

## A. Training Hyperparameters

The hyperparameters used in different stages of training are listed in Tab. 10. We adopt TSN [70] sampling for all the videos as previous methods [39, 40, 71]. For both Stage1 and Stage2, we employ large-scale image and video caption data, as outlined in the main manuscript. During Stage3, we make use of diverse instruction data and incorporate LoRA modules [24] into the LLM with a rank of 16, an alpha value of 32, and a dropout rate of 0.1. We apply flash attention [12] to expedite the training process.

config	Stage1	Stage2	Stage3
input frame	4	4	8
input resolution	224	224	224
max text length	32	32	512
optimizer	AdamW		
optimizer momentum	$\beta_1, \beta_2=0.9, 0.999$		
weight decay	0.02		
learning rate schedule	cosine decay		
learning rate	1e-4	1e-4	2e-5
batch size	2048	512	128
warmup epochs	1	0.2	0.6
total epochs	10	1	3
backbone drop path		0	
QFormer drop path		0.2	
QFormer dropout	0	0.1	0.1
QFormer token	32	96	96
flip augmentation		yes	
augmentation	MultiScaleCrop [0.5, 1]		

Table 10. Training Hyperparameters for different stages.

## B. More Ablations

We have carried out further ablation studies, the results of which are displayed in Tabs. 11, 13, 12, and 14.

**QFormer.** Considering the richer information of video, we further introduce extra random-initialized queries after Stage1. Tab. 11 shows that more queries in Stage2 and Stage3 is beneficial, leading us to adopt 64 queries by default. Furthermore, inserting instructions without a question effectively steers toward more accurate responses. We argue that overly long context (“*instruction + question*”) may be difficult for information extraction of QFormer.

#Query	Instruction	Question	Avg
32 + 0	✓	✗	47.8
32 + 32	✓	✗	50.6 $\uparrow 2.8$
32 + 64	✓	✗	<b>51.1</b> $\uparrow 3.3$
32 + 96	✓	✗	50.7 $\uparrow 2.9$
32 + 64	✓	✓	50.8 $\uparrow 3.0$
32 + 64	✗	✗	50.5 $\uparrow 2.7$

Table 11. QFormer. Introducing more extra queries helps.

**Resolution & Frame.** Tab. 12 reveals that increasing resolution does not improve performance; however, aug-

menting the number of frames enhances outcomes. This suggests that our MVBenCh primarily relies on temporal understanding instead of spatial understanding capacity.

Training	Testing	Avg
8×224×224	8×224×224	50.6
	8×384×384	49.9 $\downarrow 0.7$
	16×224×224	<b>51.1</b> $\uparrow 0.5$
	32×224×224	<b>51.1</b> $\uparrow 0.5$
	64×224×224	51.0 $\uparrow 0.4$
16×224×224	16×224×224	51.0 $\uparrow 0.4$

Table 12. Resolution & Frame. Large resolution is harmful, while more frames are better for MVBenCh.

**Instruction data.** Note that there is a minimal source gap between our instruction data and MVBenCh. Specifically, the CLEVRER [88] in our instruction data has similar questions as *Moving Attribute* and *Counterfactual Inference* in MVBenCh, leading the evaluation is not strictly out-domain. And the videos of *Action Antonym* are from SthSthV2 [21], while the antonym is from PAXION [74]. We try to remove CLEVRER and SthSthV2 in the instruction data to evaluate their impact. The results outlined in Tab. 13 suggest a more pronounced influence from CLEVRER data, while SthSthV2 data appears to have less effect.

Data	Avg
ALL	<b>51.1</b>
ALL – CLEVRER [88]	49.3 $\downarrow 1.8$
ALL – SthSthV2 [21]	51.0 $\downarrow 0.1$

Table 13. Instruction Data.

**Question prompt.** During our experiments, we observed that various MLLMs often provide options along with detailed explanations. To circumvent this, we intentionally craft our question prompts to prevent such detailed outputs. Additionally, drawing inspiration from the Chain-of-Thought [76], we introduce the phrase “Let’s think step by step” into our prompts to direct the MLLMs’ reasoning process. However, as indicated by the results in Tab. 14, these tactics appear to have negative consequences.

Question Prompt	Avg
<i>Only give the best option.</i>	<b>51.1</b>
<i>Only give the best option without any explanation.</i>	50.9 $\downarrow 0.2$
<i>Let’s think step by step. Only give the best option.</i>	50.5 $\downarrow 0.6$

Table 14. Question prompt.

## C. Details of QA Generation

In Tab. 15, we present a detailed description of our data generation methodology for MVBenCh. We have designed various strategies based on different data to increase task difficulty and enhance data diversity. For those datasets


Task	Source	Domain	Data Filtration	QA Generation
Action Sequence	STAR [77]	· Real-world · Indoor · Third-person	✓ Duration $\in (5, 22)$ ✓ Data $\in$ Prediction ✗ $len(A) = 1 \vee A.split(" ") = "the"$	QA: Directly adopt
Action Antonym	PAXION [74]	· Real-world&Simulated · Indoor&Outdoor · Third-person	N/A	Q: ChatGPT generates A: GT+Antonym+"not sure"
Fine-grained Action	MiT V1 [52]	· Real-world&Simulated · Indoor&Outdoor · Third-person	N/A	Q: ChatGPT generates A: Randomly sample 4 actions from top-6 predictions of UMT-L/16 [40]
Unexpected Action	FunQA [80]	· Real-world · Indoor&Outdoor · Third-person	✓ $len(QA \in H2) = 34, len(QA \in H3) = 33$ ✓ $len(QA \in C2) = 33, len(QA \in C3) = 33$ ✓ $len(QA \in M2) = 34, len(QA \in M3) = 33$	QA: ChatGPT generates from original QA
Object Existence	CLEVRER [88]	· Simulated · Indoor	✓ Data $\in$ descriptive $\wedge$ Data $\in$ exist ✓ $len(program) < 11$	Q: ChatGPT generates A: "yes"+"no"+"not sure"
Object Interaction	STAR [77]	· Real-world · Indoor · Third-person	✓ Duration $\in (7, 20)$ ✓ Data $\in$ Interaction ✓ "object" in Q $\vee$ "to the" in Q	QA: Directly adopt
Object Shuffle	Perception Test [56]	· Real-world · Indoor · First&Third-person	✓ Data $\in$ object permanence ✓ "Where is the" in Q	QA: Directly adopt
Moving Direction	CLEVRER [88]	· Simulated · Indoor	Select videos where a certain object is either stationary or moving in a single direction	Q: ChatGPT generates A:  + "stationary"
Action Localization	Charades-STA [19]	· Real-world · Indoor · Third-person	✓ Duration <sub>entire</sub> > 15 ✓ Duration <sub>start,end,middle</sub> $\in (5, 8)$ ✗ "person they" in Q $\vee$ "person. so they" in Q	Q: ChatGPT generates A: "start"+"end"+"middle"+"entire"
Scene Transition	MoVQA [95]	· Real-world · Indoor&Outdoor · Third-person	Select videos with continuous scene labels	QA: ChatGPT generates from original QA
Action Count	Perception Test [56]	· Real-world · Indoor · First&Third-person	✓ Data $\in$ action counting	QA: Directly adopt
Moving Count	CLEVRER [88]	· Simulated · Indoor	✓ Data $\in$ descriptive $\wedge$ Data $\in$ count ✓ $len(program) < 9$	Q: ChatGPT generates A: Randomly shift original answer
Moving Attribute	CLEVRER [88]	· Simulated · Indoor	✓ Data $\in$ descriptive $\wedge$ Data $\in$ query_color ✓ Data $\in$ descriptive $\wedge$ Data $\in$ query_shape ✓ Data $\in$ descriptive $\wedge$ Data $\in$ query_material ✓ $len(program) < 13$	Q: ChatGPT generates A: Randomly select from candidates
State Change	Perception Test [56]	· Real-world · Indoor · First&Third-person	✓ Data $\in$ state recognition ✗ Q requires audio	QA: Directly adopt
Fine-grained Pose	NTU RGB+D [45]	· Real-world · Indoor · Third-person	Select videos with specific poses	Q: ChatGPT generates A: Randomly select from similar poses
Character Order	Perception Test [56]	· Real-world · Indoor · First&Third-person	✓ Data $\in$ letter ✓ "order" $\in$ Q	QA: Directly adopt
Egocentric Navigation	VLN-CE [30]	· Simulated · Indoor · First-person	✓ moving forward > 0.75m ✓ turning left/right $\in (60^\circ, 120^\circ)$ then moving forward > 0.75m ✓ stop	Q: ChatGPT generates A: "move forward"+"stop" "turn left and move forward"+ "turn right and move forward"
Episodic Reasoning	TVQA [33]	· Real-world · Indoor&Outdoor · Third-person	✓ Duration $\in (25, 40)$	QA: Directly adopt w/o subtitles
Counterfactual Inference	CLEVRER [88]	· Simulated · Indoor	✓ Data $\in$ counterfactual ✓ $len(program) < 8$	QA: Directly adopt

Table 15. More details about MVBench generation.

requiring question generation, we utilize ChatGPT [53] to generate 3 to 5 questions based on the task definitions.

## D. Results on Challenging Video QA

In Tabs. 17 and 18, we extend the evaluation of our VideoChat2 to other challenging video benchmarks *i.e.*, NEXt-QA [79], STAR [77] and TVQA [33]. Different from

the previous methods [89], which provide answers by comparing the likelihood of different options, we output the options directly, following the protocol of MVBench. Our results indicate that VideoChat2 not only holds its own against current SOTA methods [72, 89] on NEXt-QA but also markedly outperforms them on STAR and TVQA. This underscores the effectiveness and robustness of VideoChat2.

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
<i>GPT-4V take 16 frames as input, and the resolution is 512×512, while others use small resolution of 224×224.</i>																						
GPT-4V [54]	GPT-4	49.5	<b>80.0</b>	40.0	30.0	50.0	<b>60.0</b>	60.0	<b>90.0</b>	40.0	20.0	<b>60.0</b>	<b>100.0</b>	40.0	20.0	<b>50.0</b>	40.0	<b>70.0</b>	<b>50.0</b>	20.0	<b>60.0</b>	10.0
<i>Image MLLMs: Following [11], all models take 4 frames as input, with the output embeddings concatenated before feeding into the LLM.</i>																						
Otter-I [36]	MPT-7B	35.0	50.0	30.0	20.0	30.0	40.0	50.0	50.0	30.0	30.0	10.0	40.0	40.0	20.0	40.0	40.0	40.0	30.0	30.0	40.0	40.0
InstructBLIP [11]	Vicuna-7B	34.0	40.0	50.0	40.0	40.0	40.0	60.0	40.0	10.0	20.0	20.0	70.0	30.0	20.0	30.0	50.0	40.0	20.0	20.0	30.0	10.0
LLaVA [44]	Vicuna-7B	34.5	40.0	30.0	20.0	30.0	<b>60.0</b>	50.0	30.0	20.0	30.0	30.0	40.0	20.0	<b>40.0</b>	30.0	40.0	30.0	40.0	20.0	50.0	40.0
<i>Video MLLMs: All models take 16 frames as input, with the exception of VideoChatGPT, which uses 100 frames.</i>																						
VideoChatGPT [48]	Vicuna-7B	32.5	30.0	30.0	50.0	30.0	40.0	70.0	20.0	<b>50.0</b>	10.0	20.0	10.0	30.0	10.0	30.0	<b>70.0</b>	20.0	<b>50.0</b>	10.0	50.0	20.0
VideoLLaMA [94]	Vicuna-7B	34.0	30.0	20.0	50.0	40.0	20.0	50.0	40.0	0.0	30.0	10.0	70.0	40.0	30.0	30.0	50.0	30.0	<b>50.0</b>	50.0	40.0	0.0
VideoChat [39]	Vicuna-7B	36.5	50.0	10.0	50.0	20.0	50.0	60.0	70.0	40.0	<b>40.0</b>	40.0	50.0	30.0	20.0	40.0	50.0	20.0	30.0	10.0	50.0	0.0
<b>VideoChat2<sub>text</sub></b>	Vicuna-7B	35.0	40.0	40.0	30.0	30.0	30.0	40.0	20.0	<b>50.0</b>	20.0	30.0	60.0	<b>60.0</b>	30.0	20.0	60.0	20.0	<b>50.0</b>	30.0	30.0	10.0
<b>VideoChat2</b>	Vicuna-7B	<b>56.5</b>	60.0	<b>60.0</b>	<b>90.0</b>	<b>60.0</b>	<b>60.0</b>	<b>70.0</b>	80.0	30.0	20.0	40.0	<b>100.0</b>	40.0	<b>40.0</b>	40.0	60.0	50.0	40.0	<b>60.0</b>	<b>60.0</b>	<b>70.0</b>

Table 16. **Evaluations results on MVBench subset.** We randomly sample 10 multiple-choice QAs for each task due to time constraints. The results on full MVBench can be found at [https://huggingface.co/spaces/OpenGVLab/MVBench\\_Leaderboard](https://huggingface.co/spaces/OpenGVLab/MVBench_Leaderboard).

Model	Zero-shot				Fine-tuning			
	Tem.	Cau.	Des.	Avg	Tem.	Cau.	Des.	Avg
All-in-One [69]	-	-	-	-	48.6	48.0	63.2	50.6
MIST [18]	-	-	-	-	56.6	54.6	66.9	57.1
HiTeA [86]	-	-	-	-	58.3	62.4	75.6	63.1
InternVideo [73]	43.4	48.0	65.1	59.1	58.3	62.4	75.6	63.1
SEVILA [89]	61.3	61.5	75.6	63.6	69.4	74.2	81.3	73.8
VideoChat2	<b>57.4</b>	<b>61.9</b>	<b>69.9</b>	<b>61.7</b>	<b>64.7</b>	<b>68.7</b>	<b>76.1</b>	<b>68.6</b>

Table 17. **Results on NExT-QA [79].** “Tem.”, “Cau.” and “Des.” stand for “Temporal”, “Causal” and “Descriptive” respectively. SEVILA [89] is de-emphasized since it needs to train an additional localizer. For zero-shot results, we simply remove the NExT-QA in our instruction data.

## E. Comparisons with GPT-4V

We further conduct evaluations for GPT-4V [54] in Tab. 16. Given the time constraints, we randomly sample 10 multiple-choice QAs for each task. The results indicate that GPT-4V [54] achieved satisfactory performance on our MVBench, demonstrating its considerable capacity for temporal understanding. Notably, our VideoChat2 outperforms it by increasing accuracy by 7%,

## F. Leaderboards and Analyses

To facilitate a clear comparison of different open-sourced MLLMs, we present the leaderboards for different tasks on MVBench in Tab. 19. Overall, our VideoChat2 achieves the highest rank across 15 tasks.

**Action & Pose.** For tasks associated with action and pose (a)(b)(c)(d)(e)(p), our VideoChat2 and VideoChat [39] tends to outperform VideoChatGPT [48], underscoring the significance of elaborate video backbones [38, 40] for effective action and pose recognition.

**Object & Attribute.** In object-related tasks (f)(g)(h), the performance of image MLLM, *i.e.* LLaVA [44], compares favorably with our VideoChat2. It could be attributed to its potent attribute recognition capabilities, as illustrated in (n). Note that VideoChatGPT [48] is tuned from LLaVA,

Model	STAR					TVQA
	Int.	Seq.	Pre.	Fea.	Avg	
FrozenBiLM [85]	-	-	-	-	-	29.7
InternVideo [73]	43.8	43.2	42.3	37.4	41.6	35.9
SEVILA [89]	48.3	45.0	44.4	40.8	44.6	38.2
VideoChat2	<b>58.4</b>	<b>60.9</b>	<b>55.3</b>	<b>53.1</b>	<b>59.0</b>	<b>40.6</b>

Table 18. **Zero-shot results on STAR [77] and TVQA [33].** “Int.”, “Seq.”, “Pre.” and “Fea.” stand for “Interaction”, “Sequence”, “Prediction” and “Feasibility” respectively. SEVILA [89] is de-emphasized since it needs to train an additional localizer. For TVQA, we do not input subtitles.

thus achieving similar results on these tasks.

**Position & Count & Character.** In position-related tasks (i)(j), none of the models achieve satisfactory results, their performances being analogous to random guessing. For counting and character-related tasks (l)(q), our VideoChat2 performs similarly and even worse than VideoChat2<sub>text</sub> without videos (as in Tab. 2). We hypothesize that current MLLMs have difficulty generalizing to localization and counting tasks in the absence of related tuning data. Some recent studies [2, 8, 9] incorporate grounding data and tune the LLM to enhance localizing and discriminating abilities. In our future work, we will explore improvements in VideoChat2’s grounding ability.

**Scene.** As presented in Tab. 19(k), our VideoChat2 excels at scene transition tasks, significantly outperforming other models. This showcases its sensitivity to background changes, making it effective in recognizing camera movements as shown in Fig. 7.

**Cognition.** In cognition tasks (r)(s)(t), our VideoChat2 encounters difficulties with complex egocentric navigation and episode reasoning. Given the results from FrozenBiLM [85], where the performance for TVQA reasoning significantly improves with the incorporation of speech subtitles, we suggest that visual information alone may not be sufficient. The inclusion of other modalities, such as depth and audio, could prove beneficial.

Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc
1	🏠 VideoChat2	66.0	1	🏠 VideoChat2	47.5	1	🏠 VideoChat2	83.5	1	🏠 VideoChat2	49.5
2	🏠 Otter-I	34.5	2	🏠 LLaVA	39.5	2	🏠 LLaVA	63.0	2	🏠 VideoChat	33.5
3	🏠 VideoChat	33.5	3	🏠 Otter-I	32.0	3	🏠 VideoChatGPT	62.0	3	🏠 Otter-I	30.5
4	🏠 LLaVA	28.0	4	🏠 BLIP2	29.0	4	🏠 VideoChat	56.0	4	🏠 LLaVA	30.5
5	🏠 VideoLLaMA	27.5	5	🏠 LLaMA-Adapter	28.0	5	🏠 LLaMA-Adapter	51.0	5	🏠 LLaMA-Adapter	30.0
6	🏠 mPLUG-Owl-I	25.0	6	🏠 VideoChat	26.5	6	🏠 VideoLLaMA	51.0	6	🏠 VideoLLaMA	29.0
7	🏠 BLIP2	24.5	7	🏠 VideoChatGPT	26.0	7	🏠 InstructBLIP	46.0	7	🏠 mPLUG-Owl-I	27.0
8	🏠 VideoChatGPT	23.5	8	🏠 VideoLLaMA	25.5	8	🏠 mPLUG-Owl-I	44.5	8	🏠 InstructBLIP	24.5
9	🏠 LLaMA-Adapter	23.0	9	🏠 mPLUG-Owl-I	20.0	9	🏠 Otter-I	39.5	9	🏠 VideoChatGPT	22.5
10	🏠 InstructBLIP	20.0	10	🏠 MiniGPT-4	18.0	10	🏠 BLIP2	33.5	10	🏠 MiniGPT-4	21.5
11	🏠 MiniGPT-4	16.0	11	🏠 InstructBLIP	16.5	11	🏠 MiniGPT-4	26.0	11	🏠 BLIP2	17.0

(a) Action Sequence

Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc
1	🏠 VideoChat2	60.0	1	🏠 VideoChat2	58.0	1	🏠 VideoChat2	71.5	1	🏠 VideoChat2	42.5
2	🏠 InstructBLIP	46.0	2	🏠 VideoChatGPT	54.0	2	🏠 Otter-I	44.0	2	🏠 LLaVA	41.5
3	🏠 BLIP2	42.0	3	🏠 LLaMA-Adapter	53.5	3	🏠 LLaVA	41.0	3	🏠 VideoChatGPT	40.0
4	🏠 VideoChat	40.5	4	🏠 LLaVA	53.0	4	🏠 VideoLLaMA	40.5	4	🏠 VideoLLaMA	38.0
5	🏠 LLaVA	39.0	5	🏠 VideoChat	53.0	5	🏠 VideoChat	40.5	5	🏠 InstructBLIP	37.5
6	🏠 VideoLLaMA	39.0	6	🏠 BLIP2	51.5	6	🏠 LLaMA-Adapter	32.5	6	🏠 mPLUG-Owl-I	34.0
7	🏠 Otter-I	38.5	7	🏠 InstructBLIP	51.0	7	🏠 VideoChatGPT	28.0	7	🏠 LLaMA-Adapter	33.5
8	🏠 LLaMA-Adapter	33.0	8	🏠 Otter-I	48.5	8	🏠 BLIP2	26.0	8	🏠 BLIP2	31.0
9	🏠 VideoChatGPT	26.5	9	🏠 VideoLLaMA	48.0	9	🏠 InstructBLIP	26.0	9	🏠 VideoChat	30.0
10	🏠 mPLUG-Owl-I	23.5	10	🏠 mPLUG-Owl-I	36.0	10	🏠 MiniGPT-4	25.5	10	🏠 Otter-I	29.5
11	🏠 MiniGPT-4	16.0	11	🏠 MiniGPT-4	29.5	11	🏠 mPLUG-Owl-I	24.0	11	🏠 MiniGPT-4	13.0

(b) Action Prediction

Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc
1	🏠 LLaMA-Adapter	25.5	1	🏠 VideoChat	27.0	1	🏠 VideoChat2	88.5	1	🏠 InstructBLIP	42.5
2	🏠 BLIP2	25.5	2	🏠 BLIP2	26.0	2	🏠 Otter-I	55.0	2	🏠 VideoChat2	39.0
3	🏠 VideoChat	25.5	3	🏠 Otter-I	25.5	3	🏠 VideoChat	48.5	3	🏠 VideoChat	35.0
4	🏠 VideoChat2	23.0	4	🏠 mPLUG-Owl-I	24.0	4	🏠 InstructBLIP	46.5	4	🏠 mPLUG-Owl-I	34.5
5	🏠 VideoChatGPT	23.0	5	🏠 VideoChat2	23.0	5	🏠 LLaVA	45.0	5	🏠 LLaVA	34.0
6	🏠 mPLUG-Owl-I	23.0	6	🏠 InstructBLIP	23.0	6	🏠 VideoLLaMA	43.0	6	🏠 VideoLLaMA	34.0
7	🏠 LLaVA	23.0	7	🏠 VideoLLaMA	22.5	7	🏠 mPLUG-Owl-I	34.5	7	🏠 MiniGPT-4	32.5
8	🏠 VideoLLaMA	22.5	8	🏠 LLaMA-Adapter	21.5	8	🏠 BLIP2	32.5	8	🏠 VideoChatGPT	30.5
9	🏠 InstructBLIP	22.0	9	🏠 LLaVA	20.5	9	🏠 VideoChatGPT	31.0	9	🏠 LLaMA-Adapter	29.0
10	🏠 Otter-I	19.0	10	🏠 VideoChatGPT	20.0	10	🏠 LLaMA-Adapter	30.5	10	🏠 BLIP2	25.5
11	🏠 MiniGPT-4	11.5	11	🏠 MiniGPT-4	12.0	11	🏠 MiniGPT-4	9.5	11	🏠 Otter-I	20.0

(c) Action Antonym

Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc
1	🏠 VideoChat2	49.5	1	🏠 VideoChatGPT	48.5	1	🏠 VideoChat2	49.0	1	🏠 VideoChat2	58.5
2	🏠 VideoChat	33.5	2	🏠 LLaVA	47.0	2	🏠 VideoLLaMA	32.5	2	🏠 VideoChat	42.5
3	🏠 Otter-I	30.5	3	🏠 VideoChat	46.0	3	🏠 VideoChatGPT	29.0	3	🏠 LLaMA-Adapter	41.5
4	🏠 LLaVA	30.5	4	🏠 VideoLLaMA	45.5	4	🏠 Otter-I	28.0	4	🏠 InstructBLIP	40.5
5	🏠 LLaMA-Adapter	30.0	5	🏠 VideoChat2	44.0	5	🏠 BLIP2	27.0	5	🏠 BLIP2	40.0
6	🏠 VideoLLaMA	29.0	6	🏠 BLIP2	42.0	6	🏠 VideoChat	26.5	6	🏠 VideoChatGPT	39.5
7	🏠 mPLUG-Owl-I	27.0	7	🏠 mPLUG-Owl-I	40.0	7	🏠 MiniGPT-4	26.0	7	🏠 LLaVA	38.5
8	🏠 InstructBLIP	24.5	8	🏠 LLaMA-Adapter	39.5	8	🏠 InstructBLIP	25.5	8	🏠 VideoLLaMA	32.5
9	🏠 VideoChatGPT	22.5	9	🏠 Otter-I	39.0	9	🏠 LLaMA-Adapter	25.0	9	🏠 mPLUG-Owl-I	31.5
10	🏠 MiniGPT-4	21.5	10	🏠 MiniGPT-4	34.0	10	🏠 LLaVA	25.0	10	🏠 Otter-I	28.5
11	🏠 BLIP2	17.0	11	🏠 InstructBLIP	32.0	11	🏠 mPLUG-Owl-I	24.0	11	🏠 MiniGPT-4	8.0

(d) Fine-grained Action

Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc
1	🏠 VideoChat2	60.0	1	🏠 VideoChat	27.0	1	🏠 VideoChat2	88.5	1	🏠 InstructBLIP	42.5
2	🏠 InstructBLIP	46.0	2	🏠 BLIP2	26.0	2	🏠 Otter-I	55.0	2	🏠 VideoChat2	39.0
3	🏠 BLIP2	42.0	3	🏠 Otter-I	25.5	3	🏠 VideoChat	48.5	3	🏠 VideoChat	35.0
4	🏠 VideoChat	40.5	4	🏠 mPLUG-Owl-I	24.0	4	🏠 InstructBLIP	46.5	4	🏠 mPLUG-Owl-I	34.5
5	🏠 LLaVA	39.0	5	🏠 VideoChat2	23.0	5	🏠 LLaVA	45.0	5	🏠 LLaVA	34.0
6	🏠 VideoLLaMA	39.0	6	🏠 InstructBLIP	23.0	6	🏠 VideoLLaMA	43.0	6	🏠 VideoLLaMA	34.0
7	🏠 Otter-I	38.5	7	🏠 VideoLLaMA	22.5	7	🏠 mPLUG-Owl-I	34.5	7	🏠 MiniGPT-4	32.5
8	🏠 LLaMA-Adapter	33.0	8	🏠 LLaMA-Adapter	21.5	8	🏠 BLIP2	32.5	8	🏠 VideoChatGPT	30.5
9	🏠 VideoChatGPT	26.5	9	🏠 LLaVA	20.5	9	🏠 VideoChatGPT	31.0	9	🏠 LLaMA-Adapter	29.0
10	🏠 mPLUG-Owl-I	23.5	10	🏠 VideoChatGPT	20.0	10	🏠 LLaMA-Adapter	30.5	10	🏠 BLIP2	25.5
11	🏠 MiniGPT-4	16.0	11	🏠 MiniGPT-4	12.0	11	🏠 MiniGPT-4	9.5	11	🏠 Otter-I	20.0

(e) Unexpected Action

Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc
1	🏠 VideoChat2	47.5	1	🏠 VideoChatGPT	48.5	1	🏠 VideoChat2	49.0	1	🏠 VideoChat2	58.5
2	🏠 LLaVA	39.5	2	🏠 LLaVA	47.0	2	🏠 VideoLLaMA	32.5	2	🏠 VideoChat	42.5
3	🏠 Otter-I	32.0	3	🏠 VideoChat	46.0	3	🏠 VideoChatGPT	29.0	3	🏠 LLaMA-Adapter	41.5
4	🏠 BLIP2	29.0	4	🏠 VideoLLaMA	45.5	4	🏠 Otter-I	28.0	4	🏠 InstructBLIP	40.5
5	🏠 LLaMA-Adapter	28.0	5	🏠 VideoChat2	44.0	5	🏠 BLIP2	27.0	5	🏠 BLIP2	40.0
6	🏠 VideoChat	26.5	6	🏠 BLIP2	42.0	6	🏠 VideoChat	26.5	6	🏠 VideoChatGPT	39.5
7	🏠 VideoChatGPT	26.0	7	🏠 mPLUG-Owl-I	40.0	7	🏠 MiniGPT-4	26.0	7	🏠 LLaVA	38.5
8	🏠 VideoLLaMA	25.5	8	🏠 LLaMA-Adapter	39.5	8	🏠 InstructBLIP	25.5	8	🏠 VideoLLaMA	32.5
9	🏠 mPLUG-Owl-I	20.0	9	🏠 Otter-I	39.0	9	🏠 LLaMA-Adapter	25.0	9	🏠 mPLUG-Owl-I	31.5
10	🏠 InstructBLIP	18.0	10	🏠 MiniGPT-4	34.0	10	🏠 LLaVA	25.0	10	🏠 Otter-I	28.5
11	🏠 InstructBLIP	16.5	11	🏠 InstructBLIP	32.0	11	🏠 mPLUG-Owl-I	24.0	11	🏠 MiniGPT-4	8.0

(f) Object Existence

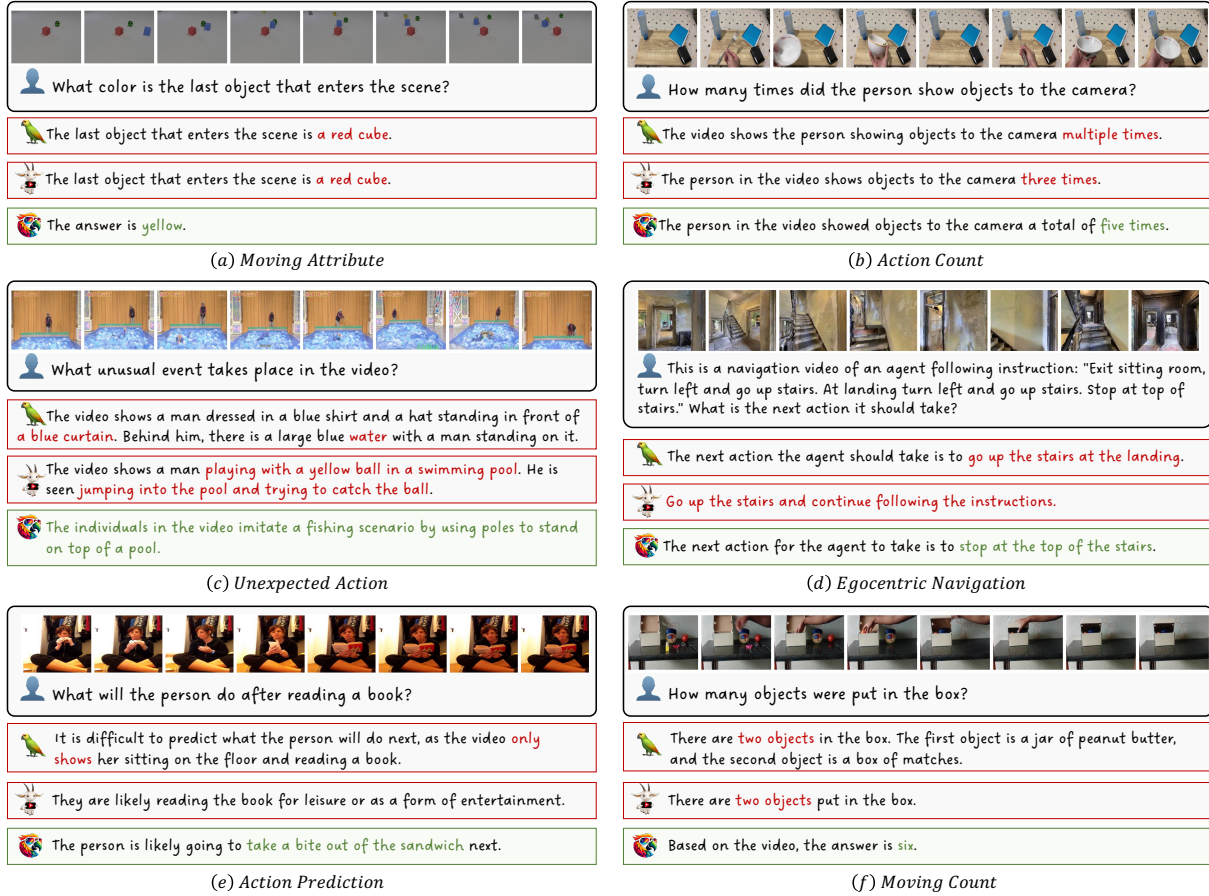
Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc
1	🏠 VideoChat2	83.5	1	🏠 VideoChat2	40.5	1	🏠 VideoChat2	40.5	1	🏠 VideoChat2	65.5
2	🏠 LLaVA	63.0	2	🏠 BLIP2	27.0	2	🏠 BLIP2	37.0	2	🏠 LLaVA	42.0
3	🏠 VideoChatGPT	62.0	3	🏠 VideoChat	26.0	3	🏠 InstructBLIP	30.5	3	🏠 InstructBLIP	38.0
4	🏠 VideoChat	56.0	4	🏠 VideoLLaMA	29.5	4	🏠 Otter-I	29.0	4	🏠 mPLUG-Owl-I	37.0
5	🏠 LLaMA-Adapter	51.0	5	🏠 LLaVA	27.0	5	🏠 LLaMA-Adapter	28.0	5	🏠 VideoLLaMA	37.0
6	🏠 VideoLLaMA	51.0	6	🏠 BLIP2	26.0	6	🏠 LLaVA	26.5	6	🏠 VideoLLaMA	37.0
7	🏠 InstructBLIP	46.0	7	🏠 mPLUG-Owl-I	25.5	7	🏠 VideoChatGPT	26.0	7	🏠 Otter-I	36.5
8	🏠 mPLUG-Owl-I	44.5	8	🏠 InstructBLIP	25.5	8	🏠 VideoChat	23.5	8	🏠 VideoChat	36.0
9	🏠 Otter-I	39.5	9	🏠 VideoChat	23.5	9	🏠 VideoChatGPT	23.5	9	🏠 VideoChatGPT	35.5
10	🏠 mPLUG-Owl-I	33.5	10	🏠 VideoLLaMA	22.5	10	🏠 mPLUG-Owl-I	21.0	10	🏠 LLaMA-Adapter	32.0
11	🏠 MiniGPT-4	26.0	11	🏠 MiniGPT-4	19.0	11	🏠 MiniGPT-4	9.9	11	🏠 BLIP2	31.0

(g) Object Interaction

Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc
1	🏠 VideoChat2	49.5	1	🏠 VideoChat2	35.0	1	🏠 VideoChat2	40.5	1	🏠 VideoChat2	65.5
2	🏠 VideoChat	33.5	2	🏠 Otter-I	32.0	2	🏠 BLIP2	37.0	2	🏠 LLaVA	42.0
3	🏠 Otter-I	30.5	3	🏠 VideoLLaMA	30.0	3	🏠 InstructBLIP	30.5	3	🏠 InstructBLIP	38.0
4	🏠 LLaVA	30.5	4	🏠 VideoChatGPT	29.5	4	🏠 Otter-I	29.0	4	🏠 mPLUG-Owl-I	37.0
5	🏠 LLaMA-Adapter	30.0	5	🏠 LLaVA	27.0	5	🏠 LLaMA-Adapter	28.0	5	🏠 VideoLLaMA	37.0
6	🏠 VideoLLaMA	29.0	6	🏠 BLIP2	26.0	6	🏠 LLaVA	26.5	6	🏠 VideoLLaMA	37.0
7	🏠 mPLUG-Owl-I	27.0	7	🏠 mPLUG-Owl-I	25.5	7	🏠 VideoChatGPT	26.0	7	🏠 Otter-I	36.5
8	🏠 InstructBLIP	24.5	8	🏠 InstructBLIP	25.5	8	🏠 VideoChat	23.5	8	🏠 VideoChat	36.0
9	🏠 VideoChatGPT	22.5	9	🏠 VideoChat	23.5	9	🏠 VideoChatGPT	23.5	9	🏠 VideoChatGPT	35.5
10	🏠 MiniGPT-4	21.5	10	🏠 LLaMA-Adapter	22.5	10	🏠 mPLUG-Owl-I	21.0	10	🏠 LLaMA-Adapter	32.0
11	🏠 BLIP2	17.0	11	🏠 MiniGPT-4	19.0	11	🏠 MiniGPT-4	9.9	11	🏠 BLIP2	31.0

(h) Object Shuffle

Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc	Rank	Model	Acc
1	🏠 VideoChat2	42.5	1	🏠 VideoChatGPT	48.5	1	🏠 VideoChat2	49.0	1	🏠 VideoChat2	58.5
2	🏠 LLaVA	41.5	2	🏠 LLaVA	47.0	2	🏠 VideoLLaMA	32.5	2	🏠 VideoChat	42.5
3	🏠 VideoChatGPT	40.0	3	🏠 VideoChat	46.0	3	🏠 VideoChatGPT	29.0	3	🏠 LLaMA-Adapter	41.5
4	🏠 VideoLLaMA	38.0	4	🏠 VideoLLaMA	45.5	4	🏠 Otter-I	28.0	4	🏠 InstructBLIP	40.5
5	🏠 InstructBLIP	37.5	5	🏠 VideoChat2	44.0	5	🏠 BLIP2	27.0	5	🏠 BLIP2	40.0
6	🏠 mPLUG-Owl-I	34.0	6	🏠 BLIP2	42.0	6	🏠 VideoChat	26.5	6	🏠 VideoChatGPT	39.5
7	🏠 LLaMA-Adapter	33.5	7	🏠 mPLUG-Owl-I	40.0	7	🏠 MiniGPT-4	26.0	7	🏠 LLaVA	38.5
8	🏠 BLIP2	31.0									


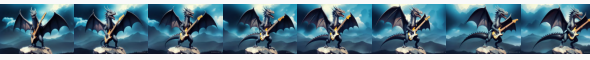
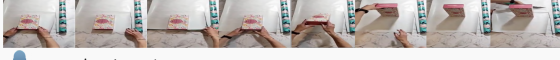
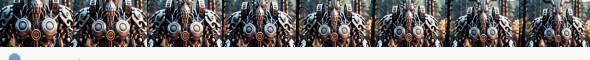
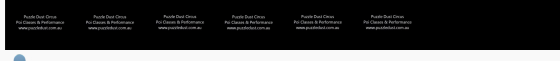
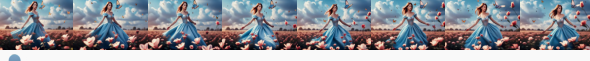


User  
 VideoChat  
 VideoChatGPT  
 VideoChat2

Figure 6. More qualitative comparisons in MVBench data. VideoChat2 handles different tasks well.

## G. Qualitative Results

Additional qualitative results can be found in Figs. 6 and 7. Compared with VideoChat [39] and VideoChatGPT [48], our VideoChat2 performs admirably across a range of tasks in MVBench. It possesses the capacity to accurately identify the properties of moving objects, recognize unforeseen actions, and predict future movements based on video context. Moreover, it exhibits robustness when dealing with both real and generated videos, adeptly providing detailed insights into human actions, camera motions, background ambiance, and character attributes.

 <p>Describe the video.</p>	 <p>Describe the video in details.</p>
<p>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.</p>	<p>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.</p>
 <p>Describe the video.</p>	 <p>Describe the video in details.</p>
<p>In the video, a woman is shown wrapping a gift box using a roll of wrapping paper. She uses a pair of scissors to cut the paper and then wraps it around the gift box. The wrapping paper has a floral design on it.</p>	<p>The video shows a robot with a futuristic design, standing in a forest. The robot has a metallic body and horns on its head. It appears to be a cyborg, with a mechanical appearance. The video is in slow motion, and the robot is shown in different angles. The forest setting adds a mysterious and eerie atmosphere to the video.</p>
 <p>Describe the video.</p>	 <p>Describe the video in details.</p>
<p>A black background with the words "puzzle dust circus" written on it. There is also a logo of the website "www.puzzledust.com" displayed on the screen. The video appears to be a promotional advertisement for the website.</p>	<p>The video shows a beautiful woman wearing a blue dress walking in a field surrounded by flowers. She is surrounded by butterflies and the sky is clear. The video is of high quality and has a serene atmosphere.</p>

(a) YouTube Videos



VideoChat2

(b) Genrated Videos

Figure 7. More descriptive examples. VideoChat2 can accurately describe the details of diverse videos.