4D-Bench: Benchmarking Multi-modal Large Language Models for 4D Object Understanding

Supplementary Material

A. License

4D-Bench is strictly for academic research purposes, and any form of commercial use is prohibited. The copyright of all 4D objects is retained by their respective owners, and proper acknowledgement will be given in the dataset. The dataset as a whole is licensed under the ODC-By v1.0 license, consistent with the licensing of Objaverse-XL[9].

B. More related work

Benchmark datasets for image and video captioning. The development of image captioning has been driven by several influential datasets[1, 6, 11]. COCO [6] stands as one of the most widely used benchmarks and covers diverse daily scenes and objects, making it a robust benchmark for evaluating captioning models. The ground-truth captions provided by early benchmark datasets typically contain limited information. Yet, recent MLLMs are capable of generating more detailed captions, making these datasets not challenging enough for evaluating modern models' capabilities of producing rich, nuanced descriptions that capture fine-grained visual details and complex relationships between objects. To fill this gap, Dong et al.[11] propose DetailCaps, a new benchmark featuring longer and more detailed captions annotated by human experts and powerful MLLMs like GPT-4V. On the other hand, several datasets[5, 14, 18, 20, 43, 45, 52] have been established for 2D video captioning. MSR-VTT[45] provides 20 descriptions per video clip for the open domain 2D video captioning. ActivityNet Captions[18] provide temporally localized multiple-sentence descriptions for video captioning. For domain-specific applications, YouCook2[52] presents task-oriented instructional cooking videos.

Reference-free captioning metrics. We use reference-based metrics [4, 10, 16, 24, 30, 33, 37, 40] in the main paper. Recently, reference-free caption metrics[15, 17, 19, 23, 34, 35] has emerged in the image and video captioning metrics field. Reference-free metrics eliminate the need for human-annotated references, reducing evaluation costs and effort. They are also ideal for open-ended scenarios, accommodating multiple valid image descriptions and overcoming the limitations of reference-based methods that rely on potentially incomplete captions. For example, CLIPScore[15] uses CLIP embeddings to compute the similarity between generated captions and their associated visual content, offering a flexible way to assess captions in open-ended settings.

C. More details about 4D-Bench

C.1. More details about 4D object representation

We chose multi-view videos as the representation for 4D objects, as we found recent advanced MLLMs [3, 7, 8, 12, 13, 21, 25, 25–29, 31, 36, 38, 39, 42, 44, 46–51] are primarily designed to take texts and 2D images/videos as inputs.

We render the multi-view videos for 4D objects collected from Objaverse-XL[9]. For each 4D object, we render a 2D video from a single view up to 125 frames and utilize pixel change detection to identify motion within the 2D video, determining the frame indices for the start and end of the motion. Based on these indices, we render videos from 23 additional views, ensuring that all 24-view videos cover the identified motion frames. The camera positions are evenly distributed around the normalized 4D object with slight jitters, the camera positions are chosen with a radius from 2.2m to 2.6m and a height from 0.8m to 1.2m.

C.2. More details about CLIP-based data curation

We propose a CLIP-based classifier to automatically select high-quality 4D objects, such that low-quality ones, such as

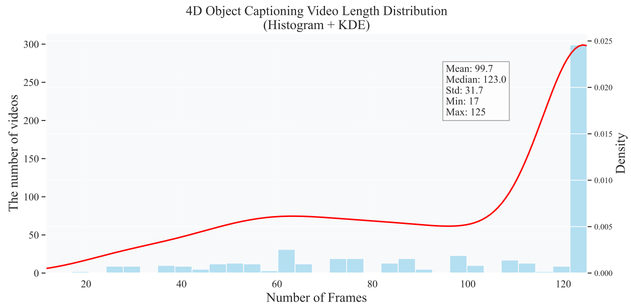


Figure I. The frame-length distribution of multi-view videos used in the 4D object captioning task

oversimplified geometry, lack of texture, and poor aesthetic quality, are removed.

To build the training dataset, we manually annotate thousands of 4D objects into three categories: high quality, textureless, and low overall quality. The "low overall quality" category typically refers to objects with significant deformation or portions that are largely outside the camera view. After that, for each object, we choose the first frame of the video from the first view and its corresponding label to build the training dataset. We build the CLIP-based classifier by adding a linear layer as the classification head to fine-tune the CLIP visual encoder, and then use this dataset to fine-tune the classifier.

During inference, we feed the first frame from 8 views of the 4D object into the CLIP-based classifier. The final label of the object is determined through majority voting across the predictions made for these eight images. Objects classified as high quality are retained, ensuring the dataset is highly usable.

C.3. Additional statistical data analysis

4D object captioning statistics. For the 4D object captioning task, we collected 580 4D objects, where each object is rendered into 24-view videos and has 5 human-annotated captions. Fig. I shows the frame-length distribution of multi-view videos, where the videos contain 99.73 frames per 4D object on average. The human-annotated captions have an average length of 19.05 words, and their length distribution is illustrated in Fig. II.

4D object question answering statistics. In the 4D object QA dataset, the multi-view videos contain an average of 101 frames per object, with the frame length distribution shown in Fig. III. Fig. V illustrates that the length distributions of the answer options are roughly similar, avoiding bias caused by answer length.

C.4. More details about evaluation metrics

Fig. VII and Fig. VIII present the prompt template designed to guide GPT-4o in assessing the correspondence between

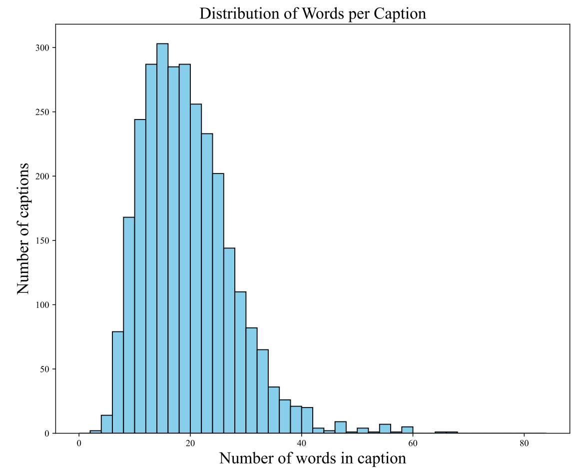


Figure II. The length distribution of ground-truth captions used in the 4D object captioning task

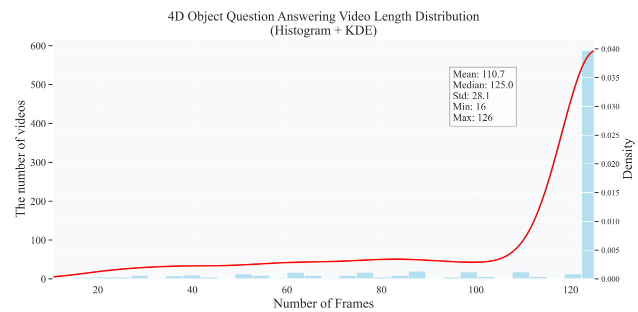


Figure III. The frame-length distribution of multi-view videos used in the 4D object question answering task

generated and human-annotated captions in terms of appearance and action descriptions. The prompt templates incorporate a comprehensive scoring rubric ranging from 0 to 5, where each score level is defined based on the accuracy and completeness of visual appearance/action descriptions. To ensure consistent evaluation, the template also provides carefully selected example pairs of human and machine-generated captions, along with their corresponding scores.

D. More experimental details on 4D-Bench

D.1. More experimental details of 4D object captioning

In the 4D object captioning experiments, all models adhere to a common function $C = M(V, t)$, where V, t, M and C denote the multi-view video input, text prompt (instruction), MLLM being tested, and generated caption respectively. The quality of generated captions is evaluated by computing various metric scores through comparison with human-annotated reference captions.

Fig. IV shows the prompt we use to prompt the MLLMs

4D Object Captioning Prompt Template

I have multiple videos of the object captured from different angles. I provide you 18 images, with every six images uniformly sampled from one video, each video captured from a different angle. Your job is to generate one fluent caption for this multi-view video in English, provide a detailed description of the object's or character's appearance, including shape, color, texture, and any notable features. Additionally, describe the actions taking place, focusing on how the object or character moves and behaves throughout the scene. The caption should not describe the background. You must strictly return in the following format: caption: caption content. Here are some examples:
Example 1: caption: A young woman with black hair wearing silver jumpsuit is lying on the floor and then gently rises.
Example 2: caption: A military infantryman in green and brown camouflage gear holds a black pistol in his left hand and dances with his arms and legs moving first to the left then to the right.
Example 3: caption: A 3D model of a fish pond with blue walls, and brown ground, a fish swims next to a creature that looks like an animal that is lying down.
Example 4: caption: 3D model of a yellow emoji with closed eyes that sticks out its red tongue and moves from right to left.
Example 5: caption: A man with brown hair, a moustache and sunglasses wears a green coat, black pants, a white shirt and a black tie walks straight then turns raising his right hand up.

Figure IV. The prompt provided to the evaluated MLLMs in the 4D object captioning task. In this prompt, we describe the video information, caption requirement, and output format. We also provide several caption examples to guide the style of captions generated by MLLMs.

to generate captions. It's notable that we give them caption examples because we found that different MLLMs may generate captions in vastly different styles when not provided with examples, which could impact the results due to stylistic variations. By providing examples, we aim to minimize the influence of different writing styles, allowing us to control experimental variables better and obtain more objective evaluation results that reflect the models' actual understanding capabilities rather than differences in writing style.

D.2. More experimental details of 4D object question answering

In the 4D object question answering experiments, all models operate under a shared function $P(A) = M(V, t, QA)$, where V, t, QA, M, A, and P represent the multi-view

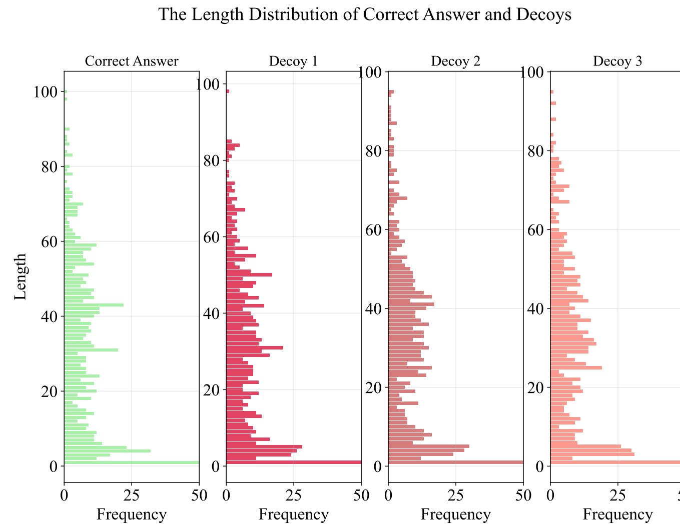


Figure V. The truncated length distribution of correct answers and decoys used in 4D object question answering dataset

video input, text prompt (instruction), question and four answer options, MLLM being tested, model output, and output processor, respectively. We add output processor to extract the selected answer option as we found that some open-source models sometimes struggled to strictly follow the prompt instructions that explicitly defined the required output format. Fig. VI shows the prompt we use to prompt the MLLMs to complete the 4D Object QA task.

4D Object QA Prompt Template

You are an excellent video analyst. I provide you 18 frames with every six images uniformly sampled from one video, each video captured from a different angle and a question and four choices. Carefully watch the provided videos and pay attention to every detail. Based on your observations, select the best option that accurately addresses the question. Here is the question and choices: <4D Object QA>. You must return only the option identifier (e.g., '(A)') without any additional text, do not add any additional analysis, just return the correct option identifier.

Figure VI. The prompt provided to the evaluated MLLMs in the 4D object QA task. In this prompt, we detailed the video information, questions and options, and the output format.

Since some open-source MLLMs may not always strictly follow the specified output format requirements, we implemented an output processor function to standardize answer extraction using the following code. This function is designed to extract a single letter answer choice (A, B, C, or D) from model responses. It first attempts to find a letter enclosed in parentheses (e.g., "(A)"). If no match is found, it looks for standalone letters that are bordered by spaces or punctuation marks to ensure consistent extraction regardless of the response format.

```
def extract_answer_option(text):
    paren_pattern = r'\(([A-D])\)'
    matches = re.findall(paren_pattern, text)
    if matches:
        return matches[0]
    isolated_pattern = r'(?:^|[\s\(\.,;:])([A-D])
        (?:[\s\)\.,;:]|$)'
    matches = re.findall(isolated_pattern, text)
    if matches:
        return matches[0]
    return None
```

E. Additional evaluation results on 4D-Bench

In this section, we first provide additional analysis for the 4D object captioning in Sec. E.1. Then, Sec. E.2 and Sec. E.3 provide additional evaluation results on the 4D object captioning and 4D object QA tasks of 4D-Bench, respectively.

E.1. Analysis for 4D object captioning evaluation

E.2. Additional qualitative results of 4D object captioning

Figs. IX, X XI and XII show 4D object captioning results of MiniGPT4-Video [2], VideoChat2-Mistral [22], Qwen2-VL-7B [41] and Gemini 1.5 Pro [32], given various 4D objects in our 4D-Bench. For example, Fig. IX illustrates MiniGPT4-Video, VideoChat2-Mistral, Qwen2-VL-7B, and Gemini 1.5 Pro achieve low GPT-Action scores.

E.3. Additional qualitative results of 4D Object questing answering

Figs. XIII, XIV, XV and XVI illustrate more 4D object QA results of advanced MLLMs. Fig. XIV shows an easy sample on the subtask of *Temporal Relationship*, where all MLLMs choose the correct answer except for GPT-4o. Fig. XV shows a more difficult example of *Temporal Relationship*, where Qwen2-VL 7B, GPT-4o and LLava-Video picks the wrong answer. Fig. XVI shows qualitative results of MLLMs on the *Object Counting* subtask, where only LLava-Video 7B answered the question correctly. Fig. XIII illustrates all MLLMs (including GPT-4o and Gemini 1.5 pro) pick the wrong option on the subtask of *Action*, indicating the limited capabilities of MLLMs in action understanding of 4D objects.

GPT-Appearance Metric Prompt Template

GPT-Appearance Metric System Prompt

You are an expert in evaluating the quality of video captions. Your task is to rate the predicted caption in terms of recall and precision of visual elements(appearance and shape) in the video with reference to the human-annotated caption. Focus only on whether the predicted caption accurately and completely contains the information from the human-annotated caption. Note you just need to focus on the visual elements. Consider synonyms or paraphrases as valid matches. Provide your evaluation as a matching score where the score is an integer value between 0 and 5. Here is the rating scale:
Score 5: The predicted caption accurately identifies the object in the video, including its appearance and shape. The caption provides a precise and complete description of the object without missing any significant visual details.
Score 4: The predicted caption mostly identifies the object accurately, with minor omissions or differences in the description of the appearance or shape. Paraphrases are acceptable, and the overall description is correct, though it may lack some finer details.
Score 3: The caption identifies some key aspects of the object but misses or incorrectly describes certain visual elements, such as the appearance or shape. There are noticeable gaps, but the overall object is still somewhat recognizable in the caption.
Score 2: The predicted caption contains several inaccuracies in describing the object's appearance or shape. While some parts may be correct, the overall description is incomplete or misleading. Precision and recall of visual elements are low.
Score 1: The caption provides an incorrect description of the object, with major inaccuracies in identifying the appearance and shape. The object is either misidentified or described in a way that does not match the video.
Score 0: The caption is entirely incorrect, failing to identify the object or its appearance and shape. No valid matches to the human-annotated caption are present.
Here are some rating examples:
Example 1:{ Human_Caption: 'A red wrecking ball with black chains swings into a big brown cube sitting on a metallic surface that scatters into smaller cubes after being hit'. Predicted Caption: 'a cube and ball connected by a chain'. Score: {'appearance_score': 1} }
Example 2:{ Human_Caption: 'A woman wearing a pair of combat pants and a tank top throwing a punch'. Predicted Caption: 'a woman in a boxing outfit, wearing a hat, hoodie, and camouflage pants, holding a gun'. Score: {'appearance_score': 3} }
Example 3:{ Human_Caption: 'Azerbaijan flag that moves with the wind'. Predicted Caption: 'the Azerbaijan flag waving in the wind and a colorful kite'. Score: {'appearance_score': 2} }
Example 4:{ Human_Caption: '3D model of arms with gray sleeves carrying a gray pistol with brown grip and gray barrel that loads it, fires two bullets, then unloads it'. Predicted Caption: 'A pair of human-like arms in a dark grey sweater holding a handgun with a brown grip and black barrel'. Score: {'appearance_score': 4} }
Example 5:{ Human_Caption: '3D model of a boy wearing glasses dancing dressed in a grey hood, black pants, gray shoes, he puts on a red cap and a blue backpack'. Predicted Caption: 'a person wearing a pink hat, holding a sword, and surrounded by a glider, bird, and windmill, all adorned with pink hats'. Score: {'appearance_score': 0} }
Example 6:{ Human_Caption: 'A 3D model of a lightsaber which is emitting blue saber'. Predicted Caption: 'light saber, and flashlight'. Score: {'appearance_score': 5} } "

GPT-Appearance Metric User Prompt

Please evaluate the following video-based captions:
Human-annotated Caption: <HUMAN CAPTION>
Predicted Caption: <PREDICTED CAPTION>
Please generate the response in the form of a dictionary string with the key 'appearance_score', where its value is the factual accuracy score in INTEGER, not STRING.
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. You must follow this command!
For example, your response should look like this: {'appearance_score': 4}.

Figure VII. Prompt used in GPT-Appearance metric

GPT-Action Metric Prompt Template

GPT-Action Metric System Prompt

You are an expert in evaluating the quality of video captions. Your task is to rate the predicted caption in terms of recall and precision of the object's actions in the video with reference to the human-annotated caption. Note you just need to focus on the action descriptions in the captions.
Consider synonyms or paraphrases as valid matches. Provide your evaluation as a matching score where the score is an integer value between 0 and 5. Here is the rating scale:
Score 5: The predicted caption accurately identifies the actions of the object in the video, including the sequence, timing, and details of the actions. Synonyms or paraphrases are valid matches. The caption provides a precise and complete description of the actions without missing any significant aspects.
Score 4: The predicted caption mostly identifies the actions accurately, with minor omissions or differences in the description of the actions. Paraphrases are acceptable, and the overall description is correct, though it may lack some finer details.
Score 3: The caption identifies some key actions but misses or incorrectly describes certain details, such as timing, order, or subtle movements. There are noticeable gaps, but the overall actions are still somewhat recognizable in the caption.
Score 2: The predicted caption contains several inaccuracies in describing the object's actions. While some parts may be correct, the overall description is incomplete or misleading. Precision and recall of actions are low.
Score 1: The caption provides an incorrect description of the object's actions, with major inaccuracies in identifying the actions or their sequence. The actions are either misidentified or described in a way that does not match the video.
Score 0: The caption is entirely incorrect, failing to identify the object's actions. No valid matches to the human-annotated caption are present.
Here are some rating examples:
Example 1:{ Human_Caption: '3D model of a woman covered in white and purple mesh is warming up and shadow boxing'. Predicted Caption:' a figure with a purple and black grid-like texture is running in place, their arms swinging at their sides and their legs lifting up alternately.' Score: {'action_score': 1} }
Example 2:{ Human_Caption: 'A white and yellow star wars sitting on his knees squatting, stretches his right arm and back'. Predicted Caption:' this is a 3d model of a clone trooper with yellow markings on his helmet, shoulders, knees, and shins. he is crouching down on one knee, wearing white armor with grey accents and a utility belt. the 327th star corps emblem is visible on his left shoulder.' Score: {'action_score': 3} }
Example 3:{ Human_Caption: 'Black puppy with white nose wiggling its tail.' Predicted Caption: 'a low-poly dog with a black body and white paws and face stands still. its tail is black, and its ears are floppy. the dog is rendered in a minimalist style. it remains stationary throughout the scene.' Score: {'action_score': 2} }
Example 4:{ Human_Caption: 'A ninja-looking robot in black and red armor with a shield and sword is jumping up, twisting and slashing the air with his sword before landing down.' Predicted Caption: 'a red and black armored warrior, adorned with a demonic mask, engages in a display of martial prowess, wielding both a gleaming sword and a circular shield with a blue emblem. they leap, twirl, and strike dynamic poses, their movements fluid and controlled.' Score: {'action_score': 4} }
Example 5:{ Human_Caption: 'A 3D model of a green turtle with a brown shell swimming.' Predicted Caption: 'a low-poly 3d model of a green sea turtle with a brown shell.' Score: {'action_score': 0} }
Example 6:{ Human_Caption: 'An animated bearded man with brown hair, red beanie and a flannel shirt is wielding an axe in his right hand while running.' Predicted Caption: 'a cartoon lumberjack with a big black beard, wearing a red beanie hat, red and black plaid shirt, blue pants, brown boots, and brown gloves runs while carrying an axe.' Score: {'action_score': 5}}

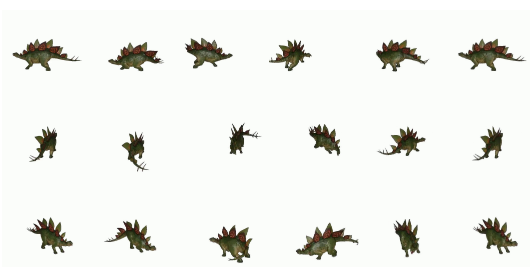
GPT-Action Metric User Prompt

Please evaluate the following video-based captions:
Human_Caption: <HUMAN CAPTION>
Predicted Caption: <PREDICTED CAPTION>
Please generate the response in the form of a dictionary string with the key 'action_score', where its value is the factual accuracy score in INTEGER, not STRING.
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. You must follow this command!
For example, your response should look like this: {'action_score': 4}.

Figure VIII. Prompt used in GPT-Action metric

Models	Generated Caption
Human	1. 3D depiction of a metal robot with a hook at the end of its arm attempting to grasp a soda can placed on a table in front of it. 2. a yellow mechanical hook attempts to grab a soda can sitting on a red and black table. …….
MiniGPT4-Video	This object has a circular base with textured surface that holds red and gold liquid, while two metallic arms attach to the sides leading towards an arm which reaches out for something. Throughout this video footage, we see movement along both horizontal axis as well as vertical but not on any rotational axis in either direction or angles shown here since these details are omitted from caption information provided. GPT-Appearance Score: 2; GPT-Action Score: 2; GPT-Eval Score: 2
VideoChat2-Mistral	a robot arm with a red base rotates in a circle. GPT-Appearance Score: 2; GPT-Action Score: 1; GPT-Eval Score: 1.5
Qwen2-VL 7B	A mechanical arm, painted in a golden hue, is seen in action, lifting a red can from one table and placing it on another. The arm moves smoothly, demonstrating its precision and functionality. GPT-Appearance Score: 3; GPT-Action Score: 2; GPT-Eval Score: 2.5
Gemini 1.5 Pro	A rusty yellow robotic arm picks up a red can and then puts it down. GPT-Appearance Score: 2; GPT-Action Score: 2; GPT-Eval Score: 2

Figure IX. Qualitative results of different MLLMs on the 4D object captioning task of 4D-Bench



Models	Generated Caption
Human	1. A large green dinosaur with a long tail and several spikes at its back, moving backward and throwing its tail to one side. 2. A green stegosaurus with brown kite-shaped upright plates along the back and two spikes on the tail is crouching and then spinning around and then getting back in the original position. 3. …….
MiniGPT4-Video	This object has a circular base with textured surface that holds red and gold liquid, while two metallic arms attach to the sides leading towards an arm which reaches out for something. Throughout this video footage, we see movement along both horizontal axis as well as vertical but not on any rotational axis in either direction or angles shown here since these details are omitted from caption information provided. GPT-Appearance Score: 2; GPT-Action Score: 1; GPT-Eval Score: 1.5
VideoChat2-Mistral	a green dinosaur with red spikes on its back is spinning in the air. GPT-Appearance Score: 4; GPT-Action Score: 3; GPT-Eval Score: 3.5
Qwen2-VL 7B	A green dinosaur with red spikes on its back and tail moves in a circular motion on a white background. GPT-Appearance Score: 3; GPT-Action Score: 5; GPT-Eval Score: 4
Gemini 1.5 Pro	A green stegosaurus with brown plates on its back and spikes on its tail is walking. GPT-Appearance Score: 3; GPT-Action Score: 2; GPT-Eval Score: 2.5

Figure X. Qualitative results of different MLLMs on the 4D object captioning task of 4D-Bench

Models	Generated Caption
Human	1. female warrior character from the game in a black bikini and black and yellow armor with a short white haircut holds a large long gray metal sword and chops with it 2. A female warrior character from the game, dressed in a black bikini and black-and-yellow armor with a short white haircut, wields a large, long gray metal sword and performs a chopping motion with it. 3. ……
MiniGPT4-Video	A powerful barbarian wields a massive sword and holds it up in the air, then slashes downward with precision. The camera captures multiple angles of this fearsome display as he continues to move around before ending abruptly by holding his arms parallel but not engaging them into violence or any action against another person onscreen. GPT-Appearance Score: 2; GPT-Action Score: 3; GPT-Eval Score: 2.5
VideoChat2-Mistral	a cartoon character with a large sword is spinning around in circles. GPT-Appearance Score: 3; GPT-Action Score: 3; GPT-Eval Score: 3
Qwen2-VL 7B	A character with blonde hair, wearing a black and yellow outfit with a red cape, wields a large sword, performing various sword moves and turns. GPT-Appearance Score: 3; GPT-Action Score: 4; GPT-Eval Score: 3.5
Gemini 1.5 Pro	A 3D model of a blonde female warrior wearing a black and gold armor set and an orange cape holds a large silver sword with her two hands and dances moving her legs and arms up and down. GPT-Appearance Score: 3; GPT-Action Score: 2; GPT-Eval Score: 2.5

Figure XI. Qualitative results of different MLLMs on the 4D object captioning task of 4D-Bench

Models	Generated Caption
Human	1. a middle-aged man in a cap, a checked green and white shirt, gray trousers and brown shoes walks straight 2. A middle-aged man with a beer belly and stubble wearing a grey flat cap, grey pants, brown shoes and a white and grey checkered shirt with sleeves rolled up is walking forward. 3. ……
MiniGPT4-Video	This man in a green and white plaid shirt with sunglasses walking quickly, swings his right arm to point up at some kind of sensor which then rotates downwards before extending forward towards the camera. GPT-Appearance Score: 3; GPT-Action Score: 2; GPT-Eval Score: 2.5
VideoChat2-Mistral	a man wearing a plaid shirt and a hat walks around in a circle. GPT-Appearance Score: 3; GPT-Action Score: 2; GPT-Eval Score: 2.5
Qwen2-VL 7B	A man with gray hair and a white cap walks back and forth, wearing a plaid shirt, jeans, and brown shoes. GPT-Appearance Score: 3; GPT-Action Score: 3; GPT-Eval Score: 3
Gemini 1.5 Pro	A man with short grey hair, wearing a grey flat cap, a black and white checkered shirt, blue jeans and brown shoes is walking in one direction. GPT-Appearance Score: 3; GPT-Action Score: 5; GPT-Eval Score: 4

Figure XII. Qualitative results of different MLLMs on the 4D object captioning task of 4D-Bench

======================= Question =======================

What action does the electric kettle perform after opening its lid in the multi-view video?

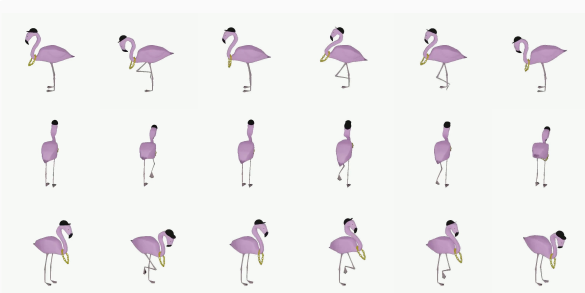
======================= Choices =======================

(A) The electric kettle slides toward the wall.
(B) The electric kettle changes color.
(C) The electric kettle lifts off the surface.
(D) The electric kettle boils water visibly.

====================== MLLM Results ======================

Gemini 1.5 Pro: **(D)** Qwen2-VL 7B: **(C)** LLava-Video 7B: **(C)** GPT-4o: **(C)**

Figure XIII. Qualitative results of different MLLMs on the 4D object question answering task of 4D-Bench



======================= Question =======================

What sequence of actions does the pink flamingo?

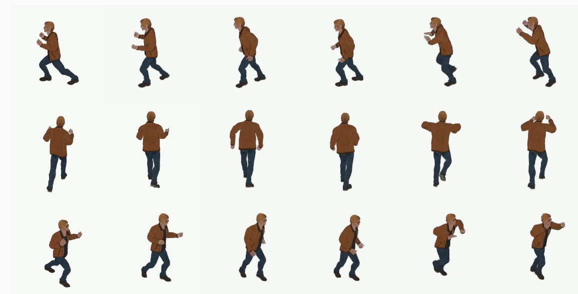
======================= Choices =======================

(A) It raises its left leg first, then lowers its head, and finally raises its right leg.
(B) It raises its right leg, lowers its head, and then raises its left leg while lifting its head.
(C) It raises its right leg, then raises both wings, and lowers its right leg.
(D) It raises its right wing, raises its left leg, and lowers its head.

====================== MLLM Results ======================

Gemini 1.5 Pro: **(B)** Qwen2-VL 7B: **(B)** LLava-Video 7B: **(B)** GPT-4o: **(A)**

Figure XIV. Qualitative results of different MLLMs on the 4D object question answering task of 4D-Bench



======================= Question =======================

During the dance, which foot does the man lift off the ground first?

======================= Choices =======================

(A) Both feet simultaneously
(B) Right foot
(C) Hops on one foot
(D) Left foot

====================== MLLM Results ======================

Gemini 1.5 Pro: **(D)** Qwen2-VL 7B: **(B)** LLava-Video 7B: **(B)** GPT-4o: **(B)**

Figure XV. Qualitative results of different MLLMs on the 4D object question answering task of 4D-Bench

======================= Question =======================

How many burnt cigarettes are there?

======================= Choices =======================

(A) Nine burnt cigarettes
(B) Zero burnt cigarettes
(C) Three burnt cigarettes
(D) Two burnt cigarettes

====================== MLLM Results ======================

Gemini 1.5 Pro: **(D)** Qwen2-VL 7B: **(D)** LLava-Video 7B: **(C)** GPT-4o: **(D)**

Figure XVI. Qualitative results of different MLLMs on the 4D object question answering task of 4D-Bench

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 1

[2] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. 4

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 1

[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*, pages 190–200, 2011. 1

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1

[7] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1

[8] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024. 1

[9] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[10] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002. 1

[11] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024. 1

[12] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024. 1

[13] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. 1

[14] Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*, 2023. 1

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 1

[16] Wooyoung Kang, Jonghwan Mun, Sungjun Lee, and Byungseok Roh. Noise-aware learning from web-crawled image-text data for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2942–2952, 2023. 1

[17] Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual information divergence: A unified metric for multimodal generative models. *Advances in Neural Information Processing Systems*, 35:35072–35086, 2022. 1

[18] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1

[19] Yebin Lee, Imseong Park, and Myungjoo Kang. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. *arXiv preprint arXiv:2406.06004*, 2024. 1

[20] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020. 1

[21] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 1

[22] 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

[23] Hao Liang, Zirong Chen, and Wentao Zhang. Evqascore: Efficient video question answering data evaluation. *arXiv preprint arXiv:2411.06908*, 2024. 1

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 1

[25] 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. 1

[26] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

[27] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024.

[28] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025.

[29] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 1

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 1

[31] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1

[32] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 4

[33] Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. Cider-r: Robust consensus-based image description evaluation. *arXiv preprint arXiv:2109.13701*, 2021. 1

[34] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6914–6924, 2023. 1

[35] Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17929–17938, 2022. 1

[36] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 1

[37] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, 2006. 1

[38] Reka Team, Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, et al. Reka core, flash, and edge: A series of powerful multimodal language models. *arXiv preprint arXiv:2404.12387*, 2024. 1

[39] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1

[40] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4

[42] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[43] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591, 2019. 1

[44] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024. 1

[45] 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. 1

[46] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 1

[47] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[48] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

[49] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.

[50] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan,

Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[51] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. 1

[52] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1