HAIC: Improving Human Action Understanding and Generation with Better Captions for Multi-modal Large Language Models

Xiao Wang[1*†] **Jingyun Hua**[2*] **Weihong Lin**[2*] **Yuanxing Zhang**[2]
Fuzheng Zhang[2] **Jianlong Wu**[1‡] **Di Zhang**[2] **Liqiang Nie**[1‡]
[1]Harbin Institute of Technology, Shenzhen [2]Kuaishou Technology
scz.wangxiao@gmail.com, huajingyun@hotmail.com, lwher1996@outlook.com, longo@pku.edu.cn
zhfzhkris@outlook.com, jlwu1992@pku.edu.cn, zhangdi08@kuaishou.com, nieliqiang@gmail.com

Abstract

Recent Multi-modal Large Language Models (MLLMs) have made great progress in video understanding. However, their performance on videos involving human actions is still limited by the lack of high-quality data. To address this, we introduce a two-stage data annotation pipeline. First, we design strategies to accumulate videos featuring clear human actions from the Internet. Second, videos are annotated in a standardized caption format that uses human attributes to distinguish individuals and chronologically details their actions and interactions. Through this pipeline, we curate two datasets, namely HAICTrain and HAICBench. **HAICTrain** comprises 126K video-caption pairs generated by Gemini-Pro and verified for training purposes. Meanwhile, **HAICBench** includes 412 manually annotated video-caption pairs and 2,000 QA pairs, for a comprehensive evaluation of human action understanding. Experimental results demonstrate that training with HAICTrain not only significantly enhances human understanding abilities across 4 benchmarks, but can also improve text-to-video generation results. Both the HAICTrain and HAICBench are released at https://huggingface.co/datasets/KuaishouHAIC/HAIC.

1 Introduction

Multi-modal large language models have notably showcased their preeminence across various video understanding tasks (Chen et al., 2024e; Li et al., 2024a; Wang et al., 2024a; Zhang et al., 2024b). Among these, human action understanding plays a critical role in many downstream applications, e.g., human-computer interaction (Hayes, 2011), autonomous driving (Xu et al., 2021), embodied intelligence (Gupta et al., 2021), and human video generation (Wang et al., 2024b).

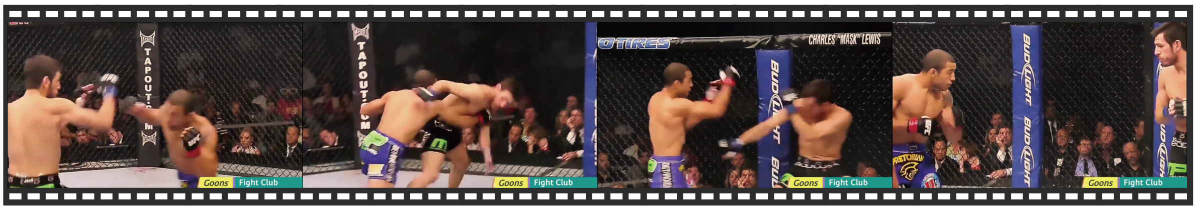
A recent study, ShareGPT4Video (Chen et al., 2024a) has demonstrated that high-quality and detailed video captions can improve MLLMs' performance in video understanding. However, most existing works (Soomro et al., 2012; Xu et al., 2016; Krishna et al., 2017; Wang et al., 2019; Chen et al., 2024d) provide only coarse captions for human actions, insufficient for understanding fine-grained behaviors. MotionLLM (Chen et al., 2024b) introduces the MoVid dataset with fine-grained action captions from MotionX (Lin et al., 2023). Nevertheless, this dataset mainly focused on single-person scenarios. For multi-person situations, MoVid only considered consistent group activities like "a group of people performing the Korean dance". A comprehensive dataset is essential to enhance MLLMs' understanding of detailed human actions and interactions in both single- and multi-person contexts, critical for tasks like emotional analysis, motivation prediction, and relationship modeling.

There are two main challenges for building such datasets: (1) *Action Video Accumulating.* How to automatically accumulate large-scale videos featuring clear actions of multiple individuals. (2) *Caption formatting.* How to define a caption format that can clearly distinguish different people and detail their behaviors and interactions respectively.

To address the above challenges, we propose a novel data generation pipeline composed of two stages. In the *video accumulation stage*, we accumulate videos from various domains featuring clear, meaningful human actions and identify their specific timestamps. This process is highly selective, with only about 1% of videos meeting our quality criteria after applying various strategies. In the *caption annotation stage*, we define a caption format that uses human attributes to distinguish individuals and chronologically annotates detailed body actions and interactions for each person (see Figure 1). With this pipeline,

*Equal contribution.
†Work done as an intern at Kuaishou Technology.
‡Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10158–10181



MSR-VTT: Two men are boxing on a stage.

LLaVA-Video: The video opens with two fighters in a fighting stance inside an octagonal cage. The fighter on the left wears black shorts with green and white accents, while the fighter on the right sports blue shorts with yellow and white details. Both fighters are equipped with gloves and appear to be in the midst of a match. The sequence shows the fighters exchanging punches and grappling near the center of the cage. The audience is visible in the background, watching the fight intently. The fighters move around the cage, with one attempting to land strikes while the other defends and counters. The video concludes with the fighters still engaged in combat, with the referee closely monitoring the action.

HAIC: The main characters are two people.

The first subject is a middle-aged man, wearing purple boxing shorts with black and green patterns and white letters, and black and red boxing gloves. Initially, he faces his opponent in an attacking stance, then bends down and rushes towards Subject 2, throwing a punch. After Subject 2 dodges, he chases Subject 2 and throws five right hooks.

The second subject is a middle-aged man, wearing black boxing shorts with green patterns and white letters, and black and blue boxing gloves. Initially, he faces Subject 1, then he bends down to dodge Subject 1's attack, stumbles, and is chased and punched by Subject 1. Finally, he straightens up, puts his hands over his head, breaks away from the attack, and the two confront each other.

Figure 1: Our standardized caption format presents each individual's detailed attributes, body actions, and interactions in chronological order, making it easier to distinguish individuals and comprehend their behaviors.

we curate two datasets: **HAICTrain** (**H**uman **A**ction and **I**nteraction **C**omprehension **Train**ing set) and **HAICBench** (a benchmark for evaluation). HAICTrain contains 126K videos accumulated from WebVid (Bain et al., 2021), annotated in our defined caption format by Gemini-1.5-Pro (Team et al., 2024). HAICBench includes 412 YouTube videos with human-annotated captions in the same format. Furthermore, we generate 2,000 multiple-choice QA pairs across five categories—human interaction, action detail, action sequence, count, and human attribute—by prompting GPT-4o (Hurst et al., 2024) and Gemini-1.5-Pro. Note that all the above machine annotation results undergo review and refinement by human annotators. Experimental results indicate that utilizing HAICTrain for training can remarkably enhance the model's human action understanding ability by 1%-2%. Additionally, in MovieGenBench (Polyak et al., 2024), our post-trained model surpasses the original model in GSB score of 2.15 and 6.81 in HunyuanVideo (Kong et al.) and Wanx2.1 (Aliyun, 2025), respectively.

Our contributions can be summarized as follows:

- We propose a novel data annotation pipeline to provide data that can facilitate human action understanding, which 1) accumulates large-scale videos with clear actions from the Internet and 2) generates standardized captions that distinguish individuals and detail their actions and interactions.
- We introduce two datasets: HAICTrain which includes 126K generated-then-verified high-quality video-caption pairs for training; and HAICBench comprising 412 human annotated video-caption pairs and 2,000 QA pairs, designed to evaluate MLLMs' human action understanding comprehensively.
- Experiments demonstrate that training with HAICTrain significantly improves human action understanding in benchmarks including MVBench (Li et al., 2024c), PerceptionTest (Patraucean et al., 2024), ActivityNet-QA (Yu et al., 2019) and HAICBench. Furthermore, HAICTrain substantially improves text-to-video generation on MovieGenBench.

2 Related Work

2.1 Video Caption Datasets

Most existing video captioning datasets prioritized general video understanding (Xu et al., 2016; Zhou et al., 2018; Wang et al., 2019; Miech et al., 2019;

Category	Example
Interaction	*What gesture does the middle-aged woman make while talking to the other man?* (A) Gestures with both hands clasped in front of her (B) Claps her hands (C) Waves her hands in the air (D) Points at the desk
Action Details	*What does the man in the black hat do with his right hand as he starts skateboarding?* (A) He waves it in greeting (B) He points down the slope (C) He places it on his hip (D) He keeps it in his pocket
Action Sequence	*What does the man in the gray cap do immediately after gripping the barbell?* (A) He looks at the camera (B) He adjusts his hat (C) He bends down and lowers the barbell (D) He walks towards his front left and rubs his hands together
Count	*How many times does the man clap his hands?* (A) Three (B) Four times (C) Once (D) Twice
Attribute	*What color is the cropped jacket worn by the young female character in the video?* (A) Pink (B) White (C) Blue (D) Black

Table 1: Task examples of HAICBench, showcasing a comprehensive human action understanding across spatial (action details), temporal (sequence, count), and multi-human interaction (interaction and attribute) aspects.

Bain et al., 2021; Zellers et al., 2021; Lei et al., 2021; Xue et al., 2022; Chen et al., 2024a,d; Wang et al., 2024d, 2025b; Xiong et al., 2024; Wang et al., 2024b). These datasets included only a subset of human videos, and the action captions tend to be coarse. Some datasets specifically focused on human-centric captioning, including ActivityNet-Captions (Krishna et al., 2017), Ego4D (Grauman et al., 2022), Kinetic-GEB (Wang et al., 2022), ActionHub (Zhou et al., 2024) and OpenHumanVid (Li et al., 2024b).

Different from these datasets, our work focuses on finer-grained human actions and interaction understanding.

2.2 Video Understanding Benchmarks

Traditional video understanding benchmarks have primarily honed in on a fixed set of classes (Soomro et al., 2012; Shahroudy et al., 2016; Carreira and Zisserman, 2017; Goyal et al., 2017). Recently, there has been a shift towards open-set video understanding, including video captioning and question answering (Xu et al., 2016; Wang et al., 2019; Xiao et al., 2021; Li et al., 2024c). While benchmarks like NeXT-QA (Xiao et al., 2021), MVBench (Li et al., 2024c), and PerceptionTest (Patraucean et al., 2023) include action-related questions, they primarily focus on general action recognition rather than fine-grained details. TemporalBench (Cai et al., 2024) evaluates event-level action understanding using a different protocol that requires distinguishing correct captions from negative ones. Action-specific benchmarks, such as ActivityNet-Captions (Krishna et al., 2017), TGIF-QA (Jang et al., 2019), MoVid-Bench (Chen et al., 2024b), EGOBody (Zhang et al., 2022), and GRAB (Taheri et al., 2020), address temporal sequences or human-object interactions but often overlook human-human interactions.

Our HAICBench provides a more comprehensive evaluation of human action understanding, covering five key aspects detailed in Table 1. In Appendix A, we compare HAICBench with three concurrent works MoVidBench (Chen et al., 2024c), MotionBench (Hong et al., 2025), and FAVORBench (Tu et al., 2025).

2.3 Multi-modal Large Language Models

MLLMs (Liu et al., 2023; Wang et al., 2024c, 2025a) generally process multi-modal input information and generate language output. Kosmos (Huang et al., 2023) introduced an end-to-end framework that integrated visual inputs with LLM from cohesive training. Flamingo (Alayrac et al., 2022) and InstructBLIP (Dai et al., 2023) merged visual and linguistic features through cross-attention and a Q-Former module, respectively. MiniGPT-4 (Zhu et al., 2024) and LLaVA (Liu et al., 2023) simplified the integration by linearly projecting visual features directly into the LLM embedding space.

Recent studies focused on different aspects to enhance the above early attempts in MLLMs. Cosmos-2 (Peng et al., 2024) and NeXT-GPT (Wu et al., 2024) have expanded MLLM applications to broader multi-modal tasks. LLaVA-1.5 (Liu et al., 2024) explored adding high-quality multi-

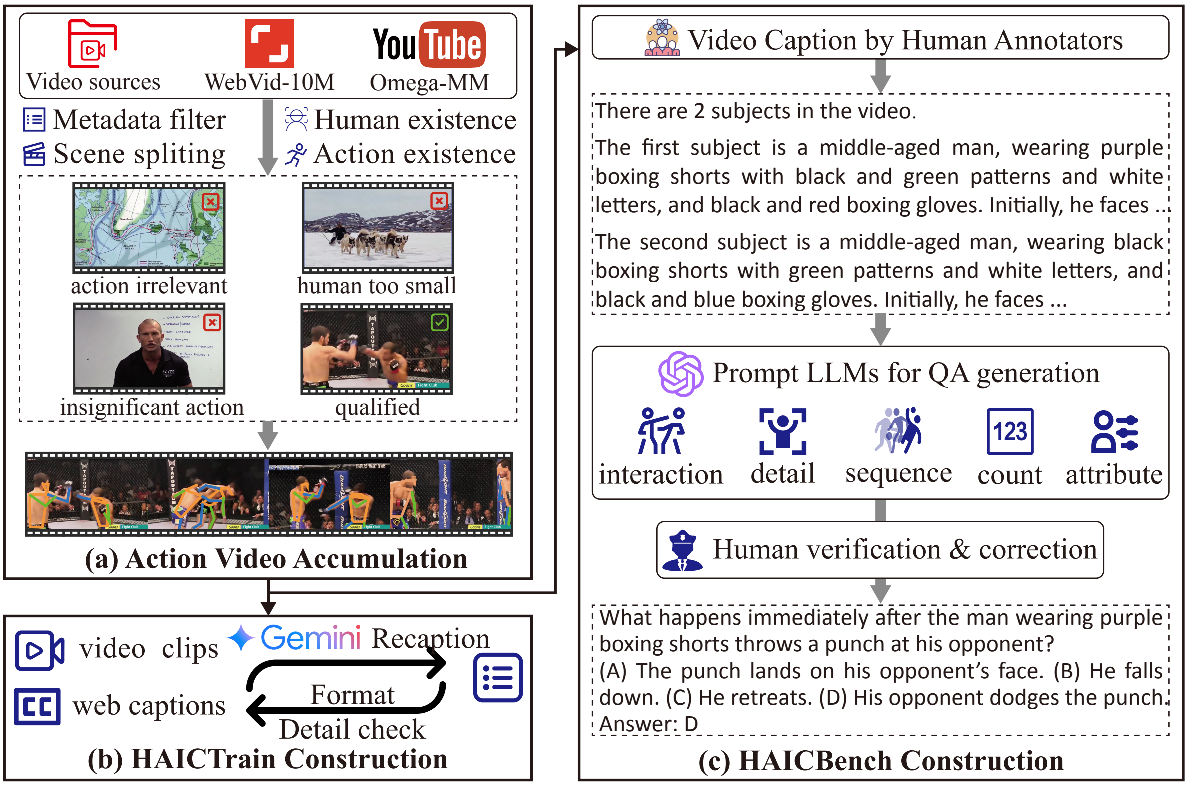


Figure 2: Our data generation pipeline. (a) The video accumulation stage collects videos featuring clear human actions from the Internet. Based on this, (b) HAICTrain is curated through Gemini-1.5-Pro re-captioning, and (c) HAICBench is created by LLM-assisted human annotation.

task training data, and scaling up the resolution and LLM size to boost MLLM performance. LLaVA-OneVision (Li et al., 2024a) explored to unify dynamic image resolution, multi-image, and video into a unified input format.

3 HAIC Data Pipeline

In this section, we detail our data generation pipeline, as illustrated in Figure 2.

3.1 Action Video Accumulation

This pipeline aims to accumulate human videos featuring clear human actions and sufficient details from large-scale videos.

Metadata Filtering. We begin by discarding low-resolution videos and those without verbs in their descriptions using spaCy (Honnibal et al., 2020). The remaining videos are split into short clips with unique scenes using SceneDetect (Castellano, 2024). We keep clips between 5 and 20s long, as actions typically occur within a single scene.

Human Existence Filtering. We uniformly sample 16 frames from each clip and use the RTMPose (Jiang et al., 2023) object detector to identify humans. Only videos where all frames contain 1-5 humans and the total bounding box area covers at least 10% of the frames are retained, ensuring sufficient human detail.

Human Action Filtering. Using RTMPose (Jiang et al., 2023), we detect human bounding boxes and 17 body keypoints at 1 fps. Tracklets are constructed based on the maximal IoU between frames. We then filter out videos with static humans by ensuring the L_1 distance between all adjacent keypoints exceeds 0.085, with keypoint coordinates normalized by the video resolution. This filtering captures clear human actions.

However, we observe that 15% of the filtered videos still contain static humans despite large keypoint L_1 distances. This often results from camera movements or image gallery videos. In these cases, we empirically find keypoints approximately follow an affine transformation (e.g., translation, scaling, rotation, and shear mapping). Based on this insight, we developed a strategy to filter these videos. Formally, let the keypoint vector in frame t be $\mathbf{P}_t \in \mathbb{R}^{3\times17}$, where 17 is the number of keypoints and 3 corresponds to homogeneous coordinates $(\text{height}, \text{width}, 1)$. We assume keypoints in these videos generally adhere to the affine transformation:

$$\begin{cases} \mathbf{P}_{t+1} = \mathbf{T}\mathbf{P}_t, \\ \mathbf{T} = \begin{bmatrix} \mathbf{A} & \mathbf{t} \\ \mathbf{O} & 1 \end{bmatrix}, \end{cases} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{2\times 2}$, $\mathbf{t} \in \mathbb{R}^{2\times 1}$ is the transformation coefficients, and $\mathbf{O} = [0, 0]$. We solve the following least squares problems:

$$\min_{\mathbf{A},\mathbf{t}} \|\mathbf{P}_{t+1} - \mathbf{T}\mathbf{P}_t\|_2, \tag{2}$$

and retain only those samples with a residual value $r > 0.0016$. A larger residual indicates a greater deviation from Equation 1, suggesting that the unwanted videos mentioned above.

Overall, the whole action video accumulation step yields 0.31% to 1.3% of human action videos, depending on the video source.

3.2 HAICTrain

We chose the WebVid-10M dataset (Bain et al., 2021) as the video source for training due to its large scale and high vision quality. Initially, we apply the action video accumulation pipeline in Section 3.1 to collect action videos from the WebVid-10M. This process results in a collection of 126K videos, representing 1.2% of the original dataset. Then, we employ Gemini-1.5-Pro (Team et al., 2024) to generate captions in the standardized format as depicted in Figure 1 referring to the videos and original captions. The specific prompts used for this process are detailed in Appendix B. We then employ an additional judgement to filter out failure cases that do not follow the pre-defined format or display low quality. This judgement ensures the quality of our data in HAICTrain.

3.3 HAICBench

We develop an LLM-assisted human annotation pipeline to create the HAICBench, which evaluates the capabilities of MLLMs in human action understanding. Human annotators first craft video captions following the format in Figure 1. To enhance question diversity, we then adopt LLMs to generate QA pairs based on these captions, which are then verified by annotators.

Video Caption Annotation. To avoid potential overlap with public benchmarks, we choose the newly proposed Omega-multimodal dataset (OMEGA Labs, 2024) as our benchmark video source, which comprises over 30 million 2-minute video clips. We apply our video accumulation stage in Section 3.1 to obtain some video clips, and manually check these clips to obtain 1,140 video clips. Then, human annotators are required to describe the number of human subjects, noting static attributes (gender, age, clothing, accessories) for subject identification, followed by body movements, action sequences, and interactions with others in chronological order.

To ensure the quality of annotated captions, we train all annotators for one week. Besides, each video is first annotated by 1 annotator and then checked and made up missing points by another 3 ones. We do not follow previous works like VATEX (Wang et al., 2019) or MSR-VTT (Xu et al., 2016) to annotate multiple captions for one video, since our caption format is an exhaustive description for human actions in a video.

QA Pair Generation. Based on captions above, we prompt GPT-4o (Hurst et al., 2024) to generate multiple-choice question-answer pairs about human interactions, action details, action sequences, count, and human attributes. The prompts are presented in Appendix D. Each question-answer pair is checked by 2 annotators and will be corrected if there are any mistakes. All options are shuffled to avoid potential bias.

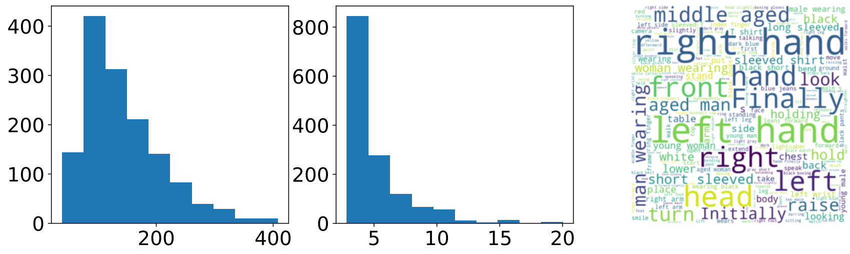
We get a total of 1,140 video-caption pairs and 7,548 QA pairs in total. Detailed statistics are presented in Figure 3. Although each video clip focuses on a single scene and is relatively short (less than 20 seconds), the captions are highly detailed, often exceeding 100 words. The word cloud analysis reveals that our captions provide comprehensive descriptions.

For evaluation, we construct the HAICBench-test (refered as HAICBench for short) by selecting 412 videos and 2,000 QA pairs, evenly distributed across all categories. Both HAICBench and the whole QA dataset will be released.

4 Experiments

4.1 Experimental Settings

Baselines. We selected LLaVA-Video-7B (Zhang et al., 2024b) as our baseline due to its minimal inductive bias in model architecture and superior performance. Furthermore, we enhanced its action understanding capability through post-training, resulting in **LLaVA-Video-ActionPro-7B**.



(a) Words in caption (b) Video duration (c) Word Cloud

Figure 3: Statistics of HAICBench. Although videos are relatively short, the video captions are of high details including various action details and sequential actions.

Model	Frames	Avg.	Action Details	Action Sequence	Interaction	Count	Attribute
Human Annotated Captions	-	96.7	100.0	94.3	95.0	95.3	99.0
Gemini-1.5-Pro (Team et al., 2024)	-	41.0	33.7	30.4	36.0	46.0	58.9
GPT-4o (Hurst et al., 2024)	50*	40.0	37.8	30.6	35.8	44.6	51.4
VideoLLaMA2-7B (Cheng et al., 2024)	64	21.0	15.7	18.9	17.3	28.9	24.2
LongVA-7B (Zhang et al., 2024a)	64	19.3	14.7	13.8	11.5	36.0	20.3
InternVL2-8B (Chen et al., 2024e)	64	28.8	**31.0**	19.9	21.5	31.5	40.2
Qwen2VL-7B (Wang et al., 2024a)	64	27.0	18.5	**22.1**	19.9	35.7	38.9
LLaVA-Video-7B (Zhang et al., 2024b)	64	29.2	23.1	15.8	23.7	35.0	48.3
LLaVA-Video-ActionPro-7B	64	**35.2**	27.5	16.9	**30.7**	**44.2**	**56.9**

Table 2: Results on HAICBench in the caption evaluation setting. We adopted MLLMs to generate video captions, based on which we prompted an LLM to answer video questions. *The GPT-4o API supports a maximum of 50 frames per video.

Training dataset. To preserve the general capability of LLaVA-Video against catastrophic forgotten, we randomly selected 200K instruction pairs from its training set LLaVA-Video-178K (Zhang et al., 2024b) for sample rehearsal (Verwimp et al., 2021). Then, this subset is combined with our HAICTrain to form a training set with 326K instruction pairs in total.

Evaluation dataset. To evaluate human action understanding ability, we executed experiments on several action-related benchmarks, including MVBench, ActivityNet-QA, PerceptionTest, and our HAICBench. Note that we focused on questions related to human actions in these benchmarks. For MVBench, we performed a comparison of all sub-tasks whose names contain "Action", resulting in 7 types: Action Antonym, Action Count, Action Sequence, Action Prediction, Action Localization, Fine-grained Action, and Unexpected Action. For PerceptionTest, we also selected all questions whose tags are related to action. More details are presented in Appendix C.

Evaluation metrics. For multiple-choice questions in MVBench, PerceptionTest, and HAICBench, we followed the prompting approach in LLaVA-1.5 (Liu et al., 2024) and used accuracy as the metric. For ActivityNet-QA which features open-ended questions, we followed the evaluation protocol in Video-ChatGPT (Maaz et al., 2024), utilizing GPT-4o-0513 to calculate accuracy. In terms of HAICBench, we implemented two evaluation settings: (1) **Standard Evaluation**, where the video and question are directly input into an MLLM to generate an answer; and (2) **Caption Evaluation**, where the model generates a caption for each video, and an LLM (Gemini-1.5-Pro) answers questions based on this caption, thereby assessing caption ability based on QA accuracy. To avoid the LLM guessing answers when captions do not mention related contents, we asked Gemini-1.5-Pro to refuse to answer those questions (see Appendix E).

Implementation Details. We used LLaVA-Video-7B as initial model and fine-tuned it on the training dataset. All parameters were updated during training. The model was fine-tuned for one epoch with a learning rate of 1e-5 for the LLM and 2e-6 for the vision encoder, with 256 batch size. We sampled 64 frames uniformly in both training and evaluation. During generation, we took the greedy search without randomness. Our experiments utilized 128 NVIDIA A800-80GB GPUs.

Model	Frames	Avg.	Action Details	Action Sequence	Interaction	Count	Attribute
Gemini-1.5-Pro (Team et al., 2024)	-	67.0	71.3	52.1	73.5	56.7	81.4
GPT-4o (Hurst et al., 2024)	50*	61.2	69.9	46.4	61.6	46.3	81.8
VideoLLaMA2-7B (Cheng et al., 2024)	64	51.4	48.5	38.8	56.3	47.0	66.3
LongVA-7B (Zhang et al., 2024a)	64	56.9	55.6	41.8	58.4	52.1	76.4
InternVL2-8B (Chen et al., 2024e)	64	58.8	64.2	43.0	59.7	50.0	76.9
Qwen2VL-7B (Wang et al., 2024a)	64	62.5	**69.4**	44.7	60.4	57.5	80.6
LLaVA-Video-7B (Zhang et al., 2024b)	64	62.0	64.3	43.0	62.5	56.8	83.5
LLaVA-Video-ActionPro-7B	64	**64.2**	65.7	**45.0**	**63.6**	**61.3**	**85.3**

Table 3: Results on HAICBench in the standard evaluation setting. *The GPT-4o API supports a maximum of 50 frames per video.

Base Model	Training Dataset				MVB	PerTest	ANetQA	HAICBench
	LLaVA-Video-178K		HAICTrain	WebVid				
	200K	126K	126K	126K				
LLaVA-Video					59.4	58.1	61.8	62.0
	✓	✓			60.0	58.3	63.8	63.1
	✓		✓		**62.1**	**59.9**	**65.2**	**64.2**
	✓			✓	58.8	55.8	62.1	60.8
LLaVA-OneVision					57.8	47.9	57.1	59.4
	✓	✓			59.3	55.7	61.8	60.7
	✓		✓		**60.6**	**56.9**	**62.7**	**61.3**
	✓			✓	58.8	55.8	60.4	59.8

Table 4: The gain from training with our high-quality human action captions is effective across several benchmarks. MVB, PerTest, and ANetQA denotes MVBench, PerceptionTest, and ActivityNet-QA, respectively.

4.2 Results on HAICBench

Caption Evaluation Setting. We follow the caption evaluation setting of HAICBench, where MLLM baselines generate action-related captions for each video, and then the captions and questions are fed into Gemini-1.5-Pro for answers. The results are presented in Table 2. Our post-trained LLaVA-Video-ActionPro-7B achieves state-of-the-art performance among open-source MLLMs, with a 6% performance gain. These results show that training with HAICTrain can significantly improve the caption quality of human actions. Furthermore, using human annotated captions as input can achieve very high accuracy, which means there is still considerable room for improvement. Note that the accuracies of some classes are lower than 25% because we prompted the model not to answer the question if no clues are found in the caption.

Standard Evaluation Setting. In this setting, videos and questions are directly fed into MLLMs for question answering. The results are presented in Table 3. Our post-trained LLaVA-Video-ActionPro-7B outperforms LLaVA-Video by 2.2%, achieving state-of-the-art performance among open-source MLLMs. On the action-related subtasks, our model surpasses the baseline LLaVA-Video-7B model by 1%-2%. These findings highlight the effectiveness of our HAICTrain dataset for enhancing human action understanding. Additionally, Gemini significantly outperforms GPT-4o, which is the reason why we leverage it to generate the captions of our HAICTrain dataset.

4.3 Effectiveness of Better Captions

In the training set, we combine our HAICTrain with 200K instruction pairs from LLaVA-Video-178K (Zhang et al., 2024b) to prevent catastrophic forgotten. To assess the contribution of 126K video-caption pairs in HAICTrain, we replace them with another 126K samples randomly selected from LLaVA-Video-178K and execute post-training on two different MLLMs. The results are presented in Table 4. With the same amount of data, training with HAICTrain consistently yields better improvements across various action question-answering benchmarks, including MVBench, PerceptionTest, ActivityNet-QA and HAICBench. We also try the same videos in HAICTain with their original captions in WebVid and find using these captions can not improve accuracy. All of these demonstrate that HAICTrain is better for the model to understand human actions.

LLaVA-Video: The video features a family in a brightly lit room, which appears to be a living area with a kitchen in the background. There are three individuals present: a man, a woman, and a young child.

1. The man is standing on the left side of the frame, wearing a blue shirt and beige pants. He is holding a red object that resembles a toy or a tool, possibly a toy hammer, and is interacting with the child.

2. The woman is seated on a couch on the right side of the frame, wearing a floral dress. She is reaching out towards the child, engaging in a playful interaction.

3. The young child, who has blonde hair, is standing on a box in the center of the frame. The child is wearing a white top with a patterned skirt. The child appears to be jumping or playing, and the woman seems to be encouraging this activity.

LLaVA-Video-ActionPro: The main characters are three people.

The first is a man, seemingly in his 30s or 40s. He has short reddish-brown hair and a beard. He is wearing a light blue long-sleeved shirt and khaki pants. He is standing behind a cardboard box, holding a red tape dispenser. Throughout the video, he is seen smiling and looking at the woman and the child.

The second is a young girl, seemingly aged between 3-6. She has blonde hair and is wearing a white long-sleeved shirt with a striped skirt. She is standing on the couch, facing the woman. Throughout the video, she is seen smiling and laughing while being lifted by the woman.

The third is a woman, seemingly in her 30s or 40s. She has long blonde hair and is wearing a floral dress. She is sitting on a couch, facing the girl. Throughout the video, she is seen smiling and lifting the girl up and down.

Figure 4: A video caption example in HAICBench.

4.4 Qualitative Analysis

We conduct a qualitative analysis of our post-trained model on HAICBench, as shown in Figure 4. The video features two human boxers fighting on the stage. As outlined in the figure, before post-training, the model can only provide a coarse description of their actions. However, after post-training with HAICTrain, our model delivers detailed descriptions of the fighters, including their hair color, clothing, and positions in the cage. Besides, it presents a clear sequence of actions, describing how the attack and defense progress over time. This also demonstrates the effectiveness of our caption format. By distinguishing the fighters based on detailed attributes, the caption facilitates better individual recognition, enhancing the viewer's ability to follow the action.

4.5 Effectiveness for Text-video Generation

To evaluate our method's effectiveness in text-to-video generation, we use the MovieGenBench dataset (Polyak et al., 2024), consisting of 1,003 videos. We first take LLaVA-Video-7B and LLaVA-Video-ActionPro-7B to generate captions, which are then fed to HunyuanVideo (Kong et al.) and Wanx2.1 (Aliyun, 2025) to generate videos. Then, five human annotators assess the semantic relevance between videos generated by two captions, classifying them as "Good", "Same", or "Bad". The annotators are asked to make decisions upon the original videos in MovieGen. We measure the GSB score, i.e., the percentage of "Good" and "Same" relative to that of "Bad" and "Same". Results show that LLaVA-Video-ActionPro-7B outperforms LLaVA-Video-7B in generating captions that lead to more semantically accurate videos. On HunyuanVideo, LLaVA-Video-ActionPro-7B achieves a GSB score of 2.15, and on Wanx2.1, 6.81. LLaVA-Video-7B often produced verbose but less accurate captions, failing to capture the original video's core actions. These findings highlight our model's superior ability to comprehend the actions in videos and relay captions that enable high-quality, semantically faithful text-to-video generation. More cases are presented in Appendix F.

5 Conclusion

This study addresses the challenge of human action understanding by introducing a novel two-stage data annotation pipeline, combining data accumulation, human-machine annotation, and human verification. This process produces two datasets: HAICTrain, which significantly enhances human action understanding and generation, and HAICBench, a comprehensive benchmark. Both datasets will be made open-source to advance research and applications in this field, promoting broader impact across human behavior understanding and generation.

6 Limitations

Although our work makes notable progress in human action understanding, it still has several limitations. Despite their comprehensive nature, our HAIC datasets may not cover the entire spectrum of human actions, particularly those involving complex interactions or cultural nuances. Furthermore, our work primarily focuses on visual and textual data, lacking integration with audio data, which could provide additional context for understanding. Future work should aim to incorporate audio and further refine the annotation process to address these limitations.

7 Acknowledgment

This work was supported in part by the National Natural Science Foundation of China(Grant Nos. 62376069 and 62236003), in part by the Young Elite Scientists Sponsorship Program by CAST (Grant No. 2023QNRC001), in part by Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515012027), in part by Jiangsu Science and Technology Major Program (Grant No. BG2024041), and in part by the Shenzhen Science and Technology Program (Grant Nos. KQTD20240729102207002 and ZDSYS20230626091203008).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.

Aliyun. 2025. Tongyi wanxiang.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *International Conference on Computer Vision*, pages 1708–1718. IEEE.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. TemporalBench: Benchmarking Fine-grained Temporal Understanding for Multimodal Video Models. ArXiv:2410.10818.

João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition*, pages 4724–4733. IEEE Computer Society.

Brandon Castellano. 2024. PySceneDetect.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. 2024a. Sharegpt4video: Improving Video Understanding and Generation with Better Captions. ArXiv:2406.04325.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. 2024b. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. ArXiv:2405.20340.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. 2024c. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint ArXiv:2405.20340*.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, and others. 2024d. Panda-70m: Captioning 70m Videos with Multiple Cross-Modality Teachers. In *Conference on Computer Vision and Pattern Recognition*, pages 13320–13331. IEEE.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi

Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024e. How Far are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. ArXiv:2404.16821.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. ArXiv:2406.07476.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *International Conference on Computer Vision*, pages 5843–5851. IEEE Computer Society.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4D: Around the World in 3, 000 Hours of Egocentric Video. In *Conference on Computer Vision and Pattern Recognition*, pages 18973–18990. IEEE.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. 2021. Embodied Intelligence via Learning and Evolution. *Nature communications*, 12:5721.

Gillian R Hayes. 2011. The Relationship of Action Research to Human-Computer Interaction. *ACM Transactions on Computer-Human Interaction*, 18.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. *arXiv preprint ArXiv:2501.02955*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint ArXiv:2410.21276*.

Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2019. Video Question Answering with Spatio-Temporal Reasoning. In *Int. J. Comput. Vis.*, volume 127, pages 1385–1412.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose. ArXiv:2303.07399.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *International Conference on Computer Vision*, pages 706–715. IEEE Computer Society.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting Moments and Highlights in Videos via Natural Language Queries. In *Advances in Neural Information Processing Systems*, volume 34. Neural Information Processing Systems Foundation.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. LLaVA-OneVision: Easy Visual Task Transfer. ArXiv:2408.03326.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. 2024b. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. *arXiv preprint ArXiv:2412.00115*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Conference on Computer Vision and Pattern Recognition*, pages 22195–22206. IEEE.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. In *Conference on Computer Vision and Pattern Recognition*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Annual Meeting of the Association for Computational Linguistics*, pages 12585–12602. Association for Computational Linguistics.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *International Conference on Computer Vision*, pages 2630–2640. IEEE.

Inc. OMEGA Labs. 2024. Omega labs bittensor subnet: Multimodal dataset for agi research. https://huggingface.co/datasets/omegalabsinc/omega-multimodal.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. 2024. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding Multimodal Large Language Models to the World. In *International Conference on Learning Representations*. OpenReview.net.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. 2024. Movie gen: A cast of media foundation models. *Preprint*, arXiv:2410.13720.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1010–1019. IEEE Computer Society.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402.

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. 2020. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint ArXiv:2403.05530*.

Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. 2025. Favor-bench: A comprehensive benchmark for fine-grained video motion understanding. *arXiv preprint ArXiv:2503.14935*.

Eli Verwimp, Matthias De Lange, and Tinne Tuytelaars. 2021. Rehearsal revealed: The limits and merits of revisiting samples in continual learning. In *International Conference on Computer Vision*, pages 9385–9394.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. ArXiv:2409.12191.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. 2024b. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *arXiv preprint ArXiv:2410.08260*.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. 2024c. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint ArXiv:2412.20504*.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. 2025a. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. *arXiv preprint ArXiv:2503.12559*.

Xiao Wang, Jianlong Wu, Zijia Lin, Fuzheng Zhang, Di Zhang, and Liqiang Nie. 2025b. Video dataflywheel: Resolving the impossible data trinity in video-language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *International Conference on Computer Vision*, pages 4580–4590. IEEE.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2024d. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*.

Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. 2022. GEB+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference of Computer Vision*, Lecture Notes in Computer Science, pages 709–725. Springer.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. NExT-GPT: Any-to-Any Multimodal LLM. In *International Conference on Machine Learning*. OpenReview.net.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *Conference on Computer Vision and Pattern Recognition*, pages 9777–9786. Computer Vision Foundation / IEEE.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. 2024. Lvd-2m: A long-take video dataset with temporally dense captions. *arXiv preprint ArXiv:2410.10816*.

Feiyi Xu, Feng Xu, Jiucheng Xie, Chi-Man Pun, Huimin Lu, and Hao Gao. 2021. Action recognition framework in traffic scene for autonomous driving system. *IEEE Transactions on Intelligent Transportation Systems*, 23.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Conference on Computer Vision and Pattern Recognition*, pages 5288–5296. IEEE Computer Society.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Conference on Computer Vision and Pattern Recognition*, pages 5026–5035. IEEE.

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

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. MERLOT: Multimodal Neural Script Knowledge Models. In *Advances in Neural Information Processing Systems*, pages 23634–23651. Neural Information Processing Systems Foundation.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long Context Transfer from Language to Vision. ArXiv:2406.16852.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. 2022. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video Instruction Tuning With Synthetic Data. ArXiv:2410.02713.

Jiaming Zhou, Junwei Liang, Kun-Yu Lin, Jinrui Yang, and Wei-Shi Zheng. 2024. Actionhub: A large-scale action video description dataset for zero-shot action recognition. ArXiv:2401.11654.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598. AAAI Press.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *International Conference on Learning*. OpenReview.net.

A Comparison with concurrent benchmarks

We compared our benchmark with three concurrent works on human action understanding—MoVidBench (Chen et al., 2024c), Motion-Bench (Hong et al., 2025), and FAVORBench (Tu et al., 2025)—across three key aspects of human actions:

- Action details: Fine-grained motion annotations (e.g., body movements, intensity, direction).
- Action transitions: Details on the changes of two actions.
- Multi-human interactions: Actions involving multiple humans.

Our benchmark is the only one covering all three aspects, as illustrated in Table 5 below.

Benchmark	Action Details	Action Transitions	Multi-Human Interactions
MoVidBench	✓	✓	None
MotionBench	×	✓	Few
FAVORBench	✓	✓	Few
HAICBench	✓	✓	Rich

Table 5: Comparison of recent human action understanding benchmarks on three aspects of human action

B Automatic Annotation

Prompt for Gemini-1.5-Pro Recpationing

Analyze the input video and describe the video according to the following details:
1. Describe the number of main characters who appear in the video. Please note that people in the background should not be included.
2. For each main character, first describe their static attributes such as gender, age, facial features, clothing, positions, etc. And then, detail all their actions, including both facial expressions and body movements, in chronological order.
Describe each main character in one paragraph. Please ignore the audio of the input video and do not mention \"audio\" in your answer.
Please note that do not mention anything which does not appear in the video footage and do not guess anything which is not clearly visable in the video footage.
Note that you are describing a video, so you should not mention the frames via \"first frame\" or \"first figure\".

Here are some examples.

Example #1:
There is 1 subject in the video.
This person is a little girl, seemingly aged between 3-6. She has brown hair styled into two twisted braids and doesn\'t have bangs. She has small gold hoop earrings and is dressed in a burgundy V-neck blouse and a blue skirt. Initially, she sits on the ground looking forward with a fearful expression. Subsequently, she quickly stands up and screams loudly, then quickly runs away.

Example #2:
There are 3 subjects in the video.
The first is a young woman positioned on the left side of the screen, holding a baby. She has wavy blonde hair, blue eyes, and fair skin. She is dressed in a blue dress and red high heels. Throughout the video, she is seen smiling and gazing at the baby in her arms.
The second is a baby who is being held by the woman. The baby is bald, undressed, with a very light complexion, and appears to be between 1-3 years old. At first, the baby looks up at the woman with a smile but then furrows its brow and eventually breaks into tears.
The third is an elderly man situated on the right side of the screen. He has white hair and is dressed in a white shirt and black pants. Initially, he is seen smiling at the baby, but his expression quickly changes to one of concern, and he extends his right hand to comfort the baby.

Figure 5: Prompt for Gemini-Pro Re-captioning.

In Section 3.2, we utilize Gemini-1.5-Pro to generate captions for HAICTrain in our defined standardized format. The prompt is detailed in Figure 5.

C Evaluation datasets

In Section 4.1, we use action subsets from MVBench and PerceptionTest as our benchmarks. Here, we explain our subset selection process in detail.

For MVBench, we choose seven categories whose names include "action": `Action Antonym`, `Action Count`, `Action Sequence`, `Action Prediction`, `Action Localization`, `Fine-grained Action`, and `Unexpected Action`.

For PerceptionTest, we select questions tagged with any of the five labels related to human actions: `Action counting`, `Action recognition`, `Adversarial action`, `Distractor action`, and `Occlusion (Occluded interactions)`.

D QA Pair Generation

As briefly discussed in Section 3.3, we utilized a large language model (LLM) to assist in generating QA pairs for HAICBench. The prompts used are as follows: for action interaction QA, see Figure 6; for action detail QA, see Figure 7; for action sequence QA, see Figure 8; for action count QA, see Figure 9; and for human attribute QA, see Figure 10.

Prompts for Interaction QA Generation

You are now a data augmentation assistant. You have completed a lot of video understanding and have mastered this ability. I will give you a video `caption' which describes the number of human, their visual attributes, and the actions they are performing.

You will help me to complete the understanding of this paragraph and generate 1 English QA pair from it. The QA pairs should focus on the interaction of human subjects in the video. If the video has only 1 subject, return `No subject interaction.'.

The form is multiple choice. Each question has 4 choices, one of which is the correct answer and the other four are interference items. The candidates should be of similar length. You need to mark the correct answer. The returned question should be of the JSON format ONLY, use the distinctive attributes to describe subjects instead of `Subject x', for example:

```json
{
    'question': 'Who does the middle-aged man wearing a light blue shirt speak to first in the video?',
    'candidates': ["A young woman wearing a black chef uniform.", "A middle-aged woman wearing a red dress.", "Both A young woman wearing a black chef uniform and a middle-aged woman wearing a red dress", "No one"],
    'answer': 'B'
}
```

Caption: {caption}

Figure 6: Prompt for action interaction QA generation.

Prompts for Detail QA Generation

video caption starts:

{caption}

video caption ends.

You are now a data augmentation assistant. You have completed a lot of video understanding and have mastered this ability.

I have given you a video caption above. All the subjects in this caption are human beings. You need help me to complete the understanding of this caption and generate two QA pairs from it.

The form is multiple choice. Each question has 4 options, one of which is the correct answer and the other three are interference items.

You need to mark the correct answer. The option is a word, phrase, or sentence of 15 words or less.

The question should be about details of expressions or body movements, including but not limited to things such as: facial reactions (smiling, frowning, surprise, etc.), gestures (raising a hand, pointing, etc.), postures (standing, sitting, leaning, etc.), specific body pars (left hand, right hand, left leg, etc.) and direction of movement (up, down, left, right, forward, backward, diagonally).

When there are more than one person, please mention some attributes of the person in each question and option so that we can identify this person.

Please note that do not ask questions which include "before" or "after".

Please note that the order of the options should be random. In creating the questions, ensure that the correct answers are distributed evenly among options 1 to 4, so that each option has an equal probability of being the correct answer.

Please note that do not mention anything like "Subject 1" in any questions or options. The result should be in JSON format. Here are some examples:

Example #1:

video catpion starts:

There are two main subjects in the video.

The first subject is a young woman, wearing a black mini dress with lace on the sleeves and edges, a gemstone ring on her right hand, and a consistently happy expression. She walks forward from the doorway towards Subject 2. Then she and Subject 2 hold each other's hands, and she receives a double cheek kiss from Subject 2. Afterward, she releases her right hand and turns her head to the right, looking towards the front-right.

The second subject is a middle-aged man, wearing a black suit, a black tie, and a white dress shirt. Initially, he holds his hands out, palms up, in a welcoming gesture. Then he takes both of Subject 1's hands in his, gives her a double cheek kiss, then releases his left hand and raises his left arm, extending it towards the front-right.

video catpion ends.

answer: [{{"question": "What gesture does the middle-aged man in the thin black suit and tie make at the start?", "opt1": "Gives a double cheek kiss", "opt2": "Holds out his hands, palms up", "opt3": "Waves his hand", "opt4": "Crosses his arms", "answer": 2}}, {{"question": "After receiving a double cheek kiss from the man, where does the woman in the black mini dress with lace look?", "opt1": "Towards the left", "opt2": "Towards the ceiling", "opt3": "Towards the floor", "opt4": "Towards the front-right", "answer": 4}}]

Figure 7: Prompt for action details QA generation.

Prompts for Action Sequence QA Generation

video caption starts:

{caption}

video caption ends.

You are now a data augmentation assistant. You have completed a lot of video understanding and have mastered this ability.

I have given you a video caption above. All the subjects in this caption are human beings. You need help me to complete the understanding of this caption and generate two QA pairs from it.

The form is multiple choice. Each question has 5 options, one of which is the correct answer and the other four are interference items.

You need to mark the correct answer. The option is a word, phrase, or sentence of 15 words or less.

The question should be about sequencing and start with "What does...do immediately after (or just before)...". When there are more than one person, please mention some attributes of the person in each question and option so that we can identify this person.

For example, you can ask "What does the woman in red shirt do immediately after smiling in the video?"

Please note that the order of the options should be random. In creating the questions, ensure that the correct answers are distributed evenly among options 1 to 5, so that each option has an equal probability of being the correct answer.

Please note that do not mention anything like "Subject 1" in any questions or options. The result should be in JSON format. Here are some examples:

Example #1:

video catpion starts:

There are two main subjects in the video.

The first subject is a young woman, wearing a black mini dress with lace on the sleeves and edges, a gemstone ring on her right hand, and a consistently happy expression. She walks forward from the doorway towards Subject 2. Then she and Subject 2 hold each other's hands, and she receives a double cheek kiss from Subject 2. Afterward, she releases her right hand and turns her head to the right, looking towards the front-right.

The second subject is s middle-aged man, wearing a black suit, a black tie, and a white dress shirt. Initially, he holds his hands out, palms up, in a welcoming gesture. Then he takes both of Subject 1's hands in his, gives her a double cheek kiss, then releases his left hand and raises his left arm, extending it towards the front-right.

video catpion ends.

answer: [{{"question": "What does the young woman do immediately after walking towards the middle-aged man?", "opt1": "She turns her head to the right", "opt2": "She holds each other's hands with the man", "opt3": "She smiles at the man", "opt4": "She receives a double cheek kiss from the man", "opt5": "She walks away from the man", "answer": 2}}, {{"question": "What does the middle-aged man do just before he gives the young woman a double cheek kiss?", "opt1": "He smiles at the woman", "opt2": "He releases his left hand", "opt3": "He raises his left arm", "opt4": "He holds out his hands, palms up, in a welcoming gesture", "opt5": "He walks towards the woman", "answer": 4}}]

Figure 8: Prompt for action sequence QA generation.

Prompts for Count QA Generation

video caption starts:

{caption}

video caption ends.

You are now a data augmentation assistant. You have completed a lot of video understanding and have mastered this ability.

I have given you a video caption above. You need help me to complete the understanding of this caption and generate at most three QA pairs from it.

The form is multiple choice. Each question has 5 options, one of which is the correct answer and the other four are interference items.

The order of the options should be random, and each option should have an equal probability of being the answer.

You need to mark the correct answer. The option is a word, phrase, or sentence of 15 words or less.

The question should be about counting and start with "How many". Subjects mentioned in the given caption are all human beings. Therefore, please use "How many people are in the video?" to ask for counting main subjects.

The result should be in JSON format. Here are some examples:

Example #1:

video catpion starts:

There is 1 subject in the video

The first subject is a young woman, wearing a green and white checkered shirt with dark green lines and a black inner layer on her upper body, and black shorts on her lower body. She wears two bracelets, one black and one white, on her right wrist, and a silver ring on her right middle finger. She bends over and stands in the water, using both hands to roll up her left pant leg while taking a step forward. Then she uses her right hand to roll up her right pant leg while also taking a step forward. She then takes two more steps forward, raises her right hand to support herself against the moss above and to her right, and stands upright with her feet together. Finally, she opens both hands, palms up, to catch the water dripping from the moss.

video catpion ends.

answer: [{{"question": "How many people are in the video?", "opt1": "Two", "opt2": "One", "opt3": "Three", "opt4": "Five", "opt5": "Seven", "answer": 2}}, {{"question": "How many steps does the woman take in the video?", "opt1": "One", "opt2": "Five", "opt3": "Two", "opt4": "Three", "opt5": "Four", "answer": 5}}]

Figure 9: Prompt for action count QA generation.

Prompts for Attribute QA Generation

You are now a data augmentation assistant. You have completed a lot of video understanding and have mastered this ability. I will give you a video 'caption' which describes the number of human, their visual attributes, and the action they are performing.

You will help me to complete the understanding of this paragraph and generate 1 English QA pair from it. The QA pairs should focus on the attributes of human subjects in the video. The form is multiple choice. Each question has 4 choices, one of which is the correct answer and the other four are interference items.

The candidates should be of similar length. You need to mark the correct answer with ONE letter.

The returned question should be of the JSON format ONLY, use distinctive attributes to describe subjects instead of `Subject x`, for example:

```json
{
    'question': 'Who does the middle-aged man wearing a light blue shirt speak to first in the video?',
    'candidates': ["A young woman wearing a black chef uniform.", "A middle-aged woman wearing a red dress.", "Both A young woman wearing a black chef uniform and a middle-aged woman wearing a red dress", "No one"],
    'answer': 'B'
}
```

Caption: {caption}

Figure 10: Prompt for human attribute QA generation.

Prompt for MLLM Caption

Please provide a description of people in the video, your description should describe their attributes and detailed actions in chronological order. Do not mention anything unrelated to people.

Prompt for Caption-based QA

{caption}
According to the video caption above, answer the following question. If the given caption doesn't provide sufficient information to answer the question, just reply, "The caption does not mention it." Question:
What color is the hat worn by the female child in the video?
Options:
A. Red
B. Pink
C. Yellow
D. Blue

Figure 11: Prompt caption evaluation setting.

E Caption Evaluation Setting

In Section 4.1, we outline our caption evaluation setup, as illustrated in Figure 11. Initially, we prompt MLLMs to generate a caption for each video using a specific prompt. We then combine the generated caption with a question to enable Gemini-1.5-Pro to produce answers.

F Case Study on Text-to-Video Generation

We used LLaVA-Video and LLaVA-Video-ActionPro to caption the reference videos (generated by MovieGen) in MovieGenBench. Then, we used these captions to generate videos, using the open source HunyuanVideo [*] and the closed source Wanx2.1 [†] respectively. In the manual evaluation, we concluded that the captions generated by LLaVA-Video-ActionPro can give a video that is more consistent and reasonable with the original video. Figure 12 is an example of a single human subject (the first sample in the dataset). It can be seen that in the caption, LLaVA-Video lost the description of details, resulting in inconsistent generated content. Figure 13 is an example of a single animated character subject (the 17th sample in the dataset). It can be seen that in the caption, LLaVA-Video incorrectly identified the number of subjects. Figure 14 is an example of multiple subjects (the 129th sample in the dataset). It can be seen that in the caption, LLaVA-Video incorrectly identified the number of subjects and missed the positional relationship. Overall, the introduction of refined and formatted action description data has greatly improved the model's ability to understand and retell movements. This is of great significance for current mainstream action understanding, anomaly recognition, motion generation, and text-to-video generation.

G Potential Risks

The source videos of our HAICBench are from YouTube. YouTube inherently contains social biases, and viewing its videos as representative of 'the world' can perpetuate hegemonic perspectives. The majority of popular YouTubers are men, and the platform's video practices often reflect gender biases. YouTube also faces issues with hate content, including radical alt-right and 'alt-lite' material. These issues are exacerbated by the platform's recommendation algorithm. Even though we downloaded videos independently, filtering them by view count still subjects us to algorithmic influence. The popularity and monetization dynamics on YouTube shape and are shaped by broader cultural trends, affecting the style and content of uploaded videos.

[*] The "infer-steps" parameter is set to 50.

[†] https://tongyi.aliyun.com/wanxiang/videoCreation

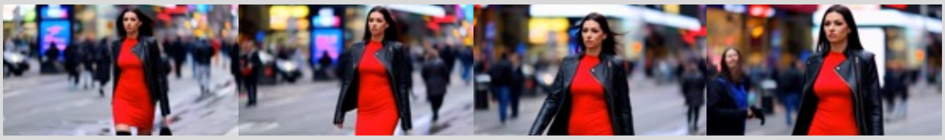
Raw video by MovieGen

LLaVA-Video: There is one primary character in the video, a woman who is walking down a busy urban street. She is wearing a bright red dress and a black leather jacket, paired with black knee-high boots. She has long dark hair and is carrying a small black handbag. The background is filled with various people walking in different directions, some of whom are blurred due to motion. The street is lined with brightly lit signs and advertisements, indicating a commercial area. The ground appears wet, suggesting recent rain. Throughout the video, the woman maintains a steady pace and a confident demeanor as she walks down the street. Her facial expression remains neutral, and she does not interact with anyone or anything around her.

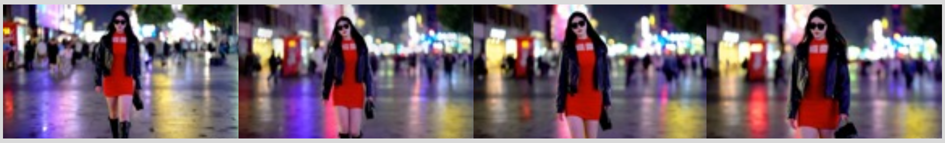
LLaVA-Video-ActionPro: The main character is one person. This person is a young woman with long, dark hair and dark sunglasses. She is wearing a red dress, a black leather jacket, and black knee-high boots. She carries a small black purse in her left hand. She walks towards the camera down a busy street at night. The street is wet, reflecting the lights from the surrounding buildings and signs. As she walks, she maintains eye contact with the camera and her expression remains neutral. Her hair moves slightly in the breeze.

Videos generated by HunyuanVideo

LLaVA-Video

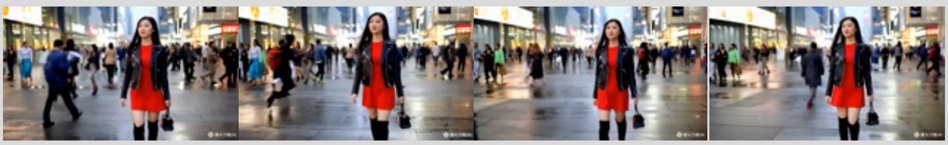


LLaVA-Video-ActionPro



Videos generated by Wanx2.1

LLaVA-Video



LLaVA-Video-ActionPro

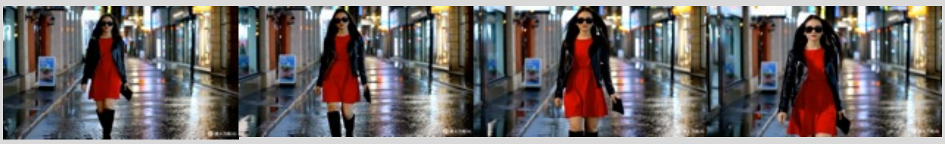


Figure 12: Videos generated by captions from LLaVA-Video and LLaVA-Video-ActionPro of the first sample in MovieGenBench. The main subject in this case is one woman walking along the street. LLaVA-Video-ActionPro provides a more detailed appearance of the woman than LLaVA-Video.

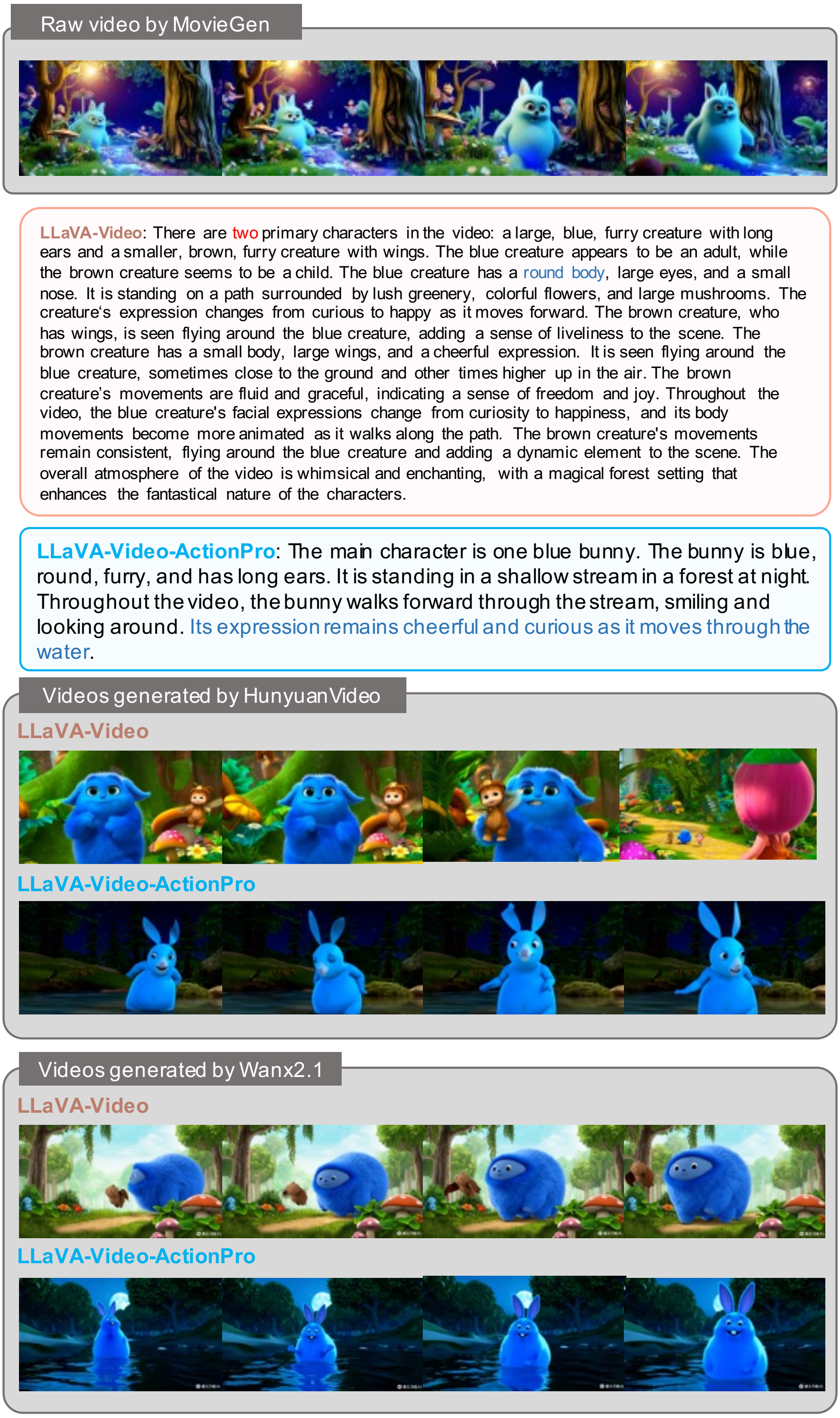
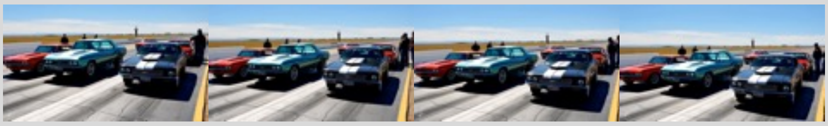


Figure 13: Videos generated by captions from LLaVA-Video and LLaVA-Video-ActionPro of the 17th sample in MovieGenBench. The main subject in this case is one blue animated character. LLaVA-Video incorrectly identifies the main subject.

Raw video by MovieGen

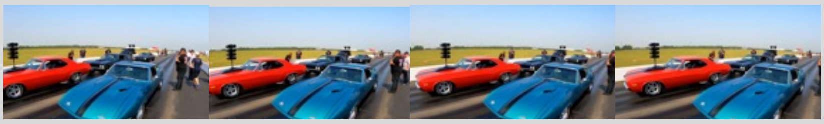


LLaVA-Video: There are no primary characters in the video; the focus is on the cars and the people around them. The main subjects are the four classic muscle cars lined up at the starting line of a drag race track, with two red cars on the left side and two cars on the right side, one black and one teal. There are also several people standing around the cars, some wearing casual clothing and others in racing attire. The cars are stationary, and there is no significant movement or action taking place. The people around the cars appear to be either preparing for the race or observing the setup. The background shows a clear sky and an open field, indicating that the event is taking place outdoors during the day.

LLaVA-Video-ActionPro: The main characters are five cars and several people. The first car is a red muscle car with a hood scoop, positioned on the left side of the frame. It remains stationary throughout the video. The second car is a teal muscle car with a hood scoop, positioned in the center of the frame. It remains stationary throughout the video. The third car is a black muscle car with white racing stripes, positioned on the right side of the frame. It remains stationary throughout the video. The fourth car is a red muscle car with a hood scoop, positioned behind the black car. It remains stationary throughout the video. The fifth car is a red muscle car with a hood scoop, positioned behind the teal car. It slightly moves forward. Several people are present in the background, some standing near the cars and others walking around. They appear to be spectators or participants in the race. They remain mostly stationary throughout the video, observing the cars.

Videos generated by HunyuanVideo

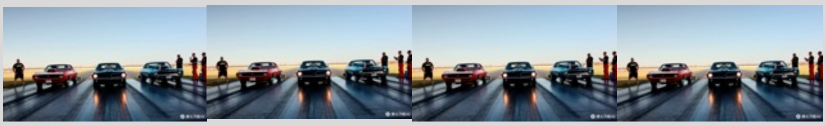
LLaVA-Video



LLaVA-Video-ActionPro

Videos generated by Wanx2.1

LLaVA-Video



LLaVA-Video-ActionPro

Figure 14: Videos generated by captions from LLaVA-Video and LLaVA-Video-ActionPro of the 129th sample in MovieGenBench. The main subjects in this case are five cars lined in two rows. LLaVA-Video incorrectly identifies the number of the main subjects.