Research Article

Tarsier2: Advancing Large Vision-Language Models from Detailed Video Descriptions to Comprehensive Video Understanding

Liping Yuan[1], Jiawei Wang[1], Haomiao Sun[1], Yuchen Zhang[1], Yuan Lin[1]

1. ByteDance Research

We introduce Tarsier2, a state-of-the-art large vision-language model (LVLM) designed for generating detailed and accurate video descriptions, while also exhibiting superior general video understanding capabilities. Tarsier2 achieves significant advancements through three key upgrades: (1) Scaling pre-training data from 11M to 40M video-text pairs, enriching both volume and diversity; (2) Performing fine-grained temporal alignment during supervised fine-tuning; (3) Using model-based sampling to automatically construct preference data and applying DPO training for optimization. Extensive experiments show that Tarsier2-7B consistently outperforms leading proprietary models, including GPT-4o and Gemini 1.5 Pro, in detailed video description tasks. On the DREAM-1K benchmark, Tarsier2-7B improves F1 by 2.8% over GPT-4o and 5.8% over Gemini-1.5-Pro. In human side-by-side evaluations, Tarsier2-7B shows a +8.6% performance advantage over GPT-4o and +24.9% over Gemini-1.5-Pro. Tarsier2-7B also sets new state-of-the-art results across 15 public benchmarks, spanning tasks such as video question-answering, video grounding, hallucination test, and embodied question-answering, demonstrating its versatility as a robust generalist vision-language model.

Corresponding author: Yuan Lin, linyuan.0@bytedance.com

1. Introduction

With the rapid advancements in large vision-language models (LVLM)[1][2][3][4][5][6], significant progress has also been made in video understanding. Leading proprietary models, such as GPT-4o[7] and Gemini-1.5-Pro[8], have achieved state-of-the-art (SOTA) performance across a variety of video understanding tasks. Additionally, several open-source models[3][9][10][11][12][13][10] also demonstrate strong performance on several video understanding benchmarks[14][15][16][17][18], although they still lag behind proprietary models,

 doi.org/10.32388/X26ILU

particularly in complex, open-ended generation tasks. Despite these advancements, current models remain behind human-level video understanding[19][20][21], mainly due to persistent challenges such as accurately perceiving temporal dynamics, spatial-temporal reasoning, and model hallucinations.

In this paper, we introduce Tarsier2, a 7B-parameter LVLM model that can outperform both GPT-4o and Gemini-1.5-Pro in generating detailed video descriptions, a fundamental challenge in video understanding. Beyond video description generation, Tarsier2 also achieves SOTA performance across various video question-answering (VQA) benchmarks at the same model size, surpassing or closely matching the performance of proprietary models on these VQA benchmarks. Figure 1 provides a comprehensive comparison between Tarsier2, GPT-4o and previous SOTA results for open-source LVLMs with the same scale. Figure 2 presents examples illustrating Tarsier2's video understanding capability across different tasks.

 doi.org/10.32388/X26ILU

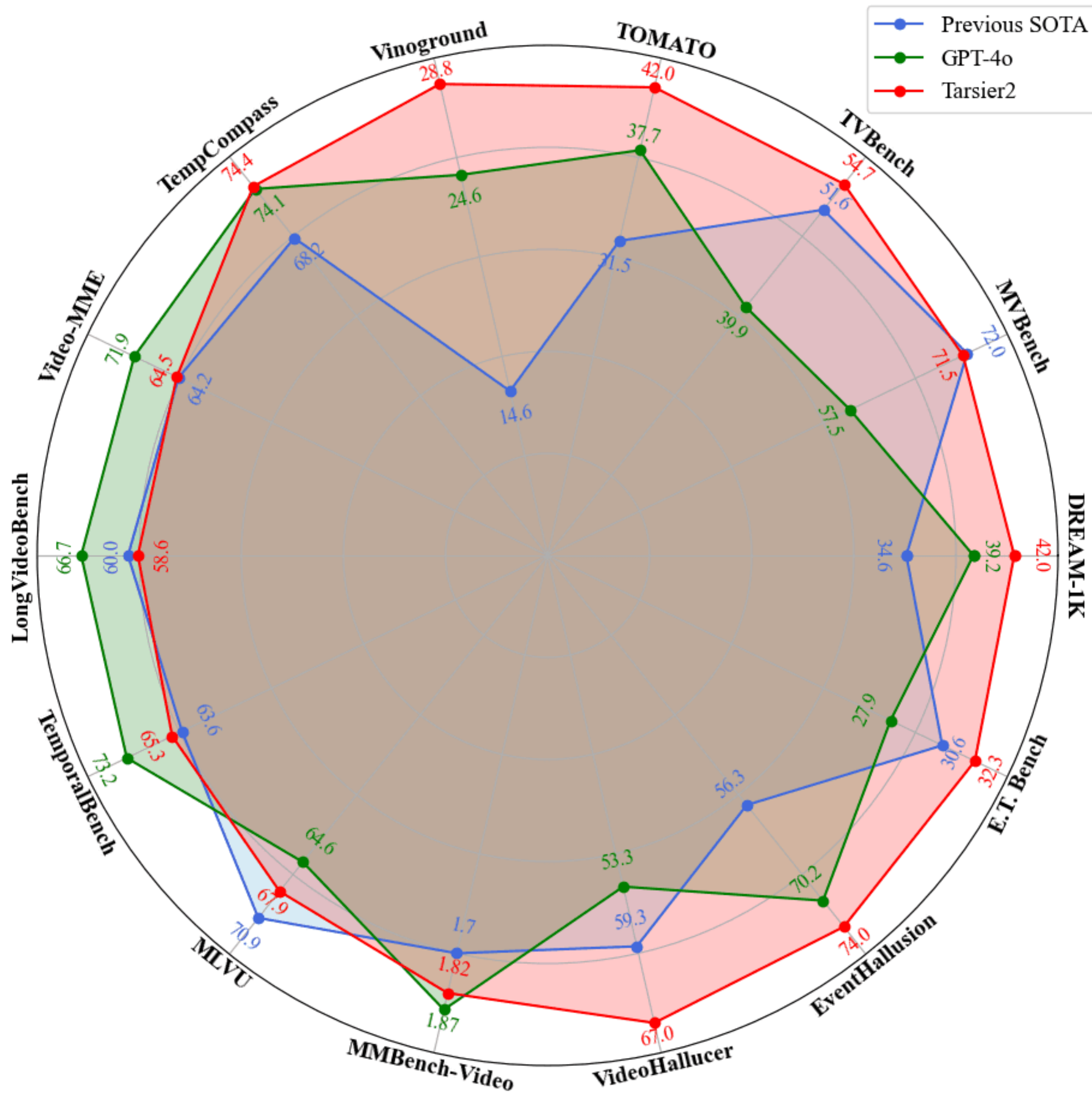


Figure 1. Performance comparison of Tarsier2 with previous SOTA models at 7B-scale and GPT-4o. We report the overall average scores for benchmarks with multiple subtasks/metrics.

Benchmark	Previous SOTA
DREAM-1K[5]	Tarsier-7B[5]
MVBench[15]	InternVL2.5-8B[12]
TVBench[14]	IXC-2.5 7B[22]
TOMATO[23]	Qwen2-VL-7B[6]

Vinoground[24]	LLaVA-OV-7B[13]
TempCompass[25]	Qwen2-VL-7B[6]
Video-MME[26]	NVILA-7B[27]
LongVideoBench[28]	Apollo-7B[29]
TemporalBench[30]	LLaVA-Video-7B[31]
MLVU[18]	InternVL2.5-8B[12]
MMBench-Video[32]	MiniCPM-V-2.6[33]
VideoHallucer[17]	Qwen2-VL-7B[6]
EventHallusion[34]	Tarsier-7B[5]
E.T. Bench[16]	E.T. Chat[16]

Figure 2. Overview of Tarsier2 capabilities. Based on its strong ability for detailed video description, Tarsier2 excels in a variety of video-centric tasks. Click the play buttons to view the videos.

doi.org/10.32388/X26ILU

Tarsier2 employs a simple model architecture consisting of a vision encoder, a vision adaptor, and a large language model (LLM). We meticulously design a three-stage training procedure: pre-training, supervised fine-tuning (SFT), and reinforcement learning (RL). In comparison with Tarsier[5], Tarsier2 features several key improvements that significantly enhance its performance:

- We scale up the pre-training dataset from 11 million to 40 million video-text pairs, addressing the challenge posed by the scarcity of high-quality video-text data. To achieve this, we implement meticulous filtering and sourcing. Specifically, we collect 11 million commentary videos, featuring explanations and analyses of movies and TV shows, providing rich contextual information to greatly enhance video understanding. Our experiments confirm that increasing the volume of pre-training data consistently improves model performance.
- We construct a video description dataset containing 150K instances, each including a detailed video description along with the specific frames corresponding to each event described. During the SFT stage, we involve this dataset to provide the model with supervision on temporal fine-grained alignment. Experimental results show that, compared with traditional video-caption alignment training, this approach significantly improves accuracy in video description and reduces the hallucinations.
- To further enhance model performance, we use the model to generate samples that automatically construct preference data for DPO training[35]. To ensure high-quality preference data, we propose two methods: a negative sampling technique that uses corrupted videos to generate negative samples for preference pairs, and a preference data filtering method that employs AutoDQ[5] to automatically filter out pairs with minimal differences. Our experiments show that DPO training on these automatically generated preference data leads to continued performance improvements over the SFT stage.

We conduct extensive experiments to evaluate Tarsier2 against both proprietary and open-source LVLMs. For video description, Tarsier2 outperforms all other models, surpassing both proprietary and open-source LVLMs in evaluations on DREAM-1K[5] and E.T. Bench-Captioning[16]. In human side-by-side evaluations, Tarsier2-7B shows a +7.8% improvement over GPT-4o and a +12.3% advantage over Gemini-1.5-Pro. It also significantly outperforms the leading open-source model, Tarsier-34B, with a +51.4% advantage. Furthermore, Tarsier2-7B proves to be a versatile generalist model, setting new SOTA results on public benchmarks for video question-answering[14][23][24], hallucination test[34], video grounding[16] and embodied QA[36]. Finally, we present extensive ablation studies to identify the key factors contributing to the model's strong performance. We also release a recaptioning dataset, Tarsier2-Recap-585K, and demonstrate its effectiveness in enhancing the capabilities of existing LVLMs for video description and general video understanding.

2. Related Work

Video-LLMs

Recently, research on Video LLMs has surged[2][37][38][39][3][40][41][9][11][4][31][6][42][43][44][45][12][29], with efforts focusing on model architectures and video-text data collection. On the architecture side, current studies emphasize visual representation[9][6][29], visual token resampling[9][12][46][47], and the integration of Vision Transformers (ViT) with LLMs[6][48][49][50]. Tarsier2 adopts a simple architecture composed of a visual encoder, a visual adaptor, and an LLM. Despite its simplicity, we demonstrate that a meticulously designed training strategy enables Tarsier2 to achieve strong video understanding capabilities.

In terms of video-text data, while many efforts aim to collect datasets for training Video LLMs, their quantity and quality remain limited. For example, LLaVA-Video[31] is trained on just 1.3 million video-text pairs, and several open-source models, such as InternVL2.5[12], Aria[42], and VILA-1.5[4], are trained on fewer than 5 million pairs. Although larger datasets like HowTo100M[51], HD-VILA[52], Panda-70M[53], and InternVid-10M[54] exist, they either cover limited domains or contain overly simplistic or low-quality text. Furthermore, some studies do not disclose the volume of video data used[6][29][43][42].

To address these challenges, our work focuses on improving the quantity and quality of video-text data. We newly collected 20 million video-text pairs, spanning a wide range of video genres. In total, 40 million pairs are used in the final pre-training stage. Additionally, we annotated 150K fine-grained video descriptions for the SFT stage.

Video Description

Video description, a foundational task in video understanding, has long been a central focus of research. Early work[55][56][57] typically involved pre-training video-language models and fine-tuning them on datasets such as MSVD[58], MSR-VTT[59], and VATEX[60], which provide single-sentence video summaries.

Recent advancements in LVLMs have improved video description, enabling more detailed outputs beyond simple summarization. However, generating comprehensive video descriptions presents challenges beyond model architecture. While multi-frame processing and temporal modeling are crucial, large-scale and rich annotated ¡video, description¿ datasets are equally important. Existing alignment datasets, such as HD-VILA[52] and HoTo100M[51], provide concise descriptions, limiting detailed video understanding. To address this, datasets such as ShareGPT4Video[61] uses a pipeline where LVLMs (e.g., GPT-V[62]) annotate frames, and LLMs (e.g., GPT-4[63]) aggregate them. This improves detail but often leads to verbosity and

doi.org/10.32388/X26ILU

hallucinations. Recent works[31][64] uses proprietary Video-LLMs, such as GPT-4o[7] and Gemini-1.5[8], for annotation, but their high cost limits application to smaller datasets.

For Tarsier2, we collect a large dataset of video-text pairs. In particular, we automatically build meaningful video-text pairs from online commentary videos. These commentaries include both low-level (atomic actions) and high-level (plot) visual elements, enhancing the model's understanding across various granularity. In addition to data collection, Tarsier2 also uses a meticulously designed three-stage training process, where DPO training after SFT further refines description accuracy and detail.

3. Approach

We initialized Tarsier with Qwen2-VL[6] weights and employed a three-stage training strategy. First, we pre-trained Tarsier2 on 40 million large-scale video-text pairs. Next, we fine-tuned the model on moderate-sized, curated, human-annotated datasets in two phases: one targeting video descriptions with fine-grained grounding and the other focusing on natural, instruction-following video descriptions. Finally, we applied Direct Preference Optimization[35] using automatically generated preference data to further enhance the quality of the video descriptions. The training process is detailed below; for a comprehensive list of hyper-parameters, please refer to Appendix A.

3.1. Pre-training

The pre-training stage encompasses a variety of tasks, including video captioning, video question answering, action recognition, action grounding, (multi-)image understanding, and text generation. The training data consists of 20 million public datasets and 20 million newly collected in-house datasets. Figure 3 illustrates the composition of the pre-training data, with a detailed breakdown presented in Appendix B. Our findings indicate that the in-house data significantly enhances model's performance, complementing the public datasets. In the following, we describe the pipeline used for in-house data collection.

Figure 3. Summary of datasets used in the pre-training stage of Tarsier2.

We collected a large group of videos from the Internet, spanning diverse genres such as animation, movies, TV series, short videos, stock footage, games and so on. The videos are categorized into three types:

- **Short videos with captions.** This category consists of 2.4 million videos directly sourced from the Internet, preserving their original video-caption pairs.
- **Commentary videos for movies or TV shows.** The videos were segmented into single-shot clips using PySceneDetect[1]. A filtering model removed static or low-quality clips. Adjacent clips were then merged to create continuous segments, ensuring final video durations ranged from 2 to 30 seconds. We utilized an internal OCR tool to extract the commentary text from the video and use it as the caption. The areas containing the commentary text in the video were obscured. To ensure relevance, we trained a lightweight BERT-style[65] model to filter out clips where the commentary lacked direct visual correspondence (e.g., character dialogues). This process produced 11.0 million video clips.

- **Other videos.** These videos were processed similarly to the commentary videos, undergoing segmentation into shorter clips, filtering out low-quality clips, and merging adjacent clips. After this, we employed a multi-modal LLM to automatically generate video captions and question-answer pairs, resulting in a total of 2.7 million clips.

Commentary videos represent a significant portion of the pre-training data. Unlike traditional video-text datasets, such as HowTo100M[51], which rely on ASR transcripts, commentary data demonstrates stronger alignment between video and text. This commentary not only describes low-level visual elements, such as atomic actions, but also highlights high-level information like plot details. This type of data can substantially enhance the model's visual understanding at varying levels of granularity.

In addition to video caption data, we incorporate large-scale synthetic datasets for tasks such as object tracking, frame order prediction, image retrieval, video question-answering, and image captioning during pre-training.

Overall, our pre-training dataset consists of 40 million samples. We trained Tarsier2 on this dataset using 128 H100 GPUs, with all components of Tarsier2 set to be trainable. For each video, we sampled between 16 and 128 frames, depending on its duration. In total, the pre-training stage of Tarsier2 processed approximately 200 billion tokens.

3.2. *Supervised fine-tuning*

During the SFT phase, our primary objectives are to further improve the model's accuracy and comprehensiveness in video descriptions and ensure the outputs are human-like: well-structured, appropriately detailed, and capable of generating accurate long-form descriptions. To achieve this, we collected 150K video clips and conducted SFT in two stages.

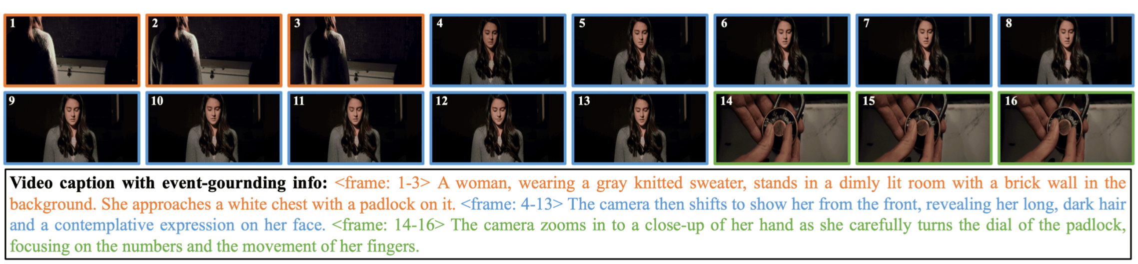


Figure 4. An example of a video description with fine-grained temporal grounding. "$<$frame: i-j $>$" indicates that the following event is inferred from frames i to j. Events are distinguished by color, with corresponding frames and descriptions marked in the same color to indicate their association.

doi.org/10.32388/X26ILU

In the first stage, each video clip in the SFT dataset is annotated with a detailed description with fine-grained temporal grounding. As shown in Figure 4, the annotations specify the frames corresponding to each event in the description. The annotation process is detailed in Appendix C. This fine-grained frame-event alignment enhances the model's ability to accurately identify and describe events by focusing on temporal and visual cues, complementing traditional video-caption alignment. Our experiments demonstrate that this approach mitigates the omission of key events in generated video descriptions.

In the second stage of SFT, we refined the model's output to achieve a more human-like style. We observed that the data used in the initial stage of SFT often fragmented complete events into multiple steps due to event-grounding requirements. For instance, the action of pouring wine might be divided into steps like opening the bottle, lifting it, and pouring. To address this, we incorporated more natural and human-like video description data. Specifically, in this stage, we designed diverse description instructions to reflect real-world variations in language, granularity, and style requirements. We then annotated each video's description to align with its corresponding instruction, as detailed in Appendix C. This data allowed the model to better interpret varying instructions and generate more accurate and diverse video descriptions.

The training data for SFT-1 contains 150k video description pairs, while SFT-2 comprises 50k diverse instructions and 150k refined video-description pairs. Each pair includes a video description aligned with one of the instructions. We trained Tarsier2 on this dataset using 32 H100 GPUs and set all components of Tarsier2 to trainable. For each video, we sampled 16 frames for training. The global training batch size was set to 64, and Tarsier2 was trained for 5000 iterations in each of the two phases. In addition, we used 2e-5 and 2e-6 as the learning rate of the model during the two-stage SFT respectively to obtain further performance improvement.

3.3. Direct Preference Optimization

In this subsection, we introduce a novel automated method for collecting preference data for video description. By performing DPO[35] training on this data, we can further improve the model's ability to generate high-quality, detailed video descriptions.

Negative sampling

Existing works often conduct multiple times sampling on the same input (video and text prompt) to acquire preference pair candidates[66][67][68]. In practice, however, we found that 1) Low-temperature sampling produces minimal variation in responses; 2) High-temperature sampling often leads to uncontrollable or abnormal generations. To address these issues, we propose a new automated preference data collection approach that enhances controllability and consistently yields high-quality preference data.

 doi.org/10.32388/X26ILU

In reinforcement learning (RL) terms, the VLM serves as a policy model π_θ, typically initialized from the SFT model. Given an input prompt x, consisting of N frames sampled from a video, π_θ generates an video description y. Then, the video frames are modified to produce a corrupted prompt $\tilde{x}$ through one of the following perturbations:

- Clip-switching: Evenly divide the video into 4 clips, then randomly choose 2 clips and swap their order.
- Clip-reversing: A random clip with $\frac{N}{2} \sim N$ frames is reversed.
- Clip-cropping: N frames are resampled from a random clip with half of the video's original duration.
- Down-sampling: Half of the N frames are randomly dropped.

The corrupted prompt $\tilde{x}$ is input into π_θ, generating a new description $\tilde{y}$. The resulting preference data is represented as $\{x, y_w = y, y_l = \tilde{y}\}$. The first two perturbations are designed to induce negative descriptions with temporal errors, while the latter two are designed to induce incomplete descriptions. Consequently, through DPO training, the model can be enhanced to produce descriptions with improved accuracy and completeness.

Figure 5. Preference data construction pipeline for DPO training.

Figure 5 provides an example to illustrate the preference data construction pipeline. From a raw video, we first generate a positive response using the current model. Next, a corrupted video, created through clip-switching, is fed into the model to obtain a negative sample, which contains two hallucinations (highlighted in red).

 doi.org/10.32388/X26ILU

Preference data filtering

Given a prompt x, response $\tilde{y}$ is generally more negative compared to y. However, an effective filter mechanism for valid preference data remains essential, as $\tilde{y}$ is not always strictly worse than y[2]. As shown on the right side of Figure 5, we utilize AutoDQ[5], an automatic method for evaluating the quality of video description, using two metrics, DQ_R and DQ_P[3]. A preference pair $\{x, y_w = y, y_l = \tilde{y}\}$ is considered valid if the following conditions are met:

$$\Delta DQ_R \geq 0 \quad \text{and} \quad \Delta DQ_P \geq 0 \quad \text{and} \quad \Delta DQ_R + \Delta DQ_P \geq \delta, \tag{1}$$

where ΔDQ_R and ΔDQ_P denotes the difference of AutoDQ recall and precision scores between the y_0 and y_1. δ serves as an adjustable threshold to fine-tune the filtering criteria.

During the DPO training phase, we utilize videos from the same training dataset, $\mathcal{D}$, as in the SFT phase, to construct preference data. The policy model is then optimized by minimizing the DPO loss, expressed as:

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right], \tag{2}$$

where π_{ref} denotes the model obtained during the SFT phase.

We conducted DPO training on a dataset with 20k preference pairs produced by the above data collection approach, with all parameters set to be trainable. For each video, we sample 16 frames as same as the SFT phase. We trained Tarsier2 for 1,000 steps in total with 64 H100 GPUs and each GPU loaded one pair at each training step, resulting in a global batch size of 64. See Appendix D for more details of DPO training.

4. Experiments

In this section, we first evaluate the model's performance on various video understanding benchmarks, comparing it to several baselines. We highlight Tarsier2's advantages not only in video description but also across other video understanding tasks. We then present an ablation study to examine key components of our approach.

4.1. Quantitative Results

4.1.1. Video Captioning

We evaluate Tarsier2 on two video captioning benchmarks: DREAM-1K[5] and E.T. Bench-Captioning[16]. DREAM-1K is a detailed video description benchmark featuring dynamic and diverse videos, assessing the model's ability to describe fine-grained actions and events. E.T Bench-Captioning is composed of four dense video captioning tasks, requiring key event localization and summary generation for segments in long-form videos.

Model	Video Categories					Overall
	Live-action	Animation	Stock	YouTube	Shorts	
Proprietary models						
GPT-4V[62]	34.8/39.2/31.3	27.4/31.9/24.0	40.7/46.7/36.1	33.8/40.1/29.2	34.8/46.1/28.0	34.4/40.8/29.7
GPT-4o[7]	39.8/42.1/37.8	35.8/39.1/33.1	44.0/46.6/41.7	35.9/41.5/31.7	39.9/47.9/34.2	39.2/43.4/35.7
Gemini-1.5-Flash[8]	34.8/36.4/33.3	29.2/32.5/26.5	39.4/39.7/39.1	34.3/38.6/30.9	35.6/42.4/30.7	34.8/37.9/32.1
Gemini-1.5-Pro[8]	36.4/36.4/36.4	30.7/31.8/29.7	42.2/40.7/43.8	34.0/36.7/31.6	37.0/42.4/32.7	36.2/37.6/34.8
Open-source models (>10B)						
PLLaVA-34B[9]	29.3/34.9/25.2	20.9/32.0/15.6	35.1/42.5/29.9	28.9/40.8/22.3	25.6/41.9/18.4	28.2/38.4/22.3
VideoLLaMA2-72B[10]	27.3/29.3/25.6	19.7/21.7/18.1	33.9/37.0/31.3	27.7/33.0/23.8	26.5/33.1/22.1	27.1/30.8/24.2
LLaVA-OV-72B[13]	31.7/32.8/30.7	27.7/30.6/25.2	38.0/39.6/36.6	34.1/34.7/33.5	33.8/41.8/28.4	33.2/35.9/30.9
LLaVA-Video-72B[31]	33.5/36.3/31.1	28.6/31.7/26.1	39.3/41.1/37.6	32.8/34.7/31.1	35.7/42.8/30.6	34.0/37.3/31.3
Qwen2-VL-72B[6]	32.1/33.7/30.6	27.6/32.6/23.9	41.1/41.2/41.1	32.0/38.1/27.7	32.1/41.0/26.4	33.2/37.3/29.9
InternVL2.5-78B[12]	25.3/31.5/21.1	21.8/28.8/17.6	33.5/38.1/29.9	31.0/38.5/25.9	31.1/41.7/24.8	28.6/35.7/23.9
Tarsier-34B[5]	38.5/39.6/37.5	32.2/35.8/29.2	41.7/46.4/37.8	34.5/41.1/29.7	34.0/44.1/27.7	36.3/41.4/32.4
Open-source models (<10B)						
Video-LLaVA-7B[3]	19.4/24.3/16.2	15.3/21.2/11.9	27.0/33.5/22.7	21.2/31.9/15.8	18.5/29.4/13.5	20.4/28.1/16.0
VideoLLaMA2-7B[10]	25.1/28.7/22.2	20.4/25.5/17.0	32.6/35.5/30.2	27.5/33.5/23.4	24.5/34.1/19.2	26.2/31.5/22.4
LLaVA-OV-7B[13]	31.2/33.2/29.3	26.8/29.0/25.0	38.1/39.1/37.1	30.6/32.1/29.2	31.4/38.3/26.6	31.7/34.3/29.4
LLaVA-Video-7B[31]	31.4/35.2/28.4	27.6/32.9/23.8	36.7/39.7/34.1	33.0/**39.5**/28.3	33.4/42.5/27.5	32.5/37.9/28.4
Qwen2-VL-7B[6]	27.7/32.5/24.2	22.2/28.0/18.4	37.0/36.1/38.0	30.7/35.5/27.0	29.1/37.6/23.8	29.6/33.9/26.3
InternVL2.5-8B[12]	26.6/32.0/22.8	21.3/28.9/16.9	32.7/37.2/29.1	27.9/35.4/23.0	28.9/39.9/22.7	27.6/34.7/22.9

doi.org/10.32388/X26ILU

Model	Video Categories					Overall
	Live-action	Animation	Stock	YouTube	Shorts	
Tarsier-7B[5]	36.6/38.5/34.8	29.3/34.6/25.5	39.6/44.7/35.5	33.0/39.2/28.4	33.6/44.6/26.9	34.6/40.3/30.2
Tarsier2-7B	44.4/41.9/47.3	39.3/39.5/39.1	45.7/45.4/46.0	36.0/38.4/33.9	43.7/48.9/39.4	42.0/42.8/41.1

Table 1. Evaluation results on DREAM-1K. We report F1/Precision/Recall scores for each category and for the overall dataset. For open-source models, all results are tested with their official checkpoint and inference code under recommended setting. SOTA results of comparable scale (<10B) are bolded and overall best results are underlined.

As shown in Table 1, Tarsier2-7B outperforms all open-source models in both precision and recall across all categories in DREAM-1K, demonstrating its ability to generate more comprehensive and less hallucinatory video descriptions. Notably, Tarsier2-7B achieved an overall F1 score of 42.0%, surpassing the strongest proprietary model, GPT-4o (39.2%). It is also the first model to exceed a 40% overall recall score, highlighting its sensitivity to dynamic actions and events.

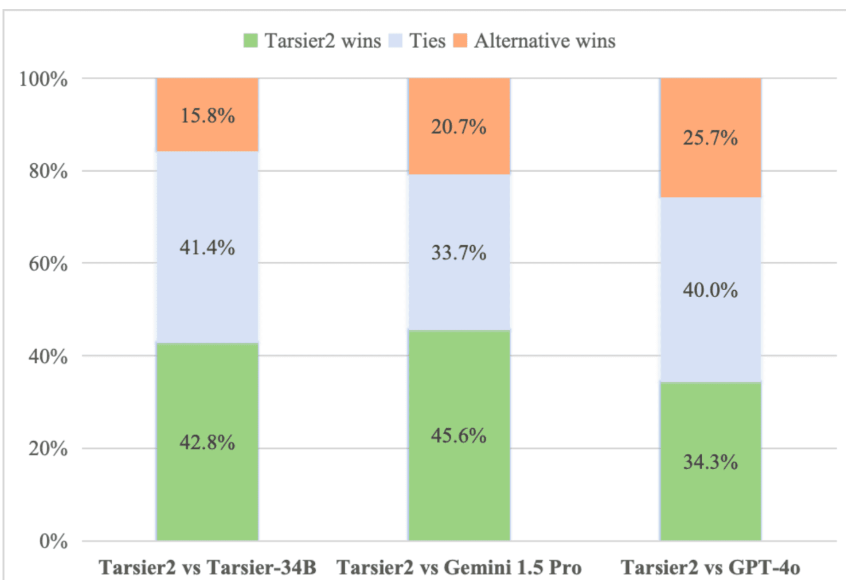


Figure 6. Human side-by-side evaluation results of Tarsier2 versus other models.

doi.org/10.32388/X26ILU

Figure 6 further presents the human side-by-side evaluation results of Tarsier2 versus the previous SOTA Tarsier-34B and two strong proprietary models, GPT-4o and Gemini 1.5 Pro. We randomly sampled 250 videos (50 videos for each category) from DREAM-1K, and asked experienced annotators to compare the descriptions generated by two different models, collecting their preferences. Each pair of descriptions was randomly shuffled to ensure that the annotators were blind to the description sources. Compared to Tarsier-34B, Tarsier2 has a slightly negative advantage rate (15.8%), but wins in a significant percentage of cases (42.8%). Compared to Gemini, Tarsier2 still maintains a significant advantage (45.6% vs 20.7%). Despite being tied with the strongest proprietary model, GPT-4o, on 40% cases, Tarsier2 still gains a slight advantage (8.6%), demonstrating the outstanding performance of Tarsier2 in detailed video description. For a comparison of generated descriptions from different models on DREAM-1K, see Appendix H.

Table 2 shows the evaluation results of dense video captioning on E.T. Bench-Captioning. Tarsier2-7B outperforms all open-source models with comparable settings (similar model scale, fine-tuned on E.T. Instruct 164K[16]) across all metrics, except for the SLC_{F1} score, which is slightly lower than Qwen2-VL-7B (24.6% vs 25.7%). These results highlight Tarsier2's strengths in generating fine-grained descriptions for short videos and providing coarse-grained summaries for long videos.

doi.org/10.32388/X26ILU

Model	E.T. Bench-Captioning[16]					
	DVC_{F1}	DVC_{Sim}	SLC_{F1}	SLC_{Sim}	Avg_{F1}	Avg_{Sim}
Proprietary models						
GPT-4V[62]	*16.1*	*19.4*	*21.9*	*13.5*	*19.0*	*16.4*
GPT-4o[7]	*46.9*	*22.3*	*23.1*	*14.9*	*35.0*	*18.6*
Gemini-1.5-Flash[8]	*31.6*	*14.9*	*16.5*	*13.3*	*24.1*	*14.1*
Gemini-1.5-Pro[8]	*24.0*	*17.5*	*5.8*	*9.8*	*14.9*	*13.7*
Open-source models (>10B)						
PLLaVA-34B[9]	13.3	10.6	9.7	11.8	11.5	11.2
LLaVA-OV-72B[13]	41.9	16.3	25.6	13.9	33.8	15.1
LLaVA-Video-72B[31]	37.0	15.7	20.4	13.5	28.7	14.6
Qwen2-VL-72B[6]	15.3	13.9	11.0	12.8	13.2	13.4
Open-source models (≤10B)						
VideoLLaMA2-7B[10]	0.6	14.5	0.0	15.2	0.3	14.8
Video-LLaVA-7B[3]	28.0	15.0	0.9	8.3	14.4	11.7
LLaVA-OV-7B[13]	22.0	15.1	9.5	10.6	15.8	12.8
LLaVA-Video-7B[31]	20.6	14.7	6.5	13.4	13.6	14.1
E.T. Chat[16]†	38.4	19.7	24.4	14.6	31.4	17.1
Qwen2-VL-7B[6]†	44.3	25.3	**25.7**	15.6	35.0	20.4
Tarsier-7B[5]†	42.8	19.1	23.7	15.2	33.2	17.1
Tarsier2-7B†	**46.5**	**28.8**	24.6	**16.4**	**35.5**	**22.6**

Table 2. Evaluation results on E.T. Bench-Captioning. Results marked in gray(italics) are tested on a subset. † denotes the model is fine-tuned on E.T. Instruct 164K. All results are transcribed from the official benchmark, except for LLaVA-OV, LLaVA-Video and Qwen2-VL, which are our evaluation using the official checkpoint and inference code.

4.1.2. Short-Video Question Answering

Model	MVBench[15]	PerceptionTest[20]	TVBench[14]	TOMATO[23]	Vinoground[24]	TempCompass[25]
	test	val	test	test	Text/Video/Group	mc/yn/cm/cg
Proprietary models						
GPT-4o[7]	57.5	-	39.6	37.7	54.0/38.2/24.6	71.0/73.7/80.8/70.8
Gemini-1.5-Pro[8]	-	-	46.5	36.1	35.8/22.6/10.2	63.9/70.3/77.5/57.9
Open-source models (>10B)						
LLaVA-OV-72B[13]	59.4	66.9	45.9	28.6	48.4/35.2/21.8	67.6/72.6/78.2/52.6
LLaVA-Video-72B[31]	64.1	74.3*	50.0	28.2	52.0/35.6/20.8	69.9/73.0/80.9/54.4
Qwen2-VL-72B[6]	73.6	66.5	52.7	37.9	50.4/32.6/17.4	76.0/75.9/84.6/58.6
Tarsier-34B[5]	67.6	60.4	53.8	34.3	37.8/32.0/15.0	69.8/74.0/73.0/60.9
Open-source models (≤10B)						
LLaVA-OV-7B[13]	56.7	57.1	45.6	25.5	41.6/29.4/14.6	64.8/69.7/73.8/49.9
LLaVA-Video-7B[31]	58.6	67.9*	45.6	24.9	36.8/29.0/12.8	56.3/68.7/76.8/53.0
Qwen2-VL-7B[6]	67.0	-	43.8	31.5	40.0/23.4/12.4	68.5/72.8/77.3/54.2
Tarsier-7B[5]	62.6	53.9	45.8	28.6	29.8/22.2/8.6	58.7/58.0/54.2/55.3
Previous SOTA	**72.0**[12]	70.0*[45]	51.6[22]	31.5[6]	41.6/29.4/14.6[11]	68.5/72.8/77.3/54.2[6]

doi.org/10.32388/X26ILU

Model	MVBench[15]	PerceptionTest[20]	TVBench[14]	TOMATO[23]	Vinoground[24]	TempCompass[25]
	test	val	test	test	Text/Video/Group	mc/yn/cm/cg
Tarsier2-7B	71.5	**71.6***	54.7	42.0	65.8/38.0/28.8	75.3/75.1/80.6/66.6

Table 3. Evaluation results on short video question answering benchmarks. * indicates that the training set has been observed in the training data mixture.

We evaluate Tarsier2-7B on several short-video question answering benchmarks to assess its ability to comprehend and reason about visual content. As shown in Table 3, Tarsier2-7B outperforms both proprietary and open-source models across various benchmarks, achieving state-of-the-art results. Tarsier2-7B exhibits exceptional performance in MVBench[15] and PerceptionTest[20], with scores of 71.5% and 71.6%, respectively.

Furthermore, Tarsier2-7B demonstrates significant performance improvements on benchmarks featuring temporal reasoning, such as TVBench[14], TOMATO[23], and Vinoground[24]. Tarsier2-7B achieves strong results with 54.7% on TVBench, 42.0% on TOMATO, and 65.8%/38.0%/28.8% on Vinoground's Text/Video/Group tasks, respectively. These results surpass both open-source and proprietary models, including GPT-4o and Gemini-1.5-Pro.

At last, Tarsier2-7B also excels on the TempCompass benchmark[25], which evaluates temporal perception in ten aspects and four task formats. Tarsier2-7B achieves impressive scores of 75.3%/75.1%/80.6%/66.6% on TempCompass' mc/yn/cm/cg tasks, respectively, outperforming both open-source models and larger proprietary models in most cases. This performance further underscores Tarsier2's advanced ability to process and interpret temporal information in video content.

 doi.org/10.32388/X26ILU

4.1.3. Long-Video Question Answering

Model	Video-MME[26]	LongVideoBench[28]	TemporalBench[30]	MLVU[18]	MMBench-Video[32]
	w/o subs	val	Binary Accuracy	M-Avg	val
Proprietary models					
GPT-4o[7]	71.9	66.7	73.2	64.6	1.87
Gemini-1.5-Pro[8]	75.0	64.0	66.4	-	1.30
Open-source models (>10B)					
VILA-1.5-40B[4]	60.1	-	-	56.7	1.61
LLaVA-Video-72B[31]	70.5	61.9	72.4	74.4	1.71
Qwen2-VL-72B[6]	71.2	-	70.2	-	1.70
InternVL2.5-78B[12]	72.1	63.6	-	75.7	1.97
Tarsier-34B[5]	52.3	54.2	66.7	58.2	1.46
Open-source models (<=10B)					
LLaVA-Video-7B[31]	63.3	58.2	63.6	70.8	1.60
Qwen2-VL-7B[6]	63.3	55.6	62.0	-	1.44
InternVL2.5-8B[12]	64.2	60.0	-	68.9	1.68
Tarsier-7B[5]	42.2	39.8	56.9	49.3	-
Previous SOTA	64.2[27]	60.0[12]	63.6[31]	70.9[29]	1.70[33]
Tarsier2-7B	**64.5** (128f)	58.6 (128f)	**65.3** (128f)	67.9 (256f)	**1.82** (128f)

Table 4. Evaluation results on long-video question answering benchmarks. We list the number of frames used for each benchmark during evaluating Tarsier2.

We evaluate Tarsier2 on long-video question answering benchmarks by uniformly sampling 128 or 256 frames, depending on the video length. Comparison results with other proprietary and open-source models are presented in Table 4. Despite our training set not including many long video data, Tarsier2, compared with others under 10 billion parameters, still achieves SOTA on three benchmarks and competitive performance on several other benchmarks.

doi.org/10.32388/X26ILU

4.1.4. Hallucination

Model	VideoHallucer[17]	EventHallusion[34]	
	Yes/No QA	Yes/No QA	Desc GPT
	Basic/Hallucinated/Overall	**Entire/Interleave/Misleading/Overall**	**Entire/Interleave/Misleading/Overall**
Proprietary models			
GPT-4o[7]	75.1/74.2/53.3	65.8/90.7/92.2/84.1	34.9/54.9/83.2/56.2
Gemini-1.5-Pro[8]	83.6/42.3/37.8	70.2/77.7/96.1/80.2	38.5/40.9/80.0/49.6
Open-Source models (>10B)			
Qwen2-VL-72B[6]	87.1/79.4/<u>70.2</u>	33.3/77.7/56.4/60.0	16.5/25.4/70.2/33.6
LLaVA-OV-72B[13]	88.3/62.6/55.2	47.4/26.9/90.1/48.3	24.8/34.7/71.3/40.7
LLaVA-Video-72B[31]	88.2/73.5/64.6	57.9/11.9/96.0/45.6	32.1/35.8/75.5/44.2
InternVL2.5-78B[12]	82.5/82.5/67.8	57.9/67.9/88.2/70.2	45.0/43.0/76.8/51.6
Tarsier-34B[5]	84.8/80.0/67.7	49.1/92.7/69.6/74.8	38.5/40.4/83.2/50.1
Open-Source models (≤10B)			
LLaVA-OV-7B[13]	81.1/69.6/53.8	46.5/67.4/86.1/66.2	22.0/26.4/73.4/36.4
LLaVA-Video-7B[31]	82.4/70.6/56.0	61.4/48.7/96.0/64.0	27.5/32.6/75.5/41.4
Qwen2-VL-7B[6]	85.0/70.8/59.3	35.1/94.3/57.4/68.6	14.7/16.1/67.0/27.8
InternVL2.5-8B[12]	72.7/78.3/53.6	46.5/69.2/90.2/68.2	23.9/20.7/60.0/31.0

doi.org/10.32388/X26ILU

Model	VideoHallucer[17]	EventHallusion[34]	
	Yes/No QA	Yes/No QA	Desc GPT
	Basic/Hallucinated/Overall	**Entire/Interleave/Misleading/Overall**	**Entire/Interleave/Misleading/Overall**
Tarsier-7B[5]	76.4/60.8/41.4	43.9/82.4/79.4/70.9	35.8/29.5/72.6/41.6
Tarsier2-7B	86.5/78.3/**67.0**	60.5/93.3/95.1/**84.6**	54.6/53.1/93.7/**63.3**

Table 5. Evaluation results on hallucination benchmarks.

We evaluate Tarsier2 on two video hallucination benchmarks: VideoHallucer[17] and EventHallusion[34]. The results are summarized in Table 5. For VideoHallucer, Tarsier2-7B achieves an overall score of 67.0%, outperforming all comparable baselines of similar model scale and even proprietary models like GPT-4o and Gemini-1.5-pro. In EventHallusion, for video question-answering task, Tarsier2-7B achieves 84.6%, surpassing GPT-4o's score of 84.1%, while outperforming all other baselines. For the detailed description matching task, which directly assesses video description hallucinations by prompting GPT-4 to answer questions based on each model's generated video description, Tarsier2-7B demonstrates superior performance, even surpassing GPT-4o by 7.1% in terms of Overall score.

 doi.org/10.32388/X26ILU

4.1.5. Video Grounding

Model	E.T. Bench-Grounding[16]					
	TVG_{F1}	EPM_{F1}	TAL_{F1}	EVS_{F1}	VHD_{F1}	$Mean_{F1}$
Proprietary models						
GPT-4V[62]	*27.0*	*1.8*	*18.0*	*28.6*	*55.1*	*26.1*
GPT-4o[7]	*40.4*	*4.5*	*20.0*	*17.6*	*56.9*	*27.9*
Gemini-1.5-Flash[8]	*43.9*	*5.4*	*27.0*	*5.4*	*60.8*	*28.5*
Gemini-1.5-Pro[8]	*43.1*	*6.2*	*33.8*	*7.9*	*47.0*	*27.6*
Open-source models (<10B)						
LITA[69]	22.2	4.6	18.0	29.7	23.9	19.7
VTG-LLM[70]	15.9	3.7	14.4	26.8	48.2	21.8
TimeChat[71]†	-	-	-	-	-	24.3
E.T. Chat[16]†	38.6	10.2	30.8	**25.4**	62.5	33.5
Tarsier-7B[5]†	39.6	9.0	25.0	**25.4**	47.6	30.9
Qwen2-VL-7B[6]†	**39.7**	7.0	26.9	17.1	**66.9**	33.5
Tarsier2-7B†	38.4	**11.0**	**31.8**	19.4	66.8	**35.5**

Table 6. Evaluation results on E.T. Bench-Grounding. Results marked in gray(italics) are tested on a subset. † denotes the model is fine-tuned on E.T. Instruct 164K.

We evaluate the video grounding capability of models on E.T. Bench-Grounding, which combines various grounding tasks from multiple datasets, including QVHighlights[72], Charades-STA[73], THUMOS'14[74], and Ego4D-NLQ[75], among others. The results, shown in Table 6, indicate that Tarsier2-7B achieves the highest mean F1 score of 35.5%, outperforming all baselines and highlighting its superior temporal perception capabilities.

doi.org/10.32388/X26ILU

4.1.6. Embodied Question Answering

Model	EgoTaskQA	Model	RoboVQA	Model	OpenEQA
	Exact Match		BLEU-1/2/3/4		GPT-4
Human	80.0	LLaMA-AdapterV2[76]	27.8/16.0/10.9/8.1	Human	86.8
HCRN[77]	42.2	LLaVA-OV-7B[13]	38.1/33.6/31.8/31.0	GPT-4V[62]	55.3
GF[78]	44.3	RoboMamba[79]	54.9/44.2/39.5/36.3	Gemini-1.5-Pro[8]	44.9
EgoVLPv2[80]	46.3	MLCD[81]	73.2/66.4/60.6/56.6	MLCD[81]	48.8
Tarsier2	**77.5**	Tarsier2	**77.1**/**67.4**/**61.5**/**56.8**	Tarsier2	**58.7**

Table 7. Evaluation results on embodied question-answering tasks, including EgoTaskQA, RoboVQA and OpenEQA.

We evaluate Tarsier2 on embodied question answering to assess its performance in real-world robotic scenarios, using three benchmarks: EgoTaskQA[82], RoboVQA[36], and OpenEQA[83]. To align with the baselines, Tarsier2 is fine-tuned on the training sets for EgoTaskQA and RoboVQA, while for OpenEQA, it is evaluated in a zero-shot setting. The results, presented in Table 7, include exact match accuracy for EgoTaskQA, BLEU score for RoboVQA, and the correctness score evaluated by GPT-4-1106-preview[63] for OpenEQA. Tarsier2 achieves top-tier performance across all three benchmarks, outperforming both generalist and specialist models. Notably, on EgoTaskQA, its performance approaches human-level accuracy, highlighting the model's significant potential in embodied intelligence.

4.2. Ablation Study

We conduct a comprehensive ablation study to evaluate key components at different stages of the training process. The study is based on three tasks: 1) **Caption**: This includes the DREAM-1K dataset, the caption generation task from TempCompass (TempCompass-cg), and the caption matching task from Vinoground (Vinoground-Text) to assess captioning performance. 2) **Video QA**: This encompasses short-video QA, measured by the average accuracy on MVBench, TVBench, and TOMATO, and long-video QA, measured by the average accuracy on Video-MME, LongVideoBench, and TemporalBench. It evaluates the model's video understanding capabilities. 3) **Hallucination**: We use the average score of two sub-tasks from EventHallusion to assess hallucination in the model. The following subsections present the results for each task, with detailed results for individual datasets provided in the Appendix E.

4.2.1. Pre-training

Model	Caption			Video QA		Hallucination
	DREAM-1K	TempCompass-cg	Vinoground-Text	Short	Long	
Tarsier1-7B	34.6	55.3	29.8	45.6	46.3	56.3
Tarsier1-7B-Qwen *upgrading model*	38.4 (↑3.8)	59.3 (↑4.0)	48.6 (↑18.8)	52.4 (↑6.8)	57.6 (↑11.3)	62.1 (↑5.8)
Tarsier2-7B *upgrading model+data*	40.8 (↑6.2)	60.1 (↑4.8)	60.2 (↑30.4)	55.3 (↑9.7)	64.1 (↑17.8)	63.5 (↑7.2)

Table 8. Results of the ablation study for pre-training. Tarsier1-7b-Qwen stands for the model where the base model is upgraded to Qwen2-VL, while the pre-training dataset remains the same as Tarsier1. Tarsier2 is trained from Qwen2-VL with an expanded pre-training dataset, growing from 13 million in Tarsier1 to 40 million samples.

In this section, we evaluate the impact of several factors during pre-training, including the base model, pre-training data and training steps. For the caption task, we report results after the SFT stage, which aligns the model's responses with the desired style. For other tasks, we report results after pre-training stage.

Compared to Tarsier1, two key improvements are made in the pre-training phase: upgrading the base model to Qwen2-VL and expanding the training dataset from 13 million to 40 million samples. Table 8 illustrates the additive contributions for each improvement, showing that both enhancements consistently and significantly boost the model's performance in caption generation, video QA, and hallucination reduction. Specifically, these enhancements lead to accuracy improvements of 9.7%, 17.8%, and 7.2% for short-video QA, long-video QA, and hallucination tests, respectively. For video description, the F1 score on the DREAM-1K dataset improves by 6.2%.

doi.org/10.32388/X26ILU

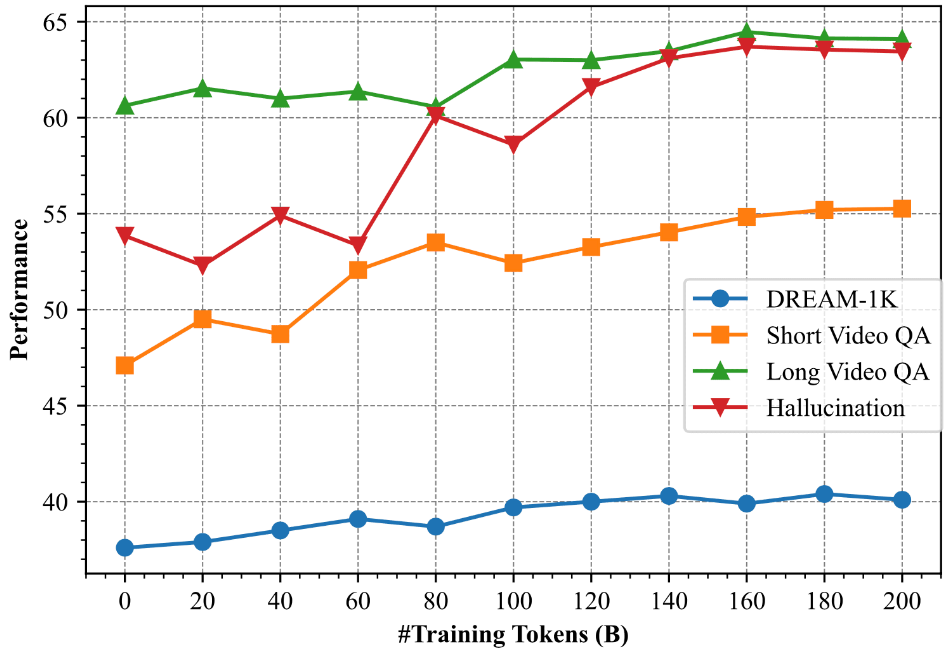


Figure 7. Model performance against training tokens. The results at the initial step reflect the performance of Qwen2-VL-7B.[4]

To better understand the effect of the number of training tokens on pre-training performance, we plot the model's performance as a function of token count during the pre-training stage, as shown in Figure 7. The results show that model performance improves with an increase in the number of training tokens, reaching convergence after 160 billion tokens. This suggests that a large volume of data is essential for optimal video understanding performance.

4.2.2. *SFT*

Model	Caption			Video QA		Hallucination
	DREAM-1K	TempCompass-cg	Vinoground-Text	Short	Long	
Tarsier2-7B-SFT	40.8	60.1	60.2	56.2	63.2	71.9
w/o *SFT*	35.2 (↓5.6)	50.5 (↓9.6)	57.2 (↓3.0)	55.3 (↓0.9)	64.1 (↑0.9)	63.5 (↓8.4)
w/o *grounding*	37.4 (↓3.4)	50.2 (↓9.9)	60.6 (↑0.4)	55.9 (↓0.3)	61.9 (↓1.3)	68.6 (↓3.3)

Table 9. Ablation study of temporal grounding dataset during the SFT phase. Tarsier2- 7B-SFT refers to the model after the SFT phase. *w/o SFT* refers to the model after pre- training; *w/o grounding* refers to the model fine-tinued without grounding information

The key factor in the SFT phase is fine-grained alignment. To investigate its impact, we conduct an ablation study, with the results presented in Table 9. When the video description data, which includes fine-grained temporal grounding information, is excluded (i.e., without grounding), model performance significantly deteriorates. Specifically, the F1 score on DREAM-1K decreases by 3.4%, accuracy on TempCompass-cg drops by 9.9%, accuracy on long-video QA falls by 1.3%, and accuracy on the hallucination test declines by 3.3%.

Furthermore, the SFT phase leads to substantial improvements, highlighting the importance of high-quality manually labeled data. It boosts the F1 score on DREAM-1K by 5.6%, accuracy on TempCompass-cg by 9.6%, accuracy on Vinoground-Text by 3.0%, and accuracy on the hallucination test by 8.4%, demonstrating the SFT phase's role in enhancing the model's fine-grained video understanding and mitigating hallucinations.

4.2.3. DPO

Model	Caption			Video QA		Hallucination
	DREAM-1K	TempCompass-cg	Vinoground-Text	Short	Long	
Tarsier2-7B	42.0	66.6	65.8	56.1	62.8	74.0
w/o DPO	40.8 (↓1.2)	62.1 (↓6.5)	60.6 (↓5.6)	56.2 (↑0.1)	63.2 (↑0.4)	71.9 (↓2.1)
w/o NS	41.5 (↓0.5)	61.1 (↓5.5)	59.8 (↓6.0)	56.1 (↓0.0)	62.8 (↓0.0)	72.9 (↓1.1)
w/o PF	40.5 (↓1.5)	65.1 (↓1.5)	67.6 (↑1.8)	56.0 (↓0.1)	62.3 (↓0.5)	74.2 (↑0.2)

Table 10. Ablation study for DPO training phase, negative sampling (NS) and preference data filtering (PF) strategies.

We conduct ablation experiments to evaluate the DPO phase, negative sampling (NS) and preference data filtering (PF) strategies. Specifically, we test the following settings: 1) ***w/o DPO***: SFT model without DPO training. 2) ***w/o NS***: Preference pairs generated by sampling the same video twice, without negative sampling. 3) ***w/o PF***: Responses from negative sampling are treated as rejections, without utilizing AutoDQ Scorer to perform preference data filtering. For a fair comparison, the training data size and hyper-parameters for the latter two settings are kept consistent with the default setting, as detailed in Appendix D.

As shown in Table 10, Tarsier2 benefits a lot from the DPO training phase with significant improvement on caption tasks, especially TempCompass-cg (6.5%) and Vinoground-Text (5.6%). The hallucination capability also drops by 2.1% without DPO, while the performance on video QA is not obviously affected. When further ablating dataset construction strategy of DPO, negative sampling plays an important role, without which the model results on most of the tasks are degraded to be almost the same as the SFT model ("*w/o DPO*"), and the hallucination capability drops by 1.1%. Additionally, preference data filtering with AutoDQ scorer has a significant impact on maintaining the quality of DPO datasets. As shown in Table 10, "*w/o PF*" leads to degradation on more than a half of the tasks, and especially the DREAM-1K F1 score is even worse than the SFT model.

4.3. Video Recaptioning using Tarsier2

Model	Caption			Video QA		Hallucination
	DREAM-1K	TempCompass-cg	Vinoground-Text	Short	Long	
Qwen2-VL-7B[6]	31.2	54.2	40.0	49.4	60.3	51.9
+ *Original FT*	35.2 (↑4.0)	49.9 (↓4.3)	39.0 (↓1.0)	46.9 (↓2.5)	55.4 (↓4.9)	43.0 (↓8.9)
+ *Recaption FT*	39.5 (↑8.3)	67.7 (↑13.5)	55.0 (↑15.0)	52.5 (↑3.1)	56.8 (↓3.5)	68.5 (↑16.6)

Table 11. The experimental results of recaptioning. "*Recaption FT*" represents fine-tune the model on the Tarsier2-Recap-585K dataset. "*Original FT*" represents fine-tune the model with the same videos as Tarsier2-Recap-585K but taking their original labels as target output.

In this section, we utilize Tarsier2 as a captioner to generate detailed descriptions for a diverse set of 1M videos sourced from public datasets, resulting in the recaptioning dataset Tarsier2-Recap-585K[5]. Details of the dataset composition are provided in Appendix F.

We fine-tune Qwen2-VL-7B[6] on Tarsier2-Recap-585K and present the evaluation results in Table 11. Fine-tuning on Tarsier2-Recap-585K significantly enhances the model's performance on detailed video description, achieving improvements in DREAM-1K (+8.3%), TempCompass-cg (+13.4%), and Vinoground-Text (+15.0%). Moreover, it achieves an improvement of 16.6% in hallucination test and an improvement of 3.1% in short video-QA.

In comparison, fine-tuning on the same 585K videos with original captions improves only the DREAM-1K F1 score (+4.0%), while other metrics show significant declines. It indicates that the performance gains from Tarsier2-Recap-585K are primarily due to its high-quality and detailed captions rather than the additional training data volume.

Table 17 in Appendix E provides detailed benchmark results corresponding to Table 11. These findings demonstrate that Tarsier2 can generate high-quality, detailed descriptions that offer fine-grained alignment information to help LVLMs to achieve significant improvements across various tasks.

 doi.org/10.32388/X26ILU

5. Conclusion

In this paper, we introduce Tarsier2, a state-of-the-art large vision-language model that outperforms existing proprietary and open-source models in generating detailed and accurate video descriptions. Furthermore, Tarsier2 sets new benchmarks across a wide range of video understanding tasks. Our ablation studies demonstrate that Tarsier2 's advancements are driven by scaling the volume and diversity of the training dataset, fine-grained temporal alignment, and DPO training.

Looking ahead, we outline several promising directions for future research. First, extending Tarsier2 to handle longer video durations by developing more efficient model architectures and expanding the training dataset. Second, enhancing real-time video processing to improve the model's ability to analyze and describe videos as they stream. Third, exploring richer interactions between video, audio, and text to create more comprehensive and context-aware video understanding systems.

Appendix A. Training hyper-parameters

Table 12 shows the training hyper-parameters in pre-training, SFT-1&2 and DPO stage. We apply a layer-wise learning rate decay of 0.9 for visual encoder training[84].

Configuration	Pre-training	SFT-1	SFT-2	DPO
VLM init.	Qwen2-VL-7B	Tarsier2-Pre-trian	Tarsier2-SFT-1	Tarsier2-SFT-2
Optimizer name	AdamW			
Optimizer β_1	0.9			
Optimizer β_2	0.999			
Optimizer eps	$1e^{-6}$			
Learning rate	$2e^{-5}$	$2e^{-5}$	$2e^{-6}$	$1e^{-6}$
Learning rate schedule	cosine			
Training steps	200,000	5,000	5,000	1,000
Warm-up steps	1,000	250	250	100
Weight decay	0.01			
Gradient clip	1.0			
Dropout rate	0.0			
Global batch size	384	64	64	64
Max pixels	460,800			
Frames per video	[8,128]	16	16	16
Numerical precision	bfloat16			

Table 12. Training hyper-parameters of Tarsier2

Appendix B. Public datasets of pre-training stage

Table 13 presents the pre-training datasets, which collectively include approximately 20 million public data and 20 million in-house data. Most of the public datasets are the same as Tarsier1, except we additionally gathered some newly released open-source data and OCR-releated data. For WebVid-10M, we used 2.9 million video-text pairs, selecting samples that are more likely to feature dynamic events. We have also incorporated some latest long video understanding datasets, such as MovieStory101[85] and LLaVA-Video-178K[31]. This greatly enhances the model's ability to understand long videos.

doi.org/10.32388/X26ILU

Video Captioning			
WebVid[86] (2.9M)	LSMDC[87] (109K)	TGIF[88] (105K)	ActivityNet[89] (38K)
Charades[90] (16K)	Charades-Ego[91] (6K)	YouCook2[92] (9K)	TACoS[93] (18K)
Ego4D[75] (1.1M)	Spoken Moments[94] (493K)	Multi-Moments[95] (997K)	TREC-VTT[96] (64K)
ShareGPT-4o-video[97] (2K)	MovieStory101[85] (11K)	GPT4o-labeled Caption† (2.5M)	Human-labeled Caption† (145K)
Film&TV Commentary† (11.5M)			
Action Recognition			
HMDB[98] (5.8K)	COIN[99] (10K)	SSV2[100] (169K)	Kinetics-700[101] (537K)
FineAction[102] (82K)	RareAct[103] (2K)	20BN-jester[104] (46K)	
Video QA			
CLEVRER[105] (83K)	TGIF-QA[106] (72K)	EgoQA[107] (5K)	VideoInstruct[37] (89K)
LLaVA-Video-178K[31] (165K)	M4-Instruct-video[11] (255K)	GPT4o-labeled QA† (16.2K)	
Grounding			
DiDeMo[108] (82K)	AVA[109] (28K)	E.T. Instruct 164K[16] (147K)	Object Tracking†(745K)
Video Self-Supervised Training			
Frame Order Prediction† (825K)			
Intent Recognition			
Oops![110] (15K)			
Multi-Image Understanding			
VIST[111] (38K)	MMDU[112] (45K)	M4-Instruct-image[11] (616K)	Image Retrival†(533K)
Single-Image Understanding			
ShareGPT4V[113] (95K)	LLaVA-1.5[114] (643K)	ShareGPT-4o-image[97] (57K)	MS COCO[115] (566K)

doi.org/10.32388/X26ILU

Flicker[116] (145K)	LLaVA-ReCap-CC3M[11] (2.9M)	Visual Genome[117] (759K)	SBU Captions[118] (860K)
GPT4o-labeled Caption† (1.13M)			
Image OCR			
RCTW-17[119] (8K)	LSVT[120] (430K)	ReCTS[121] (20K)	Art[122] (5.6K)
COCOTextV2[123] (16K)	CORD-v2[124] (1K)	HierText[125] (10K)	MSRA-TD500[126] (465)
IC03[127] (499)	SynthDoG-en[128] (100K)	SynthDoG-zh[128] (100K)	
Text Generation			
OpenOrca[129] (995K)	ShareGPT[130] (80K)		

Table 13. Datasets and their sizes used in Tarsier2 pre-training. † indicates in-housedatasets.

Appendix C. Annotation process for SFT data

In the first stage of SFT, we annotated each video clip with detailed descriptions that included fine-grained temporal grounding. Each clip first underwent manual annotation, where annotators described dynamic information such as character actions, events, scene transitions, and camera movements, while avoiding unnecessary static elements. Annotators are also required to map the dynamic information in their descriptions to the corresponding frame numbers. We performed quality inspections on the annotated data and returned any data not meeting quality standards for re-annotation. We discarded any data that might involve copyright risks.

In the second stage of SFT, we utilized GPT-4o to generate a variety of instruction tuning samples based on manual annotations. We provided GPT-4o with 16 uniformly sampled frames from the video and the original manual annotations. Figure 8 shows the prompt for re-annotation in this stage.

doi.org/10.32388/X26ILU

The re-annotation prompt for diverse instruction data (SFT-2).

Character
You are an excellent video analyst. Utilizing your incredible attention to detail, you provide clear, sequential descriptions for video. You excel in identifying and conveying changes in actions, behaviors, environment, states and attributes of objects, and camera movements between video frames.
Prompt
Here are 16 frames from a video and a short video caption in Chinese. You need to process a two step tasks: First, establish a set of guiding principles to control the style of the video description. These principles should include one or more of the following aspects:

1. Specify the length constraints of the description, including the number of paragraphs and total word count.
2. Define the level of detail for human or creature appearance, non-creature appearance, and background.
3. Determine the granularity of the event information.
4. Decide on the output format, such as plain text, JSON, lists, narrative, poetry, etc.
5. Choose the output language, such as Chinese, English, Japanese, French, and so on.
6. Decide on the text style, such as fluent, concise, professional, or just using simple words and phrases.

Next, generate the corresponding video description based on these guiding principles and the input video clip, and rephrase the guiding principles into natural language as part of the output question.
Input
Origin Short Video Caption in Chinese: {Manual Labeled Chinese Caption}
Requirement
Return in JSON format: {“qustion”: xxx,“answer”: xxx}

Figure 8. The re-annotation prompt in SFT-2.

Appendix D. Detail setting of DPO training

As a default setting, we leveraged the negative sampling and preference pair filtering strategy as introduced in Section 3.3 to construct the DPO training set. We set top_p as 0.7 and temperature as 0.7 when running both positive sampling and negative sampling on our 150K SFT dataset. The threshold δ of preference pair filtering was set as 0.3. We finally randomly sampled 20K preference pairs for DPO training. For the “w/o NS” setting, we kept other parameters and process unchanged but replaced the negative sampling with an additional positive sampling. For the “w/o PF” setting, we omitted the process of preference pair filtering and directly sample 20K pairs from all preference pair candidates. We utilized the vanilla DPO training objective (Equation 2), and set β as 0.1. See the “DPO” column of Table 12 for all the other hyper-parameters.

Appendix E. Detailed results of individual datasets at different stages

In this section, we provide detailed results for individual datasets in our ablation study. Table 14, 15 and 16 list the results for pre-training, SFT and DPO respectively. Table 17 lists the results for the recaptioning experiment. We report F1/Precision/Recall for DREAM-1K and accuracy for other benchmarks.

Capability	Benchmark	Tarsier1-7B	Tarsier1-7B-Qwen	Tarsier2-7B
Caption	DREAM-1K	34.6/30.2/40.3	38.4/40.6/36.4	40.8/42.5/39.3
	TempCompass-cg	55.3	59.3	60.1
	Vinoground-Text	29.8	48.6	60.2
Video QA Short	MVBench	62.6	69.8	72.8
	TVBench	45.8	51.0	53.5
	TOMATO	28.6	36.5	39.5
Video QA Long	Video-MME	42.2	58.9	65.3
	LongVideoBench	39.8	52.1	58.3
	TemporalBench	56.9	61.9	68.7
Hallucination	EventHallusion-Y/N	70.9	75.6	77.8
	EventHallusion-Desc	41.6	48.6	49.1

Table 14. Detailed results of the ablation study for pre-training. For the captioning task, results are reported after the SFT stage. For other tasks, results are reported after the pre-training stage.

doi.org/10.32388/X26ILU

Capability	Benchmark	pre-train	SFT w/o grounding	SFT
Caption	DREAM-1K	35.2/36.8/33.7	37.4/38.6/36.3	40.8/42.5/39.3
	TempCompass-cg	50.5	50.2	60.1
	Vinoground-Text	57.2	60.6	60.2
Video QA Short	MVBench	72.8	71.9	72.5
	TVBench	53.5	54.5	54.2
	TOMATO	39.5	41.3	41.9
Video QA Long	Video-MME	65.3	64.0	64.7
	LongVideoBench	58.3	54.7	58.2
	TemporalBench	68.7	66.9	66.6
Hallucination	EventHallusion-Y/N	77.8	80.1	84.4
	EventHallusion-Desc	49.1	56.2	59.4

Table 15. Detailed results of the ablation study for SFT.

doi.org/10.32388/X26ILU

Capability	Benchmark	Tarsier2-7B	w/o DPO	w/o NS	w/o PF
Caption	DREAM-1K	42.0/42.8/41.1	40.8/42.5/39.3	41.5/44.5/39.0	40.5/39.9/41.1
	TempCompass-cg	66.6	60.1	62.1	65.1
	Vinoground-Text	65.8	60.2	60.6	67.6
Video QA Short	MVBench	71.5	72.5	72.2	71.7
	TVBench	54.7	54.2	54.9	54.6
	TOMATO	42.0	41.9	41.3	41.8
Video QA Long	Video-MME	64.5	64.7	64.3	64.4
	LongVideoBench	58.6	58.2	58.6	57.4
	TemporalBench	65.3	66.6	65.4	65.2
Hallucination	EventHallusion-Y/N	84.6	84.4	85.1	84.8
	EventHallusion-Desc	63.3	59.4	60.7	63.5

Table 16. Detailed results of the ablation study for DPO.

doi.org/10.32388/X26ILU

Capability	Benchmark	Qwen2-VL-7B[6]	+Original FT	+Recaption FT
Caption	DREAM-1K	29.6/33.9/26.3	35.2/44.8/29.0	39.5/41.7/37.6
	TempCompass-cg	54.2	49.9	67.7
	Vinoground-Text	40.0	39.0	55.0
Video QA Short	MVBench	67.0	59.8	66.8
	TVBench	43.8	47.2	51.1
	TOMATO	31.5	33.6	39.5
Video QA Long	Video-MME	63.3	56.1	57.0
	LongVideoBench	55.6	51.4	51.9
	TemporalBench	62.0	58.7	61.4
Hallucination	EventHallusion-Y/N	68.6	39.6	80.7
	EventHallusion-Desc	27.8	46.3	56.2

Table 17. Detailed results of the recaptioning experiment.

Appendix F. Tarsier2-Recap-585K Data Composition

Table 18 lists the data composition details of Tarsier2-Recap-585K. We mainly took video caption datasets into account when picking the target datasets, together with two action recognition datasets (Kinetics-700[101] and SSV2[100]), which contain video clips of durations of $5 \sim 10$ seconds about human actions, and a special intent recognition dataset (Oops[110]) to help models learn rare actions and unexpected events. For most of the datasets, we utilized all the original video clips of the selected splits (usually train and val set), except for:

- WebVid-10M: We sampled around 30% of the total size of Tarsier2-Recap-585K from a pre-filtered subset of WebVid-10M, which are more likely to feature dynamic events.
- Ego4D: We randomly merged multiple clips into a new one that contains multiple actions and result in around 1M merged clips in total. We sampled 50K clips from this dataset for recaptioning.
- Kinetics-700 and SSV2: We randomly sampled 50K and 10K clips from the training set of Kinetics-700 and SSV2, respectively.

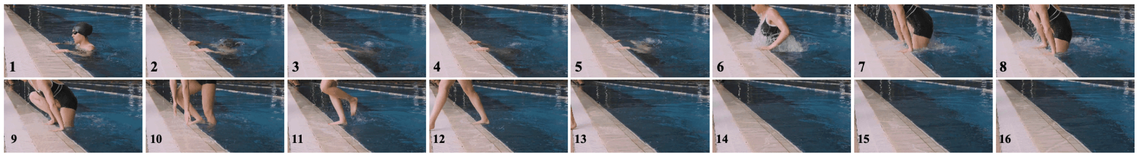
 doi.org/10.32388/X26ILU

Dataset	Original Label Type	Split	Avg Duration (s)	# Sampled Clips	Proportion (%)
WebVid-10M[86]	Video Caption	-	15.2	177,909	30.38
LSMDC[87]		train/val/test	4.1	108,271	18.49
TGIF[88]		train/test	12.3	94,775	16.18
Ego4D[75]		-	4.1	50,000	8.54
ActivityNet[89]		train/val/test	35.7	35,960	6.14
VATEX[60]		train/val/test	10.0	22,435	3.83
TREC-VTT[96]		train/val	6.3	14,199	2.42
Charades[90]		train/test	29.8	7,985	1.36
Charades-Ego[91]		train/test	30.2	6,161	1.05
Kinetics-700[101]	Action Recognition	train/val/test	8.9	50000	8.50
SSV2[100]		train/val/test	3.7	10000	1.71
Oops[110]	Intent Recognition	train/val	9.8	7,948	1.36
Sum	-	-	1,972 hours	585,643	100.00

Table 18. Data composition of Tarsier2-Recap-585K. The “Split” column lists the original dataset partitioning, and we use bold to mark the parts which we sampled the video clips from to conduct recaptioning.

Appendix G. Qualitative Comparison of the SFT Process

Figure 9 illustrates a qualitative comparison of our model at different stages, where we mark the differences in the prediction results of different models. From these differences, it can be seen that introducing temporal localization information in the first SFT stage significantly reduces the problem of hallucination in the model. However, the introduction of temporal localization information may also result in certain events being subdivided into finer actions. To address this issue, the second stage of training further improved the accuracy of the model description and optimized the output style.



Reference: A person stands in the swimming pool, with both hands braced against the edge of the pool. With a forceful leap, they step onto the ground and finally stand up and walk forward out of the frame.

Pretrain: A woman wearing a black swim cap and goggles is seen in a swimming pool. She is holding onto the edge of the pool with both hands. The pool has clear blue water with lane dividers visible in the background. The poolside is lined with a beige-colored deck. The woman then releases her grip on the pool edge, submerges herself underwater, and swims forward. After swimming, she emerges from the water, places her hands on the pool edge, and begins to climb out of the pool. She steps onto the deck, leaving the pool area.

SFT-1: <frame: 1> A woman is seen in a swimming pool, holding onto the edge with both hands. <frame: 2-5> She begins to submerge herself into the water, still holding the edge. As she goes underwater, her hands remain on the edge for support. <frame: 6> She then starts to push herself up from the water, preparing to stand. <frame: 7-8> The woman uses her hands to push herself up onto the pool edge, with water splashing around her. <frame: 9-10> She continues to lift her body out of the water, placing her feet on the edge. <frame: 11-12> The woman stands up on the pool edge, with water dripping from her body. <frame: 13-16> She then walks away from the pool, leaving the edge behind.

SFT-2: A woman in a black swimsuit and swim cap is in a swimming pool, holding onto the edge with both hands. She submerges her head underwater while still holding the edge. She then pushes herself up from the water, using her hands for support. The woman stands up in the shallow water, water dripping from her body. She steps out of the pool, walking along the edge. The pool area is empty after she leaves.

Figure 9. Qualitative comparison of our model at different stages.

Appendix H. DREAM-1K cases

Figures 10-14 display the detailed video descriptions generated by Tarsier2-7B and other models (GPT-4o, Gemini-1.5-Pro and LLaVA-Video-7B-Qwen2) for different video categories in DREAM-1K. Click the play button on the first frames to view the raw video. The correct descriptions of key objects/actions/events are marked in green, and the incorrect descriptions are marked in red.

doi.org/10.32388/X26ILU

GPT-4o: A person is seen in a dimly lit room filled with glass bottles and various objects. The person lights a match and ignites a substance, causing a brief flame. They then use a syringe to mix a green glowing liquid in a bowl. The person retrieves tools from a mirrored tray. The video focuses on the person's precise actions of lighting, mixing, and handling objects in the room.

Gemini-1.5-Pro: A young woman wearing a black dress stands at a bar, listening to someone out of frame. She then uses a long metal tool to scoop burning rose petals from one small bowl to another, creating a shower of sparks. As the sparks fall into the second bowl, they transform into a bubbling, bright green liquid. The woman uses a syringe to draw some of the liquid out of the bowl before placing the syringe onto a tray beside a pair of metal tongs.

LLaVA-Video-7B-Qwen2: The video opens with a shot of a person wearing a black sleeveless top and multiple necklaces, standing in front of a bar counter. The background is dimly lit, with shelves filled with various bottles and glasses. The scene transitions to a close-up of a hand holding a lighter, igniting a flame that is used to light a bowl containing red rose petals. The next frame shows the same person, now holding a small object that emits a spark, possibly a lighter or a matchstick. Following this, there is a close-up of a bowl with a green liquid that has a glowing effect, being stirred with a metal tool. The final frames show a hand placing a metallic object, which appears to be a syringe, on a reflective surface, possibly a mirror.

Tarsier2-7B: A woman with braided hair and wearing a black outfit with multiple necklaces and bracelets looks at a person in front of her in a room filled with various bottles and glassware. She lights a match and uses it to ignite rose petals in a small bowl, creating a burst of sparks. She then uses a syringe to draw a green liquid from a bowl and places the syringe on a reflective tray along with some metal tools.

Figure 10. Qualitative comparative analysis of various Video-MLLMs on Dream-1K dataset (Live-action Subset).

GPT-4o: The video clip features two children on a rooftop. They are initially seen pulling on a string that is attached to a colorful kite. The kite is then launched into the sky, and the children look up as it ascends. The background shows a cityscape with buildings. The children continue to hold the kite string and watch the kite as it flies higher into the sky.

Gemini-1.5-Pro: A boy and a girl are standing on a rooftop preparing to launch a kite. With a powerful throw, the kite soars high into the air, resembling a team of three dogs pulling a sled. The children watch with delight as their kite dances in the wind. However, their joy turns to concern as they notice the kite string has snapped, leaving them staring at the broken connection with worry.

LLaVA-Video-7B-Qwen2: The video opens with two children standing on a rooftop, one holding a kite string and the other holding a kite. The child with the string is wearing a striped shirt and shorts, while the other child is in a green dress with star patterns. They are both looking up at the sky, which is a clear blue with a few clouds. The kite is colorful, predominantly pink with yellow accents and has a long tail. As the video progresses, the children appear to be struggling to control the kite, which is flying erratically. The background shows a cityscape with buildings and trees below them. The video ends with the children looking at each other, seemingly discussing their next move.

Tarsier2-7B: Two children run to the edge of a rooftop and throw a colorful kite into the sky. The kite starts to fly, but it soon begins to descend. The children, holding the kite string, look surprised and lower their hands. They then look at the kite string and start talking. Suddenly, the kite returns, and the children react to its sudden return.

Figure 11. Qualitative comparative analysis of various Video-MLLMs on Dream-1K dataset (Animation Subset).

GPT-4o: A person in white clothing approaches a bar counter and places two cocktails on it. The cocktails are in clear, patterned glasses, one with a slice of orange and the other with a slice of lime. Two other individuals, one with a visible wristwatch and the other with painted nails, reach for the cocktails. The person with painted nails picks up the cocktail with the lime, gestures with it, and then places it back down on the counter. Both individuals hold the cocktails in their hands momentarily before setting them back on the bar counter. The background consists of a variety of bottles and dim lighting, typical of a bar setting.

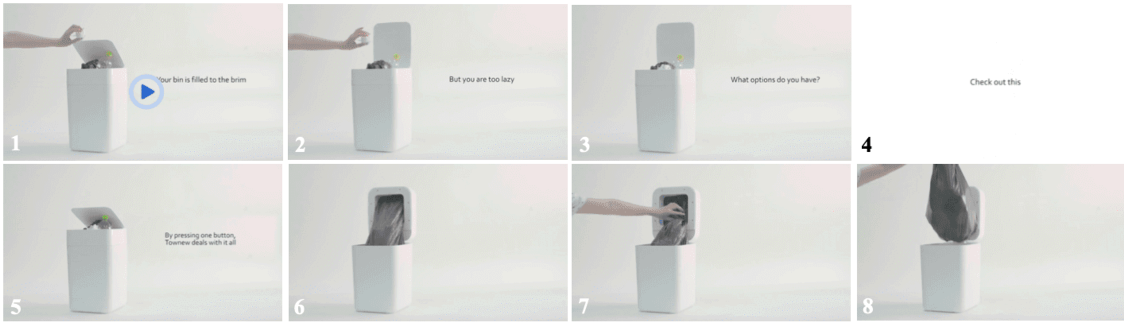
Gemini-1.5-Pro: A blurry shot of a brightly lit bar with numerous bottles and glasses provides a lively backdrop to the scene. A bartender, dressed in white, leans into the frame and places two colorful cocktails on the counter. The cocktails, a pale yellow drink garnished with an orange slice and a vibrant blue drink with a lime wedge, are each adorned with two black straws. A man and woman, seated at the bar, reach out, the woman taking the blue drink and the man selecting the yellow. They bring their glasses together in a toast, the clinking sound barely audible above the implied music and chatter of the bar. The hands retract, leaving the cocktails standing on the bar as the focus shifts entirely to the vibrant ambiance of the setting.

LLaVA-Video-7B-Qwen2: The video opens with a close-up of a bar counter, where the focus is on the gleaming surface and the blurred background filled with colorful bottles. The lighting is dim, creating a moody atmosphere typical of a bar setting. A bartender, dressed in a white shirt, appears holding two glasses filled with drinks—one with ice and an orange slice garnish, the other with a dark liquid and a lime wedge. The bartender places the drinks on the counter, and a customer's hand reaches out to take one of the drinks. The scene transitions to a similar close-up of the bar counter, where the same two drinks are now being held by the customer's hands. The customer takes a sip from the drink with the orange slice, and the bartender is seen in the background, partially obscured by the drinks. The final scene shows the bar counter again, this time with the two drinks placed side by side. The customer's hand is no longer visible, but the drinks remain untouched, emphasizing their presentation.

Tarsier2-7B: The bar counter is shown with a blurred background of various bottles and colorful lights. A bartender approaches the counter, holding two drinks. The bartender places the two drinks on the counter, one with an orange slice and the other with a lime slice, both with black straws. Two hands, one from the left and one from the right, reach for the drinks. The person on the left picks up the drink with the orange slice, and the person on the right picks up the drink with the lime slice. The counter is left empty momentarily. The person on the right places the drink with the lime slice back on the counter, followed by the person on the left placing the drink with the orange slice back on the counter. Both drinks are now back on the counter, ready to be enjoyed.

Figure 12. Qualitative comparative analysis of various Video-MLLMs on Dream-1K dataset (Stock Subset).

doi.org/10.32388/X26ILU



GPT-4o: A person is seen holding a small item over a filled trash bin. The bin is shown with various items including bottles. Text on screen indicates the bin is filled, and the person appears to be contemplating options. The person then presses a button on the bin. The trash bin begins to compact the trash inside and seals the trash bag automatically. The sealed trash bag is then lifted out from the bin. The video demonstrates the functionality of an automatic trash bin that can seal and lift the trash bag with the press of a button.

Gemini-1.5-Pro: The video starts with a hand putting trash into a bin that's already full. The text on the screen highlights the dilemma of a full trash bin. Then, it introduces a solution: a smart trash bin called "Townew". By pressing a single button, the Townew bin automatically seals the full trash bag and replaces it with a new one. A person's hand is shown pulling out the sealed bag from the bin.

LLaVA-Video-7B-Qwen2: The video begins with a close-up of a white trash bin filled to the brim with various types of waste. A hand is seen holding a small white object, presumably a button or sensor, above the bin's lid. The text on the screen reads "Your bin is filled to the brim." The scene transitions to the same bin with the text changing to "But you are lazy." The next frame shows the bin without any text, followed by the text "What options do you have?" The video then cuts to a plain white background with the text "Check out this" in the center. The final frame returns to the trash bin, now with a new feature: a blue-lit compartment at the top. The text on the screen reads "By pressing one button, Townew deals with it all." The video continues with a close-up of the same white trash bin, now with the blue-lit compartment visible. The bin is shown from a side angle, and the compartment appears to be in operation, as indicated by the blue light. The background remains plain white, emphasizing the bin and its features.

Tarsier2-7B: A hand holding a small object approaches an overflowing trash bin, with the text 'Your bin is filled to the brim' appearing on the screen. The hand drops the object into the bin, and the text changes to 'But you are too lazy.' The hand retracts, and the text 'What options do you have?' appears. The screen briefly shows 'Check out this' before transitioning to 'By pressing one button, Tounew deals with it all.' The hand presses a button on the bin, causing the lid to close and the trash to be compressed. The bin's mechanism retracts a black garbage bag, and the text disappears. A hand reaches in and pulls out the full garbage bag from the bin.

Figure 13. Qualitative comparative analysis of various Video-MLLMs on Dream-1K dataset (Youtube Subset).

doi.org/10.32388/X26ILU

GPT-4o: A person is involved in processing and crafting shells. Initially, shells are collected in a basket and washed thoroughly in water. The person then uses a foot-powered saw to cut the shells into strips. The saw is operated by moving a foot pedal up and down, which in turn moves the blade. The cut shell strips are then sorted and further refined by using a smaller tool to make precise cuts. The person carefully holds the shells and ensures they are properly cut and shaped.

Gemini-1.5-Pro: A man is cleaning an iridescent abalone shell. He then uses a foot-powered saw to cut another abalone shell into thin strips. He secures the shell to a wooden frame with twine and carefully saws through it, creating a pile of shimmering strips.

LLaVA-Video-7B-Qwen2: The video begins with a close-up of a basket filled with various seashells, some of which are open to reveal their interiors. The shells have a glossy appearance and come in different colors, including shades of green, brown, and white. The next scene shows hands holding the basket, which is placed on a surface that appears to be wet, possibly indicating it was recently submerged in water. Following this, there is a shot of water being poured over a surface, creating ripples and bubbles, suggesting a natural setting like a stream or river. The subsequent frames show hands handling a large, rough-textured shell, possibly preparing it for further processing. The hands are seen using a tool, likely a knife, to cut or scrape the shell. The final scenes depict hands arranging thin, elongated pieces of shell on a flat surface, possibly for drying or further processing.

Tarsier2-7B: A basket filled with several shells is placed in water, and a pair of hands lifts the basket out of the water. The basket, now containing various shells, is placed on a surface. A person, wearing dark shoes, uses a foot to hold the shell in place while sawing it with a handsaw. The person continues to saw the shell into strips, which are then collected and tied together with a string. The person then uses the saw to further refine the shell strips.

Figure 14. Qualitative comparison of different Video-MLLMs on Dream-1K dataset (Shorts Subset).

Footnotes

[1] https://www.scenedetect.com/

[2] An obvious counter example is that a low-dynamic video will not be significantly affected by the down-sampling perturbation.

[3] Given a reference description (d_{ref}) and a description to be assessed (d_{pred}), AutoDQ scorer outputs the recall score (DQ_R: the ratio of events in d_{ref} that are entailed by d_{pred}) and the precision score (DQ_P: the ratio of events in d_{ref} that are entailed by d_{pred}).

[4] For consistency across all checkpoints, we evaluate the Qwen2-VL-7B model using the same frame sampling strategy applied to other checkpoints. This may differ from the official sampling strategy in some benchmarks. For instance, the official setting of Video-MME uses 768 frames, while we sample 128 frames.

[5] Tarsier2-Recap-585K is available on HuggingFace.

doi.org/10.32388/X26ILU

References

1. [^] *Chen Z, Wang W, Tian H, Ye S, Gao Z, Cui E, Tong W, Hu K, Luo J, Ma Z, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821. 2024.*
2. [a, b] *Li KC, He Y, Wang Y, Li Y, Wang W, Luo P, Wang Y, Wang L, Qiao Y (2023). "Videochat: Chat-centric video understanding". arXiv preprint arXiv:2305.06355. Available from: https://arxiv.org/abs/2305.06355.*
3. [a, b, c, d, e] *Lin B, Zhu B, Ye Y, Ning M, Jin P, Yuan L (2023). "Video-llava: Learning united visual representation by alignment before projection". arXiv preprint arXiv:2311.10122. Available from: https://arxiv.org/abs/2311.10122.*
4. [a, b, c, d] *Ji Lin, Yin H, Ping W, Molchanov P, Shoeybi M, Han S. "Vila: On pre-training for visual language models." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 26689-26699.*
5. [a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s] *Wang J, Yuan L, Zhang Y, Sun H (2024). "Tarsier: Recipes for training and evaluating large video description models."*
6. [a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y] *Wang P, Bai S, Tan S, Wang S, Fan Z, Bai J, Chen K, Liu X, Wang J, Ge W, Fan Y, Dang K, Du M, Ren X, Men R, Liu D, Zhou C, Zhou J, Lin J (2024). "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution". arXiv preprint arXiv:2409.12191.*
7. [a, b, c, d, e, f, g, h] *Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, Ostrow AJ, Welihinda A, Hayes A, Radford A, et al. GPT-4o system card. arXiv preprint arXiv:2410.21276. 2024. Available from: https://arxiv.org/abs/2410.21276.*
8. [a, b, c, d, e, f, g, h, i, j, k, l] *Gemini Team, Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, Tanzer G, Vincent D, Pan Z, Wang S, Mariooryad S, Ding Y, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.*
9. [a, b, c, d, e, f] *Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, Jiashi Feng (2024). "Pllava: Parameter-free llava extension from images to videos for video dense captioning". arXiv preprint arXiv:2404.16994. Available from: https://arxiv.org/abs/2404.16994.*
10. [a, b, c, d, e] *Cheng Z, Leng S, Zhang H, Xin Y, Li X, Chen G, Zhu Y, Zhang W, Luo Z, Zhao D, Bing L (2024). "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms".*
11. [a, b, c, d, e, f] *Li B, Zhang H, Zhang K, Guo D, Zhang Y, Zhang R, Li F, Liu Z, Li C (2024). "Llava-next: What else influences visual instruction tuning beyond data."*
12. [a, b, c, d, e, f, g, h, i, j, k, l, m, n] *Chen Z, Wang W, Cao Y, Liu Y, Gao Z, Cui E, Zhu J, Ye S, Tian H, Liu Z, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271. 2024.*

13. a, b, c, d, e, f, g, h, i, j, k *Li B, Zhang Y, Guo D, Zhang R, Li F, Zhang H, Zhang K, Zhang P, Li Y, Liu Z, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326. 2024.*
14. a, b, c, d, e *Cores D, Dorkenwald M, Mucientes M, Snoek CGM, Asano YM (2024). "Tvbench: Redesigning video-language evaluation". arXiv preprint arXiv:2410.07752. arXiv:2410.07752.*
15. a, b, c, d *Li K, Wang Y, He Y, Li Y, Wang Y, Liu Y, Wang Z, Xu J, Chen G, Luo P, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 22195-22206.*
16. a, b, c, d, e, f, g, h, i, j, k, l *Liu Y, Ma Z, Qi Z, Wu Y, Chen CW, Shan Y. E.t. bench: Towards open-ended event-level video-language understanding. In: Neural Information Processing Systems (NeurIPS); 2024.*
17. a, b, c, d *Wang Y, Wang Y, Zhao D, Xie C, Zheng Z (2024). "Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models". arXiv preprint arXiv:2406.16338. arXiv:2406.16338.*
18. a, b, c *Zhou J, Shu Y, Zhao B, Wu B, Xiao S, Yang X, Xiong Y, Zhang B, Huang T, Liu Z (2024). "Mlvu: A comprehensive benchmark for multi-task long video understanding". arXiv preprint arXiv:2406.04264. Available from: https://arxiv.org/abs/2406.04264.*
19. ^ *Mangalam K, Akshulakov R, Malik J (2023). "Egoschema: A diagnostic benchmark for very long-form video language understanding". Advances in Neural Information Processing Systems. **36**: 46212–46244.*
20. a, b, c *Patraucean V, Smaira L, Gupta A, Recasens A, Markeeva L, Banarse D, Koppula S, Malinowski M, Yang Y, Doersch C, et al. Perception test: A diagnostic benchmark for multimodal video models. Advances in Neural Information Processing Systems. **36**, 2024.*
21. ^ *Chen X, Lin Y, Zhang Y, Huang W (2023). "Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering". arXiv preprint arXiv:2311.14906. Available from: https://arxiv.org/abs/2311.14906.*
22. a, b *Zhang P, Dong X, Zang Y, Cao Y, Qian R, Chen L, Guo Q, Duan H, Wang B, Ouyang L, Zhang S, Zhang W, Li Y, Gao Y, Sun P, Zhang X, Li W, Li J, Wang W, Yan H, He C, Zhang X, Chen K, Dai J, Qiao Y, Lin D, Wang J (2024). "Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output."*
23. a, b, c, d *Shangguan Z, Li C, Ding Y, Zheng Y, Zhao Y, Fitzgerald T, Cohan A (2024). "Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models". arXiv preprint arXiv:2410.23266. arXiv:2410.23266.*
24. a, b, c, d *Zhang J, Cai M, Lee YJ (2024). "Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos". arXiv preprint arXiv:2410.02763. Available from: https://arxiv.org/abs/2410.02763.*
25. a, b, c *Liu Y, Li S, Liu Y, Wang Y, Ren S, Li L, Chen S, Sun X, Hou L (2024). "Tempcompass: Do video llms really understand videos?" arXiv preprint arXiv:2403.00476. Available from: https://arxiv.org/abs/2403.00476.*

26. [a, b] *Fu C, Dai Y, Luo Y, Li L, Ren S, Zhang R, Wang Z, Zhou C, Shen Y, Zhang M, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075. 2024.*

27. [a, b] *Liu Z, Zhu L, Shi B, Zhang Z, Lou Y, Yang S, Xi H, Cao S, Gu Y, Li D, Li X, Fang Y, Chen Y, Hsieh CY, Huang DA, Cheng AC, Nath V, Hu J, Liu S, Krishna R, Xu D, Wang X, Molchanov P, Kautz J, Yin H, Han S, Lu Y. Nvila: Efficien t frontier visual language models, 2024.*

28. [a, b] *Wu H, Li D, Chen B, Li J (2024). "Longvideobench: A benchmark for long-context interleaved video-languag e understanding". arXiv preprint arXiv:2407.15754. Available from: https://arxiv.org/abs/2407.15754.*

29. [a, b, c, d, e] *Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia (2024). "Apollo: An explora tion of video understanding in large multimodal models."*

30. [a, b] *Mu Cai, Tan R, Zhang J, Zou B, Zhang K, Yao F, Zhu F, Gu J, Zhong Y, Shang Y, et al. Temporalbench: Benchm arking fine-grained temporal understanding for multimodal video models. arXiv preprint arXiv:2410.10818, 20 24.*

31. [a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q] *Zhang Y, Wu J, Li W, Li B, Ma Z, Liu Z, Li C (2024). "Video instruction tuning with synthetic data". arXiv preprint arXiv:2410.02713. Available from: https://arxiv.org/abs/2410.02713.*

32. [a, b] *Fang X, Mao K, Duan H, Zhao X, Li Y, Lin D, Chen K (2024). "Mmbench-video: A long-form multi-shot benc hmark for holistic video understanding". arXiv preprint arXiv:2406.14515. arXiv:2406.14515.*

33. [a, b] *Yao Y, Yu T, Zhang A, Wang C, Cui J, Zhu H, Cai T, Li H, Zhao W, He Z, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800. 2024. Available from: https://arxiv.org/abs/2408.01800.*

34. [a, b, c, d] *Zhang J, Jiao Y, Chen S, Chen J, Jiang YG (2024). "Eventhallusion: Diagnosing event hallucinations in vid eo llms". arXiv preprint arXiv:2409.16597. Available from: https://arxiv.org/abs/2409.16597.*

35. [a, b, c] *Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C (2024). "Direct preference optimization: Y our language model is secretly a reward model". Advances in Neural Information Processing Systems. **36**.*

36. [a, b] *Sermanet P, Ding T, Zhao J, Xia F, Dwibedi D, Gopalakrishnan K, Chan C, Dulac-Arnold G, Maddineni S, Josh i NJ, Florence P, Han W, Baruch R, Lu Y, Mirchandani S, Xu P, Sanketi P, Hausman K, Shafran I, Ichter B, Cao Y (2023). "Robovqa: Multimodal long-horizon reasoning for robotics". arXiv preprint arXiv:2311.00899. Availabl e from: https://arxiv.org/abs/2311.00899.*

37. [a, b] *Maaz M, Rasheed H, Khan S, Khan FS (2023). "Video-chatgpt: Towards detailed video understanding via lar ge vision and language models". arXiv preprint arXiv:2306.05424. arXiv:2306.05424.*

38. [^] *Luo R, Zhao Z, Yang M, Dong J, Qiu M, Lu P, Wang T, Wei Z (2023). "Valley: Video assistant with large languag e model enhanced ability". arXiv preprint arXiv:2306.07207. Available from: https://arxiv.org/abs/2306.07207.*

39. ^Zhang H, Li X, Bing L (2023). "Video-llama: An instruction-tuned audio-visual language model for video understanding". arXiv preprint arXiv:2306.02858. Available from: https://arxiv.org/abs/2306.02858.

40. ^Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, Mohamed Elhoseiny (2024). "Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens". arXiv preprint arXiv:2404.03413.

41. ^Wang H, Wang Y, Ye Y, Nie Y, Huang C (2024). "Elysium: Exploring object-level perception in videos via mllm". arXiv preprint arXiv:2403.16558. arXiv:2403.16558.

42. a, b, cLi D, Liu Y, Wu H, Wang Y, Shen Z, Qu B, Niu X, Wang G, Chen B, Li J (2024). "Aria: An open multimodal native mixture-of-experts model". arXiv preprint arXiv:2410.05993. Available from: https://arxiv.org/abs/2410.05993.

43. a, bDubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783. 2024.

44. ^Agrawal P, Antoniak S, Bou Hanna E, Bout B, Chaplot D, Chudnovsky J, Costa D, De Monicault B, Garg S, Gervet T, et al. Pixtral 12b. arXiv preprint arXiv:2410.07073. 2024. Available from: https://arxiv.org/abs/2410.07073.

45. a, bLiu Z, Dong Y, Liu Z, Hu W, Lu J, Rao Y (2024). "Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution". arXiv preprint arXiv:2409.12961. arXiv:2409.12961.

46. ^Xu M, Gao M, Gan Z, Chen HY, Lai Z, Gang H, Kang K, Dehghan A (2024). "Slowfast-llava: A strong training-free baseline for video large language models". arXiv preprint arXiv:2407.15841. arXiv:2407.15841.

47. ^Li Y, Wang C, Jia J. "Llama-vid: An image is worth 2 tokens in large language models." In: European Conference on Computer Vision. Springer; 2025. p. 323-340.

48. ^Li J, Li D, Savarese S, Hoi S. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." In: International Conference on Machine Learning. PMLR; 2023. p. 19730-19742.

49. ^Liu J, Wang Y, Ma H, Wu X, Ma X, Wei X, Jiao J, Wu E, Hu J (2024). "Kangaroo: A powerful video-language model supporting long-context video input". arXiv preprint arXiv:2408.15542. Available from: https://arxiv.org/abs/2408.15542.

50. ^Bai J, Bai S, Yang S, Wang S, Tan S, Wang P, Lin J, Zhou C, Zhou J (2023). "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond". arXiv preprint arXiv:2308.12966. **1** (2): 3.

51. a, b, cMiech A, Zhukov D, Alayrac JB, Tapaswi M, Laptev I, Sivic J (2019). "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips". Proceedings of the IEEE/CVF International Conference on Computer Vision. 2630–2640.

52. a, bXue H, Hang T, Zeng Y, Sun Y, Liu B, Yang H, Fu J, Guo B (2022). "Advancing high-resolution video-language representation with large-scale video transcriptions". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 5036–5045.

 doi.org/10.32388/X26ILU

53. ^Chen TS, Siarohin A, Menapace W, Deyneka E, Chao HW, Jeon BE, Fang Y, Lee HY, Ren J, Yang MH, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 13320-13331.
54. ^Wang Y, He Y, Li Y, Li K, Yu J, Ma X, Chen X, Wang Y, Luo P, Liu Z, Wang Y, Wang L, Qiao Y (2023). "Internvid: A large-scale video-text dataset for multimodal understanding and generation". arXiv preprint arXiv:2307.06942. Available from: https://arxiv.org/abs/2307.06942.
55. ^Xu H, Ye Q, Yan M, Shi Y, Ye J, Xu Y, Li C, Bi B, Qian Q, Wang W, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In: International Conference on Machine Learning. PMLR; 2023. p. 38728-38748.
56. ^Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, Jiahui Yu (2022). "Videococa: Video-text modeling with zero-shot transfer from contrastive captioners". arXiv preprint arXiv:2212.04979.
57. ^Chen S, Li H, Wang Q, Zhao Z, Sun M, Zhu X, Liu J (2023). "Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset". Advances in Neural Information Processing Systems. 36: 72842–72866.
58. ^Chen D, Dolan W. Collecting highly parallel data for paraphrase evaluation. In: Lin D, Matsumoto Y, Mihalcea R, editors. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics; 2011. p. 190-200.
59. ^Xu J, Mei T, Yao T, Rui Y (2016). "MSR-VTT: A large video description dataset for bridging video and language". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5288–5296.
60. a, bWang X, Wu J, Chen J, Li L, Wang YF, Wang WY (2019). "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research". Proceedings of the IEEE/CVF International Conference on Computer Vision. 4581–4591.
61. ^Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions, 2024.
62. a, b, c, d, eApplin S, Adesso G, Ashfaq R, Bai M, Brammer M, Fecht E, Goodman A, Grossman S, Groh M, Kirk HR, et al. Gpt-4v (ision) system card. 2023.
63. a, bAchiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al. GPT-4 technical report. arXiv preprint arXiv:2303.08774. 2023. Available from: https://arxiv.org/abs/2303.08774.
64. ^Tang C, Li Y, Yang Y, Zhuang J, Sun G, Li W, Ma Z, Zhang C (2024). "Enhancing multimodal llm for detailed and accurate video captioning using multi-round preference optimization."
65. ^Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. Minneapolis, Minnesota; 2019. p. 2.

66. ^ *Xiong T, Wang X, Guo D, Ye Q, Fan H, Gu Q, Huang H, Li C (2024). "Llava-critic: Learning to evaluate multimodal models."*
67. ^ *Zhang R, Gui L, Sun Z, Feng Y, Xu K, Zhang Y, Fu D, Li C, Hauptmann A, Bisk Y, Yang Y (2024). "Direct preference optimization of video large multimodal models from language model reward."*
68. ^ *Tang C, Li Y, Yang Y, Zhuang J, Sun G, Li W, Ma Z, Zhang C (2024). "Enhancing multimodal llm for detailed and accurate video captioning using multi-round preference optimization."*
69. ^ *Huang DA, Liao S, Radhakrishnan S, Yin H, Molchanov P, Yu Z, Kautz J. "Lita: Language instructed temporal-localization assistant". In: ECCV, 2024.*
70. ^ *Guo Y, Liu J, Li M, Tang X, Chen X, Zhao B (2024). "Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding". arXiv preprint arXiv:2405.13382.*
71. ^ *Ren S, Yao L, Li S, Sun X, Hou L (2023). "Timechat: A time-sensitive multimodal large language model for long video understanding". ArXiv. abs/2312.02051.*
72. ^ *Lei J, Berg TL, Bansal M (2021). "Detecting moments and highlights in videos via natural language queries". Advances in Neural Information Processing Systems. **34**: 11846–11858.*
73. ^ *Gao J, Sun C, Yang Z, Nevatia R (2017). "Tall: Temporal activity localization via language query". Proceedings of the IEEE International Conference on Computer Vision. 5267–5275.*
74. ^ *Idrees H, Zamir AR, Jiang YG, Gorban A, Laptev I, Sukthankar R, Shah M (2017). "The THUMOS challenge on action recognition for videos 'in the wild'." Computer Vision and Image Understanding. **155**: 1–23.*
75. a, b, c *Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, Hamburger J, Jiang H, Liu M, Liu X, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:18995-19012.*
76. ^ *Gao P, Han J, Zhang R, Lin Z, Geng S, Zhou A, Zhang W, Lu P, He C, Yue X, et al. Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010. 2023. Available from: https://arxiv.org/abs/2304.15010.*
77. ^ *Le TM, Le V, Venkatesh S, Tran T (2020). "Hierarchical conditional relation networks for video question answering". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 9972–9981.*
78. ^ *Bai Z, Wang R, Chen X (2024). "Glance and focus: Memory prompting for multi-event video question answering". Advances in Neural Information Processing Systems. **36**.*
79. ^ *Liu J, Liu M, Wang Z, Lee L, Zhou K, An P, Yang S, Zhang R, Guo Y, Zhang S (2024). "Robomamba: Multimodal state space model for efficient robot reasoning and manipulation". arXiv preprint arXiv:2406.04339. Available from: https://arxiv.org/abs/2406.04339.*

80. ^Pramanick S, Song Y, Nag S, Lin KQ, Shah H, Shou MZ, Chellappa R, Zhang P (2023). "Egovlpv2: Egocentric video-language pre-training with fusion in the backbone". Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5285–5297.

81. a, bXiang A, Yang K, Dai X, Feng Z, Deng J. "Multi-label cluster discrimination for visual representation learning." In: European Conference on Computer Vision. Springer; 2025. p. 428-444.

82. ^Jia B, Lei T, Zhu SC, Huang S (2022). "Egotaskqa: Understanding human tasks in egocentric videos". The 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

83. ^Majumdar A, Ajay A, Zhang X, Putta P, Yenamandra S, Henaff M, Silwal S, Mcvay P, Maksymets O, Arnaud S, et al. Openeqa: Embodied question answering in the era of foundation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16488–16498, 2024.

84. ^Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, Zhong M, Zhang Q, Zhu X, Lu L, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 24185-24198.

85. a, bHe Y, Lin Y, Wu J, Zhang H, Zhang Y, Le R (2024). "Storyteller: Improving long video description through global audio-visual character identification". arXiv preprint arXiv:2411.07076. arXiv:2411.07076.

86. a, bBain M, Nagrani A, Varol G, Zisserman A (2021). "Frozen in time: A joint video and image encoder for end-to-end retrieval". Proceedings of the IEEE/CVF International Conference on Computer Vision. pages 1728–1738.

87. a, bRohrbach A, Torabi A, Rohrbach M, Tandon N, Pal C, Larochelle H, Courville A, Schiele B (2017). "Movie description". International Journal of Computer Vision. **123**: 94–120.

88. a, bLi Y, Song Y, Cao L, Tetreault J, Goldberg L, Jaimes A, Luo J (2016). "Tgif: A new dataset and benchmark on animated gif description". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pages 4641–4650.

89. a, bKrishna R, Hata K, Ren F, Li Fei-Fei, Niebles JC (2017). "Dense-captioning events in videos". International Conference on Computer Vision (ICCV). 2017.

90. a, bSigurdsson GA, Varol G, Wang X, Farhadi A, Laptev I, Gupta A (2016). "Hollywood in homes: Crowdsourcing data collection for activity understanding". In: Computer Vision--ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11--14, 2016, Proceedings, Part I 14. Springer; 2016. p. 510–526.

91. a, bSigurdsson GA, Gupta A, Schmid C, Farhadi A, Alahari K (2018). "Charades-ego: A large-scale dataset of paired third and first person videos". arXiv preprint arXiv:1804.09626. Available from: https://arxiv.org/abs/1804.09626.

92. ^Zhou Y, Wang Z, Fang C, Bui T, Berg TL (2018). "Visual to sound: Generating natural sound for videos in the wild". Proceedings of the IEEE conference on computer vision and pattern recognition. pages 3550–3558.

93. ^Regneri M, Rohrbach M, Wetzel D, Thater S, Schiele B, Pinkal M (2013). "Grounding action descriptions in videos". Transactions of the Association for Computational Linguistics. 1: 25–36.
94. ^Monfort M, Jin S, Liu A, Harwath D, Feris R, Glass J, Oliva A (2021). "Spoken moments: Learning joint audio-visual representations from video descriptions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 14871–14881.
95. ^Monfort M, Pan B, Ramakrishnan K, Andonian A, McNamara BA, Lascelles A, Fan Q, Gutfreund D, Feris RS, Oliva A (2021). "Multi-moments in time: Learning and interpreting models for multi-action video understanding". IEEE Transactions on Pattern Analysis and Machine Intelligence. 44 (12): 9434–9445.
96. a, bAwad G, Curtis K, Butt A, Fiscus J, Godil A, Lee Y, Delgado A, Godard E, Diduch L, Liu J, et al. An overview on the evaluated video retrieval tasks at trecvid 2022. arXiv preprint arXiv:2306.13118. 2023.
97. a, bCui E, He Y, Ma Z, Chen Z, Tian H, Wang W, Li K, Wang Y, Wang W, Zhu X, Lu L, Lu T, Wang Y, Wang L, Qiao Y, Dai J. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o. https://sharegpt4o.github.io/, 2024.
98. ^Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. "HMDB: a large video database for human motion recognition". In: 2011 International Conference on Computer Vision. IEEE; 2011. p. 2556-2563.
99. ^Tang Y, Ding D, Rao Y, Zheng Y, Zhang D, Zhao L, Lu J, Zhou J (2019). "Coin: A large-scale dataset for comprehensive instructional video analysis". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1207–1216.
100. a, b, cGoyal R, Ebrahimi Kahou S, Michalski V, Materzynska J, Westphal S, Kim H, Haenel V, Fruend I, Yianilos P, Mueller-Freitag M, et al. The "something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE international conference on computer vision. 2017:5842-5850.
101. a, b, cCarreira J, Zisserman A (2017). "Quo vadis, action recognition? a new model and the kinetics dataset". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308.
102. ^Liu Y, Wang L, Wang Y, Ma X, Qiao Y (2022). "Fineaction: A fine-grained video dataset for temporal action localization". IEEE Transactions on Image Processing. 31: 6937–6950.
103. ^Miech A, Alayrac JB, Laptev I, Sivic J, Zisserman A (2020). "Rareact: A video dataset of unusual interactions". arXiv preprint arXiv:2008.01018. Available from: https://arxiv.org/abs/2008.01018.
104. ^Materzynska J, Berger G, Bax I, Memisevic R (2019). "The jester dataset: A large-scale video dataset of human gestures". Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0.
105. ^Yi K, Gan C, Li Y, Kohli P, Wu J, Torralba A, Tenenbaum JB (2019). "Clevrer: Collision events for video representation and reasoning". arXiv preprint arXiv:1910.01442. arXiv:1910.01442.
106. ^Jang Y, Song Y, Yu Y, Kim Y, Kim G (2017). "Tgif-qa: Toward spatio-temporal reasoning in visual question answering". Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2758–2766.

107. ^Chenyou Fan. "Egovqa-an egocentric video question answering benchmark dataset". In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019: 0–0.

108. ^Hendricks LA, Wang O, Shechtman E, Sivic J, Darrell T, Russell B (2017). "Localizing moments in video with natural language". Proceedings of the IEEE international conference on computer vision. pp. 5803–5812.

109. ^Gu C, Sun C, Ross DA, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 6047-6056.

110. a, b, cEpstein D, Chen B, Vondrick C (2020). "Oops! predicting unintentional action in video". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 919–929.

111. ^Huang TH, Ferraro F, Mostafazadeh N, Misra I, Agrawal A, Devlin J, Girshick R, He X, Kohli P, Batra D, et al. Visual storytelling. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies. 2016. p. 1233-1239.

112. ^Liu Z, Chu T, Zang Y, Wei X, Dong X, Zhang P, Liang Z, Xiong Y, Qiao Y, Lin D, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. arXiv preprint arXiv:2406.11833. 2024.

113. ^Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, Dahua Lin (2023). "Sharegpt4v: Improving large multi-modal models with better captions". arXiv preprint arXiv:2311.12793. Available from: https://arxiv.org/abs/2311.12793.

114. ^Liu H, Li C, Li Y, Lee YJ (2023). "Improved baselines with visual instruction tuning". arXiv preprint arXiv:2310.03744. arXiv:2310.03744.

115. ^Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In: Computer Vision--ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer; 2014. p. 740-755.

116. ^Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S (2015). "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models". Proceedings of the IEEE International Conference on Computer Vision. pp. 2641–2649.

117. ^Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision. 123:32–73, 2017.

118. ^Ordonez V, Kulkarni G, Berg T (2011). "Im2text: Describing images using 1 million captioned photographs". Advances in Neural Information Processing Systems. 24.

119. ^Shi B, Yao C, Liao M, Yang M, Xu P, Cui L, Belongie S, Lu S, Bai X (2017). "ICDAR2017 competition on reading Chinese text in the wild (RCTW-17)". In: 2017 14th IAPR International Conference on Document Analysis and R

ecognition (ICDAR), volume 1, pages 1429–1434. IEEE.

120. ^Sun Y, Ni Z, Chng CK, Liu Y, Luo C, Ng CC, Han J, Ding E, Liu J, Karatzas D, et al. ICDAR 2019 competition on large-scale street view text with partial labeling-RRC-LSVT. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE; 2019. p. 1557-1562.
121. ^Zhang R, Zhou Y, Jiang Q, Song Q, Li N, Zhou K, Wang L, Wang D, Liao M, Yang M, et al. ICDAR 2019 robust reading challenge on reading Chinese text on signboard. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE; 2019. p. 1577-1581.
122. ^Bhagavatula C, Le Bras R, Malaviya C, Sakaguchi K, Holtzman A, Rashkin H, Downey D, Yih SW, Choi Y (2019). "Abductive commonsense reasoning". arXiv preprint arXiv:1908.05739. Available from: https://arxiv.org/abs/1908.05739.
123. ^Veit A, Matera T, Neumann L, Matas J, Belongie S (2016). "Coco-text: Dataset and benchmark for text detection and recognition in natural images". arXiv preprint arXiv:1601.07140. Available from: https://arxiv.org/abs/1601.07140.
124. ^Park S, Shin S, Lee B, Lee J, Surh J, Seo M, Lee H (2019). "Cord: a consolidated receipt dataset for post-ocr parsing". Workshop on Document Intelligence at NeurIPS 2019. 2019.
125. ^Long S, Qin S, Panteleev D, Bissacco A, Fujii Y, Raptis M. Towards end-to-end unified scene text detection and layout analysis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
126. ^Yao C, Bai X, Liu W, Ma Y, Tu Z (2012). "Detecting texts of arbitrary orientations in natural images". In: 2012 IEEE conference on computer vision and pattern recognition. IEEE; 2012. p. 1083–1090.
127. ^Lucas SM, Panaretos A, Sosa L, Tang A, Wong S, Young R, Ashida K, Nagai H, Okamoto M, Yamamoto H, et al. ICDAR 2003 robust reading competitions: entries, results, and future directions. International Journal of Document Analysis and Recognition (IJDAR). 7: 105–122, 2005.
128. a, bKim G, Hong T, Yim M, Nam J, Park J, Yim J, Hwang W, Yun S, Han D, Park S (2022). "Ocr-free document understanding transformer". European Conference on Computer Vision (ECCV). 2022.
129. ^W Lian, B Goodson, E Pentland, et al. Openorca: An open dataset of gpt augmented flan reasoning traces, 2023.
130. ^Chiang WL, Li Z, Lin Z, Sheng Y, Wu Z, Zhang H, Zheng L, Zhuang S, Zhuang Y, Gonzalez JE, Stoica I, Xing EP (2023). "Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality". March 2023.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.