MVBench

A Comprehensive Multi-modal Video Understanding Benchmark

Kunchang Li[1,2,3], Yali Wang[1,3♡], Yinan He[3], Yizhuo Li[4,3], Yi Wang[3], Yi Liu[1,2,3], Zun Wang[3], Jilan Xu[5.3], Guo Chen[6.3], Ping Luo[4,3], Limin Wang[6,3♡], Yu Qiao[3,1♡]

(♡ corresponding authors)

[1] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

[2] University of Chinese Academy of Sciences [3] Shanghai AI Laboratory

[4] The University of Hong Kong [5] Fudan University [6] Nanjing University

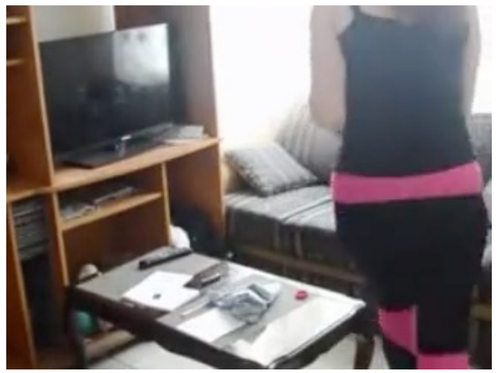
How to evaluate MLLM?

🤔 **How about video understanding?**

💡 **The tasks hard to be solved with single frame!**

Task definition

(1) ***Action Sequence:*** Retrieve the events occurring before or after a specific action.



Q: What happened after the person took the food?

(A) Ate the medicine.
(B) Tidied up the blanket.
(C) Put down the cup/glass/bottle.
(D) Took the box.

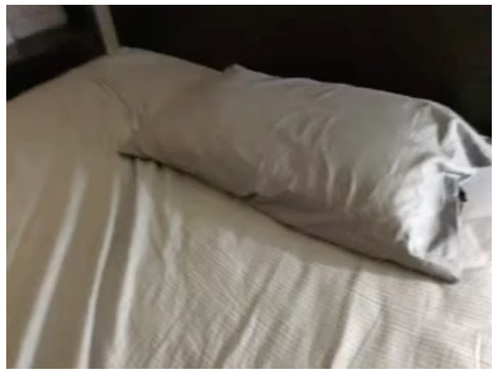
(2) ***Action Prediction:*** Infer the subsequent events based on the current actions.

Q: What will the person do next?

(A) Put down the pillow.
(B) Open the door.
(C) Take the book.
(D) Open the closet/cabinet.

Task definition

(3) ***Action Antonym:*** Distinguish the correct action from two inversely ordered actions.



Q: What is the action performed by the person in the video?

(A) Not sure.
(B) Scattering something down.
(C) Piling something up.

(4) ***Fine-grained Action:*** Identify the accurate action from a range of similar options.

Q: Which one of these descriptions correctly matches the actions in the video?

(A) bathing
(B) watering
(C) washing
(D) bubbling

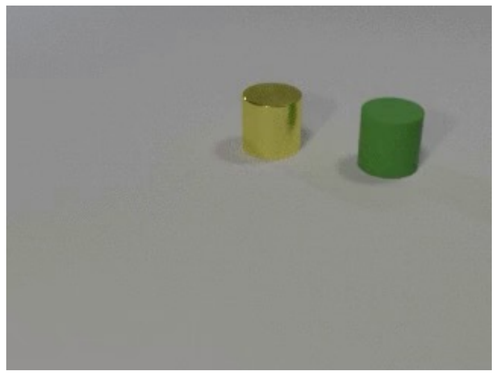
Task definition

(5) ***Unexpected Action:*** Detect surprising actions in videos characterized by humor, creativity, or magic.

Q: What surprising action did the dog do to the snake in the video?

(A) The dog barked at the snake, startling it.
(B) The dog pulled the snake towards itself, making it jump.
(C) The dog bit the snake, causing it to hiss.
(D) The dog stepped on the snake, making it squirm.

(6) ***Object Existence:*** Determine the existence of a specific object during a particular event.



Q: Are there any moving green objects when the video ends?

(A) not sure
(B) yes
(C) no

Task definition

(7) ***Object Interaction:*** Identify the object that participates in a particular event.

Q: Which object was taken by the person?

(A) The dish.
(B) The box.
(C) The blanket.
(D) The paper/notebook.

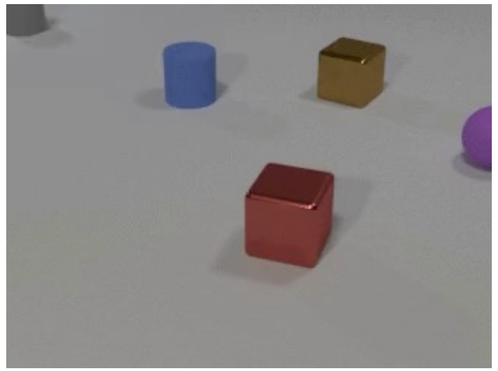
(8) ***Object Shuffle:*** Locate the final position of an object in an occlusion game.

Q: Where is the hidden object at the end of the game from the person's point of view?

(A) Under the first object from the left.
(B) Under the third object from the left.
(C) Under the second object from the left.

Task definition

(9) ***Moving Direction:*** Ascertain the trajectory of a specific object's movement.



Q: What direction is the gray cylinder moving in within the video?

(A) Up and to the right.
(B) Up and to the left.
(C) The object is stationary.
(D) Down and to the right.

(10) ***Action Localization:*** Determine the time period when a certain action occurs.

Q: During which part of the video does the action **'person sitting on a couch'** occur?

(A) In the middle of the video.
(B) At the end of the video.
(C) Throughout the entire video.
(D) At the beginning of the video.

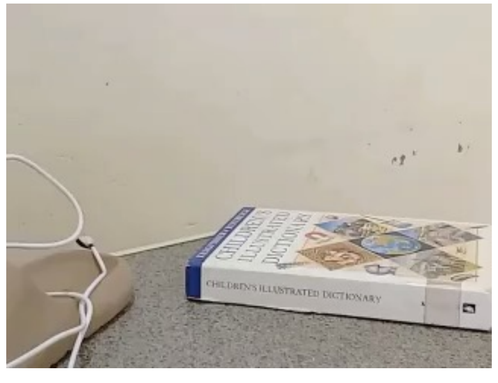
Task definition

(11) ***Scene transition:*** Determine how the scene transitions in the video.

Q: What's the right option for how the scenes in the video change?

(A) From the garden to the mall.
(B) From the playground to the office.
(C) From the beach to the mountaintop.
(D) From the creek to the stairs.

(12) ***Action Count:*** *Calculate how many times a specific action has been performed.*

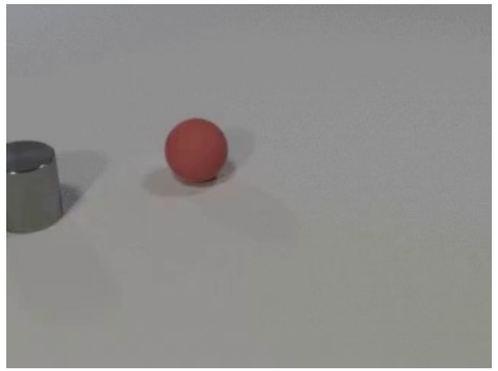


Q: How many times did the person launch objects on the table?

(A) 3
(B) 2
(C) 4

Task definition

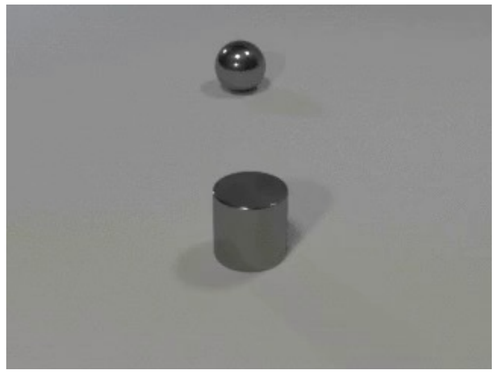
(13) ***Moving Count:*** Calculate how many objects have performed a certain action.



Q: How many red objects are moving?

(A) 3
(B) 5
(C) 4
(D) 2

(14) ***Moving Attribute:*** Determine the appearance of a specific moving object at a given moment.



Q: What color is the object that is stationary?

(A) gray
(B) green
(C) yellow
(D) blue

Task definition

(15) ***State Change:*** Determine whether the state of a certain object changes throughout the video.

Q: What can you say about the temperature of the water being poured?

(A) No water was poured.
(B) The water seems hot.
(C) The water seems cold.

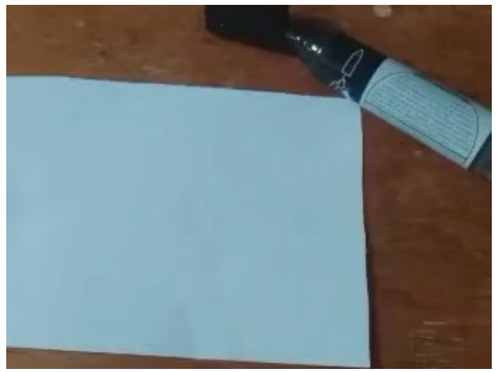
(16) ***Fine-grained Pose:*** Identify the accurate pose category from a range of similar options.

Q: What is the pose performed by the person in the video?

(A) hopping
(B) pick up
(C) stand up
(D) sit down

Task definition

(17) ***Character Order:*** Determine the order in which the letters appear.



Q: What letter did the person write first on the paper?

(A) l
(B) v
(C) e

(18) ***Egocentric Navigation:*** *Forecast the subsequent action, based on an agent's current navigation instructions.*

Q: For an agent following instruction: **"Go up the stairs. Take a left at the top of the stairs. Go into the bedroom on the left. Stop in the doorway."** What is the next action it should take?

(A) Turn right and move forward
(B) Turn left and move forward
(C) Move forward
(D) Stop

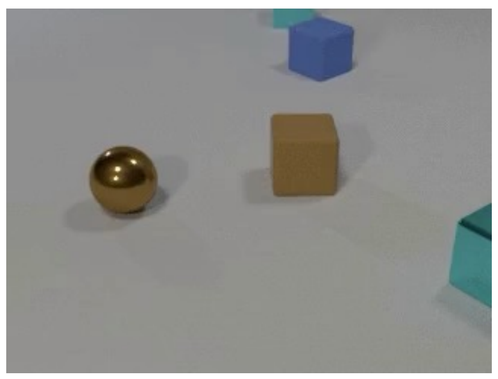
Task definition

(19) ***Episodic Reasoning:*** Perform reasoning on the characters, events, and objects within an episode of a TV series.

Q: Why did Castle dress like a fairy when he was speaking to Emily?

(A) To get her to trust him
(B) He secretly loved fairies
(C) He lost a bet with Emily
(D) It was dress like a fairy day at school
(E) Mrs Ruiz made him dress up

(20) ***Counterfactual Inference:*** *Consider what might happen if a certain event occurs.*



Q: Which of the following will happen if the cylinder is removed?

(A) The cyan rubber object and the blue cube collide
(B) The brown cube collides with the metal cube
(C) The cyan rubber object and the metal cube collide
(D) The cyan rubber cube collides with the sphere

Automatic QA Generation

Task Selection

Public video datasets with high-quality annotations

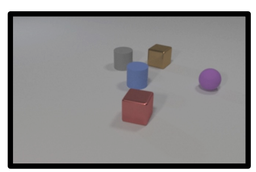
Various scenes

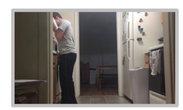

Perception Test


MiT V1

Charades-STA, NTU RGB+D PAXION, FunQA, ...


TVQA


CLEVRER


STAR

Object

Position

Appearance

Moving Direction

Action Localization

Data Filtration

Video Diversity — Each QA pair corresponds to a distinct video

Temporal Sensitivity
- ✗ Too short: minimal movement
- Intermediate duration
- ✗ Too long: complicated context

Question Difficulty
- ✗ Too easy: indistinguishable
- Proper question
- ✗ Too hard: inseparable

QA Generation

Have options?
- *yes* → Directly adopt QA
- *no* → Generate QA with video annotations

LLM asks **question** ⇐ task definition

What direction is the gray cylinder moving wiin the video?

Template-based **option** candidates

Up and to the left; Up and to the right; Down and to the left; Down and to the right; The object is stationary.

Option Processing

Order Shuffle — Options are randomly selected and shuffled

Length Check — *LLM* — Different options should have similar and reasonable text lengths

Evaluation: *Prompt Design*

Q: *What direction is the gray cylinder moving within the video?*
(A) *Up and to the right.*
(B) *Up and to the left.*
(C) *The object is stationary.*
(D) *Down and to the right.*

(D)

VideoChat, mPLUG-Owl, LLaVA, Otter

System Prompt: Consider temporal evolution

> *Carefully watch the video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons. Based on your observations, select the best option that accurately addresses the question.*

Answer Prompt: Must output option

> *Best Option: (*

MVBench Evaluation

Model	LLM	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
Image MLLMs: *Following [11], all models take **4** frames as input, with the output embeddings concatenated before feeding into the LLM.*																						
mPLUG-Owl-I [88]	LLaMA-7B	29.4	25.0	20.0	44.5	27.0	23.5	36.0	24.0	34.0	23.0	24.0	34.5	34.5	22.0	31.5	40.0	24.0	37.0	25.5	21.0	37.0
LLaMA-Adapter [96]	LLaMA-7B	31.7	23.0	28.0	51.0	30.0	33.0	53.5	32.5	33.5	25.5	21.5	30.5	29.0	22.5	41.5	39.5	25.0	31.5	22.5	28.0	32.0
BLIP2 [38]	FlanT5-XL	31.4	24.5	29.0	33.5	17.0	42.0	51.5	26.0	31.0	25.5	26.0	32.5	25.5	30.0	40.0	42.0	27.0	30.0	26.0	37.0	31.0
Otter-I [37]	MPT-7B	33.5	34.5	32.0	39.5	30.5	38.5	48.5	44.0	29.5	19.0	25.5	55.0	20.0	32.5	28.5	39.0	28.0	27.0	32.0	29.0	36.5
MiniGPT-4 [97]	Vicuna-7B	18.8	16.0	18.0	26.0	21.5	16.0	29.5	25.5	13.0	11.5	12.0	9.5	32.5	15.5	8.0	34.0	26.0	29.5	19.0	9.9	3.0
InstructBLIP [11]	Vicuna-7B	32.5	20.0	16.5	46.0	24.5	46.0	51.0	26.0	37.5	22.0	23.0	46.5	**42.5**	26.5	40.5	32.0	25.5	30.0	25.5	30.5	38.0
LLaVA [45]	Vicuna-7B	36.0	28.0	39.5	63.0	30.5	39.0	53.0	41.0	41.5	23.0	20.5	45.0	34.0	20.5	38.5	47.0	25.0	36.0	27.0	26.5	42.0
Video MLLMs: *All models take **16** frames as input, with the exception of VideoChatGPT, which uses **100** frames.*																						
Otter-V [37]	LLaMA-7B	26.8	23.0	23.0	27.5	27.0	29.5	53.0	28.0	33.0	24.5	23.5	27.5	26.0	28.5	18.0	38.5	22.0	22.0	23.5	19.0	19.5
mPLUG-Owl-V [88]	LLaMA-7B	29.7	22.0	28.0	34.0	29.0	29.0	40.5	27.0	31.5	**27.0**	23.0	29.0	31.5	27.0	40.0	44.0	24.0	31.0	26.0	20.5	29.5
VideoChatGPT [49]	Vicuna-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
VideoLLaMA [95]	Vicuna-7B	34.1	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0
VideoChat [40]	Vicuna-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$_{\text{text}}$	Vicuna-7B	34.7	24.5	27.0	49.5	27.0	38.0	53.0	28.0	40.0	25.5	27.0	38.5	41.5	27.5	32.5	46.5	26.5	36.0	33.0	32.0	40.0
VideoChat2	Vicuna-7B	**51.1**	**66.0**	**47.5**	**83.5**	**49.5**	**60.0**	**58.0**	**71.5**	**42.5**	23.0	23.0	**88.5**	39.0	**42.0**	**58.5**	44.0	**49.0**	36.5	**35.0**	**40.5**	**65.5**

Why Video MLLMs are unsatisfactory?

① Limited instruction tuning data

② Lack of strong video encoder

VideoChat2: Robust Baseline

Diverse Instruction Data

Conversation	#Num
LLaVA	56,681
VideoChat	13,884
VideoChatGPT	13,303

Classification	#Num
ImageNet	30,000
COCO-ITM	29,919
Kinetics-710	40,000
SthSthV2	40,000

Detailed Caption	#Num
MiniGPT-4	3,362
LLaVA	23,240
Paragraph Captioning	14,575
VideoChat	6,905

Reasoning	#Num
LLaVA	76,643
CLEVR	30,000
VisualMRC	15,000
NExTQA	34,132
CLEVRER_QA*	40,000
CLEVRER_MC	42,620

Simple Caption	#Num
COCO	566,747
TextCaps	97,765
WebVid	400,000
YouCook2	8,760
TextVR	39,648

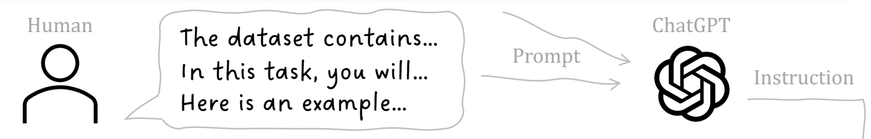
VQA	#Num
VQAv2	29,903
GQA	30,001
OKVQA	8,990
A-OKVQA	17,056
ViQuAE	1,152
OCR-VQA	11,414
TextVQA	27,113
ST-VQA	26,074
DocVQA	39,463
TGIF-Frame	39,149
TGIF-Transition	52,696
WebVidQA	10,000
EgoQA	7,813

Instruction Generation

You are professional in video understanding and instruction design. I will give you the description of video dataset and task, and one instruction example.

DATASET DESCRIPTION: {dataset_descrption}
TASK DESCRIPTION: {task_description}
INSTRUCTION EXAMPLE: {instruction_example}

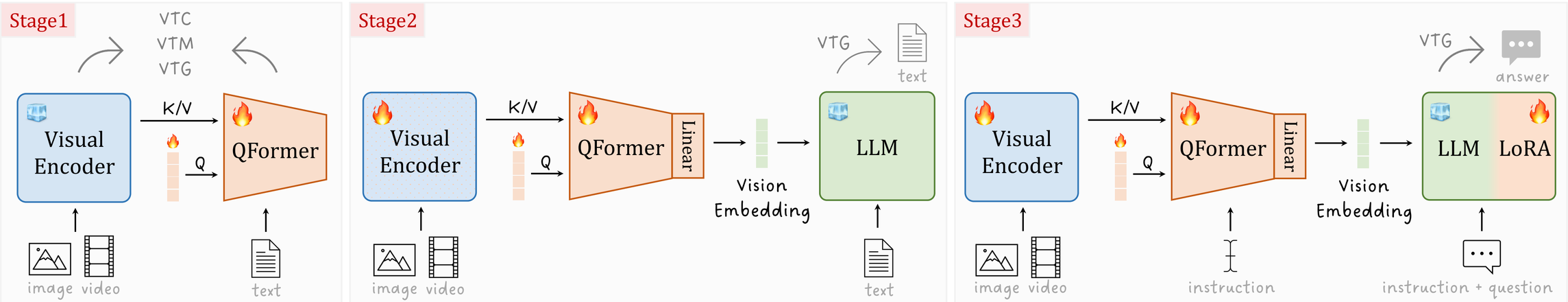
Based on the above message, you need to help me generate 10 instructions for handling the video tasks.



Data Example

```
# video data path
'video': '023601_023650/1023815317.mp4',
# conversion tasks have multiple QA
'QA': [{
    # instruction as task guidance
    'i': "Go through the video, taking into account key aspects, and respond to the question.",
    # no question for caption tasks
    'q': "What color cliff is the hindu temple on?",
    # short answer may be phrased
    'a': "The Hindu temple in the video is situated on a green cliff."
}]
```

Progressive Training



Vision-Language Alignment

Vision-Language Connection

Instruction Tuning

Rank1 at 15 Tasks on MVBench

Rank	Model	Acc
1	**VideoChat2**	**42.0**
2	**Otter-I**	**32.5**
3	**BLIP2**	**30.0**
4	InstructBLIP	26.5
5	VideoChatGPT	25.5
6	VideoLLaMA	22.5
7	LLaMA-Adapter	22.5
8	mPLUG-Owl-I	22.0
9	LLaVA	20.5
10	VideoChat	20.5
11	MiniGPT-4	15.5

(m) *Moving Count*

Rank	Model	Acc
1	**VideoChat2**	**58.5**
2	**VideoChat**	**42.5**
3	**LLaMA-Adapter**	**41.5**
4	InstructBLIP	40.5
5	BLIP2	40.0
6	VideoChatGPT	39.5
7	LLaVA	38.5
8	VideoLLaMA	32.5
9	mPLUG-Owl-I	31.5
10	Otter-I	28.5
11	MiniGPT-4	8.0

(n) *Fine-grained Pose*

Rank	Model	Acc
1	**VideoChatGPT**	**48.5**
2	**LLaVA**	**47.0**
3	**VideoChat**	**46.0**
4	VideoLLaMA	45.5
5	VideoChat2	44.0
6	BLIP2	42.0
7	mPLUG-Owl-I	40.0
8	LLaMA-Adapter	39.5
9	Otter-I	39.0
10	MiniGPT-4	34.0
11	InstructBLIP	32.0

(o) *Moving Attribute*

Rank	Model	Acc
1	**VideoChat2**	**49.0**
2	**VideoLLaMA**	**32.5**
3	**VideoChatGPT**	**29.0**
4	Otter-I	28.0
5	BLIP2	27.0
6	VideoChat	26.5
7	MiniGPT-4	26.0
8	InstructBLIP	25.5
9	LLaMA-Adapter	25.0
10	LLaVA	25.0
11	mPLUG-Owl-I	24.0

(p) *State Change*

Rank	Model	Acc
1	**VideoChat**	**41.0**
2	**VideoLLaMA**	**40.0**
3	**mPLUG-Owl-I**	**37.0**
4	VideoChat2	36.5
5	LLaVA	36.0
6	VideoChatGPT	33.0
7	LLaMA-Adapter	31.5
8	BLIP2	30.0
9	InstructBLIP	30.0
10	MiniGPT-4	29.5
11	Otter-I	27.0

(q) *Character Order*

Rank	Model	Acc
1	**VideoChat2**	**35.0**
2	**Otter-I**	**32.0**
3	**VideoLLaMA**	**30.0**
4	VideoChatGPT	29.5
5	LLaVA	27.0
6	BLIP2	26.0
7	mPLUG-Owl-I	25.5
8	InstructBLIP	25.5
9	VideoChat	23.5
10	LLaMA-Adapter	22.5
11	MiniGPT-4	19.0

(r) *Egocentric Navigation*

Rank	Model	Acc
1	**VideoChat2**	**40.5**
2	**BLIP2**	**37.0**
3	**InstructBLIP**	**30.5**
4	Otter-I	29.0
5	LLaMA-Adapter	28.0
6	LLaVA	26.5
7	VideoChatGPT	26.0
8	VideoChat	23.5
9	mPLUG-Owl-I	21.0
10	VideoLLaMA	21.0
11	MiniGPT-4	9.9

(s) *Episodic Reasoning*

Rank	Model	Acc
1	**VideoChat2**	**65.5**
2	**LLaVA**	**42.0**
3	**InstructBLIP**	**38.0**
4	mPLUG-Owl-I	37.0
5	VideoLLaMA	37.0
6	Otter-I	36.5
7	VideoChat	36.0
8	VideoChatGPT	35.5
9	LLaMA-Adapter	32.0
10	BLIP2	31.0
11	MiniGPT-4	3.0

(t) *Counterfactual Inference*

Rank	Model	Acc
1	**LLaMA-Adapter**	**25.5**
2	**BLIP2**	**25.5**
3	**VideoChat**	**25.5**
4	VideoChat2	23.0
5	VideoChatGPT	23.0
6	mPLUG-Owl-I	23.0
7	LLaVA	23.0
8	VideoLLaMA	22.5
9	InstructBLIP	22.0
10	Otter-I	19.0
11	MiniGPT-4	11.5

(i) *Moving Direction*

Rank	Model	Acc
1	**VideoChat**	**27.0**
2	**BLIP2**	**26.0**
3	**Otter-I**	**25.5**
4	mPLUG-Owl-I	24.0
5	VideoChat2	23.0
6	InstructBLIP	23.0
7	VideoLLaMA	22.5
8	LLaMA-Adapter	21.5
9	LLaVA	20.5
10	VideoChatGPT	20.0
11	MiniGPT-4	12.0

(j) *Action Localization*

Rank	Model	Acc
1	**VideoChat2**	**88.5**
2	**Otter-I**	**55.0**
3	**VideoChat**	**48.5**
4	InstructBLIP	46.5
5	LLaVA	45.0
6	VideoLLaMA	43.0
7	mPLUG-Owl-I	34.5
8	BLIP2	32.5
9	VideoChatGPT	31.0
10	LLaMA-Adapter	30.5
11	MiniGPT-4	9.5

(k) *Scene transition*

Rank	Model	Acc
1	**InstructBLIP**	**42.5**
2	**VideoChat2**	**39.0**
3	**VideoChat**	**35.0**
4	mPLUG-Owl-I	34.5
5	LLaVA	34.0
6	VideoLLaMA	34.0
7	MiniGPT-4	32.5
8	VideoChatGPT	30.5
9	LLaMA-Adapter	29.0
10	BLIP2	25.5
11	Otter-I	20.0

(l) *Action Count*

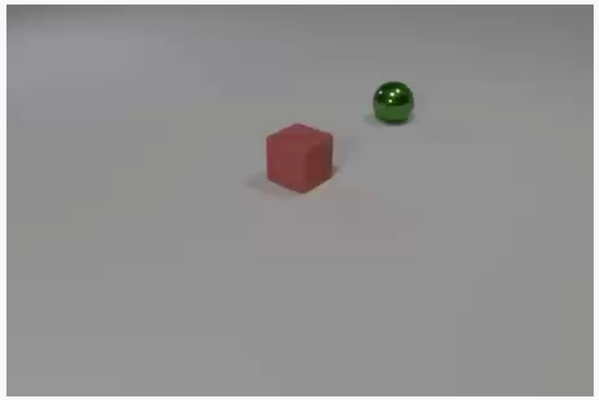
Performs Well on Other Benchmarks

Conversation, Question Answering, Reasoning…

- State of the Art | Zero-Shot Video Question Answer on ActivityNet-QA
- Ranked #3 | Zero-Shot Video Question Answer on MSRVTT-QA
- Ranked #3 | Zero-Shot Video Question Answer on MSVD-QA
- State of the Art | Zero-Shot Video Question Answer on NExT-QA
- State of the Art | Zero-Shot Video Question Answer on STAR: Situated Reasoning
- Ranked #2 | Zero-Shot Learning on TVQA
- Ranked #3 | Video Question Answering on NExT-QA
- Ranked #2 | Video-based Generative Performance Benchmarking on VideoInstruct
- State of the Art | Video-based Generative Performance Benchmarking (Consistency) on VideoInstruct
- State of the Art | Video-based Generative Performance Benchmarking (Contextual Understanding) on VideoInstruct
- State of the Art | Video-based Generative Performance Benchmarking (Correctness of Information) on VideoInstruct
- Ranked #2 | Video-based Generative Performance Benchmarking (Temporal Understanding) on VideoInstruct
- Ranked #3 | Video-based Generative Performance Benchmarking (Detail Orientation)) on VideoInstruct

Qualitative Comparisons

Moving Attribute



What color is the last object that enters the scene?

The last object that enters the scene is a red cube. ✗

The last object that enters the scene is a red cube. ✗

The answer is yellow. ✓

Qualitative Comparisons

Action Count

How many times did the person show objects to the camera?

The video shows the person showing objects to the camera multiple times. ✗

The person in the video shows objects to the camera three times. ✗

The person in the video showed objects to the camera a total of five times. ✓

Qualitative Comparisons

Unexpected Action

What unusual event takes place in the video?

The video shows a man dressed in a blue shirt and a hat standing in front of a blue curtain. Behind him, there is a large blue water with a man standing on it. ✗

The video shows a man playing with a yellow ball in a swimming pool. He is seen jumping into the pool and trying to catch the ball. ✗

The individuals in the video imitate a fishing scenario by using poles to stand on top of a pool. ✓

Qualitative Comparisons

Egocentric Navigation

This is a navigation video of an agent following instruction: "**Exit sitting room, turn left and go up stairs. At landing turn left and go up stairs. Stop at top of stairs.**" What is the next action it should take?

The next action the agent should take is to go up the stairs at the landing. ✗

Go up the stairs and continue following the instructions. ✗

The next action for the agent to take is to stop at the top of the stairs. ✓

Qualitative Comparisons

Action Prediction

What will the person do after reading a book?

It is difficult to predict what the person will do next, as the video only shows her sitting on the floor and reading a book. ✗

They are likely reading the book for leisure or as a form of entertainment. ✗

The person is likely going to take a bite out of the sandwich next. ✓

Describe Video Details

A young girl wearing a pink jacket and jeans walking down the street. As she walks, the camera follows her, capturing her movements and the surroundings. The video ends with the girl walking away from the camera.

The video shows a dragon playing a guitar. The dragon is shown in different angles and positions while playing the guitar. The video is of high quality and the dragon's movements are smooth and fluid.

Open Source

VideoChat2 Code:
https://github.com/OpenGVLab/Ask-Anything/tree/main/video_chat2

VideoChat2 Demo:
https://huggingface.co/spaces/OpenGVLab/VideoChat2

VideoChat2 Instruction Data:
https://huggingface.co/datasets/OpenGVLab/VideoChat2-IT

MVBench Data:
https://huggingface.co/datasets/OpenGVLab/MVBench