

VAST: Video Ability-Stratified Taxonomy for Data-Efficient Video Reasoning

Zhongan Wang^{1*} Xiaoyu Wen^{2,3*} Lingxiao Du⁴ Kun Li⁵
 Zhiliang Wu¹ Xingcheng Xu² Qiaosheng Zhang^{2†} Chaochao Lu² Hehe Fan^{1,6†}
¹CCAI, Zhejiang University ²Shanghai AI Laboratory
³Shanghai Jiao Tong University ⁴Fudan University ⁵United Arab Emirates University
⁶State Key Laboratory of CAD&CG, Zhejiang University

Abstract

Reinforcement learning (RL) has emerged as an effective approach for improving video reasoning in multimodal large language models (MLLMs). However, existing methods remain inefficient for two reasons. First, training data are typically organized by task formats rather than underlying reasoning abilities, creating a mismatch that encourages models to learn task-specific patterns instead of transferable capabilities. As a result, improving ability generalization requires broad coverage over many ability-task combinations, making RL training costly. Second, prior work often compensates for this inefficiency with increasingly complex algorithmic designs, such as specialized temporal architectures or multi-objective reward frameworks, which further complicate training. To address these issues, we present VAST, a cognitive taxonomy and data organization framework that structures video understanding into three layers: Perception, Reasoning, and Cognition. Based on this taxonomy, we construct VAST-15K for training and VAST-Bench for evaluation. We further introduce VideoVAST, a reinforcement learning framework that uses consistency rewards to encourage alignment between reasoning traces and final answers, without architectural modifications. Experiments show that VideoVAST achieves 66.3% accuracy on MVBench and 57.4% on VAST-Bench, compared with 62.7% and 54.3% for Video-R1, while using approximately 72% fewer GPU hours and 96% fewer training samples under the same training settings. Code and resources are available at <https://zhongan-wang.github.io/VAST>.

1. Introduction

Reinforcement learning (RL) has improved the reasoning capabilities of large language models (LLMs) [15, 66]. Following the success of reasoning-oriented models such as

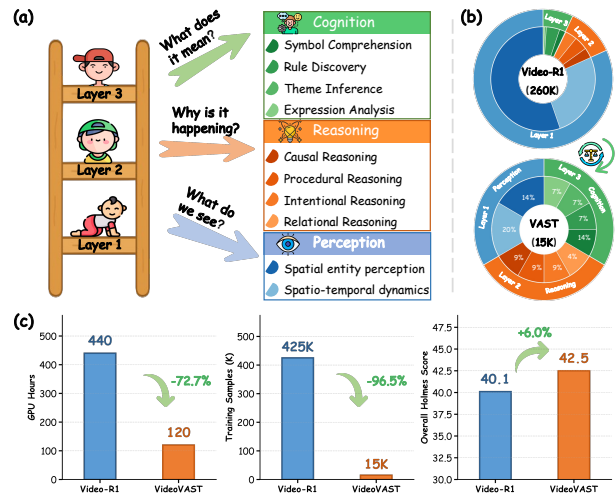


Figure 1. **Motivation and overall VAST framework.** (a) Cognitive development perspective that motivates a progression from perception to reasoning and cognition in video understanding. (b) Comparison of layer distributions between existing video reasoning datasets and VAST-15K. (c) Training efficiency and performance of VideoVAST and Video-R1 under the same settings.

DeepSeek-R1 [15], RL-based approaches have been extended to visual domains. Image-based methods [41, 43, 51] have shown notable gains on visual question answering and image reasoning, enhancing multi-step deduction and reducing hallucinations. Recent work further applies RL to video understanding [7, 21, 55], enabling models to integrate information across frames for complex reasoning about entities, events, and causal relationships.

However, endowing MLLMs with strong video reasoning capabilities remains challenging. Compared with static image reasoning, video reasoning requires models to integrate evidence across many frames and identify the most informative ones. This substantially expands the reasoning space and makes learning more difficult. Most existing methods address this challenge through a two-stage pipeline. They first perform supervised fine-tuning (SFT)

*Equal contribution.

†Corresponding author.

on chain-of-thought (CoT) annotations to imitate explicit step-by-step rationales [40, 53, 57], and then apply RL with rule-based rewards for answer accuracy and format consistency [11, 66]. Although effective, this paradigm requires large amounts of high-quality CoT data. In video reasoning, however, such data are expensive to obtain and therefore scarce, which severely limits scalability. This raises a key question: *Can we achieve comparable reasoning performance with significantly fewer training samples?*

Motivated by this question, we revisit how training data are organized for video reasoning. Existing methods typically structure datasets by annotation formats (e.g., multiple-choice questions, open-ended QA) rather than by the underlying abilities required to solve them. This format-centric organization creates a fundamental mismatch. A single ability (e.g., causal reasoning) can appear across many formats, while a single format conflates multiple abilities [2, 56]. As a result, models tend to learn superficial task cues instead of transferable abilities [12, 29]. In the video domain, this problem becomes even more severe. Frame-level visual cues introduce additional surface correlations, which further encourage shortcut exploitation rather than reasoning development. Consequently, achieving robust reasoning requires extensive data coverage across diverse task-ability combinations [5, 62]. Prior work attempts to alleviate this inefficiency by resorting to complex algorithmic designs, such as specialized temporal architectures [23, 58] or multi-objective reward frameworks [7, 55]. However, these designs typically come at the cost of greater training complexity.

To address these issues, we adopt a joint perspective from both the data and method sides. On the data side, we propose VAST (Video Ability-Stratified Taxonomy), a cognitive taxonomy and data organization framework for video understanding. Our taxonomy is inspired by Piaget’s cognitive development theory [45, 46], which states that intelligence develops through hierarchical stages from concrete perception to abstract reasoning. This hierarchical view provides a useful lens for organizing video understanding abilities: perceptual grounding supports reasoning, and reasoning enables higher-level cognition. We organize training data into three layers. Layer 1 (Perception) focuses on spatial entities and spatio-temporal dynamics. Layer 2 (Reasoning) covers causal, intentional, procedural, and relational reasoning [8, 34]. Layer 3 (Cognition) targets symbol comprehension, rule discovery, theme inference, and expression analysis [27, 47] (Figure 1 (a)). Based on this taxonomy, we construct **VAST-15K**, an ability-stratified dataset with balanced coverage across the three layers (Figure 1 (b)), and **VAST-Bench**, a benchmark for comprehensive evaluation. On the method side, we introduce VideoVAST, a reinforcement learning framework that builds on GRPO [48] and uses lightweight consistency rewards to

align the model’s reasoning with its final predictions.

Our experiments demonstrate that ability-stratified data organization yields substantial efficiency gains. VideoVAST achieves 66.3% accuracy on MVBench, outperforming Video-R1 (62.7%), and 57.4% accuracy on VAST-Bench, compared with 54.3% for Video-R1. Moreover, under the same training settings, VideoVAST uses approximately 72% fewer GPU hours and 96% fewer training samples than Video-R1, while maintaining competitive performance (Figure 1 (c)). In summary, we make three key contributions:

- A cognitive taxonomy that organizes video understanding into three capability layers spanning Perception, Reasoning, and Cognition, refined into ten sub-abilities. Based on this taxonomy, we construct VAST-15K and VAST-Bench for ability-stratified training and evaluation.
- We propose VideoVAST, a data-efficient reinforcement learning framework that employs consistency rewards over ability-stratified training data.
- Comprehensive experiments demonstrating substantial efficiency gains with competitive performance across multiple video reasoning benchmarks.

2. Related Work

Datasets and benchmarks for video reasoning. Current video reasoning datasets are typically organized by task format rather than cognitive capabilities [7, 55, 61]. Early benchmarks focused on action recognition, motion analysis, and VideoQA [14, 24, 28, 32, 52, 54, 59, 65]. Diagnostic datasets such as CLEVRER [63], NEX-T-QA [59], and AGQA [13] evaluate temporal and causal reasoning, but also reveal reliance on language priors or single-frame cues [29]. More recent benchmarks broaden evaluation to MVBench [31], VideoMME [9], TempCompass [39], MMVU [70], and Video-Holmes [3]. Related work has also expanded to facial video understanding, event streams, multi-context grounding, multimodal composition, adaptive visual tokenization, and visual anomalies [19, 33, 38, 60, 68, 69]. However, most benchmarks remain task-oriented rather than capability-stratified, making them less suitable for diagnosing how training data composition affects reasoning abilities.

Despite this progress, training data remain largely task-centric. For example, in Video-R1-260K, 82% of samples are perception-level, whereas only 18% target higher-order reasoning or cognition [7]. Prior work also shows that single-frame analysis can match multi-frame VideoQA models, indicating shortcut exploitation rather than robust temporal reasoning [12, 29]. Without a unified capability taxonomy, scaling data yields diminishing returns [10, 17]. We instead use Piaget’s theory of hierarchical cognitive development [16, 18, 45, 46] to organize video understanding into perception, reasoning, and cognition.

Reinforcement learning for video reasoning. Following progress in text reasoning [15, 51] and image understanding [21, 41, 44, 50], RL has recently been extended to video. Existing methods mainly follow two paradigms: supervised fine-tuning followed by reinforcement learning with task-specific reward functions [5, 25], exemplified by VideoRFT [55] and Video-R1 [7]; and reinforcement learning without supervised initialization, inspired by DeepSeek-R1-Zero [15], as in Video-R1-zero [7]. Yet their analyses are still centered on benchmark-level averages, leaving capability-level effects of post-training data composition largely underexplored.

Despite their effectiveness, both paradigms remain constrained by the inefficiencies of task-centric data organization, often resulting in limited coverage, high training cost, and increasingly complex reward designs for temporal consistency and cross-frame coherence [7, 42, 64, 71]. In contrast, we show that ability-stratified data organization enables more efficient RL training for video reasoning.

3. VAST

In this section, we introduce VAST, a data-efficient selection framework that organizes training data into a hierarchical structure. We first introduce the three-layer taxonomy (Sec. 3.1), then describe how we construct VAST-15K using this taxonomy (Sec. 3.2), then analyze the resulting dataset quality (Sec. 3.3), and finally present VAST-Bench for capability-level evaluation (Sec. 3.4).

3.1. VAST Taxonomy

Piaget’s cognitive development theory [45, 46] proposes that intelligence develops through qualitatively distinct stages, each building upon the previous one. This hierarchical progression closely parallels video understanding capabilities: models first require perceptual grounding to reason about relationships, and abstracting general principles further depends on well-established relational understanding. This dependency structure motivates the three-layer VAST taxonomy design, as illustrated in Figure 2.

Layer 1 - Perception (inspired by the perceptual stage, 0–2 years). In this stage, infants learn about the world through direct sensory input and motor actions. A key achievement is object permanence, the understanding that an object continues to exist even when it is not visible. This layer corresponds to foundational video skills and focuses on two core abilities. *Spatial entity perception* analyzes a scene at a single moment through object recognition, localization, pose, and layout. *Spatio-temporal dynamics understanding* captures changes over time by tracking identities, attributes, and actions, and by recognizing event boundaries. Together, these abilities answer the question of **what we see**.

Layer 2 - Reasoning (inspired by the concrete stage, 2–11 years). In this stage, children develop symbolic thought and begin to reason logically about concrete events, grasping concepts such as classification and conservation. This layer reflects the ability to move beyond static perception toward interpreting interactions and dependencies. It encompasses four types of reasoning. (i) *Causal reasoning* uncovers physical or logical chains of events. (ii) *Intentional reasoning* infers goals and plans from behavior. (iii) *Procedural reasoning* understands multi-step sequences and conditional dependencies. (iv) *Relational reasoning* recognizes roles, status, and social structures emerging from interactions. Altogether, this layer answers the question of **why it is happening**.

Layer 3 - Cognition (inspired by the abstract stage, 11 years+). In this stage, adolescents develop the capacity for abstract and hypothetical reasoning, enabling them to think beyond what is directly observed. They can identify general principles, imagine alternative scenarios, and reason systematically. This layer represents the ability to move beyond concrete reasoning toward abstract video comprehension. It includes four types of tasks. (i) *Rule discovery* identifies underlying patterns and detects exceptions. (ii) *Symbol comprehension* interprets figurative and cultural meanings. (iii) *Theme inference* captures the central idea or message conveyed by a storyline. (iv) *Expression analysis* understands how video elements like color, layout, and camera movement convey emotion, mood, and style. Collectively, this layer answers the question of **what it means**.

3.2. VAST-15K Construction

To construct a capability-balanced training set, we curate VAST-15K from existing large-scale video reasoning datasets, primarily Video-R1-260K [7]. The construction process consists of three stages (Figure 3 (left)). First, we classify each sample into capability layers using the VAST taxonomy. Second, we filter out trivial or ambiguous instances based on difficulty assessment. Third, we apply stratified selection to achieve balanced representation across all capability levels.

Classification Process. We classify each sample into one of ten sub-abilities across three capability layers based on the required reasoning skills. We employ a two-stage annotation process. In the first stage, we use multiple state-of-the-art models, including Qwen2.5-VL-72B [1], GPT-4o [22], and Gemini 2.5 Pro [6], as automatic classifiers for capability assessment. Each model is provided with the VAST taxonomy definitions and asked to assign every sample to one of the ten sub-abilities; the full prompt is in the supplement. To enhance robustness, the models are required to generate explicit evidence supporting their predictions, and their outputs are aggregated through majority voting. Labels with fully consistent voting results are regarded as re-

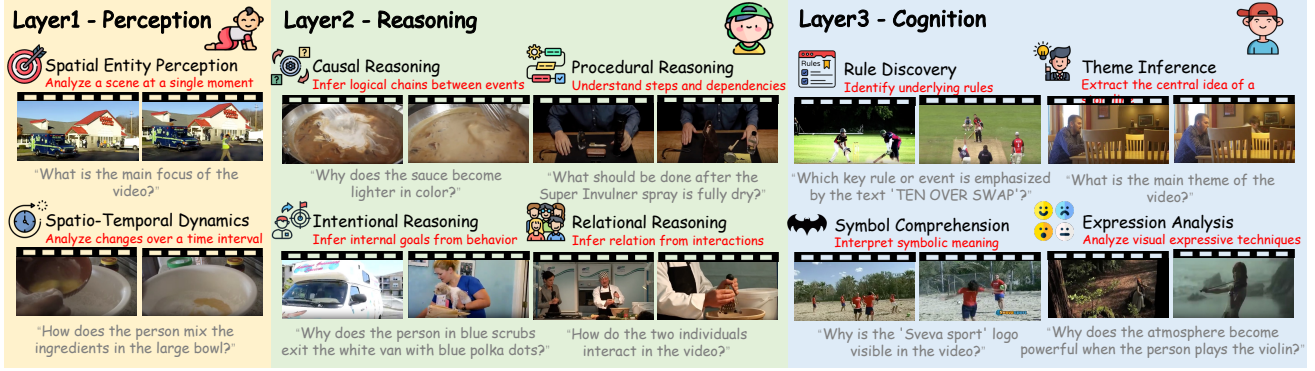


Figure 2. **Cognitive taxonomy in VAST.** VAST organizes video understanding into three cognitive layers: Perception, Reasoning, and Cognition, which are further divided into ten fine-grained abilities. Layer 1 (Perception) captures spatial entity perception and spatio-temporal dynamics. Layer 2 (Reasoning) covers causal, intentional, procedural, and relational reasoning. Layer 3 (Cognition) includes rule discovery, symbol comprehension, theme inference, and expression analysis.

liable. In the second stage, we manually review a subset of samples to verify the correctness of both the assigned labels and their associated evidence, ensuring annotation quality and interpretability. Across all samples, 78.3% receive fully consistent labels (3/3 agreement) and 94.2% reach majority agreement ($\geq 2/3$); we further manually audited $\sim 6\%$ of samples and applied corrections to $\sim 5\%$ of the audited subset.

Difficulty-Based Filtering. We apply a difficulty-based filtering procedure to exclude samples that are either insufficiently challenging or excessively ambiguous for effective learning. We use the smaller Qwen2.5-VL-7B model for this assessment, as it provides a more practical difficulty proxy aligned with typical model capabilities. Each sample is evaluated 16 times with the same model. Samples with perfect consistency across all evaluations are discarded. Specifically, samples where the model answers correctly in all 16 attempts are considered trivial, while samples with all 16 incorrect answers are deemed ambiguous or unanswerable. We retain samples with mixed results, as they represent an appropriate difficulty level that encourages meaningful learning.

Balanced Selection. To mitigate the pronounced capability imbalance in existing datasets, we apply stratified selection to achieve balanced representation. We select 5K instances from each of the three layers (Perception, Reasoning, and Cognition) to maintain equal proportions. Within each layer, we further balance across sub-abilities to ensure comprehensive coverage. Detailed statistics are in the supplement.

3.3. VAST-15K Analysis

To validate the quality of VAST-15K, we compare it with the representative dataset Video-R1-260K across two key metrics: *type-token ratio* (TTR) and *text length distribution*.

TTR measures lexical diversity by computing the ratio of unique words to total words. A higher TTR reflects greater diversity in question formulations and task descriptions, indicating that the dataset covers a broader range of reasoning scenarios instead of repetitive patterns. Figure 4 (a) shows that VAST-15K exhibits a TTR of 8.93 compared to 7.17 for Video-R1-260K, demonstrating that our capability-based curation captures more varied reasoning challenges. Figure 4 (b) shows that the text length distribution of Video-R1-260K [7] is heavily concentrated in the 0-200 token range with a long-tail pattern, while VAST-15K shifts toward longer texts with increased coverage in the 200-400 token range. This shift indicates that VAST-15K contains more complex reasoning problems requiring richer descriptions, particularly for developing Reasoning and Cognition capabilities.

3.4. VAST-Bench

To evaluate capabilities across the cognitive hierarchy, we introduce VAST-Bench, an evaluation suite that extends existing benchmarks with capability-level annotations.

Benchmark Construction. VAST-Bench aggregates samples from established video understanding benchmarks including VideoMME [9], TempCompass [39], MVBench [31], and MMVU [70]. We apply the same annotation protocol used for VAST-15K construction (Sec. 3.2) to classify each sample according to the VAST taxonomy. To ensure balanced evaluation, we select 300 samples per capability layer, resulting in 900 samples in total.

Evaluation Metrics. VAST-Bench reports performance across three hierarchical levels. At the finest granularity, we report accuracy for each of the ten sub-abilities (two under Perception, four under Reasoning, and four under Cognition). Additionally, we report an overall score averaged across all layers.

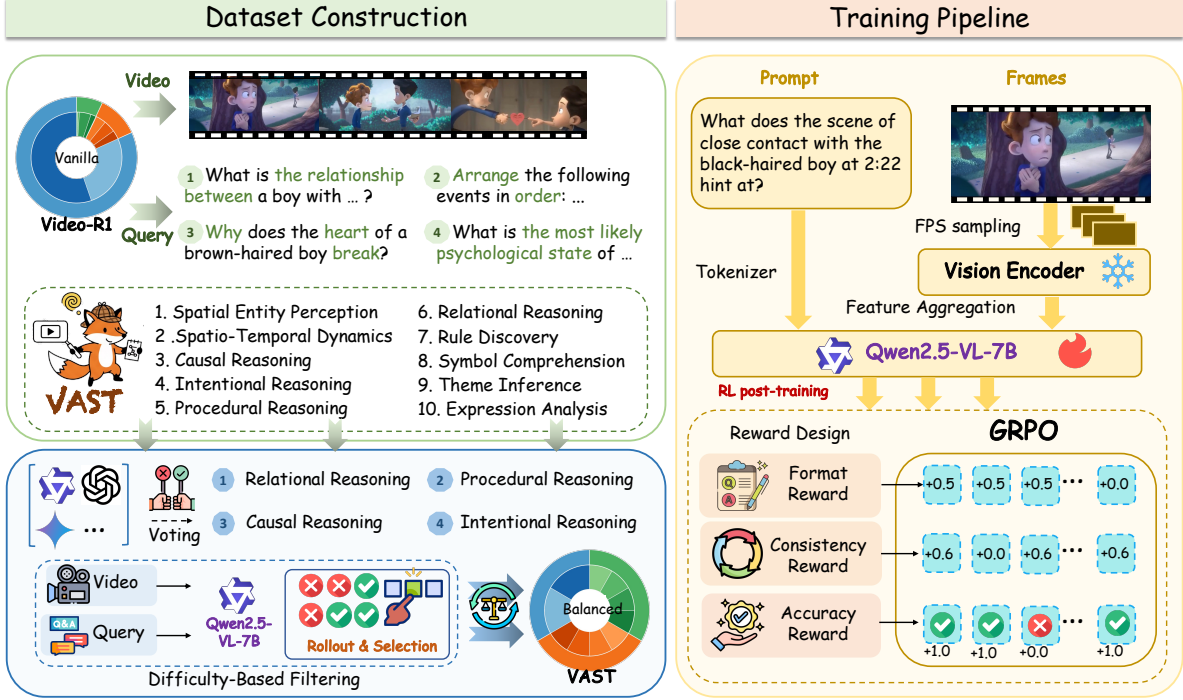


Figure 3. **VideoVAST training pipeline guided by VAST.** Left: ability-stratified dataset construction with capability classification, difficulty-based filtering, and balanced selection. Right: reinforcement learning with GRPO and a consistency reward applied to reasoning traces.

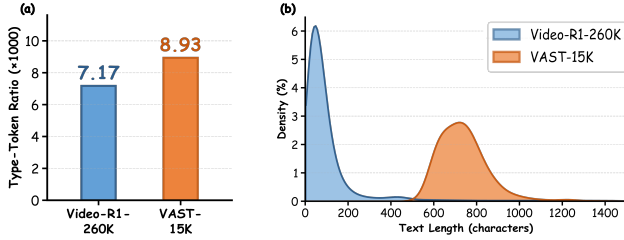


Figure 4. **Comparison of VAST-15K and Video-R1-260K.** (a) Type-token ratio (TTR), where higher values indicate greater lexical diversity. (b) Distribution of answer lengths (in characters).

Design Advantages. Existing video benchmarks typically report aggregate scores or task-specific metrics without exposing capability-level performance patterns. VAST-Bench provides a unified evaluation framework grounded in cognitive development theory. This design enables researchers to identify whether improvements stem from enhanced perception, reasoning, or cognition, facilitating more targeted model development. Moreover, the three-level hierarchy supports both coarse-grained comparison across cognitive layers and fine-grained analysis within specific sub-abilities, enabling interpretable progress measurement and targeted capability diagnosis.

4. VideoVAST

Based on the VAST-15K dataset constructed in Sec. 3.2, we propose VideoVAST, a data-efficient reinforcement learning framework for video reasoning. Figure 3 (right) illustrates our overall training strategy. We introduce our RL algorithm in Sec. 4.1, followed by our *consistency-regularized reward design* in Sec. 4.2.

4.1. Group Relative Policy Optimization

We adopt Group Relative Policy Optimization (GRPO) [48] as our reinforcement learning algorithm, following prior work on RL with rule-based rewards. GRPO replaces the critic network with a group-based relative comparison, where the model generates G responses for each input and computes their relative advantages.

For each query x with responses $\{y^{(1)}, y^{(2)}, \dots, y^{(G)}\}$, the relative advantage of the i -th response is:

$$A^{(i)} = \frac{r^{(i)} - \text{mean}(\{r^{(j)}\}_{j=1}^G)}{\text{std}(\{r^{(j)}\}_{j=1}^G)}, \quad (1)$$

where $r^{(i)}$ is the reward for $y^{(i)}$. We then update the policy using a PPO-style clipped objective with a KL penalty. We perform token-level loss aggregation over all responses and

tokens following VERL [49]:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q, \{o_i\}} \frac{1}{G} \sum_{i=1}^G \left[\min \left(\frac{\pi_{\theta}(o_i | q)}{\pi_{\theta_{\text{old}}}(o_i | q)} A^{(i)}, \right. \right. \\ \left. \left. \text{clip} \left(\frac{\pi_{\theta}(o_i | q)}{\pi_{\theta_{\text{old}}}(o_i | q)}, 1 - \epsilon, 1 + \epsilon \right) A^{(i)} \right) \right. \\ \left. - \beta \mathbb{D}_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right]. \quad (2)$$

This formulation enables learning from the *relative quality* of responses within each group, favoring better-performing trajectories without requiring explicit value function estimation.

While KL regularization provides theoretical stability guarantees, it presents challenges for video reasoning tasks. The reference model π_{ref} may not capture the diversity of valid reasoning paths in video understanding, where the same question can be approached through multiple cognitive routes. Standard GRPO includes a KL divergence term $\beta \mathbb{D}_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}})$ to prevent excessive deviation from the reference distribution. We set $\beta = 0$ and introduce a consistency reward R_{cr} (Sec. 4.2) that encourages structural alignment between reasoning and answers rather than constraining the policy distribution. This design choice is validated by supplementary ablations, where the consistency reward outperforms traditional KL regularization.

4.2. Consistency-regularized reward

To maintain training stability without KL regularization, we introduce a *consistency-regularized reward* as a lightweight structural constraint. Rather than constraining the policy distribution, our *consistency reward* evaluates whether the reasoning process logically supports the final prediction. Our reward function comprises three components: a format reward R_{format} that enforces structured reasoning within `<think>` tags, an accuracy reward R_{acc} based on answer correctness, and a *consistency reward* R_{cr} that provides a bonus for reasoning-answer alignment.

The total reward is defined as:

$$R = R_{\text{format}} + R_{\text{acc}} + R_{\text{cr}}, \quad (3)$$

where $R_{\text{format}} = 0.5$ if the response is properly formatted within `<think>` tags and 0 otherwise, $R_{\text{acc}} = 1.0$ for correct answers and 0 otherwise, and $R_{\text{cr}} = \lambda \times s_{\text{consistency}}$. The consistency signal $s_{\text{consistency}} \in \{0, 1\}$ indicates whether intermediate reasoning aligns with the final prediction and is computed via an external consistency model detailed in the supplement. We set $\lambda = 0.7$ based on supplementary validation.

With GRPO’s relative advantage mechanism, this reward design establishes a preference ranking where consistent correct responses ($R = 2.2$) receive the highest reinforcement, followed by inconsistent correct responses ($R = 1.5$),

consistent incorrect responses ($R = 1.2$), and inconsistent incorrect responses ($R = 0.5$). The dominance of R_{acc} prevents the policy from promoting self-consistent errors, while the consistency term differentiates among correct responses based on reasoning quality.

5. Experiment

5.1. Benchmarks

We evaluate our model on seven video benchmarks covering spatial reasoning, knowledge-intensive video question answering, temporal logic, and general video comprehension. These benchmarks include VideoMMMU [20], MMVU [70], MVBench [31], TempCompass [39], VSI-Bench [61], VideoMME [9], and Video-Holmes [3]. Following standard evaluation protocols, we use the multiple-choice subset of MMVU and evaluate VideoMME without subtitles. We report average accuracy across all benchmarks.

5.2. Implementation Details

We build VideoVAST on Qwen2.5-VL-7B [1] as the base model for video understanding. For scalability and reproducibility, we implement the reinforcement learning stage using the VERL framework. During training, we sample 16 frames of resolution $256 \times 28 \times 28$, where 28×28 is the size of each image patch and 256 denotes the number of patches. We use Qwen2.5-72B [1] as the consistency model for computing the consistency reward, served via vLLM [26]. For this stage, we employ the AdamW optimizer with an initial learning rate of 1×10^{-6} and a batch size of 128. We train for 2 epochs with 8 rollout responses generated per prompt for advantage estimation.

5.3. Main Results

Using 16 frames, VideoVAST-7B achieves the highest Video-Holmes score (42.5) among all evaluated models (Table 1). Compared to Video-R1-7B under the same frame setting, VideoVAST improves performance across most benchmarks, including MVBench (+3.6), Video-Holmes (+2.4), MMVU (+1.9), and TempCompass (+1.6). Compared to the Qwen2.5-VL-7B base model, VideoVAST achieves gains of +8.9 on MVBench, +2.8 on VideoMMMU, and +5.8 on VSI-Bench through ability-stratified RL training.

On VAST-Bench (Table 2), which provides balanced coverage across Perception, Reasoning, and Cognition tasks, VideoVAST achieves 57.4 overall, outperforming Video-R1 (54.3) and VideoRFT (55.3). Performance gains are particularly strong on Reasoning tasks, with +8.7 on Causal Reasoning and +7.7 on Relational Reasoning compared to Video-R1. VideoVAST also excels in Cognition tasks, achieving the best performance on Theme Inference

Table 1. **Performance comparison on general-purpose video understanding benchmarks.** The best results are highlighted in **bold**, and the second-best results are underlined.

Models	Pub.	RL Only	Frames	Video MMU	MMVU (mc)	MV Bench	Temp Compass	VSI-Bench	VideoMME (wo sub)	Video Holmes
GPT-4o [22]	–	-	-	61.2	75.4	-	-	34.0	71.9	42.0
LLaMA-VID [35]	ECCV’24	✗	-	-	-	41.9	45.6	-	-	-
VideoLLaMA 2 [4]	arXiv’24.06	✗	-	-	44.8	54.6	-	-	47.9	-
LongVA-7B [67]	arXiv’24.06	✗	-	23.9	-	-	56.9	29.2	52.6	-
VILA-1.5-8B [36]	CVPR’24	✗	-	20.8	-	-	58.8	28.9	-	-
VILA-1.5-40B [36]	CVPR’24	✗	-	34.0	-	-	-	31.2	60.1	-
Video-UTR-7B [64]	ICLR’25	✗	-	-	-	58.8	59.7	-	52.6	-
LLaVA-OneVision-7B [30]	arXiv’24.08	✗	-	33.8	49.2	56.7	-	32.4	58.2	-
Kangaroo-8B [37]	arXiv’24.08	✗	-	-	-	61.1	62.5	-	56.0	-
Qwen2.5-VL-7B (CoT) [1]	arXiv’25.02	✗	16	47.8	59.2	57.4	72.2	27.7	53.1	27.8
Qwen2.5-VL-7B (CoT) [1]	arXiv’25.02	✗	32	48.1	60.0	59.0	72.6	30.1	56.6	-
Qwen2.5-VL-7B (CoT) [1]	arXiv’25.02	✗	64	<u>50.4</u>	60.0	59.2	72.9	31.4	<u>59.6</u>	-
Video-R1-7B [7]	NeurIPS’25	✗	16	49.8	64.2	62.7	72.6	34.6	57.4	<u>40.1</u>
Video-R1-7B-zero [7]	NeurIPS’25	✓	16	49.5	63.8	60.4	70.9	31.8	53.8	-
Video-RFT-7B [55]	NeurIPS’25	✗	16	48.2	<u>66.0</u>	<u>64.0</u>	<u>73.5</u>	34.6	55.9	36.4
VideoVAST-7B (Ours)	–	✓	16	50.6	66.1	66.3	74.2	33.5	57.7	42.5

(70.3) and Expression Analysis (66.7, +6.7 over Video-R1). These results demonstrate that ability-stratified training enhances abstract reasoning capabilities across both causal logic and high-level comprehension.

Training efficiency. VideoVAST achieves significant efficiency gains through ability-stratified data organization. Compared to Video-R1, VideoVAST reduces training data from 425K to 15K samples (96% reduction) and GPU hours from approximately 440 to 123 (72% reduction) while maintaining competitive performance. This efficiency stems from two key factors: (i) eliminating the supervised fine-tuning initialization stage entirely by training directly with reinforcement learning on VAST-15K, and (ii) achieving balanced capability coverage with significantly fewer samples through strategic curation. Similarly, compared to VideoRFT’s 412K training samples, VideoVAST requires 96% fewer data points. These results demonstrate that ability-driven data organization outperforms massive scaling with task-centric datasets. See supplement for details.

5.4. Ablation Studies

We ablate three key factors using VideoVAST-7B. We vary the number of input frames, SFT initialization, and consistency-regularized reward (CR). Results are summarized in Table 3.

Reinforcement learning without supervised initialization. Ability-stratified data enables direct reinforcement

Table 2. **Performance comparison on VAST-Bench.** Columns correspond to perception (SEP: Spatial Entity Perception; STD: Spatio-Temporal Dynamics), reasoning (CR: Causal Reasoning; IR: Intentional Reasoning; PR: Procedural Reasoning; RR: Relational Reasoning), and cognition (RD: Rule Discovery; SC: Symbol Comprehension; TI: Theme Inference; EA: Expression Analysis).

Models	Perception		Reasoning				Cognition				Overall
	SEP	STD	CR	IR	PR	RR	RD	SC	TI	EA	
Video-R1 [7]	54.6	64.2	55.6	47.5	42.5	38.5	48.1	<u>62.3</u>	<u>69.5</u>	60.0	54.3
VideoRFT [55]	<u>53.9</u>	64.2	<u>55.7</u>	49.5	49.3	38.5	53.7	58.0	67.2	<u>63.3</u>	<u>55.3</u>
VideoVAST (Ours)	56.0	60.4	64.3	47.5	49.3	46.2	<u>51.9</u>	62.4	70.3	66.7	57.4

learning without supervised fine-tuning initialization, simplifying the training pipeline. Table 3 shows that training with reinforcement learning alone achieves competitive performance across all frame settings. At 16 frames, reinforcement learning with consistency reward (55.8) outperforms supervised fine-tuning followed by reinforcement learning (55.2), demonstrating that supervised initialization is unnecessary when data are organized by cognitive capabilities. This eliminates the supervised fine-tuning stage entirely, saving both data annotation costs and computational overhead. In contrast, Video-R1 and VideoRFT require 165K and 102K annotated CoT samples for supervised fine-tuning initialization, as shown in the supplement. Our results demonstrate that strategic data curation can replace

costly multi-stage training pipelines.

Table 3. Ablation study of supervised fine-tuning (SFT) and consistency-regularized reward (CR) under different numbers of frames used at inference.

Frames	SFT	CR	Video MMMU	MMVU (mc)	MV Bench	Temp Compass	VSI- Bench	VideoMME (wo sub)	Video Holmes	Overall
16	✗	✗	49.1	66.1	65.2	74.7	34.8	57.2	41.5	55.5
	✓	✗	47.3	63.0	66.0	73.0	33.5	57.2	39.8	54.3
	✗	✓	50.6	66.1	66.3	74.2	33.5	57.7	42.5	55.8
	✓	✓	49.9	63.2	65.5	73.0	36.5	58.0	40.4	55.2
32	✗	✗	48.7	65.4	66.6	73.2	37.4	60.6	38.6	55.8
	✓	✗	49.9	62.4	66.9	73.5	35.0	59.9	37.7	55.0
	✗	✓	51.2	67.0	67.4	74.9	34.3	59.4	39.7	56.3
	✓	✓	50.1	63.0	66.9	73.6	37.6	60.8	41.6	56.2
64	✗	✗	50.3	66.9	66.8	73.1	38.8	62.2	38.6	56.7
	✓	✗	50.1	62.6	68.2	73.4	37.1	61.6	39.7	56.1
	✗	✓	51.4	67.0	68.2	74.7	36.5	62.4	37.0	56.7
	✓	✓	50.6	62.7	68.0	73.8	39.5	62.1	39.2	56.6

Consistency-regularized reward. The consistency-regularized reward (CR) component improves performance for both SFT and RL models across most frame settings. Comparing rows with and without CR in Table 3, we observe that CR provides consistent gains at 16 and 32 frames, with notable improvements on Video-Holmes (up to +3.9 at 32 frames). Data stratification alone does not explicitly enforce reasoning quality. With CR, the model learns to reason coherently, as the reward enforces alignment between intermediate reasoning and final predictions, which proves beneficial for complex reasoning tasks.

Data scale and curation. Table 4 examines dataset scale and curation strategy. Scaling from 1.5K to 15K samples improves results on benchmarks such as VideoMMMU (+2.4) and MVBench (+1.2), with the overall score rising from 53.5 to 55.8. The Random 15K baseline performs substantially worse than VAST-15K (47.6 vs. 55.8 overall), with a large gap on MMVU (34.4 vs. 66.1).

Curriculum learning. We investigate several training strategies for ability-stratified data, including Sequential, Two-Stage, Smooth, and Uniform joint training. Among them, Uniform joint training achieves the best overall performance (55.8). See the supplementary material for details.

Table 4. Performance comparison across VAST data scales and curation strategies.

VAST Scale	Video MMMU	MMVU (mc)	MV Bench	Temp Compass	VSI- Bench	VideoMME (wo sub)	Video Holmes	Overall
VAST 1.5K	48.2	63.0	65.1	73.0	34.7	55.0	35.3	53.5
VAST 5K	<u>49.4</u>	<u>65.6</u>	<u>65.9</u>	<u>73.8</u>	33.2	<u>57.0</u>	<u>37.8</u>	<u>54.7</u>
VAST 15K	50.6	66.1	66.3	74.2	<u>33.5</u>	57.7	42.5	55.8
Random 15K	48.9	34.4	63.1	73.8	22.4	56.3	34.7	47.6

5.5. Further Analysis of VAST

Q1. Does ability stratification account for the gains?

To isolate the effect of ability-stratified curation from data scale, we compare VAST-15K with Random 15K under

identical training conditions. VAST-15K substantially outperforms Random 15K (55.8 vs. 47.6 overall, Table 4), with especially large gains on reasoning-intensive benchmarks such as MMVU (66.1 vs. 34.4) and Video-Holmes (42.5 vs. 34.7). Because the two sets have the same size but different capability distributions, this comparison highlights the importance of data organization rather than scale alone. VAST-15K is balanced across the three layers, whereas Random 15K remains heavily skewed toward Perception (78%, versus 15% Reasoning and 7% Cognition). As a result, Random 15K biases training toward perceptual shortcuts, while VAST-15K provides more uniform learning signals across capability levels. The same trend persists across training configurations (Table 3), supporting the conclusion that the gains are driven by ability stratification.

Q2. Does VAST reduce shortcut reliance? Supplementary qualitative cases further illustrate this trend. Across tasks including violence identification, symbolic interpretation, motivation inference, and narrative understanding, VideoVAST consistently relies on temporally and contextually grounded evidence rather than surface cues. For example, it infers harmful intent instead of treating a violent scene as an accident, preserves narrative coherence in symbolic interpretation, reasons about the shooter’s cover-up intention, and abstracts thematic meaning beyond local visual evidence. These cases suggest that ability-stratified training, together with the consistency reward, encourages cross-frame reasoning and more coherent causal interpretation. As a result, the model becomes less reliant on shortcuts such as surface feature matching and contextual misalignment.

6. Conclusion

We presented VAST, a cognitive taxonomy and data organization framework that structures video understanding into three layers: Perception, Reasoning, and Cognition. Building on this taxonomy, we constructed VAST-15K for training, introduced VAST-Bench for capability-level evaluation, and developed VideoVAST, a reinforcement learning framework with a consistency reward. Our results show that ability-stratified data organization enables strong video reasoning performance with substantially lower training cost, reducing GPU hours by 72% and training samples by 96% relative to prior methods. These results suggest that efficient video reasoning depends not only on model or reward design, but also on whether the training data provide balanced coverage across cognitive capabilities. They also highlight the value of capability-level evaluation, which helps reveal where improvements arise and where current models still fall short. We hope VAST provides a useful foundation for future work on data-efficient video understanding and reasoning.

Acknowledgments

This work was supported in part by “Pioneer” and “Leading Goose” R&D Program of Zhejiang (No. 2025C02032), the National Natural Science Foundation of China (92570101, 62502447), the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant No.25GA3200200 and Shanghai Artificial Intelligence Laboratory.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 6, 7
- [2] Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36:36000–36040, 2023. 2
- [3] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. 2, 6
- [4] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 7
- [5] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025. 2, 3
- [6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3
- [7] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-rl: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 1, 2, 3, 4, 7
- [8] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025. 2
- [9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 2, 4, 6
- [10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datcomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 2
- [11] Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115*, 2024. 2
- [12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2
- [13] Madeleine Grundle-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 2
- [14] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6238–6252, 2024. 2
- [15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-rl incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 1, 3
- [16] Aref Hakimzadeh, Yanbo Xue, and Peyman Setoodeh. Interpretable reinforcement learning inspired by piaget’s theory of cognitive development. *arXiv preprint arXiv:2102.00572*, 2021. 2
- [17] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022. 2
- [18] Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Transformation driven visual reasoning. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 6903–6912, 2021. 2
- [19] Wenjin Hou, Wei Liu, Han Hu, Xiaoxiao Sun, Serena Yeung-Levy, and Hehe Fan. Seeing is believing? a benchmark for multimodal large language models on visual illusions and anomalies. *arXiv preprint arXiv:2602.01816*, 2026. 2
- [20] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. 6
- [21] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-rl: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 1, 3
- [22] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3, 7

- [23] Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, et al. Storm: Token-efficient long video understanding for multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5830–5841, 2025. 2
- [24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2
- [25] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025. 3
- [26] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023. 6
- [27] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. 2
- [28] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1369–1379, 2018. 2
- [29] Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507, 2023. 2
- [30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 7
- [31] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2, 4, 6
- [32] Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. Mmad: Multi-label micro-action detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13225–13236, 2025. 2
- [33] Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan S Kankanhalli. Improving context understanding in multimodal large language models via multimodal composition learning. In *ICML*, page 7, 2024. 2
- [34] Xuchen Li, Xuzhao Li, Shiyu Hu, Kaiqi Huang, and Wentao Zhang. Causalstep: A benchmark for explicit stepwise causal reasoning in videos. *arXiv preprint arXiv:2507.16878*, 2025. 2
- [35] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 7
- [36] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024. 7
- [37] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoli Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024. 7
- [38] Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29139–29149, 2025. 2
- [39] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 2, 4, 6
- [40] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*, 2023. 2
- [41] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 1, 3
- [42] Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See Kiong Ng, and Luu Anh Tuan. Video-language understanding: A survey from model architecture, model training, and data perspectives. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3636–3657, 2024. 3
- [43] Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyi-dan Xie, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025. 1
- [44] 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 llms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025. 3
- [45] Jean Piaget. *The origins of intelligence in children*. International Universities Press New York, 1952. 2, 3
- [46] Jean Piaget. The role of action in the development of thinking. In *Knowledge and Development: Volume 1 Advances in Research and Theory*, pages 17–42. Springer, 1977. 2, 3

- [47] Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106, 2025. 2
- [48] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 5
- [49] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025. 6
- [50] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025. 3
- [51] 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. *arXiv preprint arXiv:2501.12599*, 2025. 1, 3
- [52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2
- [53] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 2717–2739, 2023. 2
- [54] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 2
- [55] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorf: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025. 1, 2, 3, 7
- [56] Zhihu Wang, Shivan Zhao, Yu Wang, Heyuan Huang, Sitao Xie, Yubo Zhang, Jiaxin Shi, Zhixing Wang, Hongyan Li, and Junchi Yan. Re-task: Revisiting llm tasks from capability, skill, and knowledge perspectives. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4925–4936, 2025. 2
- [57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2
- [58] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024. 2
- [59] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2
- [60] Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17675–17687, 2025. 2
- [61] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 2, 6
- [62] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025. 2
- [63] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 2
- [64] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025. 3, 7
- [65] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuet-ing Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 2
- [66] Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024. 1, 2
- [67] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 7
- [68] Yue Zhang, Hehe Fan, Wei Ji, Yongkang Wong, Roger Zimmermann, and Yi Yang. Prompt-aware adapter: Learning adaptive visual tokens for multimodal large language models. *IEEE Transactions on Artificial Intelligence*, 2025. 2
- [69] Fufangchen Zhao, Ming Li, Linrui Xu, Wenhao Jiang, Jian Gao, and Danfeng Yan. Favchat: Unlocking fine-grained facial video understanding with multimodal large language models. *arXiv preprint arXiv:2503.09158*, 2025. 2
- [70] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025. 2, 4, 6
- [71] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025. 3