

STEP: Enhancing Video-LLMs’ Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training

Supplementary Material

Overview

In this supplementary material we present:

- The detailed prompt for STSG construction and QRA refinement (Section A)
- The specifics of the backbone models (Section B)
- Implementation details (Section C)
- The evaluation details for each datasets (Section D)
- Additional experimental results (Section E)
- More Qualitative Visualizations (Section F)

A. Detailed Prompt Design

A.1. Prompt for STSG Construction

Here, we provide detailed prompts as well as their inputs and outputs, for each step of the STSG Construction.

Semantics Parsing

- **Step-1** : For each keyframe, first get its scene narrative, then extract the object information.

Step-1: Object Nodes Parsing

► **Input1:**

[INST]

Describe the keyframe of the video in one sentence, ensuring to include all main objects within the scene.

[/INST]

<Image>  </Image>

► **Output1:** [In the keyframe, a man and a woman carrying a red box are on a dock.]

► **Input2:**

[INST]

From the given sentence, the task is to extract meaningful objects. Note that the object is the entity or noun that being described in the sentence. Let’s take a few examples to understand how to extract meaningful objects.

Question: Given the sentence “Children play with red toys in the yard,” extract meaningful objects.
Answer: Meaningful objects are [children, toys, yard].

Question: Given the sentence “A beautiful woman walking a dog on top of a beach,” extract meaning-

ful objects.

Answer: Meaningful objects are [woman, dog, beach].

Question: Given the sentence “A mix of black and white cat jumps over a old fence,” extract meaningful objects.

Answer: Meaningful objects are [cat, fence].

Please answer the following question.

Question: Given the sentence “[A man and a woman carrying a red box are on a dock],” extract meaningful objects.

Answer:

[/INST]

► **Output2:**

[Meaningful objects are [man, woman, box, dock].]

► **Input3:**

[INST]

Given the object name: [man], if there is more than one, please provide the quantity and describe each instance of [man] with a corresponding unique attribute. The unique attribute should be something that only one instance has while the others do not. Use the following format for the output:

Object: [man]

Quantity: [number]

1. [man] 1: [distinct unique attribute]

2. [man] 2: [distinct unique attribute]

...

[/INST]

<Image>  </Image>

► **Output3:**

[Object: man

Quantity: 1

1. man1: wearing a light-colored shirt and shorts]

► **Input4:**

Given the object: [man], determine if it is a static object or a dynamic object. Note that static objects refer to background objects that may not move or change observably within a short time while dynamic objects are the opposite. If the object is static,

print 'Static', otherwise print 'Dynamic', no explanation is needed.

<Video>  </Videos>

► **Output4:** [Dynamic]

► **Step-2 :** For each object, we request a detailed narrative to extract its fine-grained attributes.

Step-2: Attribute Nodes Parsing

► **Input1:**

[INST] Given the object: [man], describe its attributes in the scene using one or two sentences that primarily include adjectives or present participles (-ing forms). Avoid any irrelevant details or complex explanations.

[/INST]

<Image>  </Image>

► **Output1:** [The man is standing on a wooden dock, wearing a white shirt and gray shorts and looking out over a serene marshy area.]

► **Input2:**

[INST]

For the given sentence, the task is to extract meaningful attributes for the given object. Each attribute should include its type and value. Let's take a few examples to understand how to extract meaningful attributes.

Question: Given the sentence "A woman with long and blonde hair is sitting on a park bench," extract meaningful attributes for the object "woman".

Answer: [{"object": "woman", "attributes": [{"hair": "long and blonde", "status": "sitting"}]}

Question: Given the sentence "A tall man in a red jacket is walking down the street," extract meaningful attributes for the object "man".

Answer: [{"object": "man", "attributes": [{"height": "tall", "clothing": "red jacket", "action": "walking"}]}

Question: Given the sentence "The blue car is parked," extract meaningful object label and attributes for the object "car"

Answer: [{"object": "car", "attributes": [{"color": "blue", "status": "parked"}]}

Please answer the following question.

Question: Given the sentence "[The man is standing on a wooden dock, wearing a white shirt and gray shorts and

looking out over a serene marshy area]." Extract meaningful attributes for the object "[man]"

Answer:

[/INST]

► **Output2:**

```
[{"object": "man", "attributes": [{"clothing": "white shirt", "clothing": "gray shorts", "status": "standing"}]}
```

► **Step-3 :** For each pair of object nodes, we construct < subject, predicate, object > triples to capture their relational correspondence.

Step-3: Relation Edges Parsing

► **Input1:**

[INST]

Given the object: [man] in the provided image, describe all spatial or contact relationship between [man] and other objects [woman, dock, box] in the image. Ensure the subject of every sentence is [man], and do not include detailed descriptions of other objects.

<Image>  </Image>

[/INST]

► **Output1:**

[The man is standing at the end of a dock, waiting for the woman.]

► **Input2:**

[INST] From the given sentence, the task is to extract meaningful triplets formed as <subject, predicate, object>. Note that the subject must be the given object, and the object is the entity or noun that is affected by the action or is receiving the action. The predicate is a verb or adjective without auxiliary verb, and is represented without the tense (e.g., are, being). Let's take a few examples to understand how to extract meaningful triplets.

Question: Given the sentence "a slice of bread is covered with a sour cream and quacamole," extract meaningful triplets for the given object "bread".

Answer: Meaningful triplets are <bread, covered with, sour cream>, <bread, covered with, guacamole>.

Question: Given the sentence "A clock sitting on a floor next to a woman's feet," extract meaningful triplets for the given object "clock".

Answer: Meaningful triplets are <clock, sitting on, floor>, <clock, next to, feet>.

Question: Given the sentence “One person sits in a chair looking at her phone while another rests on the couch,” extract meaningful triplets for the given object “person”.

Answer: Meaningful triplets are <person, sits in, chair>, <person, looking at, phone>, <person, rests on, couch>.

Please answer the following question.

Question: Given the sentence “[The man is standing at the end of a dock, waiting for the woman],” extract meaningful triplets for the given object [man].

Answer:

► **Output2:**

[Meaningful triplets are <man, standing at, the end of dock>, <man, waiting for, woman>]

- **Verification** : Filter the hallucinated nodes.

Step-1: Identify the Cross-frame Object Nodes

► **Input1:**

[INST]

For the given video, determine if [the box whose color is red] truly exists in the given video.

[/INST]

<Video>  </Videos>

► **Output1:** [Yes]

Dynamic Merging

- **Step-1** : Determine if the nodes in different keyframes are identical.

Step-1: Identify the Cross-frame Object Nodes

► **Input1:**

[INST]

For the given video, determine if [woman walking on the dock] in the first half and [woman kneeling on the dock] in the second half are the same object. Just output ‘Yes’ or ‘No’.

[/INST]

<Video>  </Videos>

► **Output1:** [Yes]

- **Step-2** : If the two identical object nodes are static, just merge them into a single unified node, dynamic, the prompt should be:

Step-2: Connect the Dynamic Nodes

► **Input1:**

Given the same object [woman walking on the dock] in the first half video and [woman kneeling on the dock] in the second half video, describe the action of this object in the video.

► **Output1:** [The woman fell down on the dock.]

► **Input2:**

For the given sentence, the task is to extract meaningful action for the given object. The action should be described using a verb or verb phrase. Let’s take a few examples to understand how to extract.

Question: Given the sentence “The man walked on the street,” extract meaningful action for the object “man”.

Answer: [“action”:“walking on the street”]

Question: Given the sentence “The cat jumped on the bed,” extract meaningful action for the object “cat”.

Answer: [“action”:“jumping on the bed”]

Question: Given the sentence “The dog chased the ball in the park,” extract meaningful action for the object “dog”.

Answer: [“action”: “chasing the ball in the park”]

Please answer the following question.

Question: Given the sentence “[The woman fell down on the dock],” extract meaningful action for the object “[woman]”.

Answer:

► **Output2:**

[["action": "falling down on the dock"]]

► **Input3:**

For the given action, extract the predicate and object. The predicate is a verb or verb phrase, and the object is the entity or noun that is affected by the action. Let’s take a few examples to understand how to extract.

Question: Given the action “walking”, extract the

predicate and object.

Answer: [{"predicate": "walking", "object": "None"}]

Question: Given the action "jumping on the bed", extract the predicate and object.

Answer: [{"predicate": "jumping on", "object": "bed"}]

Question: Given the action "sitting on the chair", extract the predicate and object.

Answer: [{"predicate": "sitting on", "object": "chair"}]

Please answer the following question.

Question: Given the action "[falling down on the dock]", extract the predicate and object.

Answer:

► Output3:

```
[{"predicate": "falling down on",
"object": "dock"}]
```

Cross-clip Bridging

► **Step-1** : Determine if the nodes in different clips represent the same object. If they do, connect them with a reference edge.

Step-1: Identify the Cross-clip Object Nodes

► Input1:

[INST]

For the given video, determine if [woman walking on the dock] in the first clip and [woman sitting on the blanket] in the second clip are the same object. Just output 'Yes' or 'No'.

[/INST]

<Video>  </Videos>

► Output1:

[Yes]

A.2. Prompt for QRA refinement

Step-1: Identify the Object Nodes

► Input1:

[INST]

I have the question "[q]", the rationale "[r]", and the answer "[a]" about the video. Due to template generation, these may contain grammatical errors, logical inconsistencies, or unclear semantics.

Please correct them and return the improved version in JSON format : {"Q": "question", "R": "rationale", "A": "answer"}

[/INST]

B. Backbone Models Specifications

B.1. VideoChat2 (7B)

VideoChat2 is an advanced Video-LLM designed to handle complex video-language understanding tasks [5]. Its architecture integrates a visual encoder, a Q-Former module for compressing and aligning visual data, and a large language model (LLM) for advanced reasoning and response generation. The training process follows a progressive three-stage paradigm: vision-language alignment to link visual and textual features, vision-language connection to integrate visual embeddings with the LLM, and instruction tuning to enhance multi-modal instruction-following capabilities.

The version we use is VideoChat2 with Mistral 7B. To endow the initial model with a baseline level of instruction-following capability, we selected 60K samples from the instruction-tuning dataset of VideoChat2 for foundational instruction tuning on the pre-trained (vision-language aligned) VideoChat2 model, which results in VideoChat2*. The details are presented in Table 1.

Modality	Dataset	Description	# Samples
Image-Text	CoCo	Captioning	5K
	VQAv2	General	10K
Video-Text	VideoChat	Captioning	5K
	SthSthV2	Classification	5K
	VideoChatGPT	General	10K
	NExTQA	Reasoning	10K
	CLEVRER	Reasoning	5K
	Tgif	Reasoning	10K
Total			60K

Table 1. Data mixture of initial instruction tuning

B.2. VILA (3B)

VILA is a Video-LLM evolved from an Image-LLM [6]. It employs an autoregressive architecture that integrates tokenized images with text inputs. The training process includes three stages: projector initialization for modality alignment, visual-language pre-training on interleaved datasets to enhance zero-shot and in-context learning, and visual instruction-tuning to optimize task-specific performance. In the instruction-tuning stage, VILA incorporates video-based instruction datasets, extending its capabilities to spatio-temporal reasoning as a lightweight Video-LLM.

Since VILA does not provide checkpoints after the visual-language pre-training stage, we are unable to perform instruction-tuning with a small amount of data. Instead, we directly use the fully instruction-tuned model as the initial model in our **STEP** framework.

C. Implementation Details

STEP method first constructs STSGs for diverse videos in the raw video datasets during the symbolic structure induction phase. It then utilizes stepwise graph-driven rationale learning to systematically extract reasoning-rich training data and perform explicit supervision for the reasoning process. This process is iterative rather than a one-step procedure, to better leverage the model’s enhanced capabilities in each cycle, enabling progressive improvement in reasoning through repeated data generation and training, ultimately forming a self-enhancing closed-loop mechanism.

We selected three video datasets: ActivityNet [2], Ego4D [3], and Charades [8] as the raw video sources. ActivityNet tracks human actions in videos, Ego4D captures life from a first-person perspective, and Charades records home-based daily tasks. The diversity in video nature—ranging from third-person action tracking to first-person experiences and everyday household activities—demonstrates how **STEP** can effectively adapt to different contexts, enabling robust symbolic structure induction and reasoning across a broad spectrum of real-world scenarios. The total number of QRAs extracted from these three raw video datasets is as follows:

Raw Video Sources	# QRA Samples
ActivityNet	15K
Ego4D	5K
Charades	10K
Total	30K

Table 2. Number of QRAs extracted from three raw video datasets

For the **STEP** model, we perform explicit rationale supervision using the derived QRA training samples. Additionally, to mitigate catastrophic forgetting and enhance overall video comprehension, the training data is augmented by incorporating previously utilized instruction-tuning data from VideoChat* and VILA, with $L_{\text{rationale}} = 0$ to retain only L_{answer} for these datasets.

For the *Instruct* model, we utilize the annotated labels from the same video datasets—ActivityNet, Ego4D, and Charades—to generate a annotated training set. The number of QA samples is the same as Table 2. Since these traditional datasets only contain questions and answers, conventional training is applied with $L_{\text{rationale}} = 0$, focusing solely on answer-based supervision.

For the *Distillation* model, the raw videos and training sample size are consistent with those used for the **STEP** model, ensuring that the distillation process is based on the same foundational data.

D. Evaluation Details

For the compositional reasoning benchmarks AGQA [4] and STAR [10], and the standard QA datasets MSVD-QA [12], MSRVT-QA [13], and ActivityNet-QA [2], we apply the same evaluation protocol as [1], which reports two key metrics: accuracy (the percentage of correctly answered questions) and the average score (where ChatGPT rates each response on a scale of 1-5 and calculates the mean of these scores). To ensure consistent comparisons, all evaluations utilize the same GPT model, “gpt-3.5-turbo” [11]. All evaluation prompt is “Based on the video, respond to this question with a short answer: {question}”.

For the comprehensive understanding benchmarks MVBench [5] and TempCompass [7], the long video understanding benchmarks MovieChat-1K [9] and MLVU [14], we adhere strictly to the evaluation methodologies specified in their respective frameworks.

Across all evaluations, each video is represented by 8 frames as input, with a resolution of 224×224 . This setup ensures consistency in video processing and facilitates a fair comparison of model performance across diverse benchmarks.

Type	Dataset	Split	Metrics
Compositional Reasoning	AGQA	test	Accuracy, Score
	STAR	val	Accuracy, Score
Standard QA	MSVD-QA	test	Accuracy, Score
	MSRVTT-QA	test	Accuracy, Score
	ActivityNet-QA	test	Accuracy, Score
Comprehensive Understanding	TempCompass	-	VQA Accuracy
	MVBench	-	VQA Accuracy
Long video Understanding	MovieChat-1K	-	Accuracy, Score
	MLVU	-	VQA Accuracy

Table 3. Evaluation metrics for each evaluation benchmarks

E. Additional Experimental Results

E.1. TempCompass

The TempCompass benchmark [7] encompasses four task types: Multi-Choice QA, Yes/No QA, Caption Matching, and Caption Generation. The detailed results can be found in Tables 4, 5, 7.

The results underscore that **STEP** achieves substantial advancements across all task categories in TempCompass.

Method	Action		Direction		Speed		Event	Attribute Change				Avg
	fine	coarse	object	camera	absolute	relative	order	color	size	combined	other	
Random	28.4	29.3	28.3	26.3	30.6	33.0	32.2	30.1	28.9	26.4	25.9	29.9
VideoChatGPT(7B)	28.8	62.2	33.7	26.1	28.2	28.5	37.1	26.9	31.1	35.0	33.3	35.2
Video-LLaVA(7B)	54.9	83.2	31.7	33.7	46.0	33.2	41.4	39.7	40.2	35.0	55.6	44.7
VideoChat2(7B)*	73.9	90.8	37.4	45.7	54.8	32.6	37.7	34.6	42.4	46.7	27.8	49.0
VideoChat2* <i>Instruct</i>	82.4	95.6	37.4	44.8	54.8	35.3	40.7	43.6	42.4	45.0	38.9	51.8
VideoChat2* <i>Distillation</i>	84.0	94.8	40.7	43.3	58.9	34.2	43.1	48.7	44.8	50.7	38.8	53.6
VideoChat2* STEP	85.7	95.7	41.2	43.5	60.1	33.7	46.4	51.3	51.5	53.5	38.9	55.4
VILA(3B)	74.3	93.5	33.7	37.0	52.4	41.9	40.4	62.8	46.2	38.7	50.0	51.4
VILA <i>Instruct</i>	78.9	94.8	34.8	38.9	47.8	43.2	39.7	65.7	51.5	38.9	51.7	52.6
VILA <i>Distillation</i>	80.4	94.8	37.0	38.0	50.2	36.8	43.0	74.8	50.0	41.0	51.6	53.5
VILA STEP	83.0	95.7	38.0	40.2	47.8	34.2	43.4	78.2	56.8	41.0	52.1	54.4

Table 4. Model performance comparison on multi-choice QA for TempCompass on VideoChat2* and VILA

Method	Action		Direction		Speed		Event	Attribute Change				Avg
	fine	coarse	object	camera	absolute	relative	order	color	size	combined	other	
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
VideoChatGPT(7B)	50.2	54.5	50.0	50.0	49.7	49.4	51.0	50.0	50.0	50.0	50.0	50.7
Video-LLaVA(7B)	58.4	87.6	51.4	52.7	50.8	50.0	49.2	52.2	50.0	53.0	45.8	56.4
VideoChat2(7B)*	57.6	89.2	50.3	50.7	65.1	49.4	53.4	55.9	52.1	54.0	50.0	58.3
VideoChat2* <i>Instruct</i>	58.0	90.4	53.5	50.0	66.4	49.1	53.6	58.1	53.7	56.0	50.0	59.4
VideoChat2* <i>Distillation</i>	59.2	91.6	53.5	52.1	68.3	51.1	54.0	58.6	54.8	57.0	49.8	60.4
VideoChat2* STEP 60.1	91.7	52.9	52.0	68.1	51.3	56.4	59.3	54.5	57.7	50.3	60.9	
VILA(3B)	65.8	87.4	52.3	52.7	52.9	50.6	54.2	63.2	61.2	56.0	62.5	60.1
VILA <i>Instruct</i>	67.8	90.5	52.1	53.0	53.2	54.2	55.4	63.8	64.7	57.3	62.9	61.8
VILA <i>Distillation</i>	68.3	91.5	53.2	52.9	53.8	53.7	55.9	64.7	65.0	58.6	63.5	62.3
VILA STEP	69.1	91.7	53.6	53.0	53.7	53.9	56.4	65.3	65.5	59.2	63.9	62.7

Table 5. Model performance comparison on yes/no QA for TempCompass on VideoChat2* and VILA

Specifically, in the Multi-Choice QA and Yes/No QA tasks, **STEP** exhibits superior performance in fine-grained dimensions such as fine action recognition, event order comprehension, and attribute change detection. These dimensions necessitate sophisticated spatio-temporal reasoning and a precise understanding of video semantics. Furthermore, **STEP** reduces hallucinations by incorporating multi-granular details and reasoning steps aligned with the query, enabling it to interpret complex temporal-spatial dependencies and tackle challenging reasoning tasks with greater accuracy and interpretability.

E.2. MVBench

MVBench [5] is a comprehensive video benchmark designed to evaluate temporal reasoning in dynamic tasks. It includes 20 tasks that assess various temporal skills, from perception to cognition, aiming to test the real-world applicability of Video-LLMs. MVBench enhances static tasks

by incorporating dynamic elements, such as transforming “object location” into “motion direction,” and uses an automated annotation pipeline to ensure fairness and accuracy.

The performance metrics across cognitive and visual tasks are as follows: Action Sequence (AS) measures accuracy in identifying action sequences in videos. Action Prediction (AP) evaluates the model’s ability to predict subsequent actions. Action Antonym (AA) assesses accuracy in distinguishing opposite actions, while Fine-grained Action (FA) gauges precision in recognizing subtle action variations. Unexpected Action (UA) tests the detection of abnormal actions. Object Existence (OE) evaluates object presence or absence recognition, and Object Interaction (OI) assesses understanding of object relationships. Object Shuffle (OS) measures object tracking through rearrangements. Moving Direction (MD) evaluates movement direction recognition, and Action Localization (AL) measures spatial-temporal action pinpointing. Scene Transition (ST)

Model	Avg	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI
Random	27.3	25.0	25.0	33.3	25.0	25.0	33.3	25.0	33.3	25.0	25.0	25.0	33.3	25.0	33.3	33.3	25.0	33.3	25.0	20.0	30.9
VideoChatGPT(7B)	32.7	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	48.5	29.0	33.0	29.5	26.0	35.5
VideoChat(7B)	35.5	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	26.5	41.0	23.5	23.5	36.0
VideoChat2(7B)*	47.0	63.5	47.0	66.5	43.0	56.0	51.5	68.0	35.0	20.0	32.5	82.5	38.5	36.5	54.5	45.0	51.0	42.5	28.5	38.0	39.0
VideoChat2* <i>Inst</i>	47.9	64.0	49.0	68.5	43.0	58.0	53.5	68.5	36.5	22.0	33.5	83.5	39.0	36.0	54.5	46.5	52.0	42.0	31.0	38.5	39.5
VideoChat2* <i>Dist</i>	48.5	65.0	49.5	69.0	45.5	57.0	54.5	70.0	36.5	21.5	33.5	85.0	39.0	36.0	55.0	46.5	52.5	43.0	32.0	41.0	38.5
VideoChat2* STEP	49.2	67.0	49.5	69.0	46.5	57.0	54.5	71.5	36.5	21.0	34.0	85.5	39.5	37.5	56.0	46.5	53.5	43.5	33.0	42.5	39.5
VILA(3B)	43.0	56.0	48.5	57.0	39.0	51.5	52.5	54.0	26.0	27.5	28.0	74.5	37.0	30.5	53.5	39.0	40.0	47.0	31.5	37.0	30.0
VILA <i>Instruct</i>	44.2	58.0	50.5	59.0	41.5	52.5	53.5	55.5	28.0	27.5	29.5	76.5	38.5	30.0	53.5	39.0	41.0	48.0	32.0	38.5	32.0
VILA <i>Distillation</i>	44.8	60.5	50.5	60.0	43.5	54.0	54.5	55.5	28.5	27.0	29.0	76.5	38.5	30.5	53.0	39.0	41.0	49.0	32.5	40.5	33.0
VILA STEP	45.6	60.5	52.5	62.0	44.0	53.5	54.0	57.5	32.0	26.5	30.0	78.5	39.5	30.0	53.5	40.0	42.5	49.0	33.5	41.5	32.0

¹VideoChat2* *Inst* refers to VideoChat2* *Instruct*, VideoChat2* *Dist* refers to VideoChat2* *Distillation*.

Table 6. Model performance comparisons on various dimensions in MVBench

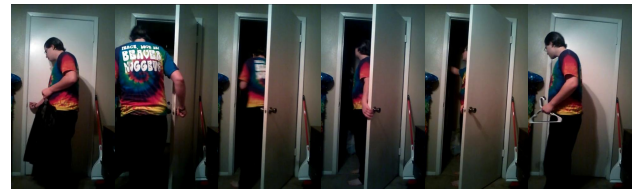
gauges setting changes recognition, Action Count (AC) tests action counting accuracy, and Moving Count (MC) focuses on counting moving objects. Moving Attribute (MA) assesses recognition of motion-related traits like speed or trajectory. State Change (SC) measures detection of object state changes, and Fine-grained Pose (FP) evaluates precise posture recognition. Character Order (CO) assesses narrative sequence recognition, Egocentric Navigation (EN) tests first-person spatial understanding, Episodic Reasoning (ER) measures episodic inference capacity, and Counterfactual Inference (CI) evaluates reasoning in hypothetical scenarios. Avg represents the average performance across tasks.

From the Table 6, it is evident that the **STEP** method leads to a significant improvement in the performance of MVBench for both models, particularly in tasks such as Action Sequence, Character Order, Fine-grained Action, and Episodic Reasoning. This underscores its enhanced ca-

pability in capturing and reasoning over complex spatio-temporal dependencies.

F. More Qualitative Visualizations

We present additional qualitative visualizations to illustrate how **STEP** enhances spatio-temporal reasoning and fine-grained detail capture across a range of diverse scenarios, using VideoChat2* as the baseline.



Question: What will the person do next?

- A. Tidy up the table.
- B. Open the box. VideoChat2*
- C. Tidy up the closet/cabinet. VideoChat2* STEP
- D. Hold the dish.

Figure 1. Qualitative examples of action prediction



Question: What happened after the person held the dish?

- A: Sat on the bed.
- B: Ate the sandwich.
- ✔ Put down the food. VideoChat2* STEP
- ✗ Took the food. VideoChat2*

Figure 2. Qualitative examples of action sequence

	Caption Matching	Caption Generation
Random	50	30
VideoChatGPT (7B)	51.8	31.8
Video-LLaVA(7B)	63.7	34.8
VideoChat2 (7B)*	67.9	45.3
VideoChat2* <i>Instruct</i>	68.6	47.8
VideoChat2* <i>Distillation</i>	69.4	48.5
VideoChat2* STEP	70.3	49.7
VILA (3B)	65.2	39.1
VILA <i>Instruct</i>	66.2	40.6
VILA <i>Distillation</i>	66.7	40.8
VILA STEP	67.0	40.8

Table 7. Model performance comparison on Caption Matching and Caption Generation for TempCompass on VideoChat2* and VILA.

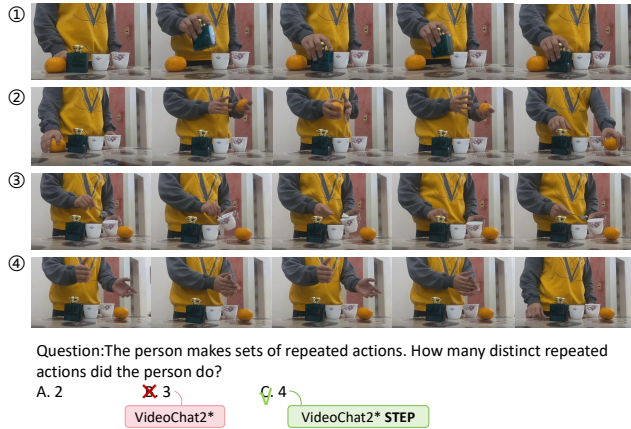


Figure 3. Qualitative examples of action count

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5
- [2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015. 5
- [3] 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*, pages 18995–19012, 2022. 5
- [4] 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*, pages 11287–11297, 2021. 5
- [5] 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. 4, 5, 6
- [6] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 4
- [7] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompas: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 5
- [8] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 5
- [9] Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 5
- [10] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 5
- [11] Wenhao Wu. Freeva: Offline mllm as training-free video assistant. *arXiv preprint arXiv:2405.07798*, 2024. 5
- [12] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 5
- [13] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5
- [14] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 5