TTA-VID: GENERALIZED TEST-TIME ADAPTATION FOR VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent video reasoning models have shown strong results on temporal and multimodal understanding, yet they depend on large-scale supervised data and multi-stage training pipelines, making them costly to train and difficult to adapt to new domains. In this work, we leverage the paradigm of Test-Time Reinforcement Learning on video-language data to allow for adapting a pretrained model to incoming video samples at test-time without explicit labels. The proposed test-time adaptation for video approach (TTA-Vid) performs step-by-step reasoning at inference time on multiple frame subsets. We then use a batch-aware frequency-based reward computed across different frame subsets as pseudo ground truth to update the model. It shows that the resulting model trained on a single batch or even a single sample from a dataset, is able to generalize at test-time to the whole dataset and even across datasets. Because the adaptation occurs entirely at test time, our method requires no ground-truth annotations or dedicated training splits. Additionally, we propose a multi-armed bandit strategy for adaptive frame selection that learns to prioritize informative frames, guided by the same reward formulation. Our evaluation shows that TTA-Vid yields consistent improvements across various video reasoning tasks and is able to outperform current state-of-the-art methods trained on large-scale data. This highlights the potential of test-time reinforcement learning for temporal multimodal understanding. [1]

1 INTRODUCTION

Understanding the content of long videos and reasoning over it remains a fundamental challenge in video comprehension and multimodal learning. Recent progress in large Vision Language Models (VLMs) such as the InternVL (Chen et al., 2024; Zhu et al., 2025; Wang et al., 2025a) or QwenVL (Bai et al., 2023; 2025b) series, and video reasoning models such as Video-R1 (Feng et al., 2025), VideoRTS (Wang et al., 2025d), Video-RFT (Wang et al., 2026), VideoChat-R1 (Li et al., 2025a), VideoChat-R1.5 (Yan et al., 2026), and Video-R2 (Maaz et al., 2025) has brought remarkable advances in e.g. captioning or question answering tasks. However, when applied to lengthy, highly structured, and conceptually rich videos, these generic MLLMs struggle to produce coherent reasoning and accurate answers. Instructional videos have become particularly important in evaluating MLLM performance in this domain as they often feature visual and verbal content that is inherently structured over time, with one information building up on another. Building models that can capture such content provides a compelling test of whether AI systems can capture relevant information over time and reason and generalize over it. Despite their potential, existing video reasoning models face critical limitations in this context, as they typically adopt multi-stage pipelines that involve supervised finetuning followed by reinforcement learning or reward model optimization. These procedures are not only computationally intensive, especially in the case of video, but can also limit the model's ability to adapt to new domains without retraining. Second, as those models need a significant amount of training data, they leverage large-scale datasets such as ActivityNet-QA (Yu et al., 2019), which often contain shorter clips. As a result, the models struggle to capture longer, more complex reasoning structures, as can e.g. be found in real-world content. To address those limitations, we propose a test-time adaptation framework for video reasoning. Unlike few-shot learning, which relies on a small labeled support set, test-time adaptation operates entirely on unlabeled data without any annotations.

[1] All code, data, and checkpoints will be made available.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

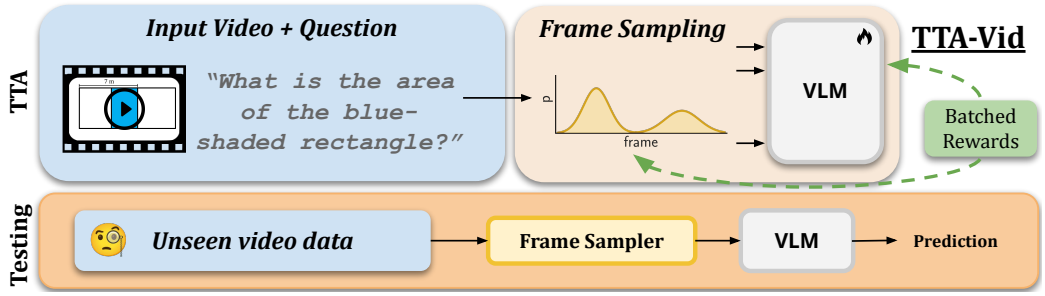


Figure 1: TTA-Vid adapts an MLLM at test-time by sampling multiple frame subsets, enforcing majority-consistency among generated answers, and updating a frame-importance distribution via a multi-armed bandit. This allows a test-time adaptation without labels.

Test-time Adaptation for Video: Given an instructional video and a question, the proposed method samples multiple sets of frames and generates multiple candidate reasoning traces and answers. We compute a batch-wide frequency reward based on the empirical probability of each answer for all generated outputs, combined with an entropy-based confidence penalty, allowing the model to reinforce frequently occurring answers while suppressing high-uncertainty generations. While related test-time RL methods such as TTRL (Zuo et al., 2025) and TTRV (Singh et al., 2026) target language models and single-image tasks, respectively, TTA-Vid exploits a fundamental property of video: videos contain far more frames than a model can process at once. Sampling different frame subsets naturally produces diverse rollouts with different visual evidence, and the resulting reward signal can be leveraged to learn which frames contribute most. An overview is shown in Figure 1.
Generalized Test-time Adaptation: We found that models adapted on a single batch generalize to the full dataset and even across datasets. Thus, training on a single batch of 32 unlabeled samples is sufficient to improve performance on the entire unseen test set, and the gains transfer across datasets and even improve compared to a test-time training on every batch. This indicates that the model acquires transferable reasoning capabilities rather than overfitting to dataset-specific patterns.
Adaptive Frame Selection: We introduce a frame distribution learning mechanism based on the multi-armed bandit formulation. Instead of relying on fixed frame sampling, our method learns a distribution that assigns importance scores to frames based on their contribution to successful reasoning. This distribution is optimized using test-time reinforcement learning signals derived from a batch-wide frequency reward that combines empirical answer frequencies with an entropy-based confidence penalty, gradually learning which frames matter most for answering each question. As test-time training progresses, the model converges toward a video-specific sampling policy, highlighting frames that are semantically important. This frame selection process benefits the final systems in two ways: (i) the model can focus on better frames during the training can help to generate better rollouts and thus more traces relevant for the reward, and (ii) the learned sampling mechanism generalizes beyond the adaptation batch. Rather than memorizing video-specific frame locations, the averaged distribution over adaptation batch acts as a coarse temporal prior that captures general information-density patterns across videos. As a result, the temporal prior transfers to unseen test samples, outperforming standard baselines such as uniform and CLIP-based matching.

We evaluate the proposed method on five benchmarks, VideoMMMU (Hu et al., 2026), MMVU (Zhao et al., 2025), SciVideoBench (Deng et al., 2026), VideoMME (Fu et al., 2025), and LongVideoBench (Wu et al., 2024), and two distinct vision-language backbones, InternVL-3 (Zhu et al., 2025) and Qwen2.5-VL (Bai et al., 2025b). The evaluation shows that the proposed test-time adaptation approach consistently improves answer consistency and is able to outperform current methods trained on large-scale ground-truth data based on only a single batch of 32 test samples without any labels. It further shows that the improvement is not only limited to the respective test dataset, but also extends to unseen test data, allowing for a generalized test-time adaptation.

We summarize the contributions as follows: (a) TTA-Vid: A test-time reinforcement learning framework that adapts vision-language models to video reasoning tasks without any labeled data or pre-generated reasoning traces, yet outperforms video reasoning models trained on large-scale supervised datasets. (b) A multi-armed bandit strategy that learns frame distributions for efficient and interpretable reasoning using the test-time training reward signals. (c) An extensive evaluation that focuses on generalized test-time adaptation across multiple instructional video question-answering datasets and model backbones, showing state-of-the-art performance even compared to models trained on large-scale data.

2 Related Work

Video Reasoning Models. The success of large language models (LLMs) has motivated the extension of their capabilities to multimodal tasks (Brown et al., 2020; Team, 2024; Luo et al., 2024; Ouyang et al., 2022; Touvron et al., 2023; Chiang et al., 2023; Radford et al., 2019) . This lead to the emergence of vision-language models allowing models to consider visual content as well as video models which interpret dynamic visual content (Huang et al., 2023; Driess et al., 2023; Liu et al., 2023; Bai et al., 2023; Li et al., 2023; 2022; Zhu et al., 2025; Bai et al., 2025b; An et al., 2025; Zhang et al., 2024c; Li et al., 2024a; Wang et al., 2024; Bai et al., 2025a; Zhang et al., 2024a). However, most models like LLaMA-VID (Li et al., 2024b), VideoLLaMA2 (Cheng et al., 2024), LongVA (Zhang et al., 2024b), VISA (Yan et al., 2024) among others focus on video perception tasks. On the other hand, works inspired by reasoning in language models (Guo et al., 2025; Jaech et al., 2024; Shao et al., 2024; Liu et al., 2025; Yu et al., 2025; Team et al., 2025) such as (Xu et al., 2025; Thawakar et al., 2025; Deng et al., 2025; Yang et al., 2025; Chen et al., 2025; Shen et al., 2025) target image-based reasoning using hand-crafted CoT structures. Several recent works, (Meng et al., 2025; Wang et al., 2025b; Sun et al., 2025; Zhang et al., 2025; Dang et al., 2026; Wang et al., 2026) have extended vision-language reasoning to the video domain. Video-R1 (Feng et al., 2025) introduces a T-GRPO algorithm, specifically designed to handle temporal information in videos. It utilizes two datasets: Video-R1-CoT-165k for supervised finetuning (SFT) and Video-R1-260k for reinforcement learning (RL) training. Video-RFT (Wang et al., 2026) proposes a multi-expert driven, cognition-inspired CoT curation pipeline. In this framework, an LLM first generates preliminary CoTs based on rich, structured, and literal representations of video content. A VLM then refines these CoTs by conditioning them on the actual video input. This process results in two datasets: VideoRFT-CoT-102K for SFT and VideoRFT-RL-310K for RL training. In contrast, Video-RTS (Wang et al., 2025e) presents a different approach by combining efficient RL with a video-adaptive test-time scaling (TTS) strategy. All these methods depend on ground truth annotations and large, high-quality CoT and RL data. Compared to that, test-time reinforcement learning approach leverages the majority answer as a reward signal allowing a training without annotation.

Test-Time Training and Adaptation. Various LLM and VLM approaches have explored leveraging unlabeled data through test-time adaptation and unsupervised learning. Parameters of the models are adjusted at inference or the methods learn from external unlabeled datasets by optimizing objectives such as RL rewards, entropy minimization, auxiliary self-supervised loss among others (Sun et al., 2019; 2024; Behrouz et al., 2025; Akyürek et al., 2025; Zuo et al., 2025; Sun et al., 2020; 2024; Mirza et al., 2023). TTRL (Zuo et al., 2025) utilizes repeated sampling strategy during the rollout phase to accurately estimate the labels, followed by a majority voting reward applied on the unlabeled data, and TTRV (Singh et al., 2026) extends it by combining the frequency-based rewards with entropy regularization on vision tasks such as classification and VQA. Complementary to these, recent work has shown that reinforcement learning can generalize from remarkably few examples. 1-shot RLVR (Wang et al., 2025c) demonstrates that RL with verifiable rewards on a single training example is sufficient to substantially improve mathematical reasoning in LLMs, while LIMR (Li et al., 2025b) shows that a strategically selected subset of training samples can match or outperform full-scale RL datasets, suggesting that sample selection matters more than data scale. These findings motivate our extension to the video domain, where labeled data is particularly scarce. Building on these ideas, our method extends test-time adaptation to video understanding tasks, specifically for educational videos containing reasoning based questions. We incorporate an adaptive frame sampling strategy based on a multi-armed bandit problem, which complements the test-time adaptation, resulting in a novel framework for video reasoning on educational and lecture video content.

3 TTA-Vid

Our method proposes a framework that adapts a pretrained model at test time without ground-truth labels. While the adaptation is performed on a small batch of unlabeled samples, the resulting improvements generalize beyond the training batch to unseen test data and even transfer across datasets. It is composed of two components that work simultaneously: (i) The **Test-Time Adaptation (TTA)** procedure updates the model parameters using a reinforcement learning objective guided by a batch-aware reward signal. Instead of relying on a single view of a video, we sample multiple subsets (or "views") of frames from the same video and generate multiple candidate outputs for each subset. Ag-

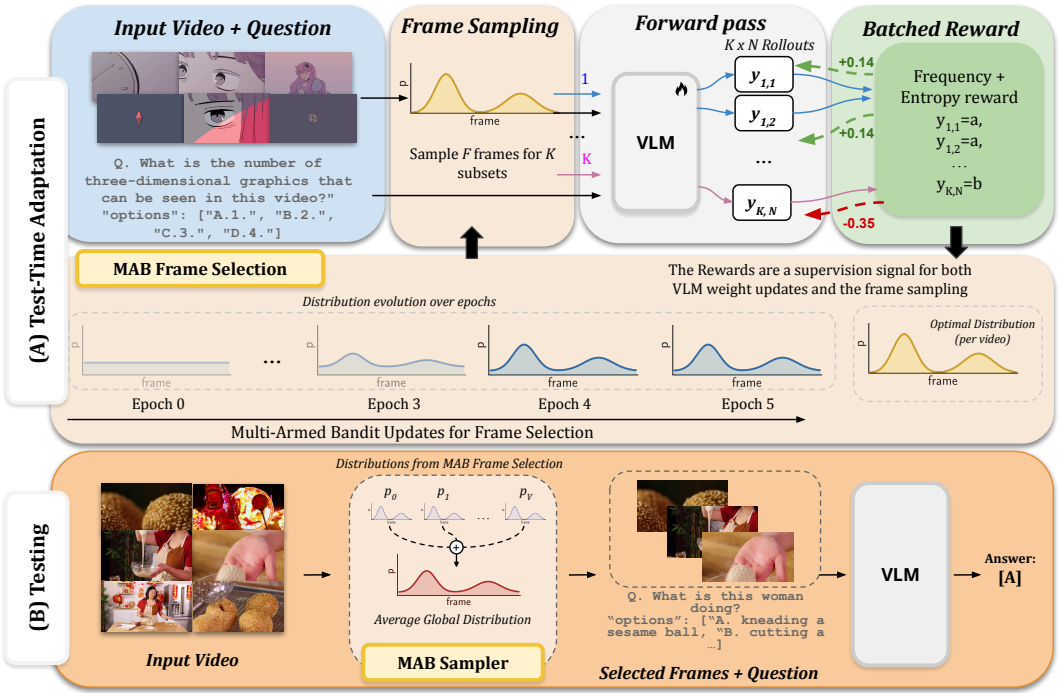


Figure 2: Overview of TTA-Vid: TTA-Vid adapts model parameters at test-time through a batch-aware reinforcement learning objective, and it adaptively selects the most informative frames using a multi-armed bandit approach. Both components leverage the reward signal computed on multiple video frame subsets, enabling the model to jointly learn what to predict and which frames to attend.

gregating these outputs allows us to estimate a batch-level answer distribution that provides a stable reward signal, encouraging the model to produce consistent predictions across different frame subsets. (ii) We leverage an **Adaptive Frame Selection**, formulated as a multi-armed bandit problem. During adaptation, we sample different frame subsets based on a sampling distribution over frames. We use the reward signal formulated for the TTA component to update the sampling distribution and to give more weight to frames from successful rollouts. After each step, these distributions are updated to guide frame sampling for the next step. Figure 2 provides an overview of the approach. In the following, we first describe the test-time adaptation procedure and batch reward formulation, followed by the inference protocol using the adapted model.

3.1 Test-Time Adaptation with Batch-Wide Frequency Reward

We begin by representing an input video $\mathcal{V}$ as an ordered sequence of frames: $\mathcal{V} = (f_1, f_2, \ldots, f_T)$. From this video, we sample K subsets $\{S_1, S_2, \ldots, S_K\}$ with F frames $S_k = (\hat{f_{t_1}}, \ldots, \hat{f_{t_F}}) \in V$ with $(t_1, \ldots, t_F) \in T$. Given a prompt x as the combination of a question and the frame subset S_k, the model, parameterized by θ, generates an output y_k from its policy $\pi_\theta(y_k|x, S_k)$. Note that practically, we pool a new set of subsets at each epoch, following the learned distribution. For simplicity of notation, we assume the notation to refer to a single epoch.

To construct the reward signal, for each of the K subsets we generate N rollouts resulting in N outputs, creating a total pool of $K \times N$ output candidates, with a single candidate denoted as $\{y_{k,n}\}$ for $k = 1, \ldots, K$ and $n = 1, \ldots, N$, and all samples of the pool relating to prompt x, thus the same question and the same video, and corresponding to n rollouts for a single subset k.

Our reward formulation extends the frequency-based reward concept from TTRV (Singh et al., 2026) which estimates the empirical probability of an answer to be correct by calculating frequency within the answers extracted from the rollouts for a single image. Compared to that, we estimate the empirical probability of an answer based on all N rollouts across all K subsets. This approach leverages the diversity of views of a video to create a more stable and robust reward signal. We use a fixed answer extractor function, h (e.g., a regular expression), to parse the specific answer from each generated output string $y_{k,n}$. Let $\mathcal{A}$ be the set of all unique answers. We define the counts and

the empirical frequency for each unique answer $a \in \mathcal{A}$ as:

$$c(a) = \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbf{1}\big[h(y_{k,n}) = a\big], \quad p(a) = \frac{c(a)}{\sum_{a' \in \mathcal{A}} c(a')} \tag{1}$$

To promote convergence and control for diversity (Singh et al., 2026), we incorporate an entropy-based confidence weight. We calculate the normalized entropy of the answer distribution p:

$$H(p) = -\sum_{a \in \mathcal{A}} p(a) \log p(a), H_{\text{norm}}(p) \quad = \frac{H(p)}{\log |\mathcal{A}|} \in [0, 1]. \tag{2}$$

The final reward for an individual output $y_{k,n}$ combines the frequency score with this entropy penalty as $r(y_{k,n}) = p\big(h(y_{k,n})\big) - \alpha \cdot H_{\text{norm}}(p)$, where α is a hyperparameter that controls the strength of the entropy penalty. This reward structure encourages outputs that are frequent across the batch while simultaneously penalizing the model for high uncertainty (high entropy) in its overall answer distribution, pushing it toward more confident predictions. The reinforcement learning objective is to maximize the expected reward, and the model parameters θ are updated accordingly as $\theta \leftarrow \theta + \eta \nabla_\theta \mathbb{E}_{S_k} \mathbb{E}_{y \sim \pi_\theta(\cdot|S_k)} \big[r(y) \big]$, where η is the learning rate. Following (Zuo et al., 2025), we implement this using Group Relative Policy Optimization (GRPO) (Shao et al., 2024).

3.2 Multi-Armed Bandit Adaptive Frame Selection

While the TTA process adapts the model's parameters, we further incorporate an adaptive frame selection that learns an optimal policy for sampling video subsets. We treat each of the T frames as an "arm" in a contextual multi-armed bandit problem (Auer et al., 2002), where the goal is to learn a probability distribution that favors frames that are most informative for the given task.

We maintain nonnegative *weights* $\mathbf{w} = [w_1, w_2, \ldots, w_T]$ for the T frames, initializing $w_t = c$ for all t, where c is the normalized CLIP score obtained based on the similarity for the T frames and the question. From these, we define a learnable probability distribution over frames as $\mathbf{p} = [p_1, p_2, \ldots, p_T]$, where

$$p_t = \frac{w_t}{\sum_{j=1}^{T} w_j}. \tag{3}$$

At each epoch, the indices $(t_1, \ldots, t_F) \in T$ for the frames of each subset S_k are sampled stochastically according to $\mathbf{p}$.

The reward signal calculated for the TTA component (see Eq. **??**) is repurposed to guide the frame selection. For each subset S_k, we calculate its average reward by averaging the rewards of the N outputs generated from it:

$$\bar{r}_k = \frac{1}{N} \sum_{n=1}^{N} r(y_{k,n}). \tag{4}$$

This score, $\bar{r}_k$, reflects how informative the frames in subset S_k were in contributing to high-frequency, high-confidence answers. To update the frame distribution, we use a multiplicative weights algorithm. We first establish a baseline reward, $\bar{r}_{\text{baseline}} = \frac{1}{K} \sum_{k=1}^{K} \bar{r}_k$, which represents the average performance across all subsets. The probabilities of frames in subsets that performed better than this baseline are increased, while those in underperforming subsets are decreased.

The update for the weight w_t of each frame t is given by:

$$w_t^{\text{new}} = w_t \cdot \exp\left(\eta_{fs} \sum_{k=1}^{K} (\bar{r}_k - \bar{r}_{\text{baseline}}) \cdot \mathbf{1}[t \in S_k] \right), \tag{5}$$

where η_{fs} is the frame selection learning rate and $\mathbf{1}[t \in S_k]$ is an indicator function. We then form the new sampling probabilities by normalizing:

$$p_t^{\text{new}} = \frac{w_t^{\text{new}}}{\sum_{j=1}^{T} w_j^{\text{new}}}. \tag{6}$$

Model	#params	#frames	VideoMMMU	MMVU	SciVideoBench	VideoMME	LongVideoBench	Avg.
Random	-	-	14.00	20.00	10.00	25.00	20.00	17.80
Proprietary Models								
GPT-4o Hurst et al. (2024)	–	–	61.22	75.40	24.90	71.90	66.70	60.02
Gemini 1.5 Flash Team et al. (2024)	–	–	49.78	58.80	–	70.30	61.60	–
Non-Reasoning Models								
LLaVA-OneVision Li et al. (2024a)	7B	64	33.89	49.20	18.80	58.20	56.30	43.27
Qwen-2.5-VL Bai et al. (2025b)	7B	<= 768	47.44	59.20	16.40	56.00	65.10	48.82
InternVL-3 Zhu et al. (2025)	8B	32	49.33	60.80	26.50	61.18	59.08	51.37
Reasoning Models								
Video-RTS Wang et al. (2025e)	7B	128	52.70	66.40	–	63.00	56.60	–
Video-R1 Feng et al. (2025)	7B	128	47.00	64.00	26.80	64.30	57.60	51.94
VideoChat-R1 Li et al. (2025a)	7B	128	52.00	64.80	26.50	64.10	54.30	52.34
VideoChat-R1.5 Yan et al. (2026)	7B	128	50.00	67.00	25.80	64.80	53.60	52.24
Video-RFT Wang et al. (2026)	7B	128	48.10	66.70	25.70	64.10	57.00	52.32
Video-R2 Maaz et al. (2025)	7B	128	50.80	67.40	28.40	63.80	59.20	53.92
Test-Time Adaptation	(Ours)							
*TTRV** (w Qwen2.5-VL) Singh et al. (2026)	7B	32	45.89	64.48	25.50	59.26	57.07	50.46
TTA-Vid (w Qwen2.5-VL)	7B	32	49.44	65.12	25.70	61.14	57.81	51.84
TTA-Vid (w InternVL-3)	8B	32	**53.66**	65.60	**29.80**	**65.11**	**61.48**	**55.13**

Table 1: **Performance Comparison.** We compare TTA-Vid with the state-of-the-art methods on two instructional video question-answering tasks. The best results are highlighted in **bold**. For fair comparison, we repurpose TTRV with by training it in multi-frame setup on four frames.

3.3 Generalization of the Adapted Model

After the test-time adaptation stage, we evaluate the adapted vision-language model on the full test dataset. During adaptation, the model parameters θ are optimized together with the individual probability distribution for a set of videos $(\mathcal{V}_1, \ldots, \mathcal{V}_V)$ resulting in a set of probability distribution $\mathbf{p}_1 \ldots \mathbf{p}_V$, with each training video yielding a learned frame distribution $\mathbf{p}_i$. To obtain a global frame distribution, we average these final distributions across the samples used for test-time adaptation $\bar{p} = \frac{1}{V} \sum_{i=1}^{V} p_i$. V denotes the number of videos used during adaptation. The resulting distribution $\bar{p}$ acts as a generalized frame sampling prior. Rather than memorizing video-specific frame locations, $\bar{p}$ acts as a coarse temporal prior that captures broad information patterns inherent to instructional videos (e.g., favoring structurally informative introductory or summary segments). During inference on unseen videos, sampling according to $\bar{p}$ provides an improved temporal heuristic over uniform or random sampling.

4 Experiments

4.1 Benchmarks

We evaluate the proposed method on five challenging and diverse video question-answering benchmarks: VideoMMMU (Hu et al., 2026), MMVU (Zhao et al., 2025), SciVideoBench (Deng et al., 2026), VideoMME (Fu et al., 2025), and LongVideoBench (Wu et al., 2024). Among these, VideoMMMU, MMVU and SciVideoBench are instructional videos, while VideoMME focuses on generalized understanding and LongVideoBench focuses on long video reasoning problems. **VideoMMMU** (Hu et al., 2026) is a multimodal, multi-disciplinary benchmark that assesses LMM's ability to acquire and utilize knowledge from videos. The dataset has three splits: Perception, Comprehension, and Adaptation. Each split contains 300 QA pairs, and a total of 900 QA pairs in the full benchmark. Perception questions assess the ability to perceive information from videos, Comprehension questions assess the ability to understand knowledge presented in videos, and Adaptation questions assess the ability to adapt video knowledge to new scenarios. **MMVU** (Zhao et al., 2025) is a comprehensive expert-level, multi-discipline benchmark which contains expert-annotated questions spanning 27 subjects across four core disciplines: Science, Healthcare, Humanities, Engineering, and Social Sciences. We test on the val split of MMVU on a multiple-choice QA format, which contains 625 QA pairs that require expert-level reasoning on complex videos. **SciVideoBench** (Deng et al., 2026) is a benchmark with consists of 1000 MCQ questions from scientific experimental videos which span over 25 specialized academic subjects. **VideoMME** (Fu et al., 2025) is a comprehensive evaluation benchmark consisting of videos from short to long videos (avg 17 min). We use the standard split of VideoMME which contains 2700 expert-labeled QA pairs designed for both perception and reasoning tasks. **LongVideoBench** (Wu et al., 2024) is a videoqa benchmark that

TTRL	MAB Rewards	Gen. Frame Sampling	VideoMMMU				MMVU
			Perception	Comprehension	Adaptation	Avg	
×	×	×	66.33	42.33	39.33	49.33	60.80
✓	✓	×	73.33	46.66	37.00	52.33	64.80
✓	✓	✓	74.33	48.66	38.00	53.66	65.60

Table 2: **Component ablation on InternVL3-8B.** Models are trained on subsets of each dataset for multiple epochs and evaluated on the full test sets, with results averaged across runs. Combining all components (highlighted) achieves the best overall performance across benchmarks.

highlights referred reasoning questions, which are dependent on long frame inputs. We test on the validation split with 1337 reasoning questions.

4.2 Implementation Details

For the test-time adaptation, we consider a batch of randomly sampled 32 video question-answer pairs of a dataset. For training, we sample $K = 4$ subsets from each video question-answer pair, consisting of $F = 4$ frames, and do a step-by-step reasoning over $N = 8$ rollouts per subset using a temperature of 1.0. We set the maximum prompt length to 7524 and the maximum response length to 1024 tokens. We set $\alpha = 0.75$ in the final reward function (Equation **??**). We set the frame selection learning rate $\eta_{fs} = 3$ (Equation 5) for all our experiments. For hyperparameters, we use cosine learning rate schedule with a peak value of 5×10^{-7} and adopt the AdamW optimizer for the policy model. We train on this batch for five epochs. At test time, we test the model on the full test-set. The learned frame distribution for the samples in training set are averaged and used as global frame distribution. For the final inference step, we sample 32 frames from each video and generate the answer based on the updated model. All experiments were conducted on 4 × NVIDIA A100 40GB GPUs.

4.3 Comparison to State-of-the-Art

In Table 1, we compare TTA-Vid with proprietary MLLMs, open-source general-purpose multimodal models, and recent video reasoning LLMs across five benchmarks. When applied to InternVL3-8b and Qwen2.5VL-7b, TTA-Vid consistently improves the base backbones, achieving 55.13 and 51.84 average accuracy respectively. In particular, TTA-Vid (InternVL3-8b) improves over its base model by +3.76 points (51.37 → 55.13), with gains across all the datasets. Compared to the open-source non-reasoning models such as LLaVA-OneVision, and InternVL, our method demonstrates clear advantages while maintaining the parameter scale. Although proprietary models like GPT-4o (Hurst et al., 2024) remains stronger overall, TTA-Vid substantially narrows the gap. We further compare against video specific reasoning models, including Video-R1, VideoChat-R1, VideoChat-R1.5, Video-RFT, Video-R2 and VideoRTS. These approaches adopt multi-stage SFT+RL pipelines with substantial supervision: for example, VideoRFT was trained with over 100K CoT samples followed by more than 300K RL samples. Despite such extensive supervision, these models achieve average scores between 51.94 and 53.92, inferior compared to TTA-Vid (55.13). Our method requires only 32 unlabeled samples at test time, without any CoT/RL annotations, demonstrating the effectiveness of test-time adaptation and even surpass heavily trained video reasoning models. We repurpose TTRV which is trained with four frames, and is tested with 32 frames. TTRV improves the baseline scores on datasets like MMVU (avg. score from 48.82 → 50.46). On the other hand, TTA-Vid with batch-wide frequency reward-based RL, and MAB further boost the performance, achieving the best scores compared to the baseline and TTRV on all the datasets (avg. score from 48.82 → 51.84).

4.4 Ablation Studies

In this section, we present a comprehensive set of ablation studies to analyze the behavior and robustness of our method. We perform the following ablations: component-wise analysis with frame sampling (Section 4.4), full TTA vs. generalized setting (Section 4.4), cross-dataset generalization (Section 4.4), varying training sizes and sampling strategies (Sections 4.4–4.4), and ground-truth vs. majority-vote training (Section 4.4).

Setup	VideoMMMU				MMVU
	Perception	Comprehension	Adaptation	Avg	
InternVL3-2B	47.33	28.67	28.33	34.77	50.72
Full TTA	48.23	29.27	31.14	36.21	53.45
Generalized TTA	**48.33**	**31.33**	**32.00**	**37.22**	**54.40**

Table 3: **Full vs. generalized test-time adaptation.** TTA-Vid is trained on multiple subsets of 32 samples and evaluated on the same batch (Per-batch Evaluation) or on unseen batches (Generalized). Results are reported for InternVL3-2B.

Train ↓ / Test →	VideoMMMU	MMVU	SciVideoBench	LongVideoBench
Base InternVL3-8b	49.33	60.80	26.50	59.08
VideoMMMU	51.48	62.08	30.20	60.81
MMVU	53.66	65.60	28.20	60.95
SciVid	50.89	61.28	28.70	59.98
LVB	50.55	61.60	28.80	61.48

Table 4: **Cross-dataset generalization of InternVL3-8B** (32 frames, global averaging). Rows denote training dataset and columns denote evaluation dataset. Best indicates highest performance per test set. Diagonal indicates in-domain results.

Component Ablation. We first evaluate the contribution of both major components: the test-time adaptation via batched rewards and frame sampling. Table 2 shows the performance as we ablate both components, comparing the baseline performance, as well as TTRL with batched rewards, and the full method combining both batched rewards with adaptive frame sampling. It shows that each component yields distinct gains across different benchmarks. Namely, TTRL with batched rewards provides an average improvement of $+3\%$ on the VideoMMU and 4% on the MMVU dataset. Furthermore, adding frame sampling boosts the performance by $+4.33\%$ on VideoMMMU and $+4.8\%$ on MMVU.

Full vs Generalized TTA. We evaluate TTA-Vid under two settings. In the first setting (Full TTA), both training and testing are performed on the same batch of samples. In the second setting (Generalized TTA), the model is trained on a single batch and then evaluated on the entire test set. For the full method, we use the average frame distribution learned across videos during training. We use InternVL3-2B for this ablation with respect to computational runtime. The results are reported in Table 3. We observe that per-batch training and testing already lead to consistent improvements over the baseline, indicating that the RL-based reward signal, together with our frame-sampling strategy, provides meaningful optimization even within a single batch. But interestingly, we achieve the strongest overall performance for the generalized case. This suggests that the adaptation process does not overfit to the specific batch used for training, but instead improves the model in a way that generalizes effectively to unseen data.

Cross-dataset generalization. In Table 4, we present the cross-dataset generalization results for InternVL3-8B. In this setup, each model is trained on 32 samples from a single dataset and then evaluated on all datasets, including its own. The diagonal entries denote in-domain performance, while off-diagonal entries reflect cross-dataset transfer. Overall, a model trained on any individual dataset achieves competitive and in most cases stable performance across other datasets, indicating strong generalization. Notably, several off-diagonal results surpass the corresponding in-domain (diagonal) scores (example: Perception (72.67 $\rightarrow$ 74.67) with Comprehension and Adaptation model, SciVideoBench (28.70 $\rightarrow$ 30.20) with Perception model), showing that the learned model is not limited to dataset-specific patterns.

Different number of samples in the training set. In this experiment, Table 5, we analyze the effect of varying the number of training samples $V = \{1, 4, 8, 16, 32\}$ on model performance. Specifically, we train QwenVL-7B using different numbers of samples per training setup. We show results on the VideoMMMU-Perception split and the MMVU dataset to assess consistency across benchmarks. It can be seen that even with a batch size of one, $V = 1$, the base model's accuracy increases

#Samples	Accuracy (%)	
	Perception	MMVU
Qwen2.5-VL	58.33	59.20
$V = 1$	67.00	61.28
$V = 4$	67.33	61.60
$V = 8$	67.67	61.28
$V = 16$	67.67	62.60
$V = 32$	69.00	65.12

Table 5: Effect of training sample size on model performance. Evaluation is conducted on the VideoMMMU-Perception split and the MMVU benchmark with TTA-Vid+Qwen2.5VL-7b.

Type	Accuracy (%)	
	Perception	MMVU
InternVL3-8b	66.33	60.80
Random	73.33	64.80
CLIP	73.00	65.28
(Ours)	74.33	65.60

Table 6: Effect of different generalized frame sampling strategies on VideoMMMU-Perception and MMVU with TTA-Vid+InternVL3-8b. The learned distribution performs slightly better than other baselines.

from 58.33% to 67.00%. Adding more samples to the training set furthermore, increases accuracy, with $V = 32$ obtaining 69.00% accuracy. A similar trend is also observed for MMVU dataset.

Generalized frame sampling. In Table 6, we show the results where all the experiments are conducted on the same model, i.e. trained with adaptive frame selection but with different frame sampling strategies during testing which are random sampling, CLIP score-based sampling, and our learned generalized frame sampling. We ablate with the two datasets: VideoMMMU-Perception split and MMVU dataset. It can be seen that CLIP based sampling helps MMVU (65.28%) when compared to random sampling. However, the best scores are obtained when we use the learned frame distribution 74.33% and 65.60% on VideoMMMU-Perception and MMVU datasets, respectively. This indicates that learning a coarse temporal prior over information-density patterns captures structural temporal characteristics more effectively than static visual-matching baselines like CLIP.

Majority answer vs Ground Truth answer vs Self-consistency. Finally, we replace the majority-vote answer used for reward computation with the ground-truth answer during training on the MMVU dataset to assess the robustness of the majority-based reward formulation. We train the Qwen2.5VL-7B model under both settings. The model trained using ground-truth answers achieves an accuracy of 60.96%, while the model trained with majority-vote answers attains 65.11% accuracy. These results suggest that majority-vote rewards lead to better generalization, rather than overfitting to ground-truth answers, and contribute to improved test-time adaptability. To compare the proposed method against a common test-time scaling method, we test self-consistency on MMVU dataset. It achieves a score of 62.72%, whereas our proposed method reaches 65.60%, outperforming self-consistency. Moreover, self-consistency requires an increasing number of forward passes as the dataset size grows, while the proposed method needs to be trained only once for a batch.

5 Conclusion

In this work, we introduced TTA-Vid, the first test-time reinforcement learning framework for long instructional videos, where rewards are extracted on-the-fly from unlabeled test data. Our approach addresses the challenge of adapting vision-language models to notoriously difficult instructional video content at inference time by leveraging consistency-based self-reinforcement signals: we compute a frequency-based reward from agreement among generated answers across diverse frame subsets, and optimize a multi-armed bandit formulation using these test-time reinforcement learning signals to learn a coarse temporal prior for frame distributions. Our extensive evaluation across five different datasets demonstrates consistent improvements over strong baseline models, with gains up to 4.33% on VideoMMMU. Beyond empirical gains, TTA-Vid enhances video understanding abilities without annotations, showing that test-time optimization through reinforcement learning can be a powerful paradigm for bridging pre-trained vision-language models and downstream instructional video understanding.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing. The experiments, the results and their interpretation are the authors' own. We have not used generative AI tools for any other tasks. We take responsibility for the final content of this work

REPRODUCIBILITY STATEMENT

All code and scripts will be published together with the paper to ensure full reproducibility.

REFERENCES

Ekin Akyürek, Mehul Damani, Adam Zweiger, Linlu Qiu, Han Guo, Jyothish Pari, Yoon Kim, and Jacob Andreas. The surprising effectiveness of test-time training for few-shot learning. In *ICML*, 2025.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *NeurIPS*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Meng Cao, Pengfei Hu, Yingyao Wang, Jihao Gu, Haoran Tang, Haoze Zhao, Chen Wang, Jiahua Dong, Wangbo Yu, Ge Zhang, et al. Video simpleqa: Towards factuality evaluation in large video language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2616–2624, 2026.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *Transactions on Machine Learning Research*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

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.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org*, 2023.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. In *ECCV*, 2026.

Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. Scivideobench: Benchmarking scientific video reasoning in large multimodal models. *ICML*, 2026.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. In *NeurIPS*, 2025.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. In *NeurIPS*, 2025.

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 *CVPR*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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. In *Proceedings of the Association for Computational Linguistics*, 2026.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *NeurIPS*, 2023.

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.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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*, 2024a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025a.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*, 2025b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *COLM*, 2025.

Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. LLM as dataset analyst: Subpopulation structure discovery with large language model. In *ECCV*, 2024.

Muhammad Maaz, Hanoona Rasheed, Fahad Shahbaz Khan, and Salman Khan. Video-r2: Reinforcing consistent and grounded reasoning in multimodal language models. *arXiv preprint arXiv:2511.23478*, 2025.

Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, et al. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *arXiv preprint arXiv:2506.01725*, 2025.

Muhammad Jehanzeb Mirza, Pol Jané Soneira, Wei Lin, Mateusz Kozinski, Horst Possegger, and Horst Bischof. Actmad: Activation matching to align distributions for test-time-training. In *CVPR*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogerio Feris, Sivan Doveh, James Glass, and M Jehanzeb Mirza. Ttrv: Test-time reinforcement learning for vision language models. In *CVPR*, 2026.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn-o1: Reasoning-enhanced audio-visual large language model. *arXiv preprint arXiv:2502.11775*, 2025.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A Efros, and Moritz Hardt. Test-time training for out-of-distribution generalization. *arXiv preprint arXiv:1909.13231*, 2019.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

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

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.

The Llama Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Advances in neural information processing systems*, 38:4350–4376, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025b.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025c.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *EMNLP*, 2025d.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *EMNLP*, 2025e.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 2024.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *ICCV*, 2025.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *ECCV*, 2024.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

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*, 2019.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *EMNLP*, 2024a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024c.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan 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*, pp. 8475–8489, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. In *NeurIPS*, 2025.