compositional actions. The benchmark provides 77.1K high-quality temporal alignment annotations, with average video and moment durations of 134.1 and 24.3 seconds, respectively. Up to this point, it is a novel benchmark with compositional, controllable, and unbiased temporal distributions.

### 3.1. Problem Formulation

We define visual-language temporal alignment as the task of synchronizing video and language in the temporal domain, aiming to identify the timestamps of video moments that most closely match the semantics of the corresponding sentences. In particular, we denote an untrimmed video as  $V = \{v_i\}_{i=1}^M$  and a sentence in the language as  $L =$

Table 1. Comparison of SVLTA and existing benchmarks for Vision-Language Temporal Alignment. SVLTA is a synthetic benchmark with controllable, compositional, and unbiased data distributions. N/A: not available.

Benchmark	Dataset Statistics			Dataset Characteristics				
	# Videos / # Annotations	# Actions	Avg. Video / Moment Duration (s)	Scalable	Controllable	Synthetic	Compositional	Unbiased
TACoS [59]	0.1K / 18.8K	60	287.1 / 27.9	✗	✗	✗	✗	✗
ActivityNet Captions [28]	14.9K / 54.9K	N/A	117.6 / 37.1	✗	✗	✗	✗	✗
Charades-STA [14]	6.7K / 16.1K	157	30.0 / 8.1	✗	✗	✗	✗	✗
DiDeMo [2]	10.5K / 40.5K	N/A	30.0 / 6.5	✗	✗	✗	✗	✗
TVR [31]	21.8K / 109K	N/A	76.1 / 9.1	✗	✗	✗	✗	✗
MAD [63]	0.7K / 384.6K	N/A	6646.2 / 4.1	✓	✗	✗	✗	✗
Ego4D [16]	1.6K / 19.2K	N/A	495.3 / 11.2	✗	✗	✗	✗	✗
Ego4D Goal-Step [64]	0.8K / 48K	N/A	1560.0 / 32.5	✗	✗	✗	✗	✗
E.T.Bench [45]	7K / 7.3K	N/A	129.0 / 11.0	✗	✗	✗	✗	✗
SVLTA (ours)	25.3K / 77.1K	96	134.1 / 24.3	✓	✓	✓	✓	✓

$\{l_j\}_{j=1}^N$ , where  $v_i$  is a frame in the video and  $l_j$  means a word in the sentence. The goal of the visual-language temporal alignment task is to build a model  $f$  with the input  $V$  and  $L$  that can correctly predict the start time  $t_s$  and the end time  $t_e$  of the moment, which are formulated as follows:

$$[t_s, t_e] = f(V, L; \theta) \quad (1)$$

where  $M$  and  $N$  are respective length of video and text, and  $\theta$  is the model parameter.

### 3.2. Temporal Distribution Analysis

**Temporal Distribution in Decomposition Perspective** We first explore multiple levels of temporal alignments in the existing video benchmarks and validate if they are appropriate to the vision-language temporal alignment evaluations. Inspired by several works [7, 34], we take a decomposition perspective and treat a situation as comprising actions, with each action containing a verb-noun structure. Thus, the semantic constituents and video segments are associated on multiple levels. As illustrated in Figure 2, the visualizations reveal that temporal distributions are influenced by several biases, ranging from global to local levels. The **process temporal bias (video-level)** indicates limitations in overall data selection, while the **composition temporal bias (action-level)** and **entity temporal bias (verb/object-level)** result in evaluations focusing only on particular positions of temporal segments in videos, neglecting others.

**Quantitative Comparison via Temporal Divergence** To effectively analyze the imbalance of temporal distributions within datasets, quantifiable metrics are necessary. Drawing inspiration from prior research [11, 74] that designed metrics to measure class imbalance in classification benchmarks, we propose a new metric, Temporal Jensen–Shannon Divergence (TJSD), to measure the differences between the target distribution and the ideal distribution. The target distribution means the temporal distribution of the current dataset and the ideal distribution denotes the uniform distribution. To address the problem that time is

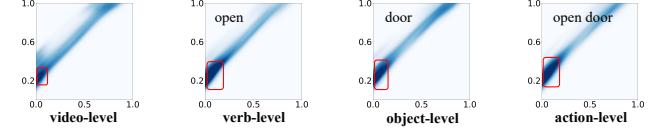


Figure 2. Multiple Levels of Temporal Distributions. We sample decomposed semantic constituents in the Charades-STA. The color darkness represents the sample density. The horizontal and vertical axes represent the normalized start and end time points.

continuous without natural categorization, we first divide the video into  $n$  equal moments to discrete time, leading to  $\frac{n(n+1)}{2}$  different temporal bins, each bin represents a temporal class, and then we assign the timestamps into these bins. Therefore, the target temporal distribution can be represented by the number of samples in these bins and the ideal distribution means that the number of samples in each bin is the same. Finally, the Jensen–Shannon divergence is utilized to calculate the difference between the target and uniform temporal distribution. The detailed TJSD equation is in *Supplementary*. The statistics of existing datasets are shown in Table 2, demonstrating that although some benchmarks have smaller process temporal bias, they ignore other types of temporal bias when collecting the videos.

### 3.3. Benchmark Generation

#### 3.3.1. Synthetic Situation Generation

**Situation Component Initialization** To create a temporal alignment dataset with high-quality and diverse temporal alignment samples, we generate the synthetic videos by defining compositional situation components and functional programs via a human-centric simulator Virtual-Home [38, 55, 56]. To begin with, we establish the diverse action, situation, and agent spaces for video generation, as shown in Figure 3 (a). We select 96 meaningful actions that can be executed within the simulator, 7 scenes that serve as the environments for these actions, and 6 alternative agents who act as characters to carry out the pre-defined actions. Executing functional programs will trigger an agent to per-

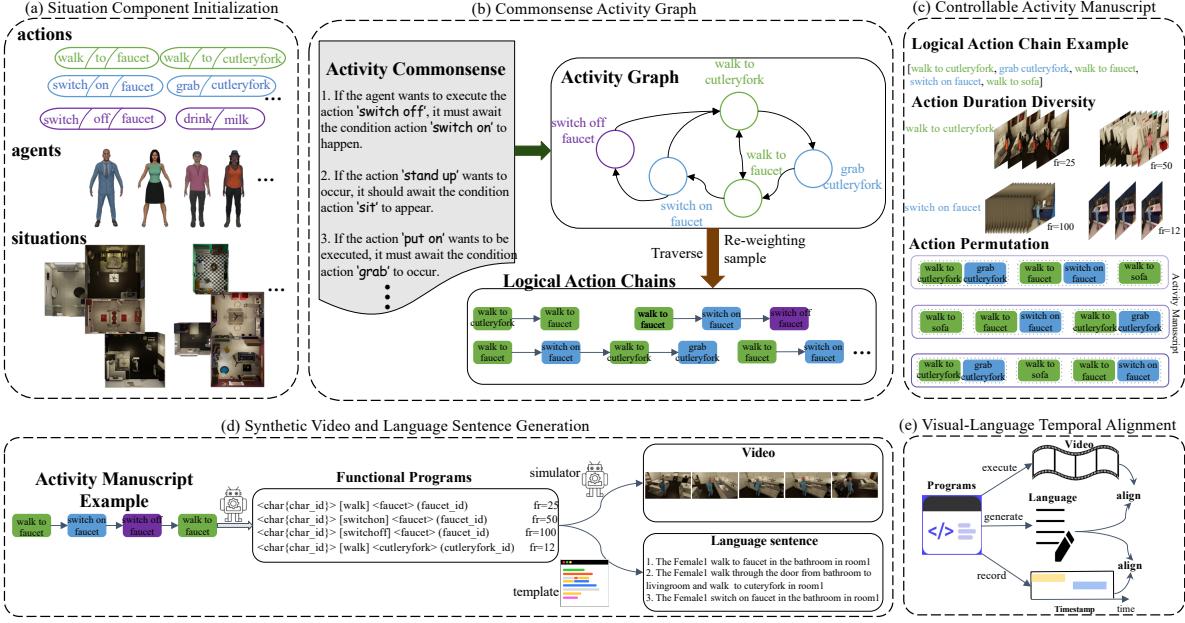


Figure 3. Overview of the benchmark generation process, which contains (a): Situation Component Initialization defines a series of compositional elements, which includes diverse actions, agents, and situations, (b): Commonsense Activity Graph builds a graph on the activity commonsense and then use the traversal algorithm and re-weighting sampling to acquire various and meaningful logical action chains, (c): Controllable Activity Manuscript operates the actions in logical action chains through different framers and permutations to obtain the final activity manuscript, thereby balancing the temporal distribution, (d): Synthetic Video and Language Sentence Generation convert the generated activity manuscript to the functional programs and utilize it to generate synthetic videos and sentences, and (e): Visual-Language Temporal Alignment automatically associates the timestamps with the action in the sentence to obtain high-quality annotations.

form the pre-defined activity in a virtual home scene.

**Commonsense Activity Graph** After initializing component spaces, a conventional strategy is to generate synthetic activity videos by selecting several actions. However, random selection would lead to meaningless combinations of actions, making it challenging to accomplish our objective. VirtualHome provides basic logical rules between paired actions, i.e. some actions must wait until the conditional actions are completed before they can happen. But longer action sequences would encounter unreasonable compositions accidentally. For instance, if four actions *walk to fridge*, *open fridge*, *close fridge*, and *grab sandwich* are selected, only the first three actions can be performed correctly and the last one is meaningless since *grab sandwich* and *walk to fridge* are inconsistent. Therefore, we manually check the defined rules and only keep the reasonable and executable action relations and agents in situations, which consistent with human commonsense knowledge in the real world. Based on this, we can get all potential relationships between these actions, and an activity graph is built upon to generate diverse and reasonable action compositions. Several recent works adopt novel ideas to enhance the generation quality [25, 71, 73]. Our activity graph inspired from them and specialized in following aspects: 1) our activity graph is built on commonsense knowledge with pre-defined action sets while others design the graph based on real-world

videos, 2) our graph aims to generate new synthetic videos yet others are used to produce novel questions, answers, and annotations. Then, we adopt the graph traversal algorithm (DFS [12] or BFS [12]) to generate logical action chains by traveling action nodes in activity graphs with given lengths. However there are different levels of constraint between the actions in the activity commonsense, i.e., some actions have fewer conditional actions and they do not need to wait for other actions to happen, thus the node degree is imbalanced in the activity graph. To solve this problem, a re-weighting sampling strategy is proposed to ensure that all candidate actions have a uniform probability of being selected in each traversal. The complete process is illustrated in Figure 3 (b) and the strategy details are in *Supplementary*.

**Controllable Activity Manuscript** Directly utilizing logical action chains to generate videos can introduce potential temporal bias since the action positions and durations are uncontrollable. Previous work [63] attempted to mitigate process imbalances in temporal distribution by collecting long-term videos. However, this approach fails to address other types of temporal bias within the video and does not provide essential solutions to this critical issue. We propose two strategies to produce better temporal distribution by controlling the positions and durations of actions: Action Duration Diversity (ADD) and Action Permutation (AP). As shown in Figure 3 (c), the idea is to ensure each action

can appear at any position in the video and have diverse durations. Specifically, AP permutes the actions in chains so that each action will appear in as many positions as possible while satisfying the activity commonsense. ADD enables the diverse action durations by adopting varied video framerates. We employ ADD and AP to create controllable activity manuscripts for generating synthetic videos.

**Synthetic Video Generation** We randomly choose an agent and a situation from their corresponding spaces. Then we execute functional programs with the activity manuscript, agent, and situation in the simulator to create situation videos, as shown in Figure 3 (d).

### 3.3.2. Language Sentence Generation

For text generation, previous works [14, 19, 28, 31, 63] produced the language sentence by crowdsourcing annotation or an ASR model. These methods may have two drawbacks: 1) ambiguity problem, since different human annotators would write different semantic text for a sample, this problem is also proposed in the [86], and 2) noise issue, pre-trained models sometimes provide unreliable results, which leads to incorrect generated text. These disadvantages would reduce the quality of the benchmark and increase the challenges of model training. Thus, we utilize template-based generation to create the sentence to get high-quality language queries. In detail, three templates are defined to directly convert each action in action chains into sentences with different scenes and agents based on whether the scenes change when the action occurs, as shown in *Supplementary*. We utilize all actions that happened in the videos and use templates directly to construct the sentences, which can reduce the ambiguity and noise problems in the dataset. Additionally, since the Large Language Models demonstrate superior performance in natural language generation, we use the GPT-3.5-turbo to rewrite the original template-based sentences into more natural and diverse descriptions, serving as an auxiliary resource to strengthen our benchmark. This process is exhibited in Figure 3 (d).

### 3.3.3. Vision-Language Temporal Alignment

After we derive the synthetic videos and languages by prior steps, we need to align them to produce corresponding timestamps. Previous works [14, 28, 31] let humans assign each language query to the video content, but it may cause noisy temporal labels as mentioned in [51]. Thanks to the VirtualHome, when we generate synthetic videos, it records the time and duration of each action. Consequently, we only need to associate the automatically annotated timestamp with the action in the sentence to generate high-quality temporal annotations, which is depicted in Figure 3 (e).

### 3.3.4. Inequality Constrained Global Filtering

Though we propose two strategies to control the temporal distribution, they only operate the local distribution in each

Table 2. Comparison of multi-level temporal biases.

Benchmark	Process	Entity		
		Verb	Object	Composition
TACoS [59]	0.243	0.786	0.787	0.899
ActivityNet Captions [28]	0.107	0.764	0.827	0.921
Charades-STA [14]	0.287	0.739	0.877	0.881
TVR [31]	0.229	0.779	0.84	0.914
MAD [63]	0.628	0.842	0.869	0.926
SVLTA (ours)	<b>0.073</b>	<b>0.266</b>	<b>0.101</b>	<b>0.322</b>

logical action chain, which may produce potential temporal biases from the global perspective. Therefore, a debiasing method should be utilized to balance the temporal distributions as a post-processing step. Previous research [30, 62] designed the Adversarial Filtering (AF) method to reduce the bias in the dataset to achieve the balanced distribution goal. However, AF only has sub-optimal results since it filters those samples that most influence the distribution in each iteration. To perform a better debiasing effect, we propose a novel approach Inequality Constrained Global Filtering (ICGF) to adjust the temporal distribution of each action since a video is composed of multiple different actions. The main idea of ICGF is to filter some samples to obtain a more balanced temporal distribution while not filtering too many samples. Specifically, we treat this idea as a non-linear optimization problem with inequality constraints, the optimization goal is to reduce the gap between the current distribution and the uniform distribution (we use an absolute deviation function to measure the distribution gap) and the constraint is that too many samples should not be filtered (a filtering rate is utilized to control sample size). The details of ICGF and its comparison with other methods are included in *Supplementary*.

### 3.3.5. Benchmark Comparison

To evaluate our strategies for balancing temporal distributions and the effectiveness of our data-level debiasing approach ICGF, we first compare the temporal bias in our SVLTA dataset with five mainstream datasets: MAD [63], TACoS [59], ActivityNet Captions (Anet-Captions) [28], Charades-STA [14], and TVR [31]. MAD features long-term videos (over 1 hour), while TACoS, Anet-Captions, and TVR contain medium-term videos (over 1 minute), and Charades-STA includes short-term videos (around 30 seconds). The diverse characteristics of these datasets enable a comprehensive comparison, as shown in Table 2. Our results indicate that SVLTA exhibits the least temporal bias across various metrics, highlighting the effectiveness of our synthetic generation method in creating well-controlled temporal alignments. Additionally, we plot the temporal distribution of the moment start and end times for all temporal annotations, as illustrated in Figure 4. The results show that the distribution curve of SVLTA looks flatter and has a smaller variance than other datasets, indicating the validity

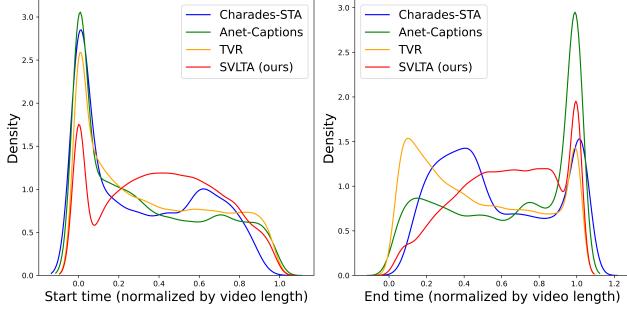


Figure 4. Temporal distributions of beginning or ending times. of our controllable strategies and filtering methods.

## 4. Experiment

We diagnose the temporal alignment ability of models from three perspectives: 1) temporal question answering, which aims to evaluate the temporal alignment of current VidLLMs, 2) distributional shift sensitiveness attempts to analyze the impact of temporal distribution shift on temporal alignment, and 3) temporal alignment adaptation explores whether the temporal alignment capability of models can be transferred to the new video situations or domains.

### 4.1. Experimental Setting

Temporal question answering utilizes the simple question-answer paradigm, which is widely used in other benchmarks [33, 43, 81] that evaluate the comprehension and reasoning ability of MLLMs. Here, we utilize the same temporal-related question prompt as used in the previous works [45, 60]. Additionally, Charades-STA is selected as the target for new video situations transferring in the temporal alignment adaptation since Charades-STA is collected in the real-world home, which has situations and actions similar to our benchmark. In distributional shift sensitiveness, a set of training, validation, and test sets are initially created with high temporal bias through long-tailed sampling, then a low temporal bias test set is constructed using ICGF from the remaining samples, and finally evaluating the difference between the results of the two test sets.

### 4.2. Evaluation Model

We evaluate both VidLLMs and specific temporal alignment models. For the temporal question answering, we analyze several VidLLMs, including open-sourced models, such as Video-LLaMA2 [8], Video-ChatGPT [46], Video-LLaVA [39], Videochat2 [36], and LLaVA-Video [85], as well as close-sourced models, such as Gemini 1.5 Pro [66], GPT-4o [50], and GPT-4o-mini [50]. Additionally, time-aware VidLLMs like TimeChat [60], VTimeLLM [22], and E.T.Chat [45] are also considered. For distributional shift sensitiveness and temporal alignment adaptation, we benchmark different specific frameworks: 1) anchor-free:

VSLNet [83] and LGI [48], 2) anchor-based: 2D-TAN [84], and 3) transformer-based: QD-DETR [47]. Meanwhile, two debiased models DCM [75] and Shuffling [20] are also diagnosed for distributional shift sensitiveness.

### 4.3. Evaluation Metric

For temporal question answering and temporal alignment adaptation, we employ the same metrics as used in prior studies [14, 79] to show their performance, namely  $R@1$ ,  $IoU = 0.1, 0.3, 0.5, 0.7, 0.9$  and mean IoU (mIoU). Furthermore, a new metric RC is developed to assess distributional shift sensitiveness. The motivation of RC is that a temporal robust model should not be easily affected by temporal bias when training, meaning it could perform reliably on test sets with varying distributions. In our setting, the RC is the difference between the results of the two test sets when training on the high temporal bias data. The higher the RC, the worse the temporal robustness of the model.

### 4.4. Results and Analysis

**Temporal Question Answering** The results of VidLLMs on SVLTA are shown in Table 3, indicating that none of the current VidLLMs can achieve satisfactory performance on our SVLTA benchmark, even some time-sensitive and close-sourced models. Specifically, the VTimeLLM only obtains the highest mIoU of 10.29 among these time-aware VidLLMs and current strong close-sourced models like Gemini 1.5 Pro and GPT-4o just get the mIoU of 12.48 and 18.90, respectively. This means that current VidLLMs do not have strong temporal alignment capabilities. Additionally, we can observe that most general open-sourced VidLLMs often have poor temporal alignment ability such as Videochat2 and Video-LLaVA merely achieve the mIoU of 0.87 and 2.59 correspondingly, demonstrating that their training stage ignore the temporal understanding capability modeling. However, the Video-LLaMA2 has a mIoU of 12.33, even higher than the time-aware VidLLMs, this is because of its temporal encoding design and high-resolution frame input. Further analysis of the visual domain gap, the number of frames, performance comparisons, and detailed question prompts is provided in *Supplementary*.

**Distributional Shift Sensitiveness** The results in Table 4 show the diagnosis of various specific temporal alignment methods in the distribution shift scenario. Notably, DCM, despite using causal inferencee [53, 54] to mitigate temporal bias effects, exhibits poorer robustness than biased methods (has the highest RC value of 17.86). This suggests that the causal-based approach may have limitations in fine-grained shifts of temporal distribution, possibly due to the imperfect disentanglement of action content and position in videos. In contrast, Shuffling demonstrates better robustness (only has the lowest RC value of 1.04), highlighting the effectiveness of using pseudo labels for video data augmentation to bal-

Table 3. The results of current popular open-sourced and close-sourced VidLLMs on SVLTA.

Method	# Frames	Size	Visual Encoder	LLM	R@1					mIoU
					IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.7		
General Open-sourced Models: All models use their default setting. Except LLaVA-Video, due to the GPU memory limits.										
LLaVA-Video [85]	16	7B	SIGLIP-SO400M	Qwen2	2.52	0.89	0.40	0.27	0.84	
Videochat2 [36]	16	7B	UMT-L/16	Vicuna-0	2.93	0.87	0.32	0.13	0.87	
Video-LLaVA [39]	8	7B	LanguageBind-ViT-L/14	Vicuna-1.5	8.22	3.19	0.96	0.23	2.59	
Video-ChatGPT [46]	100	7B	CLIP-ViT-L/14	Vicuna-1.1	10.68	3.17	0.90	0.21	2.94	
Video-LLaMA2 [8]	16	7B	CLIP-ViT-L/14	Mistral-7B	35.48	16.02	6.64	2.28	12.33	
Time-aware Open-sourced Models: All models utilize their default configuration.										
E.T.Chat [45]	1FPS	3.8B	EVA-ViT-G/14	Phi-3-Mini	17.86	8.07	3.48	1.36	6.29	
TimeChat [60]	96	7B	EVA-ViT-G/14	Llama-2	23.29	13.58	6.96	3.25	9.61	
VTimeLLM [22]	100	7B	CLIP-ViT-L/14	Vicuna-1.5	29.97	13.29	5.26	1.71	10.29	
Close-sourced Models: Evaluated on a subset with 2000 samples.										
GPT-4o-mini [50]	32	—	—	—	24.79	6.49	1.57	0.42	6.70	
Gemini 1.5 Pro [66]	1FPS	—	—	—	32.30	17.45	7.45	3.15	12.48	
GPT-4o [50]	32	—	—	—	49.54	27.38	11.69	5.62	18.90	

Table 4. The performance of distributional shift sensitiveness on SVLTA.

		R@1							
		Method	Test set	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.9	mIoU	RC ↓
Biased Models	2D-TAN [84]		high bias	93.82	87.08	72.55	35.06	76.41	
			low bias	84.40 <sup>(-9.42)</sup>	76.10 <sup>(-10.98)</sup>	60.75 <sup>(-11.8)</sup>	22.75 <sup>(-12.31)</sup>	66.66 <sup>(-9.75)</sup>	10.85
	VSLNet [83]		high bias	98.14	97.03	95.26	83.40	92.63	
			low bias	85.59 <sup>(-12.55)</sup>	83.22 <sup>(-13.81)</sup>	79.60 <sup>(-15.66)</sup>	67.34 <sup>(-16.06)</sup>	79.16 <sup>(-13.47)</sup>	14.31
	LGI [48]		high bias	97.02	94.26	87.38	56.36	85.25	
			low bias	89.70 <sup>(-7.32)</sup>	82.98 <sup>(-11.28)</sup>	68.74 <sup>(-18.64)</sup>	31.49 <sup>(-24.87)</sup>	72.67 <sup>(-12.58)</sup>	14.94
	QD-DETR [47]		high bias	98.96	98.35	96.46	82.61	93.05	
			low bias	95.59 <sup>(-3.37)</sup>	93.93 <sup>(-4.42)</sup>	90.17 <sup>(-6.29)</sup>	72.43 <sup>(-10.18)</sup>	87.72 <sup>(-5.33)</sup>	5.92
Debiased Models	DCM [75]		high bias	92.89	85.72	69.75	32.29	74.85	
			low bias	79.55 <sup>(-13.34)</sup>	68.11 <sup>(-17.61)</sup>	46.15 <sup>(-23.6)</sup>	13.49 <sup>(-18.8)</sup>	58.88 <sup>(-15.97)</sup>	17.86
	Shuffling [20]		high bias	93.78	89.43	82.25	49.63	81.62	
			low bias	93.26 <sup>(-0.52)</sup>	88.61 <sup>(-0.82)</sup>	80.23 <sup>(-2.02)</sup>	49.04 <sup>(-0.59)</sup>	80.36 <sup>(-1.26)</sup>	1.04

Table 5. The results of temporal alignment adaptation task.

Method	R@1			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
2D-TAN [84]	15.81	5.03	1.94	11.8
VSLNet [83]	28.33	8.52	3.87	19.66
LGI [48]	33.96	12.52	3.30	22.24
QD-DETR [47]	33.74	18.39	7.55	22.32

ance temporal distribution. It generally shows weak result consistency regarding biased methods due to the lack of debiasing and inadvertently learning these biases. However, QD-DETR, a transformer-based model, outperforms other biased methods in robustness (has the lowest RC value of 5.92 among the biased models), indicating superior generalization capabilities of transformer architectures.

**Temporal Alignment Adaptation** The results are illustrated in Table 5 and we can observe: 1) several frameworks of alignment models can transfer temporal knowledge (e.g., VSLNet and LGI can achieve the mIoU of 19.66 and 22.24 respectively). It means these models trained from scratch can transfer their temporal alignment ability to the new sit-

uations or domains, 2) transformer-based model has better transferability than other frameworks, it can achieve 10.52 higher mIoU than 2D-TAN and 2.66 than VSLNet, which demonstrates the advantages of transformer architectures in temporal alignment when adapting to new situations.

## 5. Conclusion

In this work, we first systematically analyze the temporal distributions for the vision-language temporal alignment problem from the decomposition aspect and introduce a new metric TJSD to examine three specific types of temporal bias related to process, entity, and composition. After that, we build a new large-scale and compositional benchmark SVLTA, using a proposed synthetic pipeline. Our approach involves activity commonsense, controllable activity manuscript, and constrained filtering to ensure it is diverse, compositional, and unbiased. The experiments reveal interesting insights for using this dataset in various diagnostic tasks, such as temporal question answering, distributional shift sensitiveness, and temporal alignment adaptation.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 2022. 1
- [2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 2017. 1, 2, 4
- [3] Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1
- [4] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the “video” in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 1
- [5] Che-Jui Chang, Honglu Zhou, Parth Goel, Aditya Bhat, Seonghyeon Moon, Samuel S Sohn, Sejong Yoon, Vladimir Pavlovic, and Mubbashir Kapadia. Learning from synthetic human group activities. *arXiv preprint arXiv:2306.16772*, 2023. 3
- [6] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018. 3
- [7] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 4
- [8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-lmms. *arXiv preprint arXiv:2406.07476*, 2024. 7, 8
- [9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Itsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1
- [10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhang-hao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yong-hao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3), 2023. 1
- [11] Edward Collins, Nikolai Rozanov, and Bingbing Zhang. Evolutionary data measures: Understanding the difficulty of text classification tasks. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 2018. 4
- [12] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022. 5
- [13] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3
- [14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, 2017. 1, 2, 4, 6, 7
- [15] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. In *International Conference on Learning Representations*, 2020. 3
- [16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4
- [17] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3
- [18] Xi Guo, Wei Wu, Dongliang Wang, Jing Su, Haisheng Su, Weihao Gan, Jian Huang, and Qin Yang. Learning video representations of human motion from synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3
- [19] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 6
- [20] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *European Conference on Computer Vision*. Springer, 2022. 7, 8
- [21] Rishi Hazra, Brian Chen, Akshara Rai, Nitin Kamra, and Ruta Desai. Egotv: Egocentric task verification from natural language task descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3
- [22] 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*, 2024. 3, 7, 8
- [23] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1
- [24] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications. *IEEE Access*, 2021. 3
- [25] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition*, 2020. 5
- [26] Ilker Keser, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. VilMA: A zero-shot benchmark for linguistic and temporal grounding in video-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [27] Yo-whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. How transferable are video representations based on synthetic data? *Advances in Neural Information Processing Systems*, 2022. 3
- [28] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 2017. 1, 4, 6
- [29] Xiaohan Lan, Yitian Yuan, Hong Chen, Xin Wang, Zequn Jie, Lin Ma, Zhi Wang, and Wenwu Zhu. Curriculum multi-negative augmentation for debiased video grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 3
- [30] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International conference on machine learning*. PMLR, 2020. 6
- [31] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020. 2, 4, 6
- [32] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34, 2021. 3
- [33] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1, 7
- [34] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yuetong Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4
- [35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 2023. 1
- [36] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7, 8
- [37] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. Lego: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024. 3
- [38] Yuan-Hong Liao, Xavier Puig, Marko Boben, Antonio Torralba, and Sanja Fidler. Synthesizing environment-aware activities via activity sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 4
- [39] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 7, 8
- [40] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shruman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3
- [41] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 3
- [42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1
- [43] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 7
- [44] 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? *arXiv preprint arXiv:2403.00476*, 2024. 3
- [45] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Et bench: Towards open-ended event-level video-language understanding. *arXiv preprint arXiv:2409.18111*, 2024. 3, 4, 7, 8
- [46] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 7, 8
- [47] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 7, 8
- [48] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 7, 8
- [49] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 3
- [50] OpenAI. Hello gpt-4o. 2024. 7, 8
- [51] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based

- video moment retrieval. *arXiv preprint arXiv:2009.00325*, 2020. 1, 3, 6
- [52] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens Continente, Larisa Markeeva, Dylan Sunil Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multi-modal video models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3
- [53] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018. 7
- [54] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 7
- [55] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 4
- [56] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B. Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human- $\{\text{ai}\}$  collaboration. In *International Conference on Learning Representations*, 2021. 2, 4
- [57] Yue Qiu, Yoshiki Nagasaki, Kensho Hara, Hirokatsu Kataoka, Ryota Suzuki, Kenji Iwata, and Yutaka Satoh. Virtualhome action genome: A simulated spatio-temporal scene graph dataset with consistent relationship labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 3
- [58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 2021. 1
- [59] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1, 2013. 1, 2, 4, 6
- [60] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7, 8
- [61] Cesar Roberto de Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel Lopez. Procedural generation of videos to train deep action recognition networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3
- [62] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 6
- [63] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 4, 5, 6
- [64] Yale Song, Gene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2, 4
- [65] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020. 1
- [66] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 7, 8
- [67] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [68] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1
- [69] Aisha Urooj, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, Chuang Gan, Niels Lobo, and Mubarak Shah. Learning situation hyper-graphs for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1
- [70] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7), 2021. 3
- [71] Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. Sokbench: A situated video reasoning benchmark with aligned open-world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5
- [72] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *arXiv preprint arXiv:2405.19209*, 2024. 3
- [73] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems*, 2021. 1, 5
- [74] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *Computer Vision–ECCV 2020*:

- 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020. 4
- [75] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. 7, 8
- [76] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1
- [77] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. 3
- [78] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019. 3
- [79] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 3, 7
- [80] Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Dataset and metric. In *Proceedings of the 2nd international workshop on human-centric multimedia analysis*, 2021. 1, 3
- [81] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 1, 7
- [82] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3
- [83] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 7, 8
- [84] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 7, 8
- [85] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 7, 8
- [86] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 6
- [87] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1