Scaling RL to Long Videos

Yukang Chen[1*] Wei Huang[1,3*] Baifeng Shi[1,4†] Qinghao Hu[2†] Hanrong Ye[1†]
Ligeng Zhu[1] Zhijian Liu[1] Pavlo Molchanov[1] Jan Kautz[1] Xiaojuan Qi[3]
Sifei Liu[1] Hongxu Yin[1] Yao Lu[1] Song Han[1,2]

[1]NVIDIA [2]MIT [3]HKU [4]UC Berkeley

Abstract

We introduce a full-stack framework that scales up reasoning in vision-language models (VLMs) to long videos, leveraging reinforcement learning. We address the unique challenges of long video reasoning by integrating three critical components: (1) a large-scale dataset, LongVideo-Reason, comprising 104K long video QA pairs with high-quality reasoning annotations across diverse domains such as sports, games, and vlogs; (2) a two-stage training pipeline that extends VLMs with chain-of-thought supervised fine-tuning (CoT-SFT) and reinforcement learning (RL); and (3) a training infrastructure for long video RL, named Multi-modal Reinforcement Sequence Parallelism (MR-SP), which incorporates sequence parallelism and a vLLM-based engine tailored for long video, using cached video embeddings for efficient rollout and prefilling. In our experiments, LongVILA-R1-7B achieves strong performance on video benchmarks, reaching 65.1% and 71.1% accuracy on VideoMME without and with subtitles, respectively, and consistently outperforming LongVILA-7B across multiple benchmarks. Moreover, LongVILA-R1-7B supports processing up to 8,192 video frames per video, and configurable FPS settings. Notably, our MR-SP system achieves up to 2.1× speedup on long video RL training. In addition, we release our training system for public availability that supports RL training on various modalities (video, text, and audio), various models (VILA and Qwen series), and even image and video generation models. On a single A100 node (8 GPUs), it supports RL training on hour-long videos (*e.g.*, 3,600 frames). Code and models are available at https://github.com/NVlabs/Long-RL and https://huggingface.co/Efficient-Large-Model/LongVILA-R1-7B.

1 Introduction

Understanding long videos requires more than simple recognition—it demands reasoning from temporal, spatial, goal-oriented, and narrative perspectives [51]. As illustrated in Figure 1, answering high-level questions often hinges on a model's ability to integrate clues distributed across time, infer hidden goals or strategies, track entities spatially, and comprehend the evolving plot. For instance, predicting the winner of a football penalty shootout involves assessing emotional cues and tactical behavior (temporal and goal reasoning), while determining the final location of a hidden ball requires precise spatial tracking. Likewise, evaluating a poker player's decision demands interpreting implicit strategies beyond surface actions (goal reasoning) and understanding a character's development or match trajectory reflects the need for plot reasoning. These examples underscore that reasoning is indispensable for long video understanding that goes beyond recognition alone. Despite the clear importance of reasoning in long video understanding, enabling such capabilities in long video VLMs poses significant challenges [8, 43, 47, 12, 37]. First, the collection of high quality long video reasoning datasets is inherently difficult. Unlike domains such as math or code reasoning, where

*Equal contribution †Core contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Figure 1: Examples of LongVILA-R1. The illustration demonstrates sample tasks and their reasoning. From left to right, the examples include predicting the results of a football match, decision-making reasoning in Texas Hold'em Poker, and trajectory for spatial dynamics of objects. Notably, the spatial tracking video involves a relatively complex dynamic moving, for which the model fails to achieve accurate reasoning until the number of input video frames increases to 128.

structured supervision and benchmarks are readily available [35, 24], long video reasoning requires annotating complex temporal dynamics, goals, spatial relations, and narrative elements—often across minutes or hours of footage [14]. This process is labor-intensive and subjective, making large-scale dataset construction slow and costly. Second, the RL training framework for long videos is challenging. Reinforcement learning, a common strategy for aligning models with complex reasoning objectives, is computationally expensive and sample-inefficient [38]. When applied to long videos, RL becomes even more burdensome due to the extended video frames, requiring more memory and longer rollout runtime. These challenges jointly hinder the development of effective long video VLMs with strong reasoning capabilities.

In this work, we introduce LongVILA-R1, a comprehensive framework exploring the reasoning capabilities for long video understandings. Firstly, we strategically construct a high quality dataset with CoT annotations for long video reasoning, named LongVideo-Reason. Leveraging a powerful VLM (NVILA-8B) [27] and a leading open-source reasoning LLM, we develop a dataset comprising 104K high quality *Question-Reasoning-Answer* pairs for long videos. We use 36K high quality samples for Long-CoT-SFT to initialize the model's reasoning and instruction-following abilities, and 68K samples with an additional 102K video data [53, 46, 31, 18, 44] for reinforcement learning. This two-stage training combines high quality reasoning annotations with reinforcement learning, enabling LongVILA-R1 to achieve superior and generalized video reasoning. We also manually curate a balanced set of 1K long video samples to build a new benchmark, LongVideo-Reason-eval, that evaluates performance from four perspectives: *Temporal*, *Goal and Purpose*, *Spatial*, and *Plot and Narrative*, for a comprehensive assessment.

Figure 2: Training efficiency comparison with MR-SP (SP degree=4) on Qwen2.5-VL-7B and LongVILA-R1-7B and a single node 8A100 GPUs. It achieves 2.1speed-up and avoids GPU OOM issue on long frames.

Figure 3: Data Distribution of LongVideo-Reason and total data in the LongVILA-R1 training framework. LongVideo-Reason comprises a total of 18K videos and 104K QAs with reasoning annotations. Additionally, we include 102K QAs from existing works [53, 46, 31, 18, 44].

Secondly, we propose a training framework for VLMs to advance long video reasoning. As illustrated in Figure 5, this framework incorporates two stages, i.e., Stage-1: Long CoT-SFT, and Stage-2: RL for long video reasoning. To address the unique challenges of long video RL, including massive visual embeddings, heavy rollouts, and long-context LLM prefilling, we develop an efficient and scalable solution, referred to as Multi-modal Reinforcement Sequence Parallelism (MR-SP). It incorporates a vLLM engine [19] tailored for LongVILA and a caching scheme for video embeddings. The MR-SP system alleviates the problem of intensive memory and facilitates RL training of long video VLMs. As shown in Figure 2, our MR-SP system achieves up to 2.1 speedup on 512-frame video RL training on 7B models and enables longer training frames without out-of-memory (OOM).

In our experiments, LongVILA-R1-7B demonstrates strong performance on video benchmarks, achieving 65.1% and 71.1% accuracy on VideoMME without and with subtitles, respectively. It consistently outperforms LongVILA-7B across a range of benchmarks, including ActivityNet-QA, LongVideoBench, PerceptionTest, NExT-QA, VNBench, and VideoMME. Additionally, LongVILA-R1 supports processing up to 8,192 video frames per video, and configurable FPS settings. On our LongVideo-Reason-eval benchmark, LongVILA-R1-7B achieves an average accuracy of 72.0%, surpassing Video-R1-7B[8], as well as proprietary models such as Gemini-1.5-Pro[34].

2 Related Work

Multi-modal reasoning models. The field of multi-modal reasoning has advanced significantly, particularly in Vision-Language Models (VLMs). GPT-4o [29] improves visual understanding through enhanced reasoning, while Gemini-1.5-Pro [34] extends context length to 1 million tokens, achieving state-of-the-art performance in VideoMME [9]. Following the substantial progress of architecture and training algorithms in open-source VLMs [56, 27, 48, 23, 39], multi-modal reasoning has been further explored in works including LMM-R1 [32] which employs a two-stage training strategy, Vision-R1 [15] that addresses post-cold-start overthinking, and Video-R1 [8] which enhances RL for video via T-GRPO in 16 frames. However, these approaches primarily focus on single images or short videos, and long video reasoning still poses great challenges.

Sequence parallelism. Training with long contexts often exceeds the memory capacity of a single device [4], necessitating efficient distribution strategies. Sequence parallelism (SP) has become a widely adopted solution [25, 16, 3, 7, 11]. For example, ring-based systems like LightSeq [20] and Ring Attention [25] use point-to-point (P2P) communication, while DeepSpeed-Ulysses [16] employs all-to-all (A2A) primitives to optimize attention computations. Additionally, USP and LoongTrain [7, 11] were introduced to integrate Ring-style SP and Ulysses SP. LongVILA [47] further proposed multi-modal SP (MM-SP), enabling vision-language models to handle long-context inputs. However, multi-modal reinforcement learning introduces additional challenges, as it requires extensive sampling from long, mixed-token sequences [38], particularly in complex group optimization tasks [35].

RL frameworks for LLMs/VLMs. Reinforcement Learning (RL) has become a key strategy for enhancing Large Language Models (LLMs), particularly through Reinforcement Learning with Human Feedback (RLHF) [30] or Direct Preference Optimization [33], which aligns model outputs with human preferences. Recent advancements demonstrate that RL significantly improves LLM reasoning abilities. For example, DeepSeek-R1 [6] utilizes the Group Relative Policy Optimization (GRPO) algorithm [36], integrating group-based sampling and rule-based rewards. On the other hand, RL [1] poses a unique challenge of heavy computational cost, especially in multi-modal settings. To address this problem, HybridFlow [38] is introduced, leveraging Ray [28] for efficient data flow and vLLM [19] for faster sampling. Nevertheless, this remains a bottleneck when processing long video sequences, with group-based sampling constrained by the high computational cost of long-context sampling and visual encoding. In this work, we propose MR-SP that provides up to 2.1× speedup.

3 LongVideo-Reason Data Construction

3.1 Overview of Data Curation

We first curate 18K long videos from the Shot2Story dataset [13] (Figure 3, left). We further incorporated 2k additional 4K-resolution videos spanning scenarios such as autonomous driving, video games, household robotics, and wildlife. We then apply a high quality automated annotation pipeline for CoT as detailed in Section 3.2, and end up with a total of 104K Question-Reasoning-Answer pairs where each sample, based on the type of question it is reasoning about, can be categorized into Temporal Reasoning, Goal and Purpose Reasoning, Spatial Reasoning, or Plot and Narrative Reasoning (Figure 3, middle). This dataset is designed to support various types of long-video reasoning tasks comprehensively.

Given the sensitivity of GRPO to batch sampling [40, 32], a data filtering approach is adopted. Specifically, we employ a test-scaling method where LongVILA [47] performs inference 10 times on the original datasets. Questions consistently answered correctly or incorrectly are labeled as easy or hard while those inducing diverse predictions are labeled as medium. We filter out too easy and too hard questions, and use remaining questions for training. The reason is that GRPO expects different rollouts of each sample to be diverse in order to have meaningful advantages, and the gradient vanishes if all the rollouts predict correct or incorrect answers. The COT-SFT subset (36K) features high quality CoT reasoning processes formatted in a standard <think></think><answer></answer> structure, providing abundant resources for warm-up training during Stage 1 of the model's reasoning capabilities. Meanwhile, the RL subset contains 68K challenging long-video Q&A, which is leveraged in Stage 2 for scaling reasoning through reinforcement learning. To further enhance RL scaling, we incorporate an additional 102K high quality open-source videos (Figure 3, right) from other datasets [53, 46, 31, 18, 44]. This combination improves the model's generalization.

3.2 Long-Video Reasoning Generation

We introduce an automated annotation pipeline (Figure 4) that generates high quality Question-Reasoning-Answer pairs from long videos. This process begins by segmenting videos into short clips (~10 seconds each), each of which is annotated using the NVILA-8B [27] model to provide descriptive captions. For the spatial reasoning category, We employed VILA3D [39] to generate object bounding boxes in video frames and constructed spatial reasoning QAs based on these boxes and corresponding captions. Leveraging the breakthrough in text-based reasoning, we then deploy a leading open-source reasoning LLM, DeepSeek-R1-671B [6], provide the captions of all the clips in each video, and prompt it to generate diverse types of Question-Reasoning-Answer pairs that involve reasoning over the content across the whole video.

Figure 4: Data generation process for the LongVideo-Reason dataset. This process begins with segmenting videos into 10-second clips and generating captions for each clip using NVILA-8B. Then based on the captions of all clips in a video, we generate question-answer pairs that involve reasoning across the content of the whole video, along with the reasoning annotations using a leading open-source reasoning LLM. Reasoning questions are categorized into Temporal, Goal and Purpose, Spatial, and Plot and Narrative. Finally, the reasoning annotations are reformatted for conciseness and alignment with video details. We present a more detailed figure of data generation process in Figure 9.

Specifically, we design four types of prompts to encourage the LLM to generate a Question-Reasoning-Answer pair that focuses on one of the four types of reasoning: Temporal Reasoning, Goal and Purpose Reasoning, Spatial Reasoning, or Plot and Narrative Reasoning. To ensure VLMs focus on visual details, we also craft the prompts with phrases such as “checking the video” and “analyzing the scene”, which guide iterative examination of visual content. Finally, an LLM is then used to refine and streamline the reasoning steps. We also manually curated 1,000 high quality complex reasoning questions across four reasoning categories to serve as a new benchmark (LongVideo-Reason-eval) for evaluating VLMs in reasoning abilities. Note that we also include open-ended questions. Among the total 104K QA pairs, approximately half are multiple-choice and hald are open-ended. This entire data procedure consumes about 80,000 H100 GPU hours.

4 LongVILA Training Pipeline

As shown in Figure 5, there are two extended training stages in LongVILA-R1, i.e., (1) warm-up for long video reasoning, utilizing 36K data with high quality CoT for SFT on the MM-SP system [47]; (2) reinforcement learning with dense frames from long videos.

4.1 Long Video CoT Supervised Fine-Tuning

Utilizing 104K high quality question-reasoning-answer pairs, we apply the data filtering method described in Section 3.1 to select 36K examples for long CoT-SFT, serving as a warm-up phase for subsequent RL. This stage equips the model with fundamental reasoning abilities and instruction-following skills for long video scenarios. To efficiently perform SFT on hundreds of frames, we adopt the MM-SP [47] training system from LongVILA. As demonstrated in Section 6.2, SFT solely on our LongVideo-Reason dataset also effectively improves the model with reasoning capabilities.

4.2 GRPO for Long Video

Building on the advancements of the GRPO [35] algorithm and prior explorations of multi-modal reasoning training [8, 32], we adhere to the standard GRPO framework to train our model. For each given question q, the policy model generates a group of candidate responses $\{o_1, o_2, ..., o_G\}$ from the old policy $\pi_{\theta_{old}}$, accompanied by their corresponding rewards $\{r_1, r_2, ..., r_G\}$, which are computed based on rule-based reward functions (format/accuracy). The model π_θ is subsequently optimized by maximizing the following objective function:

$$J(\theta) = \mathbb{E}_{q,\{o_i\}}\Big[\frac{1}{G}\sum_{i=1}^{G}\Big(\min\Big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i,\ \mathrm{clip}\Big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\epsilon, 1+\epsilon\Big)A_i\Big) - \beta D_{KL}(\pi_\theta || \pi_{ref})\Big)\Big] \quad (1)$$

where ϵ and β are hyper-parameters. G is set as 8 in our experiments, and the sampled rewards above are normalized to get the advantages (A_i) for updating the model:

Figure 5: The LongVILA-R1 training pipeline. LongVILA-R1 builds upon the base training pipeline for LongVILA. MM-SP is further employed for SFT on long video understanding tasks with long CoT. Then, reinforcement scaling learning is conducted through Multi-modal Reinforcement Sequential Parallelism (MR-SP).

Figure 6: LongVILA-R1 RL training framework. The framework integrates multi-modal reinforcement sequence parallelism (MR-SP) for scalable video frame encoding and LLM prefilling. RL utilizes a vLLM-based engine with cached video embeddings, tailored for LongVILA rollout. Rewards for accuracy and format guide policy optimization.

$$A_i = \frac{r_i - \mathrm{mean}(\{r_1, r_2, \ldots, r_G\})}{\mathrm{std}(\{r_1, r_2, \ldots, r_G\})} \tag{2}$$

However, RL for long videos presents significant challenges due to the high computational demands of processing hundreds to thousands of frames. Existing RL frameworks struggle with such long-context training in rollout and LLM prefilling. To address this, we develop the MR-SP framework (Section 5), which efficiently scales reinforcement learning for long-context video reasoning.

Considering the sensitivity of GRPO to sampling during training [40], we use the 68K filtered data for reinforcement learning as described in Section 3.1. Additionally, an extra 102K samples from other datasets [53, 46, 31, 18, 44] are incorporated to scale up the RL. This approach aims to guide the model in freely exploring and developing more effective and generalized reasoning strategies.

5 Multi-modal Reinforcement SP

Existing RL frameworks for VLMs, such as R1-V [7] and EasyR1 [55], are not designed for long videos which present unique challenges due to their high token volume. To address this, we introduce Multi-modal Reinforcement Sequence Parallelism (MR-SP), a framework for efficient RL training on long videos. MR-SP leverages sequence parallelism in both rollout and pre-filling stages, enabling long videos in RL, with reduced overhead. We show the training curve with MR-SP in Figure 8.

5.1 Stage 1 - Rollout with Paralleled Encoding

To support long-video reinforcement learning efficiently, we adopt sequence parallelism (SP) for the video encoding stage. As shown in Figure 7, the input video frames are first evenly divided across multiple GPUs (e.g., GPU 1 to GPU 3), each equipped with its own vision tower. Each GPU independently processes a slice of the video, encoding only a subset of the frames. The resulting video embeddings are then aggregated with text embeddings via an all-gather operation as indicated by the "All-Gather" arrow in the figure. This strategy distributes the encoding workload, allowing the system to handle significantly longer videos by leveraging more GPUs, while avoiding the risk of GPU memory overflow. The parallel encoding scheme ensures balanced utilization of the vision towers and enables scalable long-video processing that would otherwise be infeasible.

After the video embeddings are globally gathered, they are reused for downstream usage throughout the RL pipeline. As illustrated in Figure 7, the gathered embeddings are reused during multiple

Figure 7: The work ow of multi-modal reinforcement sequential parallel (MR-SP). To accommodate multi-modal input in reinforcement learning, we develop a custom sharding strategy to ensure balanced workload distribution and compatibility with SP communication. Ef cient video embedding reuse and vLLM rollout acceleration strategies are implemented, while meeting the demands of policy model pre lling for dense video frames.

Figure 8: Reward curve of our LongVILA-R1-7B with MR-SP. During training, LongVILA-R1-7B exhibits stable reward improvements across overall, format, and accuracy dimensions.

rollouts without recomputation. For instance, in each training step, we typically perform 8 to 16 rollouts. Without recycling, the same video would need to be re-encoded dozens of times per step, severely impacting training speed. By caching and reusing the gathered embeddings, MR-SP eliminates this redundancy and signi cantly accelerates training.

5.2 Stage 2 - Pre lling with Sequence Parallelism

For each rollout, both the reference and policy models require pre lling compute-intensively in RL for long videos. With the gathered embedding from Stage 1 reused, we parallelize the inference stage across devices using sequence parallelism. As illustrated in Figure 7, we globally gathered input embeddings are rst padded to a uniform length (Padding Sequence) and then evenly partitioned across GPUs (Sharding to Local GPU). This allows each GPU to handle only a portion of the input sequence during pre lling. This parallelism is applied to both policy and reference model pre lling. Then, each GPU locally computes logits for its token slice, pre lling in parallel.

6 Experimental Results

6.1 Main Results

Table 1 shows the performance comparison on 6 video benchmarks [49, 45, 31, 46, 54, 9]. LongVILA-R1-7B consistently outperforms LongVILA-7B across all benchmarks, with performance gaps that vary according to the complexity of the reasoning tasks. Table 3 presents the general performance of LongVILA-R1, comparing to existing advanced models [22, 52, 2, 21, 17, 26, 42, 51, 8, 57, 41, 5, 53, 27, 50] under comparable model sizes on the Video-MME [9] benchmark. The LongVILA-R1-7B is tested using 512 video frames as inputs. LongVILA-R1-7B achieves leading scores across different video lengths, obtaining scores of 65.1% and 71.1% in the settings without subtitles and with subtitles,

Table 1: Performance on ActivityNet-QA [49], LongVideoBench [45], PerceptionTest [31], NExT-QA [46], VNBench [54], and VideoMME [9]. LongVILA-R1-7B outperforms LongVILA-7B across all these benchmarks, with varying margins.

Model	Act.Net-QA test	L.V.Bench val	Per.Test val	NExT-QA mc	VNBench val	VideoMME w/o sub.	VideoMME w/ sub.
GPT-4o mini	-	56.5	-	-	-	64.8	68.9
GPT-4o	61.9	66.7	-	-	64.4	71.9	77.2
Gemini-1.5-Pro	57.5	64.0	-	-	66.7	75.0	81.3
Video-LLaVA-7B	45.3	37.6	-	-	12.4	39.9	41.6
Flash-VStream-7B	51.9	-	-	61.6	-	-	-
ShareGPT4Video-8B	50.8	41.8	-	-	-	39.9	43.6
VideoLLaMA2-7B	50.2	-	51.4	-	4.5	47.9	50.3
VideoLLaMA2.1-7B	53.0	-	54.9	-	-	54.9	56.4
Kangaroo-8B	-	54.8	-	-	-	56.0	57.6
PLLaVA-7B	56.3	39.2	-	-	-	-	-
LLaVA-OV-7B	56.7	56.4	57.1	79.4	51.8	58.2	61.5
LongVILA-7B	59.5	57.1	58.1	80.7	63.0	60.1	65.1
LongVILA-R1-7B	64.8	58.0	68.9	81.5	75.5	65.1	71.1

Table 2: Performance on LongVideo-Reason-eval. LongVILA-R1-7B achieves a strong overall score.

Model	Temporal	Goal	Plot	Spatial	Overall
Video-R1-7B [8]	61.4	85.0	62.0	58.5	68.1
Gemini-1.5-Pro [34]	65.4	81.9	67.8	53.3	69.3
LongVILA-7B	58.0	80.2	57.1	46.7	62.7
LongVILA-R1-7B	68.1	85.7	70.6	53.3	72.0

respectively. Table 2 compares the results of our LongVideo-Reason-eval benchmark. LongVILA-R1-7B model achieves a strong performance with an average score of 72.0%, surpassing Video-R1-7B [8] and slightly outperforms Gemini-1.5-Pro [34]. Qualitative results are in the appendix.

6.2 Ablation Study

Scaling video frames. The reasoning capability of LongVILA-R1 scales consistently with the number of input video frames. Specifically, Table 5 illustrates the performance of LongVILA-1.5B (grey line) and LongVILA-1.5B-R1 (red line) on the long-video reasoning benchmark under varying frame inputs. With only 16 input frames, LongVILA-R1-1.5B is close to LongVILA-1.5B. However, as the number of frames increases to 512, LongVILA-R1-1.5B consistently outperforms and eventually achieves a score of 64.3%. Notably, LongVILA-R1-1.5B demonstrates steady performance improvements throughout the scaling process. In contrast, LongVILA-1.5B hits a performance bottleneck with 256 frames, and got a degradation on 512 frames. The enhanced reasoning capabilities of LongVILA-R1-1.5B allow it to effectively integrate and infer information from long videos.

Ablation on pipeline and datasets. As shown in Table 4, we ablate the effectiveness of training stages and datasets, using LongVILA-1.5B as a starting point. The accuracies are evaluated on LongVideo-Reason-eval. 7 means skipping this stage, 3 means training this stage with our datasets, and O means training this stage with other datasets [53, 46, 31, 18, 44]. Our CoT-SFT dataset results in a better performance than that of other datasets. In addition, incorporating RL on top of the warm-up phase (CoT-SFT) yields additional improvements compared to using only SFT. We show that if we skip CoT-SFT and train our models directly with RL, the accuracy drops. If we apply Video-R1 datasets in both CoT-SFT and RL stages, the performance is inferior to using ours.

Training efficiency on MR-SP. We conduct the training efficiency comparison for our MR-SP system, one A100 node, i.e., 8xA100 (80GB) GPUs. We measure the forward time for each training

Table 3: Performance comparison on VideoMME [9] benchmark in details.

Model	w/o subtitle				w subtitle			
	Overall	Short	Medium	Long	Overall	Short	Medium	Long
Video-R1-7B [8]	61.4	-	-	-	-	-	-	-
Apollo-7B [57]	61.1	-	-	-	63.3	-	-	-
LLaVA-Video-7B [53]	63.3	-	-	-	69.7	-	-	-
NVILA-8B-Video [27]	64.2	-	-	-	70.0	-	-	-
Video-LLaVA-7B [22]	39.9	45.3	38.0	36.2	41.6	46.1	40.7	38.1
SliME-8B [52]	45.3	53.3	55.4	39.8	47.2	55.4	44.4	41.7
ShareGPT4Video-8B [2]	39.9	48.3	36.3	35.0	43.6	53.6	39.3	37.9
VideoChat2-7B [21]	39.5	48.3	37.0	33.2	43.8	52.8	39.4	39.2
Chat-Univi-v1.5-7B [17]	40.6	45.7	40.3	35.8	45.9	51.2	44.6	41.8
Kangaroo-8B [26]	56.0	66.1	55.3	46.7	57.6	68.0	55.4	49.3
ShareGemini-7B [42]	43.2	49.1	41.3	39.1	47.9	49.1	47.3	43.4
LongVA-7B [51]	52.6	61.1	50.4	46.2	54.3	61.1	53.6	47.6
VITA-1.5-7B [10]	56.1	67.0	54.2	47.1	58.7	69.9	55.7	50.4
LongVILA-7B [47]	60.1	69.0	58.3	53.0	65.1	72.9	64.9	57.4
LongVILA-R1-7B	65.1	76.8	63.2	55.2	71.1	79.2	69.7	64.3

Table 4: Ablations on frames and reasoning (R.).

Frames	16	32	64	128	256	512
w/o R.	55.7	56.4	58.1	60.5	60.7	60.2
w/ R.	55.9	56.7	61.9	62.6	64.1	64.3

Table 5: Ablations on CoT-SFT and RL.

CoT-SFT	7	3	7	O	O	3
RL	7	7	3	7	O	3
Accuracy	58.1	60.2	52.4	59.1	59.4	61.9

step. The results are obtained after 10 warming up iterations and averaged over 5 iterations. We use LongVILA-7B-R1 model with training batch size as 1 per GPU and rollout number as 5.

Figure 2 presents the training ef ciency comparison across different numbers of frames. The gure plots the runtime per step (in seconds) for three settings: the plain RL system without MR-SP, MR-SP Stage 1 only, and the full MR-SP system (Stage 1 & 2). The baseline runtime increases steeply as the frame count grows. Using only Stage 1 of MR-SP signi cantly improves ef ciency up to 512 frames but encounters GPU out-of-memory (OOM) issues beyond that point. In contrast, the full MR-SP system consistently reduces runtime, achieving up to a 2.1× speedup at 512 frames and scaling ef ciently to 1024 frames without OOM, highlighting the bene t of combining sequence reuse and sequence parallelism for long video RL training.

7 Conclusion

We present a comprehensive framework designed to fully scale VLMs for reasoning over long videos. LongVILA-R1 encompasses a meticulously constructed large-scale dataset, LongVideo-Reason, and a parallelized training framework, MR-SP. Leveraging our curated dataset of 104K long video question-reasoning-answer pairs, combined with other open-source video datasets, we adopt a two-stage training process that integrates CoT-SFT and RL. LongVILA-R1-7B demonstrates outstanding performance on mainstream video benchmarks, achieving 65.1% and 71.1% on VideoMME without and with subtitles. LongVILA-R1 supports processing up to 8,192 video frames per video, and con gurable FPS settings. Notably, our MR-SP leads to a speed up of 2.1 for long video RL training, supporting hour-level (3,600 frames) RL training on a single node of 8 A100 GPUs. In addition, we release our training system to the public, which supports RL training across multiple modalities (video, text, and audio), various models (including VILA and Qwen series), and even image and video generation models.

Limitations While our RL training system is capable of handling long video frames ef ciently (3,600 frames on a single node of A100 GPUs), scaling to signi cantly longer sequences, more modalities (like include audio for omni VLMs) or large batch sizes would require distributed training across multiple GPUs. This requires more GPUs, making it less feasible to run on limited resources.

Figure 9: Detailed data generation process for the LongVideo-Reason dataset, supplementing Figure 4.

8 Border Impacts

The development of LongVILA-R1 represents a signi cant advancement in long video reasoning, enabling real world systems to perform sophisticated temporal and compositional understanding across diverse and extended visual contexts. This progress holds transformative potential across multiple domains, including embodied AI, robotics, autonomous systems, and AR/VR applications. By equipping vision-language models (VLMs) with the ability to process and reason over long-duration video data, LongVILA-R1 lays the foundation for AI systems capable of understanding event sequences, and inferring causal and physical relationships over extended frames.

Long video reasoning technologies powered by LongVILA-R1 can signi cantly enhance embodied AI and robotics by enabling agents to sustain coherent, long-term understanding of their environment. Robots equipped with such capabilities would excel in performing complex, multi-stage tasks, adapting to dynamic contexts, and building richer world models for planning and decision-making. These advancements also promise to unlock new opportunities in education, healthcare, and entertainment. For instance, long video understanding could enable AI tutors to analyze and summarize extended instructional videos, or assist healthcare professionals in reviewing lengthy procedural recordings. Furthermore, such systems could enhance sports analytics, and other areas requiring nuanced temporal reasoning.

Figure 9 shows the the data-generation where our supervision is built from grouped short clips that form long videos, per-clip captions, and LLM-generated Q&A with Long-CoT that is formatted as multiple-choice reasoning rather than identity labels or free-form attribution. This design choice intentionally steers the supervision signal toward temporal and causal evidence (what happens, when, and why) and away from sensitive attributes (who a person is), thereby reducing privacy and pro ling risks at the data layer. Concretely, we derive questions from captions and temporally grounded observations instead of collecting personal metadata; we avoid tasks that require demographic inference, face recognition, or persistent identity tracking; and we release structured Q&A rather than user-originating personal content. In this way, the dataset construction supports the intended societal bene ts of long-video reasoning while mitigating risks such as privacy leakage, surveillance-like use, and stereotype ampli cation.

In conclusion, LongVILA-R1 demonstrates the potential of long video reasoning to drive progress across a wide range of applications, from robotics to immersive virtual environments. However, unlocking the full promise of this technology requires a steadfast commitment to ethical principles, privacy protection, and the broader goal of bene ting humanity. By addressing these challenges, the AI community can ensure that advancements in long video reasoning contribute positively to society while mitigating associated risks.

References

[1] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3, 2025. Accessed: 2025-02-02.

[2] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. NeurIPS, 37:19472–19495, 2024.

[3] Qiaoling Chen, Diandian Gu, Guoteng Wang, Xun Chen, YingTong Xiong, Ting Huang, Qinghao Hu, Xin Jin, Yonggang Wen, Tianwei Zhang, and Peng Sun. Internevo: Ef cient long-sequence large language model training via hybrid parallelism and redundant sharding. CoRR, abs/2401.09149, 2024.

[4] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Ef cient ne-tuning of long-context large language models. In ICLR, 2024.

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. CoRR, abs/2412.05271, 2024.

[6] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. CoRR, abs/2501.12948, 2025.

[7] Jiarui Fang and Shangchun Zhao. USP: A uni ed sequence parallelism approach for long context generative AI. CoRR, abs/2405.07719, 2024.

[8] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. CoRR, abs/2503.21776, 2025.

[9] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The rst-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. CoRR, abs/2405.21075, 2024.

[10] Chaoyou Fu, Haojia Lin, Xiong Wang, Yifan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. VITA-1.5: towards gpt-4o level real-time vision and speech interaction. CoRR, abs/2501.01957, 2025.

[11] Diandian Gu, Peng Sun, Qinghao Hu, Ting Huang, Xun Chen, Yingtong Xiong, Guoteng Wang, Qiaoling Chen, Shangchun Zhao, Jiarui Fang, Yonggang Wen, Tianwei Zhang, Xin Jin, and Xuanzhe Liu. Loongtrain: Ef cient training of long-sequence llms with head-context parallelism. CoRR, pdf/2406.18485, 2024.

[12] Yanan Guo, Wenhui Dong, Jun Song, Shiding Zhu, Xuan Zhang, Hanqing Yang, Yingbo Wang, Yang Du, Xianing Chen, and Bo Zheng. Fila-video: Spatio-temporal compression for ne-grained long video understanding. CoRR, abs/2504.20384, 2025.

[13] Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. CoRR, abs/2312.10300, 2023.

[14] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for ne-grained video reasoning via core frame selection. CoRR, abs/2411.14794, 2024.

[15] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. CoRR, abs/2503.06749, 2025.

[16] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. CoRR, abs/2309.14509, 2023.

[17] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Uni ed visual representation empowers large language models with image and video understanding. In CVPR, pages 13700–13710, 2024.

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, pages 1988–1997, 2017.

[19] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Ef cient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.

[20] Dacheng Li, Rulin Shao, Anze Xie, Eric Xing, Joseph E Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. Lightseq: Sequence level parallelism for distributed training of long context transformers. 2023.

[21] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. CoRR, abs/2305.06355, 2023.

[22] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In EMNLP, pages 5971–5984. ACL, 2024.

[23] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In CVPR, pages 26689–26699, 2024.

[24] Changshu Liu, Shizhuo Dylan Zhang, and Reyhaneh Jabbarvand. Codemind: A framework to challenge large language models for code reasoning. CoRR, abs/2402.09664, 2024.

[25] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-in nite context. CoRR, abs/2310.01889, 2023.

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

[27] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. NVILA: ef cient frontier visual language models. CoRR, abs/2412.04468, 2024.

[28] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. Ray: A distributed framework for emerging fAIg applications. In OSDI, pages 561–577, 2018.

[29] OpenAI. Gpt-4o. 2025.

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. NeurIPS, 35:27730–27744, 2022.

[31] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In NeurIPS, 2023.

[32] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. LMM-R1: empowering 3b lmms with strong reasoning abilities through two-stage rule-based RL. CoRR, abs/2503.07536, 2025.

[33] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In NeurIPS, 2023.

[34] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. CoRR, abs/2403.05530, 2024.

[35] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. CoRR, abs/2402.03300, 2024.

[36] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. CoRR, abs/2402.03300, 2024.

[37] Yunhang Shen, Chaoyou Fu, Shaoqi Dong, Xiong Wang, Peixian Chen, Mengdan Zhang, Haoyu Cao, Ke Li, Xiawu Zheng, Yan Zhang, et al. Long-vita: Scaling large multi-modal models to 1 million tokens with leading short-context accuray. CoRR, abs/2502.05177, 2025.

[38] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybrid ow: A exible and ef cient RLHF framework. In EuroSys, pages 1279–1297. ACM, 2025.

[39] Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, and Hongxu Yin. Scaling vision pre-training to 4k resolution. CoRR, abs/2503.19903, 2025.

[40] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. CoRR, abs/2501.12599, 2025.

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

[42] Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Retake: Reducing temporal and knowledge redundancy for long video understanding. CoRR, abs/2412.20504, 2024.

[43] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Ef cient long video understanding via large language models. In ECCV, pages 453–470. Springer, 2024.

[44] Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In NeurIPS - Datasets and Benchmarks, 2021.

[45] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. CoRR, abs/2407.15754, 2024.

[46] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In CVPR, pages 9777–9786, 2021.

[47] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. In ICLR, 2025.

[48] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. CoRR, abs/2412.15115, 2024.

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

[50] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. CoRR, abs/2501.13106, 2025.

[51] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. CoRR, abs/2406.16852, 2024.

[52] Yifan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. CoRR, abs/2406.08487, 2024.

[53] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. CoRR, abs/2410.02713, 2024.

[54] Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in A video haystack: A scalable synthetic framework for benchmarking video mllms. CoRR, abs/2406.09367, 2024.

[55] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An ef cient, scalable, multi-modality rl training framework, 2025.

[56] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. CoRR, abs/2504.10479, 2025.

[57] 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. Apollo: An exploration of video understanding in large multimodal models. In CVPR, pages 18891–18901, 2025.

NeurIPS Paper Checklist

1. Claims

   Question: Do the main claims made in the abstract and introduction accurately re ect the paper's contributions and scope?

   Answer: [Yes]

   Justi cation: The main claims made in the abstract and introduction accurately re ect the paper's contributions and scope, as they align with the theoretical and experimental results presented in the paper and provide a clear understanding of the paper's goals.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and re ect how much the results can be expected to generalize to other settings.
   - It is ne to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justi cation: The paper acknowledges the limitations of the work performed by the authors.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-speci cation, asymptotic approximations only holding locally). The authors should re ect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should re ect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should re ect on the factors that in uence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational ef ciency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be speci cally instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justi cation:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justi cation: The paper fully discloses all the information needed to reproduce the main experimental results, ensuring that the main claims and conclusions can be independently veri ed. This includes providing relevant details, methodologies, and any necessary parameters or con gurations for conducting the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or veri able.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suf ce, or if the contribution is a speci c model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

   Question: Does the paper provide open access to the data and code, with suf cient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justi cation: The paper will provide open access to the data and code necessary to reproduce the main experimental results. It also includes suf cient instructions in the supplemental material on how to faithfully replicate the experiments conducted in the paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips:cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips:cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justi cation: The paper speci es all the training and test details necessary to understand the results. This includes information on data splits, hyperparameters, the methodology for selecting hyperparameters, the type of optimizer used, and any other relevant details that are crucial for replicating and comprehending the reported results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical signi cance

   Question: Does the paper report error bars suitably and correctly de ned or other appropriate information about the statistical signi cance of the experiments?

   Answer: [Yes]

   Justi cation: The paper provides suitable information about the statistical signi cance of the experiments. This indicates that the authors have appropriately addressed the need for statistical analysis and have reported the relevant measures to support the reliability and signi cance of their experimental ndings.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, con - dence intervals, or statistical signi cance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not veri ed.
- For asymmetric distributions, the authors should be careful not to show in tables or gures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding gures or tables in the text.

8. Experiments compute resources

   Question: For each experiment, does the paper provide suf cient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justi cation: The paper provides suf cient information on the computer resources required to reproduce each experiment. This includes details such as the type of compute workers used, the amount of memory required, and the time taken for the execution of the experiments. This information allows for accurate replication of the experiments and provides transparency regarding the computational requirements of the study.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips:cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justi cation: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justi cation: The paper discusses both potential positive and negative societal impacts of the work performed, especially in the resource limited application.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake pro les, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact speci c groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the ef ciency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justi cation: The paper poses no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety lters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justi cation: The paper only use the opensource datasets with legal license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced in the paper are well documented and are the documentation provided alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Appendix

In this section, we selected examples of four types of reasoning tasks from the benchmark to evaluate and compare the reasoning processes and answers provided by Gemini-1.5-Pro, Video-R1-7B, and LongVILA-R1-7B. On Figure 12, a 20-minute StarCraft match is depicted, where the models analyze the players' unit compositions, strategies, and play styles to predict the potential developments on the battle field. While Gemini-1.5-Pro produced a correct prediction of the outcome, its reasoning process contained factual inaccuracies. In contrast, Video-R1-7B, influenced by the characteristics of its training data, tended to summarize answers based on options, neglecting critical video details and resulting in incorrect reasoning. LongVILA-R1-7B, however, is able to accurately analyze the players' operational styles and specific moments marked in the video, leading to a comprehensive and accurate prediction of the match's trajectory. On Figure 13, another example demonstrates the models' abilities in narrative reasoning and visual information analysis. Gemini-1.5-Pro failed to correctly infer why the man appearing for the second time in the video is not the husband. In contrast, both Video-R1-7B and LongVILA-R1-7B successfully reasoned that the man's habit of wearing a ring on his left hand is a key indicator, providing accurate answers.

Figure 14 illustrates the models' spatial perception and reasoning abilities as the camera moves through a room. Gemini-1.5-Pro effectively identified the key information within the video and provided the correct answer through straightforward reasoning. In contrast, Video-R1-7B experienced significant localization errors during the reasoning process, leading to a critical issue for reasoning models: a mismatch between the reasoning analysis and the final answer. LongVILA-R1-7B demonstrated superior performance by leveraging dense frame analysis to accurately infer the spatial relationships between rooms and furniture across different levels, ultimately delivering a coherent reasoning process and the correct answer. On Figure 15, the focus shifts to temporal analysis in a Lego video featuring diverse events and interactions. All three models successfully reasoned through the sequence of events and provided correct answers, showcasing their proficiency in temporal reasoning tasks. As a supplement to Figure 1, Figure 10 provides a more comprehensive comparison of two examples: "2022 FIFA Argentina vs. Netherlands" and "Moving the Cup and Guessing Where the Ball Is." In the football match example, while Gemini-1.5-Pro produced the correct answer, its output contained hallucinatory content influenced by biases in its pre-learned knowledge. Video-R1 not only failed to provide accurate video analysis reasoning but also made incorrect predictions. In contrast, LongVILA-R1 successfully analyzed the players' performance and emotions during the match, integrating these factors through its robust reasoning capabilities to make accurate predictions about the outcome. For the more challenging task of tracking the ball, Gemini-1.5-Pro's reasoning is inconsistent with the spatial content throughout, while Video-R1 failed to deduce the ball's final position accurately. Remarkably, LongVILA-R1 precisely analyzed the spatial transformations following the movement of the box, demonstrating superior interpretative and reasoning abilities.

Figure 10: Detailed comparisons in the football game example. The video is available at Link.

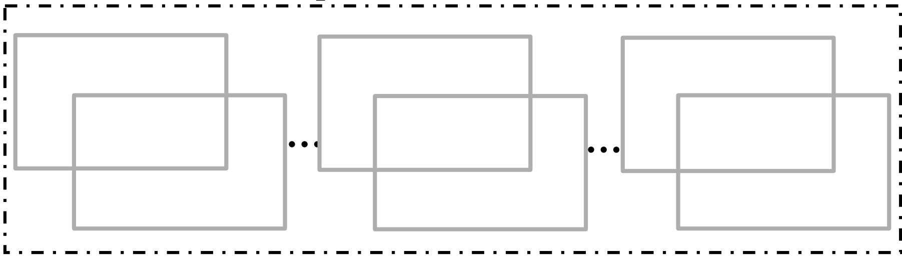
Figure 11: Detailed comparisons in the moving cup example. The video is available at Link.

Figure 12: LongVILA-R1 reasoning example in the "Goal and Purpose" category. The video is available at Link.

Figure 13: LongVILA-R1 reasoning example in the "Plot and Narrative" category. The video is available at Link.

Figure 14: LongVILA-R1 reasoning example in the "Spatial" category. The video is available at Link.

Temporal – LEGO PINATA



Question: Which cake's preparation process involved both the use of a blue hammer for smashing and a red silicone mold for shaping decorative elements, based on the temporal progression of tools and ingredients observed in the video?

A: *Green/yellow/red Lego block cakes*
B: *Red velvet cake with batter mixed in a blue bowl*
C: *Chocolate cake with parchment paper and red cake pops*
D: *Lego-themed cake smashed at 0:00:30-0:00:40 and decorated with bricks*

Gemini-1.5-Pro: The question is about the cake's preparation process shown in the video…The yellow Lego-themed cake is smashed with an orange hammer. The green Lego-themed cake is opened with a knife. 3. At 0:30-0:40, red silicone mold is used for shaping a decorative brick by pouring melted chocolate into its cases. 4. So, both tools are present only at the preparation of the red Lego-themed cake in the video.
<answer>D</answer>

Video-R1: <think>Let me think about this carefully…First, let's consider the options: A: Green/yellow/red Lego block cakes - This option mentions Lego blocks, but there's no mention of a blue hammer or a red silicon…after going through all the options, it's clear that the correct choice is D…. Let's verify one last time to make sure. Yes, the video indeed shows the use of a blue hammer and a red silicone mold in the preparation of the Lego-themed cake. So, the final answer is D</think>
<answer>D</answer>

LongVILA-R1: <think> The preparation of the Lego-themed cake involved the use of a blue hammer to smash colorful Lego blocks, which were then used as decorations. The red silicone mold was used to shape the red decorative elements for the cake. The process started with the woman using a blue hammer to smash green, yellow, and red Lego blocks, which confirmed the use of the blue hammer. The red silicone mold was explicitly mentioned when shaping the red decorative elements. The cake was then decorated with bricks, indicating the use of the red silicone mold for the red elements…The red velvet cake used a blue bowl for mixing batter but did not use the blue hammer or red silicone mold. Therefore, the correct answer is D.</think>
<answer>D</answer>

Figure 15: LongVILA-R1 reasoning example in the "Temporal" category. The video is available at Link.

Figure 16: LongVILA-R1 reasoning example in the taboo game. The video is available at Link.