LLaVA-Video: Video Instruction Tuning With Synthetic Data

Anonymous authors
Paper under double-blind review

1 Video Representations

1.1 Efficient Video Representations in LMMs

Current designs of large multimodal models (LMM) typically connect a vision encoder (Radford et al., 2021; Zhai et al., 2023) to a large language model (Yang et al., 2024) through a lightweight projector (Liu et al., 2024) or a resampler (Li et al., 2023; Alayrac et al., 2022). These components transform a set of visual representations into "visual tokens" aligned with text embeddings. In contrast to image-based LMMs, which generate only a small number of visual tokens easily managed by a standard GPU, video LMMs face challenges due to a large number of visual tokens derived from multiple video frames. The LLaVA-NeXT-Video (Zhang et al., 2024b) and PLLaVA (Xu et al., 2024a) models address this by simly considering average pooling to reduce the number of tokens representing each frame.

Following the idea of SlowFast in the traditional video understanding (Feichtenhofer et al., 2019), adaptive reductions in visual tokens are demonstrated by recent video LMMs, LITA (Huang et al., 2024) and SlowFast-LLaVA (Xu et al., 2024b). Initially, these methods represent all sampled frames with a minimal number of visual tokens (fast frame)— typically just one—by using a large pooling stride. They then switch to a smaller pooling stride for certain frames to retain more visual tokens (slow frame). Finally, they combine the visual tokens of fast frames with those of slow frames. However, this approach can lead to some frames being represented twice. In contrast, our method uses a larger pooling stride for sampled frames to maintain fewer visual tokens (fast frame) *or* a smaller stride for others to keep more (slow frame). We then arrange slow and fast frames in an interleaving pattern.

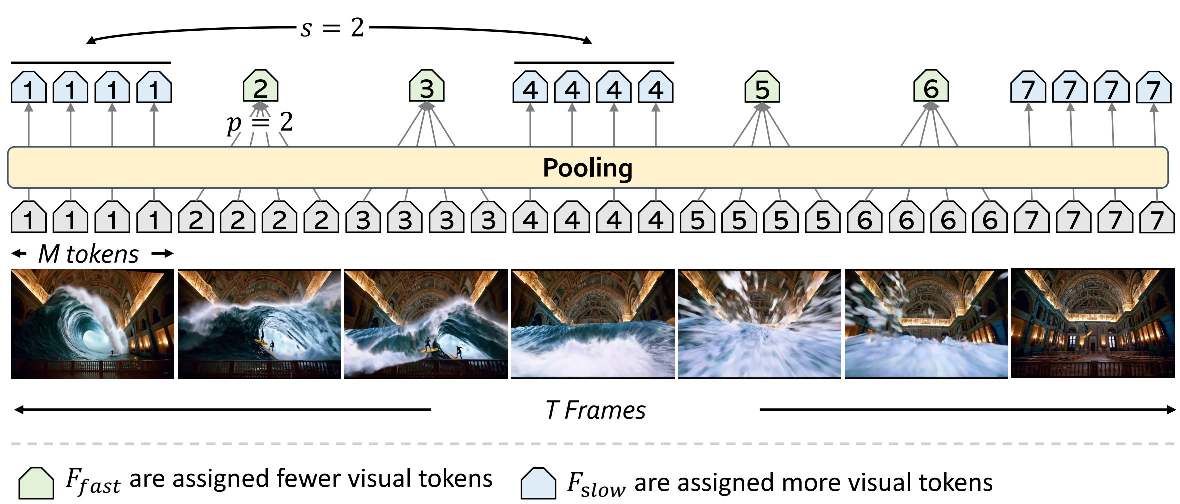


Figure 1: Video representations. A different number of tokens are utilized to represent frames.

1.2 LLaVA-Video $_{\texttt{SlowFast}}$

We represent each video as a sequence with maximum T frames. Each frame is represented in M tokens. FPS-based video representation can be considered in the future. Specifically, each frame is encoded via an image encoder and a two-layer MLP for projection. These visual tokens are concatenated with word tokens and processed by a large language model (LLM). Managing tokens for every frame can be computationally demanding. For instance, employing the SigLIP (Zhai et al., 2023) encoder for a video with $T = 100$ results in 67,600 tokens, assuming $M = 729$ tokens per frame, which often exceeds GPU memory limits. This issue is exacerbated when using large-parameter LLMs; with the Qwen2-72B model, we could only process 8 frames before maxing out the memory on 128 NVIDIA H100 GPUs. Such a limited number of frames can introduce inconsistencies in language annotations, reducing model efficacy. One strategy to incorporate more frames is by applying $p \times p$ spatial average pooling to reduce M to M/p^2, thus lowering the token count per frame as suggested by recent studies (Xu et al., 2024a; Zhang et al., 2024b). However, the number of visual tokens is crucial for preserving the informational content of each frame, which is vital for video comprehension.

In our LLaVA-Video $_{\texttt{SlowFast}}$, we categorize the frames into two groups, based on the a strike rate s, where the every s frames are uniformly selected to form the *slow* frame group, and the rest of the frames are consdiered as the *fast* frame group. Note that a special case $s = 1$ leads to only one group, reducing the SlowFast representation to the original simple representation. For each group, we apply different pooling rate using Pytorch function pooling `avg_pool2d()`. $p \times p$ pooling and $2p \times 2p$ pooling for slow and fast frames, respectively. To summarize, we paramterize the video representation configuration as $\mathcal{V} = (T, M, s, p)$. The total number of tokens is $\#tokens = \lfloor T/s \rfloor \times \lfloor M/p^2 \rfloor + (T - \lfloor T/s \rfloor) \times \lfloor M/4p^2 \rfloor$

2 Data

2.1 Video Detail Description

As discussed in Section 3.2, we show that generating *level-1 description* should consider historical context. Figure 2 illustrates the impact of excluding historical context on the quality of video descriptions. Specifically, including historical context helps accurately identify characters across different times as the same individual.

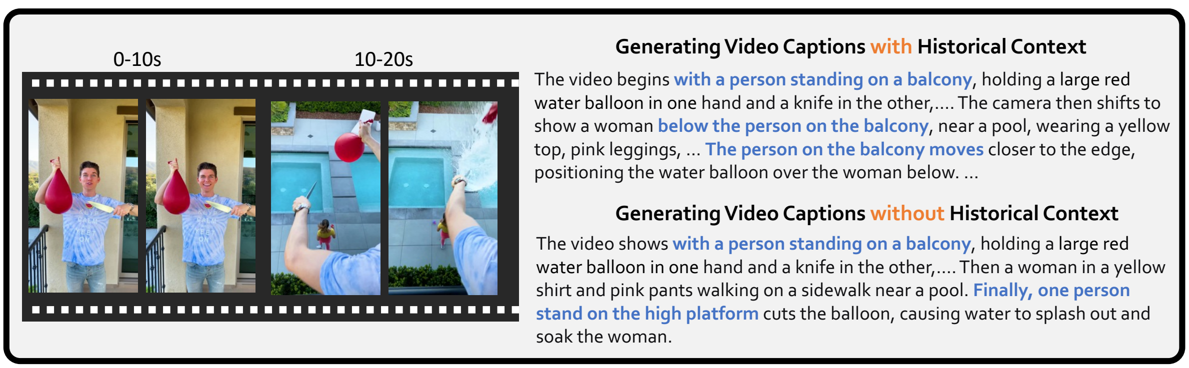


Figure 2: Generating video captions with or without historical context.

2.2 Video Question Answering

In Table 1, we list the names and descriptions of different question types and their corresponding proportions in the LLaVA-Video-178K dataset. The prompt used to generate video question-answer pairs from GPT-4O is shown in Table. 2. In Fig. ??, we show an example of a video along with its detailed description, an open-ended question, and a multiple-choice question.

2.3 Dataset Comparison

We provide a more comprehensive comparison of LLaVA-Video-178K with other video-language datasets for the video caption task and video question answer task. Specifically, we organize the table into four groups, each characterized by its method of text annotation. As shown in Table 3, unlike other datasets,

Table 1: Question types for video question answering in data creation. For each type, we provide its name, description, and the proportion it represents in the LLaVA-Video-178K.

Question type	**Description**	**Proportion**
Temporal	Designed to assess reasoning about temporal relationships between actions/events. Questions involve previous, present, or next actions.	7.2%
Spatial	Tests ability to perceive spatial relationships between observed instances in a video scene.	7.2%
Causal	Focuses on explaining actions/events, determining intentions of actions or causes for subsequent events.	7.2%
Description-Scene	Assesses ability to describe the major scene of the video, like where it takes place and the overall environment.	7.2%
Description-Human	Involves describing actions or attributes of people, such as their activities and appearances.	6.7%
Description-Object	Assesses ability to describe attributes of objects, like their appearance and function.	7.0%
Count	Tests ability to count instances of objects, people, actions, and to distinguish between old and new elements in a scene.	7.1%
Binary	Involves yes or no questions related to the video content.	7.2%
Fine Grained Action Understanding	Creates questions challenging comprehension of subtle actions.	6.5%
Plot Understanding	Challenges ability to interpret the plot in the video.	7.1%
Non-Existent Actions with Existent Scene Depictions	Assesses reasoning with introduced non-exist ent activities without changing physical details.	6.6%
Time Order Understanding	Challenges recognition of temporal sequence of activities in videos.	6.9%
Object Direction	Emphasizes perception of object movement direction.	3.8%
Camera Direction	Focuses on the direction of camera movement.	4.1%
Speed	Delves into discerning variations in speed, including absolute and relative speeds.	3.6%
Attribute Change	Centers on how attributes of objects or the entire video change over time, like size, shape, color, and more.	4.5%

LLaVA-Video-178K uniquely includes all three types of annotations: captions, open-ended questions, and multiple-choice questions.

3 Beyond Singularity: Extensive Sampling Matters

We perform experiments to explore how video representations affect the model's performance. All experiments were carried out in a video-only setting, using video data with durations from 0 to 30 seconds as our training data. We focused on evaluating how the number of frames and the number of visual tokens per frame impact model performance. Regarding the frame count, it is noteworthy that observing the effects of a high number of frames—such as over 100—does not necessarily require long videos. Our results indicate that the dynamic properties of the data render even 100 frames insufficient to fully capture the condent of a 30-second video, which typically runs at 15 FPS.

In Table 4, the first group shows an increase in the number of frames from 32 to 110. We set 110 frames as the upper limit to avoid overloading the GPU. With more frames, we see significant improvements in all datasets. While it's generally expected that using more frames boosts performance, previous studies (Luo et al., 2021; Lei et al., 2021; 2022) have noted that performance tends to plateau when training with more than 16 frames. We propose that the saturation observed in earlier studies arises due to the selection of training datasets such

```
tasks = “
# Temporal: this task is designed to assess the capability of reasoning ...<omitted>
## caption-1: The video features a child sitting in a baby chair at a dining table, creating...<omitted>
## question-1: What was the child doing as he sat on the baby chair?
## answer-1: The child was reading a book.
...
## caption-3: ...<omitted>
## question-3: ...<omitted>
## answer-3: ...<omitted>
# Spatial: this task involves creating questions that test a person’s ability...<omitted>
...<omitted> ”
system_message = “
### Task:
Given a detailed description that summarizes the content of a video, generate question-answer pairs based
on the description to help humans better understand the video. The question-answer pairs should be faithful
to the content of the video description and developed from different dimensions to promote comprehensive
understanding of the video.
Here are some question dimensions and their explanations and exampled question-answer pairs for reference:
{task_definitions}
#### Guidelines For Question-Answer Pairs Generation:
- Read the video description provided carefully, paying attention to the content, such as the scene where the
video takes place, the main characters and their behaviors, and the development of the events.
- Generate appropriate question-answer pairs based on the description. The question-answer pairs should
cover as many question dimensions and not deviate from the content of the video description.
- Generate 1 question-answer pair for each dimension.
### Output Format:
1. Your output should be formed in a JSON file.
2. Only provide the Python dictionary string.
Your response should look like:
["Dimension": <dimension-1>, "Question": <question-1>, "Answer": <answer-1>,
"Dimension": <dimension-2>, "Question": <question-2>, "Answer": <answer-2>...] ”
user_message = “
Please generate question-answer pairs for the following video description:
Description: {caption} ”

for cur_video in videos:
    sys_msg = system_messages.format(task_definitions=tasks)
    usr_msg = user_messages.format(caption=cur_video)
    response = GPT4O(sys_msg,usr_msg)
```

Table 2: We explain the process of creating prompts for GPT-4O to gather question-answer pairs from each video description. `tasks` includes the definition of all question types along with examples of question-answer pairs. We instruct GPT-4O to generate questions that cover as many question types as possible.

as MSVD (Chen & Dolan, 2011) and WebVid (Bain et al., 2021), where the video content is highly static, allowing a small number of frames to represent the entire video effectively. In contrast, the dynamic nature of the videos and the detailed nature of the annotations in LLaVA-Video-178K allow for continuous benefits from extensive sampling

The second group in Table 4 demonstrates the effects of varying the number of inference frames while keeping the number of training frames constant. A modest increase in the inference frames slightly enhances performance; however, excessively increasing the number of inference frames can degrade it.

In Table 4’s third group, we illustrates the trade-off between the number of frames and the number of tokens per frame. Configurations with fewer tokens per frame but more frames yield superior results, even with a lower total count of visual tokens (18,590 versus 21,632). This finding emphasizes that increasing the number

Table 3: **Comparison of LLaVA-Video-178K and other video-language datasets**. Average FPS represents the average number of frames per second that are used to prompt GPT-4o/GPT-4V for annotation.

	Text	#Video	Total Video Length	Average FPS	#Caption	#OE QA	#MC QA
HowTo100M (Miech et al., 2019)	ASR	136M	134.5Khr	-	136M	0	0
ACAV (Lee et al., 2021)	ASR	100M	277.7Khr	-	100M	0	0
YT-Temporal-180M (Zellers et al., 2021)	ASR	180M	-	-	180M	0	0
HD-VILA-100M (Xue et al., 2022)	ASR	103M	371.5Khr	-	103M	0	0
MSVD (Chen & Dolan, 2011)	Manual	1970	5.3h	-	1K	0	0
LSMDC (Rohrbach et al., 2015)	Manual	118K	158h	-	118K	0	0
MSR-VTT (Xu et al., 2016)	Manual	10K	40h	-	10K	0	0
DiDeMo (Anne Hendricks et al., 2017)	Manual	27K	87h	-	27K	0	0
ActivityNet (Caba Heilbron et al., 2015)	Manual	100K	849h	-	100K	0	0
YouCook2 (Zhou & Corso, 2017)	Manual	14K	176h	-	14K	0	0
TVQA (Lei et al., 2018)	Manual	21K	3.39Khr	-	0	0	152K
ActivityNet-QA (Yu et al., 2019)	Manual	5.8K	290h	-	0	58K	0
Social-IQ (Zadeh et al., 2019)	Manual	1.2K	20h	-	0	0	7.5k
NExT-QA (Xiao et al., 2021)	Manual	5.4K	66h	-	0	52K	47K
MSVD-QA (Xu et al., 2017)	Open-source Model	1.9K	5.3h	-	41K	50K	0
MSRVTT-QA (Xu et al., 2017)	Open-source Model	10K	40h	-	0	243K	0
Panda-70M (Chen et al., 2024b)	Open-source Model	70.8M	166.8Khr	-	70.8M	0	0
LLaVA-Hound (Zhang et al., 2024a)	GPT-4V	900K	3Khr	0.008	900K	900K	0
ShareGPT4Video (Chen et al., 2024a)	GPT-4V	40K	0.2Khr	0.15	40K	0	0
LLaVA-Video-178K	GPT-4o	178K	2Khr	1	178K	960K	196K

Table 4: Visual Representation Configurations and Performance Correlation. T^{train} and T^{test} are the number of frames in the training and inference stage, respectively. M/p^2: number of visual tokens per frame.

			in-domain		out-of-domain	
T^{train}	T^{test}	M/p^2	**NExT-QA** mc	**PerceptionTest** val	**EgoSchema** test	**VideoMME** wo
Training with more frames						
32	32	169	80.4	68.2	56.3	59.1
64	64	169	81.4 (+1.0)	68.3 (+0.1)	58.4 (+2.1)	59.6 (+0.5)
110	110	169	82.0 (+1.6)	68.3 (+0.1)	59.1 (+2.8)	60.4 (+1.3)
Inference with more frames						
32	32	169	80.4	68.2	56.3	59.1
32	64	169	80.7 (+0.3)	68.9 (+0.7)	56.3 (+0.0)	59.9 (+0.8)
32	110	169	80.5 (+0.1)	67.2 (-1.0)	55.2 (-1.1)	58.8 (-0.3)
Using more frames with fewer visual tokens per frame						
32	32	729	79.4	69.5	58.3	59.1
110	110	169	82.0 (+2.6)	68.3 (-1.2)	59.1 (+0.8)	60.4 (+1.3)
440	440	64	81.6 (+2.2)	67.2 (-2.3)	59.4 (+1.1)	60.2 (+1.1)

of frames, rather than the tokens per frame or the total number of tokens, enhances performance. However, a balance is necessary; as the number of frames increases to 440 and the tokens per frame decreases to 64, performance drops. This observation led us to use LLaVA-Video $_{\texttt{SlowFast}}$ for video representation.

Table 5: Comparison of different video representations. The video representation $\mathcal{V}$ is consistent in training and inference for all methods, except that SlowFast-LLaVA considers simple representation $\mathcal{V}$ in training and its specified $\mathcal{V}$ in inference.

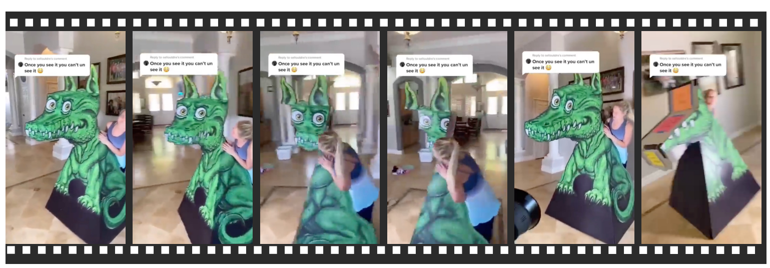
Method	$\mathcal{V} = (T, M, s, p)$	#Visual Tokens	in-domain **NExT-QA** mc	in-domain **PerceptionTest** val	out-of-domain **EgoSchema** test	out-of-domain **VideoMME** wo
Simple representation	(32, 729, 1, 2)	5,408	80.4	68.2	56.3	59.1
LLaVA-Video $_{\texttt{SlowFast}}$	(64, 729, 3, 2)	5,396	81.1	67.7	57.1	59.8
LITA	(42, 729, 2, 2)	5,313	80.8	68.3	54.3	59.1
SlowFast-LLaVA	(42, 729, 2, 2)	5,313	79.4	68.2	56.2	58.9

4 Capabilities

Beyong achieve good benchmark performance, Our observations of LLaVA-Video reveal various capabilities in video understanding. Specifically, it show a great abilities in the understanding video using real-world knowledge,including, but not limited to:

- *Optical Illusion*: As shown in Table 7, LLaVA-Video recognizes that the green dragon in the video is not a real 3D object. It appears three-dimensional due to an optical illusion that affects human perception.
- *Special Domain*: As indicated in Table 7, LLaVA-Video understands the content within special domains in the video, such as sketches and fights in video games.
- *Unusual Action*: As detailed in Table 8, LLaVA-Video identifies atypical actions in the video, such as "physical therapy" for pets, beyond ordinary activities.
- *Physical Laws*: As shown in Table 9, LLaVA-Video comprehends basic physical laws demonstrated in the video, like zero gravity in space stations, which allows objects to float without falling.

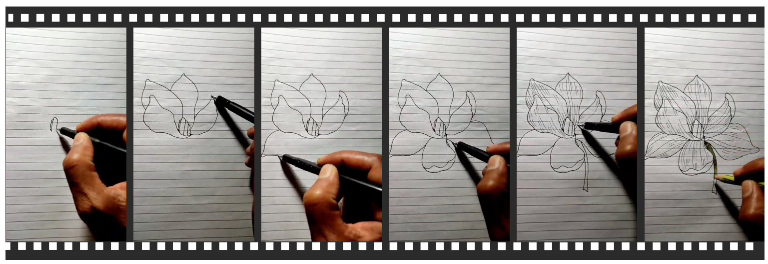
Real-world Knowledge: Illusion



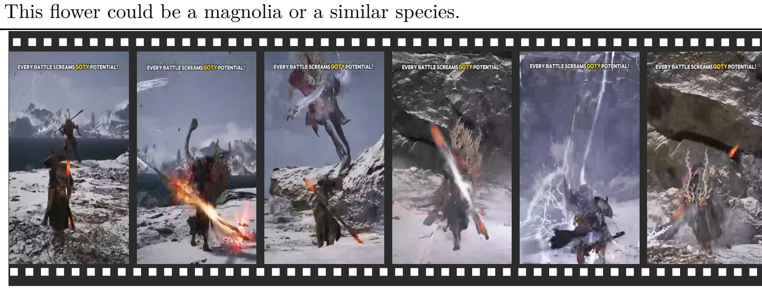
User	Please describe this video.
LLaVA-Video	The video features a person interacting with a large, cartoonish green dragon cutout in a well-lit, spacious room adorned with framed pictures on the walls. The person, dressed in a blue tank top and light blue jeans, is seen hugging and adjusting the dragon cutout, which has a wide, toothy grin and large, expressive eyes. The text overlay reads 'Reply to seloutria's comment: Once you see it you can't un see it,' suggesting that the cutout is designed to be visually striking and memorable. The person moves around the cutout, showcasing different angles and perspectives.
User	Is this green dragon a real 3D structure?
LLaVA-Video	No, the green dragon is a 2D cutout that creates an optical illusion.

Table 6: LLaVA-Video learns to understand the *optical illusion* in the video.

Real-world Knowledge: Special Domain



User	Please describe this video.
LLaVA-Video	The video begins with a blank, lined sheet of paper. A hand holding a black pen starts drawing on the paper, creating the outline of a flower with five petals. The artist adds details to the petals, including lines and shading to give them a more realistic appearance. The stem and leaves are also drawn, completing the flower illustration. The artist adds final touches to the flower drawing, enhancing the details and shading to give it a more realistic appearance. The artist then switches to a yellow pen and begins coloring the center of the flower, adding depth and contrast to the drawing. The video concludes with the artist adding final touches to the coloring, enhancing the overall appearance of the flower illustration.
User	What could this flower be?
LLaVA-Video	This flower could be a magnolia or a similar species.



User	Please describe this video.
LLaVA-Video	The video showcases an intense battle scene set in a snowy, mountainous landscape. A character dressed in dark, heavy armor with glowing orange accents wields a sword and faces off against a large, menacing creature with glowing red eyes and antlers. The text 'EVERY BATTLE SCREAMS GOTY POTENTIAL!' is prominently displayed at the top of each frame, emphasizing the high stakes and epic nature of the confrontation. The battle is fierce, with the character and the creature exchanging powerful attacks. The creature emits blue energy and lightning, while the character retaliates with fiery and icy attacks. The snowy ground is littered with debris and frost, adding to the harsh and brutal environment. The background features a large body of water and a distant lighthouse, enhancing the sense of isolation and danger.

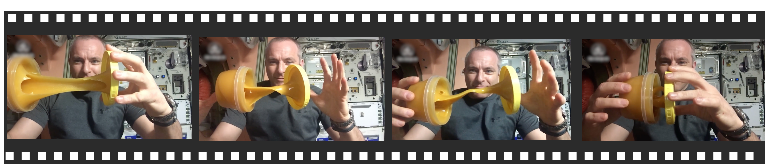
Table 7: LLaVA-Video learns to understand the the video in *special domain*.

Real-world Knowledge: Unusual Action

User	Please describe this video.
LLaVA-Video	The video begins with a person holding a tabby cat, with the text 'DOGS WITH BACK PAIN' and a crying emoji displayed at the bottom of the screen. The person, who has tattoos on their arms and is wearing a light-colored shirt, is in a room with large windows and a white wall. The scene transitions to the person petting a large, brown and white dog, who is sitting on the floor. The person then stands up and gestures with their hands, while the dog looks up at them. The next scene shows the person kneeling and petting a large, fluffy dog with a black and white coat, who is sitting on a table. The person continues to pet the dog, who appears calm and relaxed.
User	What did this man do to his pets?
LLaVA-Video	This man provided physical therapy to his pets.

Table 8: LLaVA-Video learns to understand the *unusual action* in the video.

Real-world Knowledge: Physical Laws



User	Please describe this video.
LLaVA-Video	The video features a person in a space station environment, wearing a dark gray shirt and a wristwatch. They are holding two containers filled with a thick, yellowish substance. The person tilts the containers to demonstrate the fluid's behavior, showing how it stretches and moves between the containers. The background includes various equipment and panels typical of a space station, with labels such as 'GALLEY' and 'FUSION'. The person continues to manipulate the containers, emphasizing the fluid's properties. The video concludes with the person holding one container and twisting the lid to close it, further demonstrating the fluid's properties.
User	Why doesn't this yellowish substance fall?
LLaVA-Video	This yellowish substance doesn't fall because it is in zero gravity.

Table 9: LLaVA-Video learns to understand the *physical laws* in the video.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198. 1

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017. 5

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 4

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*, pp. 961–970, 2015. 5

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011. 4, 5

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024a. 5

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024b. 5

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019. 1

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. *arXiv preprint arXiv:2403.19046*, 2024. 1

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10274–10284, 2021. 5

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 5

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7331–7341, 2021. 3

Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 3

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597. 1

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 3

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 5

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 (ICML)*, pp. 8748–8763. PMLR, 2021. 1

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3202–3212, 2015. 5

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021. 5

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 *ACM Multimedia*, 2017. 5

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*, pp. 5288–5296, 2016. 5

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024a. 1, 2

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models, 2024b. URL `https://arxiv.org/abs/2407.15841`. 1

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pp. 9127–9134, 2019. 5

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019. 5

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems*, 34:23634–23651, 2021. 5

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. 1, 2

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward, 2024a. 5

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/. 1, 2

Luowei Zhou and Jason J. Corso. Youcookii dataset. 2017. URL https://api.semanticscholar.org/CorpusID:19774151. 5