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

Supplementary Material

The Supplementary Material is organized as follows:

- Section 1 includes extended related works, which mainly discuss the relation between synthetic and real data.
- Section 2 introduces more temporal distribution visualizations in ActivityNet Captions and Charades-STA (§ 2.1), and presents the metric TJSD details (§ 2.2).
- Section 3 contains detailed component spaces (§ 3.1), commonsense activity graph rules, and re-weighting sampling strategy on the graph(§ 3.2), template-based and GPT-based sentence generation (§ 3.3), ICGF details and its comparison with other methods (§ 3.4), and more statistics and visualizations of SVLTA (§ 3.5).
- Section 4 outlines model implementation details (§ 4.1), RC metric and the statistics of different temporal bias data in the Distributional Shift Sensitiveness task(§ 4.2), and the further analysis of VidLLMs(§ 4.3).
- Section 5 provides some data examples from the SVLTA.

1. Extend related works

In this section, we mainly discuss the relationship between synthetic and real-world data. Both of them are valuable and they are complementary to achieve different goals. The real-world data has real appearances and scenarios and can well evaluate the model's ability to be applied in the real world, dominating the current benchmarks, such as ImageNet [4] and Kinetics [1]. Additionally, the model trained on the large-scale real-world benchmark usually has strong robustness and generalization [14, 35], which can be transferred to multiple scenarios due to the diverse semantics and appearances. However, with the emergence of large language models [5, 15, 40, 41] and more and more simulators [20, 33, 42], synthetic data with high-quality annotations and controllable elements is gradually gaining popularity, which can provide rigorous or fair diagnostic evaluations or benchmarks and also shown in the existing work like CLEVR [12] or CATER [8]. These synthetic benchmarks mainly provide a diagnostic framework for various tasks (e.g. understanding [39], reasoning [24, 46], and recognition or detection [7, 34, 43]), which often reveal the drawbacks of current models and facilitate some insightful conclusions for the community. The aforementioned synthetic datasets are not designed to study vision-language temporal alignment and ignore the explicit control of temporal alignment as the primary generation objective, which is different from our proposed SVLTA benchmark.

2. More Temporal Distribution Analysis

2.1. Temporal Distribution Visualizations

Here, we provide more temporal distribution visualizations related to the process (Figure 1), composition (Figure 2), and entity (Figure 3 and 4) in ActivityNet Captions (ANCaps) [16] and Charades-STA [6]. These distributions are acquired by kernel density estimation with the Gaussian kernel, consistent with previous works [28, 47, 49]. The color darkness represents the sample density, and the horizontal and vertical axes represent the normalized start and end time points respectively. We can observe that these distributions are not uniform, i.e., some temporal annotations have a high-frequency occurrence at the beginning or end of the video. We analyze the current benchmarks and derive two reasons for this unbalanced phenomenon: 1) actions cannot be guaranteed to happen in arbitrary positions in the video, and 2) human annotators overlook some of the actions in the video and do not label them. The above visualization implies that the current mainstream benchmarks are subject to multi-level temporal biases, spanning from global to local perspectives.

2.2. TJSD Metric

TJSD metric is proposed to measure the temporal bias by computing the difference between the target temporal distribution and the uniform distribution. Specifically, we first divide the video into n equal moments to discretize time, leading to $\frac{n(n+1)}{2}$ different bins, each bin means a temporal class, and then we assign the timestamps into these bins. Therefore, the target distribution can be represented by the number of samples in these bins and the uniform distribution means that the number of samples in each bin is the same, both of them are divided by all the samples in the dataset to be normalized. Finally, the Jensen–Shannon divergence calculates the difference between the above two distributions. The TJSD is formulated as follows:

$$\begin{aligned} TJSD &= JS(P||U), \\ P &= \left[\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}, \dots\right], \\ U &= \left[\frac{2}{n(n+1)}, \frac{2}{n(n+1)}, \dots\right] \end{aligned} \tag{1}$$

Where P and U denote the target temporal logits and uniform logits with $\frac{n(n+1)}{2}$ elements. N_i and N refer to the number of samples in the i-th bin and in all bins, respectively. We use the whole temporal annotations in the dataset

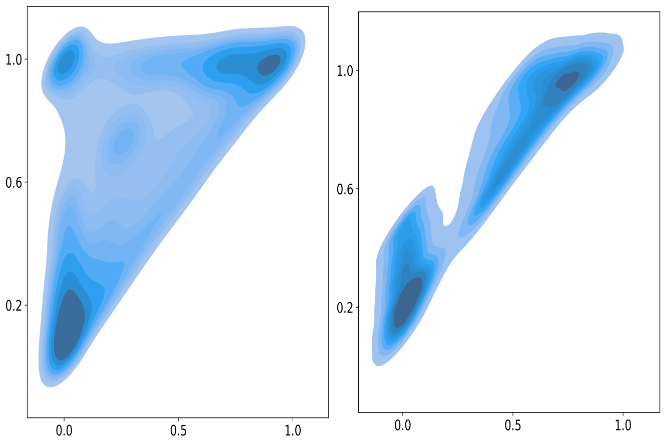


Figure 1. Temporal Distributions of the whole temporal annotations. The right column is AN-Caps, and the left column is Charades-STA.

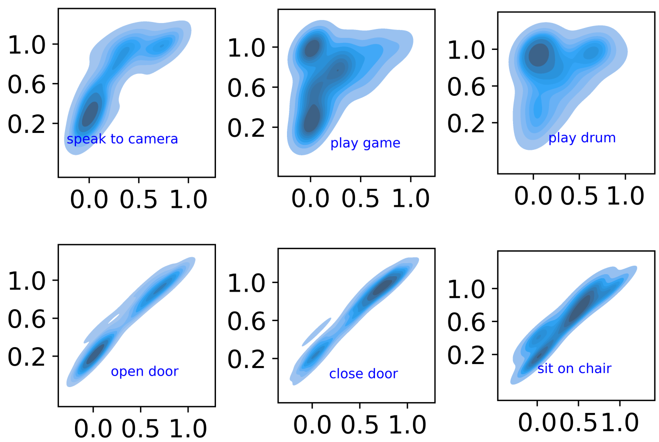


Figure 2. Temporal Distributions of Top3 actions. The top row is AN-Caps, and the bottom row is Charades-STA.

to compute the process temporal bias. For the entities or composition temporal bias, we first extract the verb (or object or action) and then utilize their corresponding temporal annotations to calculate the temporal bias separately. The smaller the value, the less temporal bias of the dataset.

3. Benchmark Details

3.1. Detailed Situation Component Spaces

We define 10 verbs and 34 objects to compose 96 diverse meaningful actions. The concept of compositional actions has already been proposed in previous works [8, 36], which is different from the verb-only actions that only focus on standalone activities without objects, such as walking or running, often seen in early action recognition benchmarks [17]. The compositional action introduces more complex interactions by combining different verbs with different objects, creating distinct actions with unique semantic meanings (such as *open fridge*, *open microwave*, and

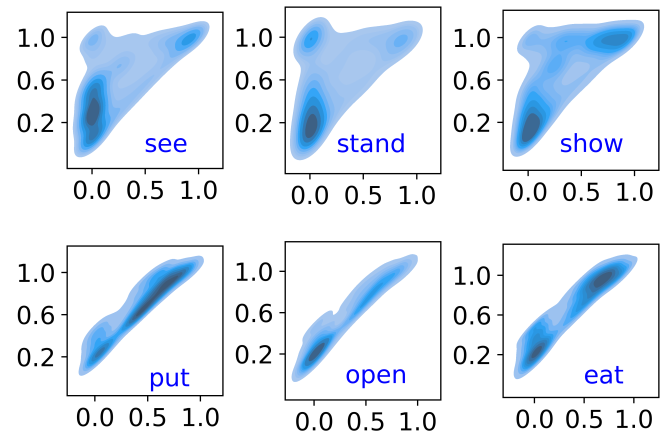


Figure 3. Temporal Distributions of Top3 verbs. The top row is AN-Caps, and the bottom row is Charades-STA.

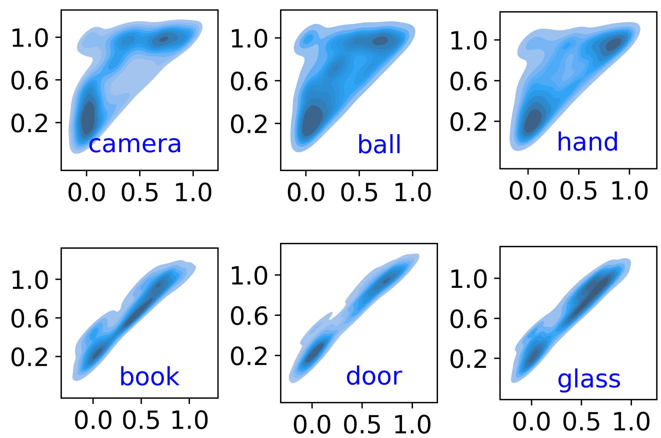


Figure 4. Temporal Distributions of Top3 objects. The top row is AN-Caps, and the bottom row is Charades-STA.

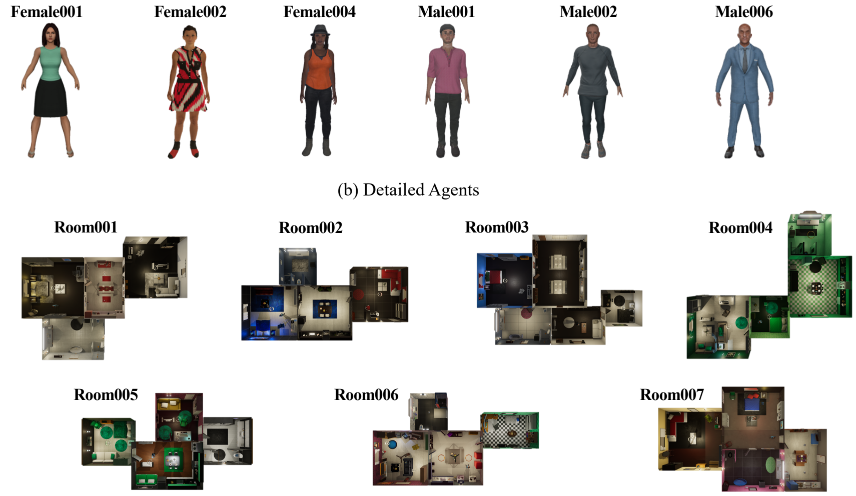
grab milk). For instance, walking to a bed and walking to a sofa may share the core verb *walk*, but they differ in terms of the object involved. In the vision-language temporal alignment, this is not only relevant to verb-only actions but also adds layers of semantic context that are crucial for the evaluations in vision-language temporal alignment tasks. We also manually check whether each action can be executed by the VirtualHome simulator [22, 31, 32] and determine the scenes and characters in which each action can be performed. All actions and objects are quite distinct in motion and shape. Figure 5 (a) shows the detailed verb, object, and action lists. We can see that these actions often occur in daily life, like *open fridge*, *switch on tv*, *grab plate*, etc. Figure 5 (b) and (c) illustrate the detailed agents and situations, which are the same in the original VirtualHome.

3.2. Commonsense Activity Graph

Activity Commonsense VirtualHome provides the basic rules between paired actions, i.e., some actions must wait until the conditional actions are completed before they can happen. However, it would have some unreasonable com-

Verb
open, close, sit, grab, stand, walk, run, put, switch, drink
Object
fridge, bathroomcabinet, kitchencabinet, cabinet, dishwasher, microwave, sofa, bed, bench, chair, waterglass, wineglass, milk, juice, cereal, chips, salmon, cutleryfork, plate, folder, remotecontrol, paper, mouse, facecream, fryingpan, bedroom, bathroom, kitchen, livingroom, tv, toaster, faucet, stove, computer
Action
open fridge, open bathroomcabinet, open kitchencabinet, open cabinet, open dishwasher, open microwave, close fridge, close bathroomcabinet, close kitchencabinet, close cabinet, close dishwasher, close microwave, sit on sofa, sit on bed, sit on bench, sit on chair, grab waterglass, grab wineglass, grab milk, grab juice, grab cereal, grab chips, grab salmon grab cutleryfork, grab plate, grab folder, grab remotecontrol, grab paper, grab mouse, grab facecream, grab fryingpan, stand up, walk to fridge, walk to bathroomcabinet, walk to kitchencabinet, walk to cabinet, walk to dishwasher, walk to microwave, walk to sofa, walk to bed, walk to bench, walk to chair, walk to waterglass, walk to wineglass, walk to milk, walk to juice, walk to cereal, walk to chips, walk to salmon, walk to cutleryfork, walk to plate, walk to folder, walk to remotecontrol, walk to paper, walk to mouse, walk to facecream, walk to fryingpan, walk to tv, walk to toaster, walk to faucet, walk to stove, walk to computer, run to bedroom, run to bathroom, run to kitchen, run to livingroom, put waterglass on the microwave, put wineglass on the microwave, put milk on the microwave, put juice on the microwave, put cereal on the microwave, put chips on the microwave, put salmon on the microwave put cutleryfork on the sofa, put plate on the sofa, put folder on the sofa, put remotecontrol on the sofa, put paper on the sofa, put mouse on the sofa, put facecream on the sofa, put fryingpan on the sofa, put milk in microwave, put salmon in microwave, put plate in microwave, put chips in fridge, put cereal in fridge, switch on tv, switch on toaster, switch on faucet, switch on stove, switch on computer, switch off tv, switch off toaster, switch off faucet, switch off stove, switch off computer, drink waterglass, drink wineglass, drink milk, drink juice

(a) Detailed Verbs, Objects, Actions



(b) Detailed Agents

(c) Detailed Situations

Figure 5. Detailed Verbs, Objects, Actions, Agents, and Situations in the SVLTA.

positions for longer action sequences when directly using these rules to sample the actions from the component space. To address this problem, a commonsense activity graph is developed. Specifically, we first manually check these rules and filter some unreasonable rules (cannot be executed) in some situations and agents to ensure that they are consistent with the normal commonsense in our lives, as shown in Table 1. Based on these filtered rules, we can obtain a whole activity-directed graph for all our compositional actions. Then, a graph traversal algorithm like DFS or BFS is utilized to traverse action nodes in activity graphs with given lengths to obtain the logical action chains.

Re-weighting Sampling Since different actions have different constraints in the activity graph, e.g., action *walk to sth* does not need to wait for other actions to happen while action *close sth* must wait for action *open sth* to happen before it can occur, which causes the degree of action node to be imbalanced in the activity graph and reduce the diversity of action compositions when directly traversing the 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. In detail, the re-weighting ratios are based on the degree of the action nodes, if the action node has a high degree, then we would give it a small value, and vice versa.

3.3. Language Sentence Generation

Template-based Generation We introduce three templates to convert each action in activity manuscripts into sentences directly, which is illustrated in Table 2. The definition of these templates depends on whether the scenes would change when the action occurs. Specifically, for the action that does not cross the room, i.e., the scene remains unchanged, the template is *The <character>* + *<action>* + *in the* + *<room>* + *in the* + *<scene>*. If the action crosses the room, it still has two cases, i.e., for the action *walk to sth*, its template is *The <character>* + *walk through the door from* + *<original room>* + *to* + *<current room>* + *and* + *<action>* + *in the* + *<current room>* + *in the* + *<scene>*, for the action *run to sth*, the template is *The <character>* + *run through the door* + *<original room>* + *to* + *<current room>* + *in the* + *<scene>*. The reason why these two action templates are different is that the action *run to sth* does not associate with some objects, but the action *walk to sth* needs to interact with some objects. Utilizing the above template to generate the language sentence would reduce the ambiguity and noise problems in the dataset, which can improve the benchmark quality.

GPT-based Generation Large Language Models (LLMs) achieve excellent results in the natural language generation area, so we also use the ChatGPT(GPT-3.5-turbo) [27] to rewrite the original template-based sentences into more natural and diverse descriptions to strengthen our benchmark. The detailed prompt is shown in Figure 6.

3.4. Inequality Constrained Global Filtering

In this section, we provide more details about our Inequality Constrained Global Filtering (ICGF) method. The ICGF is a global-level debiasing method since the two strategies ADD and AP control the temporal biases in each logical action chain from a local perspective and may produce potential temporal biases from a global perspective. 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 nonlinear optimization problem with inequality constraints, the optimiza-

Table 1. The activity commonsense in the VirtualHome. Pre-action: completed action. Post-action: pending action. Condition: The condition under which the post-action can happen if the pre-action has already occurred. $o_i \mathrel{!}= o_j$: cannot be the same object. $o_i == o_j$: must be the same object. —: do not have relation.

Pre-action	Post-action	Condition
walk to o_i	walk to o_j	$o_i \mathrel{!}= o_j$
	run to o_j	$o_i \mathrel{!}= o_j$
	open o_j	$o_i == o_j$
	grab o_j	$o_i == o_j$
	switch on o_j	$o_i == o_j$
	sit on o_j	$o_i == o_j$
	close o_j	$o_i == o_j$
	switch off o_j	$o_i == o_j$
	drink o_j	—
	put o_{j1} on o_{j2}	$o_i == o_{j2}$
	put o_{j1} in o_{j2}	$o_i == o_{j2}$
run to o_i	run to o_j	$o_i \mathrel{!}= o_j$
	walk to o_j	$o_i \mathrel{!}= o_j$
	drink to o_j	—
open o_i	close o_j	$o_i == o_j$
	walk to o_j	$o_i \mathrel{!}= o_j$
	run to o_j	$o_i \mathrel{!}= o_j$
	drink o_j	$o_i \mathrel{!}= o_j$
	put o_{j1} in o_{j2}	$o_i == o_{j2}$
grab o_i	drink o_j	$o_i == o_j$
	walk to o_j	$o_i \mathrel{!}= o_j$
	run to o_j	$o_i \mathrel{!}= o_j$
	put o_{j1} in o_{j2}	$o_i == o_{j1}$
switch on o_i	switch off o_j	$o_i == o_j$
	walk to o_j	$o_i \mathrel{!}= o_j$
	run to o_j	$o_i \mathrel{!}= o_j$
	drink o_J	—
sit on o_i	stand up	—
close o_i	walk to o_j	$o_i \mathrel{!}= o_j$
	run to o_j	$o_i \mathrel{!}= o_j$
	drink o_J	—
drink o_i	walk to o_j	$o_i \mathrel{!}= o_j$
	run to o_j	$o_i \mathrel{!}= o_j$
switch off o_i	walk to o_j	$o_i \mathrel{!}= o_j$
	run to o_j	$o_i \mathrel{!}= o_j$
	drink o_j	—
stand up	walk to o_j	—
	run to o_j	—
	drink o_j	—
put o_{i1} in o_{i2}	walk to o_j	$o_{i2} \mathrel{!}= o_j$ and $o_{i1} \mathrel{!}= o_j$
	run to o_j	$o_{i2} \mathrel{!}= o_j$ and $o_{i1} \mathrel{!}= o_j$
	drink o_j	$o_{i2} \mathrel{!}= o_j$ and $o_{i1} \mathrel{!}= o_j$
	close o_j	$o_{i2} == o_j$
put o_{i1} on o_{i2}	walk to o_j	$o_{i2} \mathrel{!}= o_j$ and $o_{i1} \mathrel{!}= o_j$
	run to o_j	$o_{i2} \mathrel{!}= o_j$ and $o_{i1} \mathrel{!}= o_j$
	drink o_j	$o_{i2} \mathrel{!}= o_j$ and $o_{i1} \mathrel{!}= o_j$

tion goal is to reduce the gap between the current distribution and the uniform distribution and the constraint is that too many samples should not be filtered. Here, we use an absolute deviation function to measure the distribution gap, and a filtering rate is utilized to control the sample size, which can be formulated as follows:

$$
\begin{aligned}
\min \quad & \sum_{j=1}^{\frac{n(n+1)}{2}} \left| \frac{1}{\frac{n(n+1)}{2}} - \frac{\alpha_{ij} N_{ij}}{\sum_j \alpha_{ij} N_{ij}} \right| - \beta \sum_j \alpha_{ij} N_{ij} \\
s.t. \quad & 0 \leq \alpha_{ij} \leq 1 \\
& \sum_j \alpha_{ij} N_{ij} \geq \gamma N_i
\end{aligned} \tag{2}
$$

Where N_i means the sample numbers of i-th action, and N_{ij}, α_{ij} denotes the sample numbers and the sampling rate in j-th temporal class of i-th action. The first term of the above objective function means we want the temporal distribution can be as balanced as possible for the i-th action, an absolute deviation function is utilized to quantify the difference between the target distribution and uniform distribution, and the second term indicates the number of samples in the i-th action after sampling. Therefore, the objective function demonstrates that we want to balance the temporal distribution of i-th action and retain more samples after sampling. The restrictions mean the sampling rate should be between $[0, 1]$, and the number of samples after sampling cannot be lower than the initialization sample number N_i with a certain ratio γ. The β and γ are hyper-parameters with $1e^{-5}$ and 0.6 in our method, respectively. To solve the aforementioned optimization problem, we employ Sequential Quadratic Programming using the SciPy tool [44].

In addition, we also compare the Adversarial Filtering (AF) method [18, 38] with our ICGF, and the results are shown in Table 3. It indicates that the ICGF has a better debiasing effect than the AF, i.e., it can make the dataset obtain the lower three types of temporal bias and achieve more balanced distributions, which demonstrates that ICGF can find the optimal solution in a global iteration, unlike AF which can only get sub-optimal in each iteration.

3.5. Dataset Statistics and Visualizations

Vocabulary Distribution The count distributions of verbs, objects, and actions are shown in Figure 7. In Figure 7 (b) and (c), the histograms refer to the count of verbs and objects in the SVLTA, which can be seen that the highest frequent verbs and objects are *walk*, *grab*, *open*, *fridge*, *microwave*, *sofa*, etc. These verbs and objects compose actions, which are shown in Figure 7 (a) and (d). It can represent the most popular actions in the SVLTA, which are *walk to fridge*, *walk to microwave*, *run to livingroom*, *open fridge*, *stand up*, and *sit on sofa*, etc. **Note:** our goal is to construct a benchmark with uniform temporal distribution, so their count distributions are not restricted and the count distributions do not affect models' visual-language temporal alignment ability, as mentioned in [28].

Temporal Distribution Here, we also visualize the verb-level, object-level, and action-level temporal distributions in the SVLTA, as shown in Figure 8, 9, and 10. For verb-level and object-level distributions, we depict the Top3 verbs and

Table 2. Three detailed sentence templates for sentence generation. Suppose we select agent *Female1* and situation *scene0* to generate the synthetic video.

Condition	Template	Action	Example
action occurs not across the room	The <agent> + <action> + in the + <room> + in the + <scene>.	open fridge (in the livingroom)	The Female1 opens fridge in the livingroom in the scene0.
action *walk to sth* occurs across the room	The <agent> + walks through the door from + <original room> + to + <current room> + and + <action> + in the + <current room> + in the + <scene>.	walk to cabinet (from bedroom to kitchen)	The Female1 walks through the door from bedroom to kitchen and walks to cabinet in the kitchen in the scene0.
action *run to sth* occurs across the room	The <agent> + runs through the door + <original room> + to + <current room> + in the + <scene>.	run to bathroom (from kitchen to bathroom)	The Female1 runs through the door from kitchen to bathroom in the scene0.

Role: system
Content: You are a rewriting expert, your task is to improve the clarity, naturalness, and conciseness of the sentence. This will involve assessing the existing content, identifying areas that can be simplified or clarified, and restructuring sentences where necessary. Make sure you maintain the original meaning while reducing verbosity and ensuring the text is easy to understand. The final product should be a clear, concise, and coherent version of the original text.
Role: user
Content: Now, I will give you a sentence: <SENTENCE>, then you should rewrite it and keep the clarity, naturalness, and conciseness of the sentence. In addition, please use the present tense.

Figure 6. The prompt used in the GPT-3.5-turbo.

Table 3. The comparison between ICGF and AF.

Benchmark	Process	Entity		Composition
		Verb	Object	
SVLTA (raw)	0.127	0.306	0.152	0.343
SVLTA (w/ AF)	0.107	0.302	0.145	0.358
SVLTA (w/ ICGF)	**0.073**	**0.266**	**0.101**	**0.322**

objects in the SVLTA, for action-level distribution, Top6 actions are illustrated. Most of these distributions look flatter and have a small variance, demonstrating the validity of our controllable strategies and constrained filtering methods.

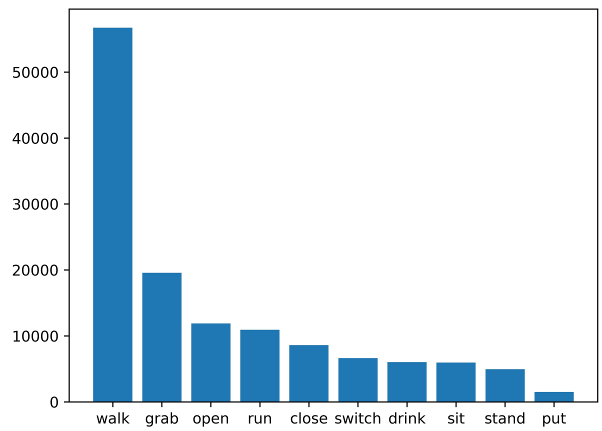
Action Duration Distribution Due to the Action Duration Diversity (ADD) strategy being used to balance the temporal distribution in the dataset, there are up to 7 different action framerates for each action in the SVLTA to make the action duration diverse. Figure 11 illustrates the action duration and frame rate distributions in the SVLTA. We can observe that 1) the action durations are diverse, there are 27.69% short-term actions, 44.43% middle-term actions, 19.63% long-term actions, and 8.25% very long-term actions and 2) the framerates are also diverse, it goes from 3 to 25, the lower the frame rate, the slower the action. **Note**: we post-process the video that contains different frame rates by sampling based on the minimum frame rate in each video, to keep the frame rate the same within a video.

4. Experiment Details

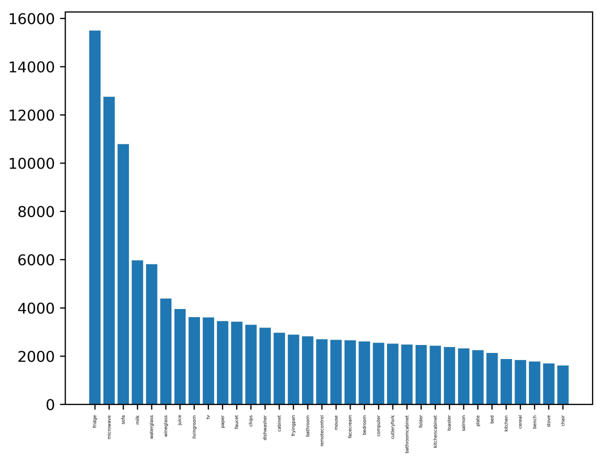
4.1. Model Implementation Details

Video Large Language Models The SVLTA has nearly 77.1K temporal annotations, directly evaluating Video Large Language Models (VidLLMs) on the whole dataset would cause redundancy. Therefore, we sample 13K+ annotations from the original SVLTA to compose the SVLTA-VLLM dataset, which contains all actions in the SVLTA and is consistent with the dataset size of mainstream multimodal evaluation benchmarks [19, 48]. Here, we employ the default configuration as those utilized in their original papers, except the LLaVA-Video [50], due to the GPU memory limits. We use the 16 frames rather than the 64 frames mentioned in the original paper. For the question prompts, due to the VTimeLLM [11] being fine-tuned on various frame-formatted instruction datasets, we ask their model to return the answer through frame numbers, for the other VidLLMs,

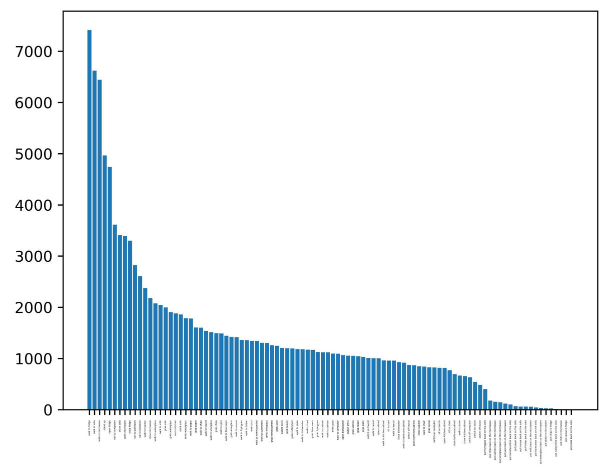
(a) Word cloud for actions in the SVLTA.



(b) The verbs statistics in the SVLTA.



(c) The objects statistics in the SVLTA.



(d) The actions statistics in the SVLTA.

Figure 7. The distribution of the vocabulary in the SVLTA.

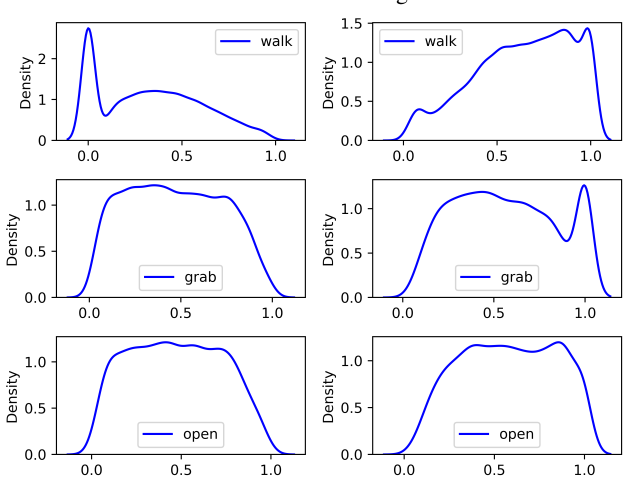


Figure 8. Temporal Distributions of Beginning and Ending Times of Top3 Verbs in SVLTA. Left is the start time and Right is the end time. All the times are normalized by video length.

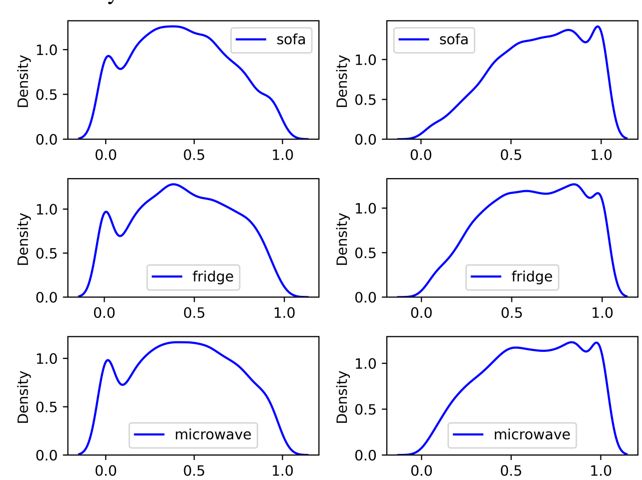


Figure 9. Temporal Distributions of Beginning and Ending Times of Top3 Objects in SVLTA. Left is the start time and Right is the end time. All the times are normalized by video length.

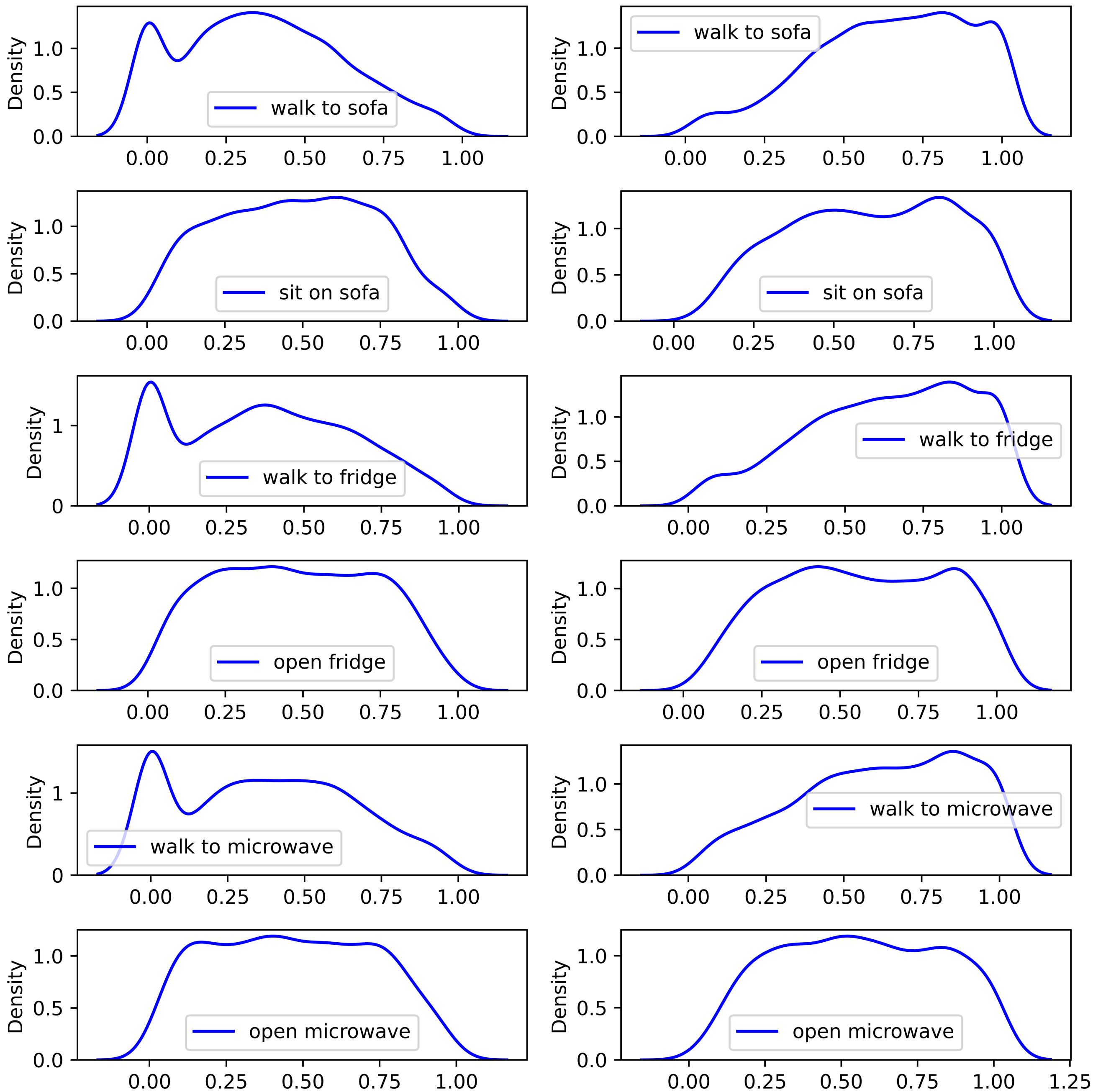
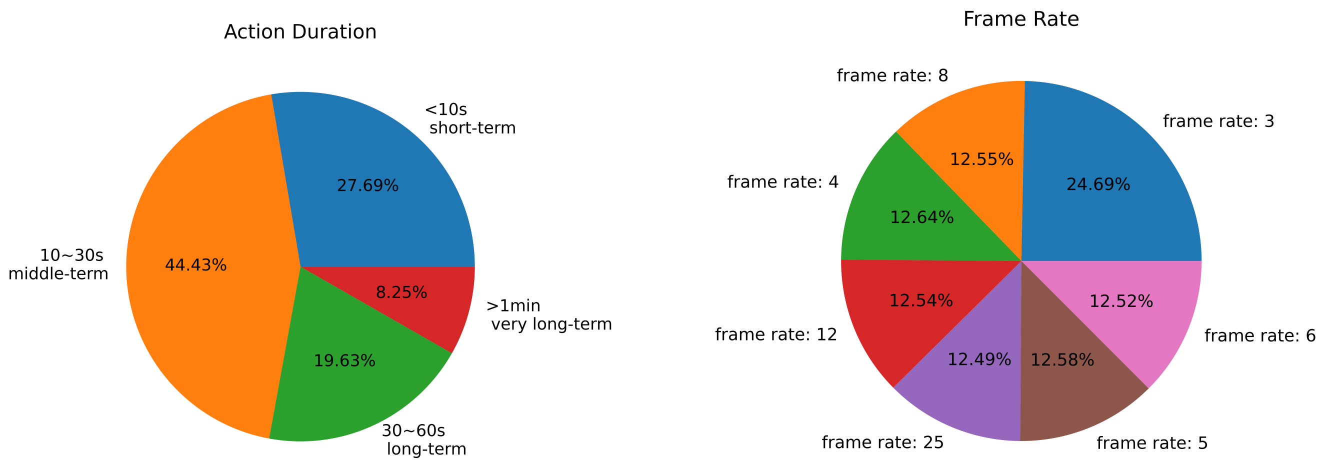


Figure 10. Temporal Distributions of Beginning and Ending Times of Top6 Actions in SVLTA. Left is the start time and Right is the end time. All the times are normalized by video length.

we ask the model returning the seconds, as shown in Figure 12. Additionally, we only use the 7B version of these models at the FP16 or BF16 precision. All experiments are conducted on a single RTX 3090 GPU with 24GB memory.

Specific Visual-Language Temporal Alignment Models All of the models we selected keep the same training settings of Charades-STA as mentioned in their original papers. For video feature extraction, we first resize the frames into 112×112 resolution and then use a pre-trained 3D ResNexT-101 [10] (on Kinetics-700 [2]) to extract the 2048-dimensional features. All sentences are represented by 300-dimensional pre-trained GolVe [30] word embedding. We conduct all experiments on a single RTX 2080 Ti GPU with 11GB memory.



(a) Action duration statistics (b) Frame rate statistics

Figure 11. The distribution of action duration in the SVLTA.

4.2. RC Metric

We introduce the RC metric to evaluate the ability of the distributional shift sensitiveness. Two test sets are constructed through different sampling strategies in our experimental setting, one has a high temporal bias, and the other has a low temporal bias. RC is the difference between the results of the two test sets when training on the high temporal bias data to assess whether the model is affected by temporal distribution shift, which can be formulated as follows:

$$RC = \frac{\sum_k \left| r^k_{low} - r^k_{high} \right|}{k} \tag{3}$$

Where k means different metrics to measure the models' performance, the k is 5, including the $R@1, IoU = 0.3, 0.5, 0.7, 0.9$, and the mIoU. r^k_{low} and r^k_{high} are the results of low-biased and high-biased test sets.

Moreover, we also give the statistics of the high temporal bias dataset and low-biased test set in the Distributional Shift Sensitiveness task, as shown in Table 4 and 5. The high-biased dataset is split into training/validation/test sets with a ratio of about 6:2:2.

Table 4. The size of two datasets.

Dataset	# Training	# Validation	# Test
High-Biased	56679	12212	12666
Low-Biased	—	—	11988

Table 5. The temporal bias of two datasets.

Dataset	Process	Entity		Composition
		Verb	Object	
High-Biased	0.331	0.610	0.469	0.683
Low-Biased	0.030	0.167	0.051	0.248

Table 6. Analysis on action recognition accuracy (%) in SVLTA.

Method	Acc.
Video-LLaVA [23]	71.06
Video-LLaMA2 [3]	77.94

4.3. Further Analysis of VidLLMs

Visual domain gap Although synthetic videos feature simpler visual appearances, their visual difference from real videos does not cause serious domain effects on most video models. To illustrate this, we conducted action recognition experiments with VidLLMs on our SVLTA. The results in Table 6 demonstrate that these models trained on real videos still perform strongly in visual perception within the synthetic domain. This means that the visual domain gap in the SVLTA has fewer effects on the temporal alignment.

The number of frames We also evaluate the effect of frame numbers in our SVLTA and take time-aware VidLLM TimeChat as an example. We consider three different frame numbers: 96 (default setting), 128, and 256. The results are in Table 8, demonstrating that frame numbers have little im-

Role: system
Content: You are able to understand the visual content that the user provides. Follow the instructions carefully and explain your answers in detail.
Role: user
Content:
Here is an illustrative example:
=== example start ===
You are given a video from a new synthetic dataset. Please find the visual event described by a sentence in the video, determining its starting and ending times. The format should be: 'The event happens at the start time-end time'. For example, The event 'person turns a light on' happens in the 24.3 - 30.4 seconds. Now I will give you the textual sentence: <SENTENCE> and the video length is <VIDEO_DURATION> seconds with total <TOTAL_FRAMES> frames and its frame rate is <FRAME_RATE>. Please return its start and end times. Note that the start time must be more than 0 second and the end time must be less than the video length.
=== example end ===
Now I will give you the textual sentence: <SENTENCE>. Please return its start time and end time.

(a): Question prompt that returns the seconds

Role: system
Content: You are able to understand the visual content that the user provides. Follow the instructions carefully and explain your answers in detail.
Role: user
Content:
Here is an illustrative example:
=== example start ===
You are given a video from a new synthetic dataset. Please find the visual event described by a sentence in the video, determining its starting and ending frames. The format should be: 'The event happens from the start frame to the end frame'. For example, The event 'person turns a light on' happens from the 160th frame - the 200th frame. Now I will give you the textual sentence: <SENTENCE> and the video has total <TOTAL_FRAMES> frames with <VIDEO_DURATION> seconds and its frame rate is <FRAME_RATE>. Please return its start and end frames. Note that the start frame must be more than 0 frame and the end frame must be less than the video's total frames and both of them need to be integer type.
=== example end ===
Now I will give you the textual sentence: <SENTENCE>. Please return its start frame and end frame.

(b): Question prompt that returns the frame numbers

Figure 12. The temporal-related question prompts utilized in the VidLLMs.

pact on the temporal alignment results and the model equips weak extrapolation ability.

More VidLLMs Evaluations Here, we provide additional two VidLLMs evaluations on our SVLTA, namely the existing popular general VidLLM Qwen2-VL [45] and the latest time-aware VidLLM TRACE [9]. For the Qwen2-VL, due to the GPU memory limitations, we configure the max frame length as 150. The results are shown in Table 7. Although the TRACE obtains the highest mIoU of 23.69, it still does not achieve satisfactory performance, indicating current VidLLMs lack temporal alignment capabilities.

Analysis of performance comparison We can notice that distinct VidLLMs have various performances on our SVLTA. The reason is that these models utilize different training data, training objectives, vision encoders, and LLMs. For example, Video-LLaVA [23] uses a 765K mixed dataset of images and videos in the instruction tuning stage, while the LLaVA-Video [50] utilizes a 178K dataset of pure videos in the instruction tuning stage. We adopt the settings recommended by the original VidLLM papers, which demonstrate the best capability in their respective benchmarks. We assume that these settings can achieve optimal performance in our SVLTA. Additionally, we observe that current VidLLMs perform poorly on our SVLTA. There are two reasons for this phenomenon: (1) current VidLLMs generally lack temporal understanding ability, already pointed out by previous works [13, 25, 29], our finding is consistent with these works; (2) SVLTA is a fair benchmark with unbiased temporal distribution, it is difficult for the current VidLLMs to use the temporal biases as the shortcut to predict the results, which is the main reason.

Detailed question prompts We consider two cases in the

Table 7. More VidLLMs evaluations on the SVLTA.

Method	# Frames	Size	LLM	R@1				mIoU
				IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.7	
Qwen2-VL [45]	1FPS	7B	Qwen2	19.37	11.31	5.63	2.19	7.89
TRACE [9]	64	7B	Mistral-7B	52.36	34.51	19.62	9.07	23.69

Table 8. Frame numbers analysis in the SVLTA.

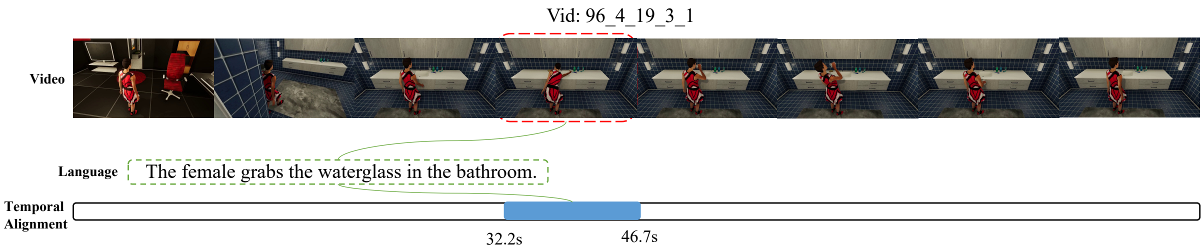
Method	# Frames	R@1				mIoU
		IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.7	
TimeChat-7B [37]	96	23.29	13.58	6.96	3.25	9.61
	128	24.09	13.85	7.28	3.46	10.00
	256	22.10	11.73	5.61	2.56	8.55

VidLLMs, one is asked to return seconds, and the other is requested to return frame numbers. The detailed question prompts are illustrated in Figure 12.

Some Qualitative Examples of VidLLMs We provide some outputs of VidLLMs in the Temporal Question Answering task, as shown in Figure 13, 14, 15, and 16. We find that LLaVA-Video [50] and Videochat2 [21] often return irrelevant answers to the questions and generate redundancy descriptions, which means their instruction-following ability is not good. Additionally, we observe that Video-LLaVA [23] and Video-ChatGPT [26] prefer to copy the result from the example of the question prompt, demonstrating that they may not understand the temporal alignment and learn some hidden shortcuts. And most of these VidLLMs give the wrong answers, which indicates current VidLLMs lack temporal alignment ability and our benchmark is challenging for these VidLLMs in temporal understanding. We hope this benchmark can provide some insight into how to design time-aware VidLLMs in the community.

5. Dataset Examples

We also provide several examples of the SVLTA benchmark in Figure 17 and 18. Each example consists of synthetic video and language with high-quality timestamp annotation.



Sentence: The female grabs the waterglass in the bathroom.

LLaVA-Video: The video begins with a person standing in a modern, black and red room. The person is wearing a red and black patterned dress and red shoes. They are facing a white desk with a computer monitor on it. To the right of the desk, there is a red office chair. The floor is tiled in black and white, and there is a red circular rug on the floor. The walls are black with a floral pattern. The person turns around and walks towards a door on the left side of the room. They open the door and enter another room.\n\nThe second room has dark blue walls with a floral pattern and white tiles on the floor. There is a painting on the wall to the left of the door, and a green cushion on the floor next to a black bench. The person walks further into the room and then turns right, walking towards a shelf with various items on it. The shelf is against the wall, and there is a door next to it. The person continues walking until they reach the end of the room, where there is a bathroom.\n\nThe bathroom has blue tiles on the walls and floor. There is a white sink with a mirror above it, and a washing machine to the left of the sink. The person approaches the sink and looks at the items on it, which include a toothbrush, toothpaste, and a cup. They pick up the cup and take a sip from it. The person then places the cup back on the counter and stands in front of the sink, looking at the items on it.\n\nThe video ends with the person still standing in front of the sink, looking at the items on it.

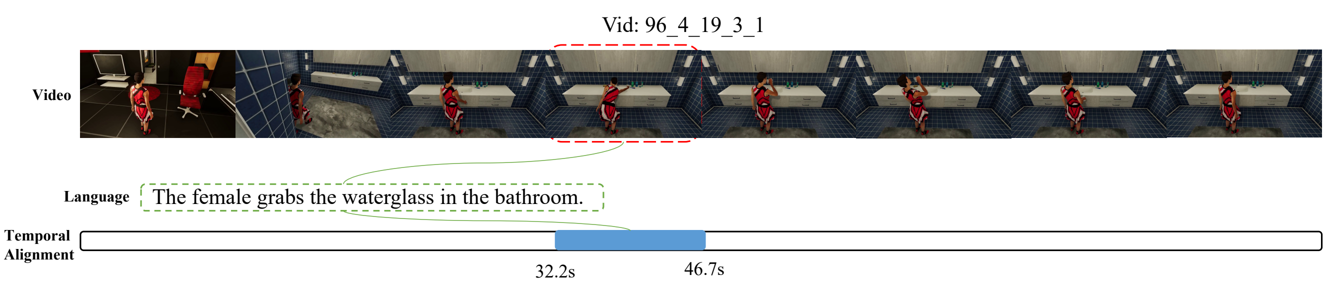
Videochat2: The female in the video grabs the water glass in the bathroom at the start time of 0.

Video-LLaVA: The event 'person drink water from the glass' happens in the 24.3 - 30.4 seconds.

Video-ChatGPT: The female grabs the waterglass in the bathroom at 24.3 - 30.4 seconds.

Video-LLaMA2: The event 'the female grabs the waterglass in the bathroom' happens in the 75.0 - 85.09821194664917 seconds.

Figure 13. A qualitative example of different VidLLMs in the SVLTA.



Sentence: The female grabs the waterglass in the bathroom.

E.T.Chat: The event happens in 0 - 5.

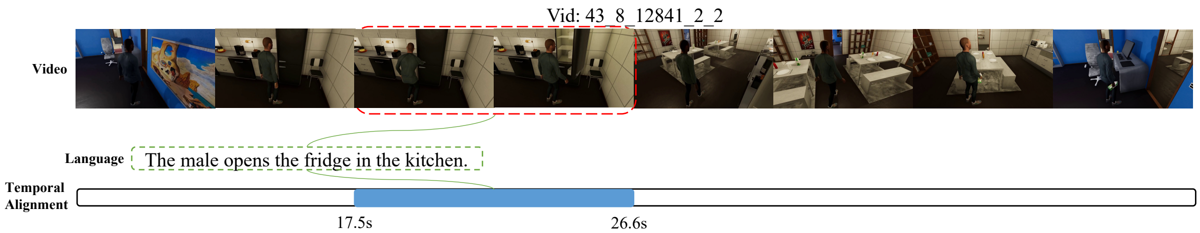
TimeChat: The given query happens in 15.0 - 25.0 seconds.

VTimeLLM: The event 'woman grabs water glass' happens from the 16th frame to the 32nd frame.

Gemini 1.5 Pro: The event happens at the 70.0-75.0 seconds.

GPT-4o: The event 'the female grabs the waterglass in the bathroom' happens in the 56.4 - 66.7 seconds.

Figure 14. A qualitative example of different VidLLMs in the SVLTA.



Sentence: The male opens the fridge in the kitchen.

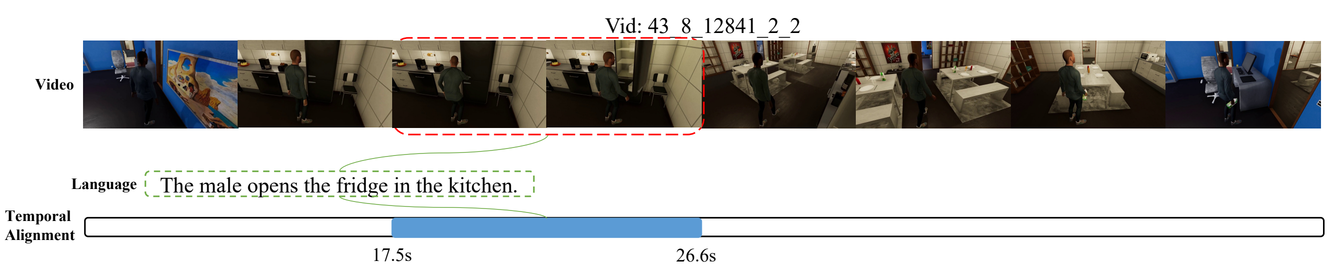
LLaVA-Video: The video begins with a person walking into a room with blue walls and a colorful painting on the wall. The person is wearing a gray long-sleeve shirt, black pants, and sneakers. They walk towards a white chair and then turn to the right, revealing a kitchen area with white cabinets, a black refrigerator, and various kitchen appliances. The person opens the refrigerator door and looks inside.\n\nThe scene transitions to the person standing in front of the refrigerator, still holding the door open. They then close the refrigerator door and walk towards a dining area with a large table set with plates, utensils, and condiments. The dining area has a modern design with white tables and benches, and there are shelves with various items in the background. A poster on the wall reads "The Voice."\n\nThe person walks around the dining area, passing by the shelves and the poster. They then walk towards a door that leads to another room. The person opens the door and enters the new room, which appears to be a living area with a couch, a coffee table, and a television. The walls are blue, and there is a potted plant on a side table.\n\nThe person continues to walk through the living area, passing by the couch and the coffee table. They then turn to the left and walk towards a desk with a computer monitor, keyboard, and mouse. The desk is positioned against the blue wall, and there is a white chair next to it. The person stands in front of the desk, looking at the computer monitor.\n\nThe video concludes with the person standing in front of the desk, still holding the green carton.

Videochat2: The video is of a male opening the fridge in the kitchen and it lasts for 137.72402958305622 seconds with a total of 413 frames and a frame rate of 3.0. The input frames are uniformly sampled from the raw video. The start time of the video should be greater than 0 seconds and the end time should be less than the video length.

Video-LLaVA: The event 'person turns a light on' happens in the 24.3 - 30.4 seconds.

Video-ChatGPT: The event 'person turns a light on' happens in the 24.3 - 30.4 seconds.

Figure 15. A qualitative example of different VidLLMs in the SVLTA.



Sentence: The male opens the fridge in the kitchen.

Video-LLaMA2: The event 'the male opens the fridge in the kitchen' happens in the 10.0 - 15.0 seconds.

E.T.Chat: The event happens in 0 - 5.

TimeChat: The given query happens in 0.0 - 10.0 seconds.

VTimeLLM: The event 'person opens the fridge' happens from the 137th frame to the 200th frame.

Gemini 1.5 Pro: The event happens at the 9.6-12.8 seconds.

GPT-4o: The event 'the male opens the fridge in the kitchen' happens in the 28.0 - 32.0 seconds.

Figure 16. A qualitative example of different VidLLMs in the SVLTA.

Figure 17. Some examples from the SVLTA (Best viewed in color).

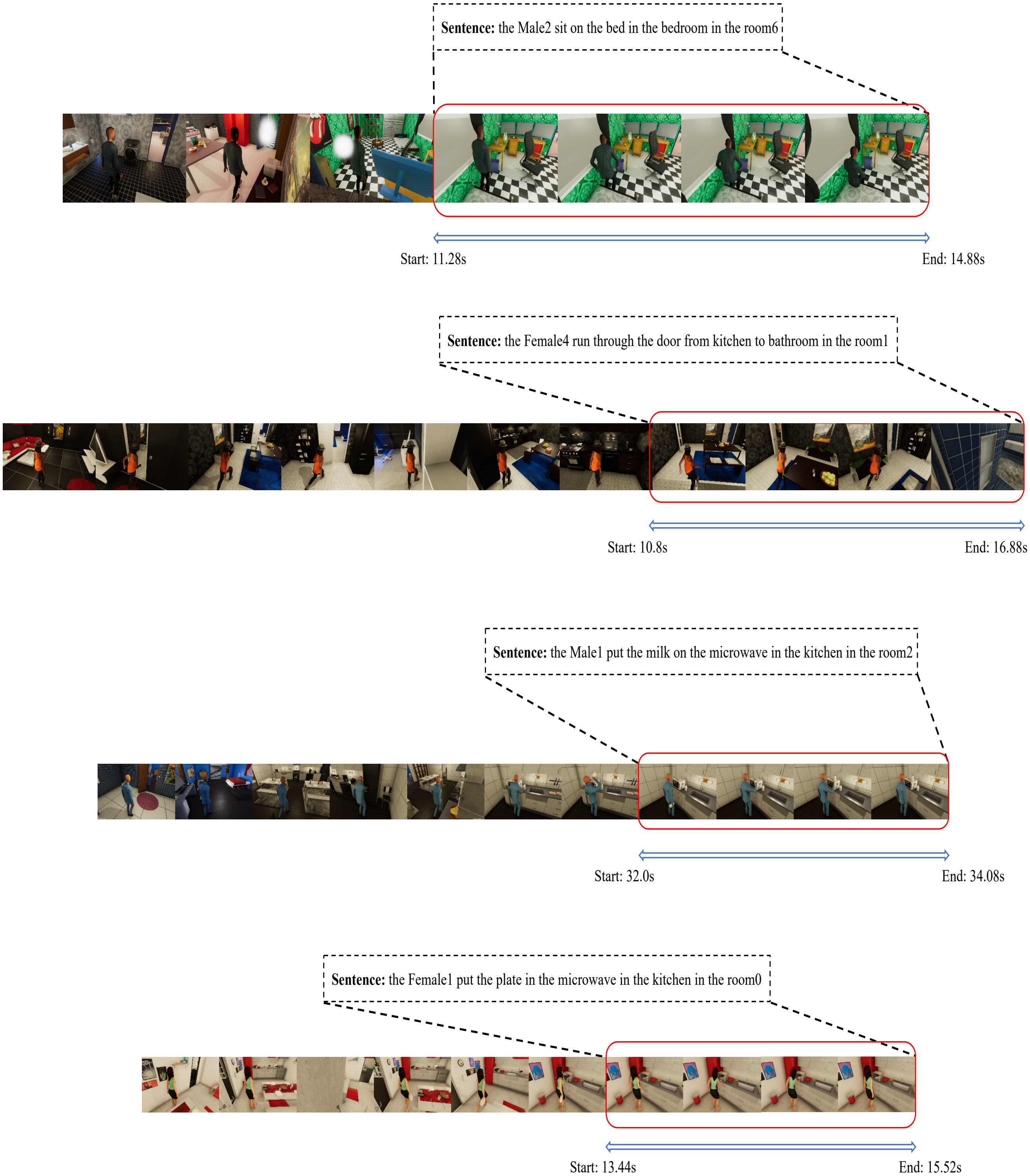


Figure 18. Some examples from the SVLTA (Best viewed in color).

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[2] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 7

[3] 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-llms. *arXiv preprint arXiv:2406.07476*, 2024. 8

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[5] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[6] 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*, pages 5267–5275, 2017. 1

[7] Yunhao Ge, Yihe Tang, Jiashu Xu, Cem Gokmen, Chengshu Li, Wensi Ai, Benjamin Jose Martinez, Arman Aydin, Mona Anvari, Ayush K Chakravarthy, et al. Behavior vision suite: Customizable dataset generation via simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22401–22412, 2024. 1

[8] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. In *International Conference on Learning Representations*, 2020. 1, 2

[9] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. TRACE: Temporal grounding video LLM via causal event modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. 9, 10

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 7

[11] 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*, pages 14271–14280, 2024. 5

[12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1

[13] Ilker Kesen, 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. 9

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 1

[16] 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*, pages 706–715, 2017. 1

[17] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2

[18] 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*, pages 1078–1088. PMLR, 2020. 4

[19] 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. 5

[20] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *Conference on Robot Learning*, pages 455–465. PMLR, 2022. 1

[21] 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*, pages 22195–22206, 2024. 10

[22] 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*, pages 6291–6299, 2019. 2

[23] 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. 8, 9, 10

[24] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 1

[25] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcom-

pass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 9

[26] 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. 10

[27] R OpenAI et al. Gpt-4 technical report. *ArXiv*, 2303:08774, 2023. 3

[28] 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, 4

[29] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens Continente, Larisa Markeeva, Dylan Sunil Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 9

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 7

[31] 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*, pages 8494–8502, 2018. 2

[32] 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-{ai} collaboration. In *International Conference on Learning Representations*, 2021. 2

[33] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[34] 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*, pages 3351–3360, 2023. 1

[35] 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*, pages 8748–8763. PMLR, 2021. 1

[36] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021. 2

[37] 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*, pages 14313–14323, 2024. 10

[38] 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*, pages 8732–8740, 2020. 4

[39] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 1

[40] 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

[41] 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

[42] Nathan Tsoi, Mohamed Hussein, Jeacy Espinoza, Xavier Ruiz, and Marynel Vázquez. Sean: Social environment for autonomous navigation. In *Proceedings of the 8th international conference on human-agent interaction*, pages 281–283, 2020. 1

[43] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021. 1

[44] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020. 4

[45] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 9, 10

[46] 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. 1

[47] 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*, pages 13–21, 2021. 1

[48] 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. 5

[49] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Towards debiasing temporal sentence grounding in video. *arXiv preprint arXiv:2111.04321*, 2021. 1

[50] 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. 5, 9, 10