# TTA-Vid: Test-Time Adaptation for Long Instructional Videos

Soumya Shamarao Jahagirdar[1*] Edson Araujo[1*] Anna Kukleva[6] M. Jehanzeb Mirza[2]
Saurabhchand Bhati[2] Samuel Thomas[3] Brian Kingsbury[3] Rogerio Feris[3,4]
James R. Glass[2] Hilde Kuehne[4,5]

[1] University of Tübingen, [2] MIT, [3] IBM Research, [4] MIT-IBM Watson AI Lab,
[5] Tuebingen AI Center, [6] Max Planck Institute for Informatics, SIC

## Abstract

*Understanding instructional videos requires both semantic alignment between visual and textual modalities as well as temporal reasoning across frames. In this work, we leverage the paradigm of Test-Time Reinforcement Learning to video-language data to allow to adapt a pretrained model to incoming video samples at test-time without explicit labels. The proposed test-time adaptation for video (TTA-Vid) combines two key components that work simultaneously: (1) a test-time adaptation that performs step-by-step reasoning at inference time on multiple frame subsets, using a batch-aware frequency-based reward computed across different frame subsets as pseudo ground truth, and (2) a multi-armed bandit strategy for adaptive frame selection that learns to prioritize informative frames, guided by the same reward formulation. Because the adaptation occurs entirely at test time, our method requires no ground-truth annotations or dedicated training splits. Our evaluation shows that TTA-Vid yields consistent improvements across instructional video reasoning tasks, on the test data as well as for the generalized case, highlighting the potential of test-time reinforcement learning for temporal multi-modal 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 [13, 55, 72] or QwenVL [4, 6] series, and video reasoning models such as Video-R1 [21] and VideoRTS [60] has brought remarkable advances in e.g. captioning or question answering tasks. However, when applied to lengthy, highly structured, and conceptually rich educational or instructional videos, these models struggle to produce coherent reasoning and accurate answers. Instructional videos can be considered particularly important and underexplored in this domain: as they are usually recorded for human learning environments, 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 some critical limitations in this context. First, they rely on common video datasets such as ActivityNet-QA [65] for training, covering only narrow domains, and as a result fail to capture more complex reasoning structures as can e.g. be found in real-world educational content. Second, most existing methods process only a small set of frames using static or random sampling, neglecting the fact that different frames contribute unequally to video understanding. As a result, models often attend to visually salient but semantically irrelevant moments, leading to inconsistent or shallow comprehension.

To overcome these limitations, we propose a new test-time learning framework for video reasoning. Orthogonal to supervised approaches, test-time adaptation methods train the model at inference time without any ground-truth labels by leveraging frequency-based self-reinforcement signals. Specifically, given an instructional video and a question, the proposed method samples multiple sets of frames and generates multiple candidate reasoning traces and answers. A batch-wide frequency reward is computed by measuring the empirical probability of each answer across all generated outputs, combined with an entropy-based confidence penalty, allowing the model to reinforce frequently occurring answers while suppressing high-uncertainty generations. This enables unsupervised adaptation during inference, effectively turning any pre-trained vision-language model into a domain-aware reasoning model.

Additionally, we introduce a novel frame importance distribution learning mechanism based on the multi-armed

[*] Equal contribution.

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

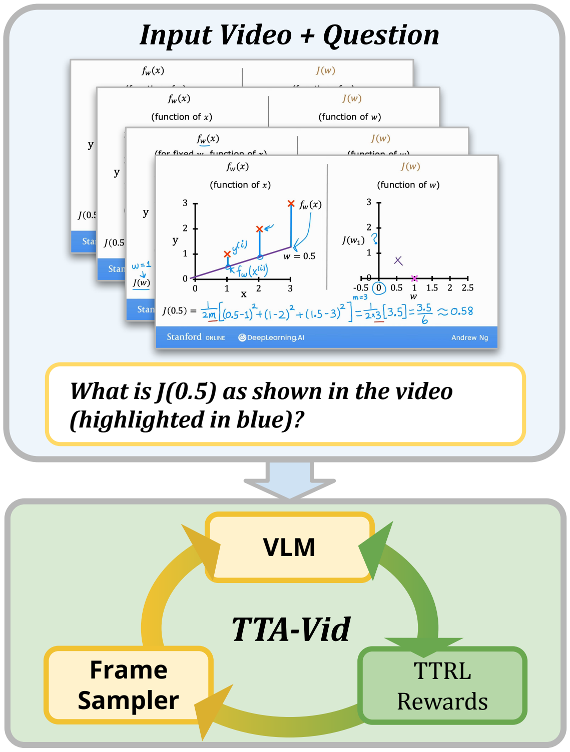


Figure 1. TTA-Vid adapts vision-language models at inference by sampling multiple frame subsets, enforcing majority-consistency among generated answers, and updating a frame-importance distribution via a multi-armed bandit. This yields video-specific adaptation without labels, improving consistency and highlighting frames most relevant for reasoning.

bandit formulation. Instead of relying on fixed frame sampling, our method maintains a learnable 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 not only improves efficiency but also provides an interpretable measure of what the model has learned to attend to during reasoning. The resulting approach requires no additional supervised data, but instead leverages the inherent redundancy in long instructional videos to build a self-supervised adaptation loop at test time, making it particularly well-suited for educational and procedural content.

We evaluate the proposed method on two instructional and educational video QA benchmarks, VideoMMMU [24] and MMVU [71], and two distinct vision-language backbones, InternVL-3 [72] and Qwen2.5-VL [6] and show that the proposed test-time adaptation approach consistently improves answer consistency. Our evaluation further shows that the improvement is not only limited to the test batch that it was applied to, but also extends to unseen test data, allowing for a generalized test-time adaptation. Our results overall show that test-time reinforcement learning can serve as a powerful and efficient tool for video reasoning, paving the way for models that can autonomously adapt and reason over long, structured video content.

The contributions of this work can be summarized as follows: (a) TTA-Vid: A new framework that enables vision-language models to adapt to long video reasoning tasks without labeled data using test-time reinforcement learning. (b) A multi-armed bandit strategy that learns frame importance distributions for efficient and interpretable reasoning using the test-time training reward signals. (c) An extensive evaluation that shows consistent performance gains for the test set as well as for the generalized case across multiple instructional video question-answering datasets and model backbones, validating the effectiveness of our approach.

## 2. Related Work

### 2.1. Video Reasoning Models

The success and applications of large language models (LLMs) [8, 15, 19, 36, 39, 40, 53] have resulted in extending their capabilities to multi-modal tasks, leading to the emergence of vision-language models (VLMs) where the models reason over the visual content and video understanding models enabling them to interpret and reason over dynamic visual content [2, 4–6, 18, 25, 29–31, 34, 57, 66, 70, 72]. However, models like LLaMA-VID[33], VideoLLaMA2 [14], LongVA [67], VISA [62] among others focus on video perception tasks. On the other hand, works inspired by reasoning in language models [22, 27, 35, 42, 51, 64] such as [12, 17, 43, 52, 61, 63] target image-based reasoning using hand-crafted CoT structures. Several recent works, [16, 37, 46, 54, 58, 69] have extended vision-language reasoning to the video domain. Video-R1 [21] 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 [54] 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 [60] 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 amounts of high-quality CoT and RL data. In this work, we propose a test-time reinforcement learning approach for long video reasoning in VLMs, which leverages the majority answer as a reward signal.

## 2.2. Instructional Video Understanding

Instructional videos serve as a valuable medium for information transfer and encompass a wide range of topics. As such, they offer a rich source of diverse challenges and tasks for research in computer vision [9, 10, 20, 23, 24, 28, 41, 68, 71]. Works such as [23, 28] leverage lecture video datasets for tasks such as temporal segmentation, figure-to-text and text-to-figure retrieval, and generation of slide explanations. Multimodal Textbook [68] is a large-scale corpus comprising a total of 22,000 class hours. It includes keyframes, texts, symbols, formulas, along with ASR transcripts, all organized in an interleaved structure, which can be used as pretraining dataset for training large VLMs. Recently, several benchmarks, including [11, 24, 41, 45, 71] have been introduced to evaluate the multimodal understanding capabilities of models on educational video content and to understand their reasoning capabilities.

## 2.3. 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 [1, 7, 38, 47–49, 49, 73]. TTRL [73] utilizes repeated sampling strategy during the rollout phase to accurately estimate the labels, followed by a majority voting reward applied on the given unlabeled data, and TTRV [44] extends it by combining the frequency-based rewards with entropy regularization on vision tasks such as classification and VQA. Building on these ideas, our method extends test-time adaptation to video understanding tasks, specifically for educational videos containing reasoning based questions. We incorporate a 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 allows a model to adapt to a specific video at test time without ground-truth labels. It is composed of two components that work simultaneously: (i) **Test-Time Adaptation (TTA)** adapts the model's parameters using a reinforcement learning paradigm guided by a novel, batch-aware reward signal, and (ii) an **Adaptive Frame Selection** learns to identify and prioritize the most informative frames in the video. To this end, we frame the selection process as a multi-armed bandit problem. The reward signal formulated for the TTA component directly supervises the frame selection mechanism. We leverage multiple subsets (or "views") of frames from the same video within a single batch, which allows the model to efficiently explore the frame sampling space and learn both what to predict and where to look. An overview of the method is shown in Figure 2. We discuss both approaches in detail in the following.

## 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 of frames $\{S_1, S_2, \ldots, S_K\}$. Given a prompt $x$ as the combination of a question and the frame subset $S_k$, the model, parameterized by $\theta$, generates an output $y_k$ from its policy $\pi_\theta(y_k|x_k)$. Note that practically, we pool a new set of subsets at each epoch, following the distribution as described in Section 3.2. For simplicity of notation, we assume the notation to refer to a single epoch.

To construct the reward signal, we generate $N$ candidate outputs for each of the $K$ subsets, 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 t 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 [44] 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 found in the batch. 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 [44], we incorporate an entropy-based confidence weight. We calculate the normalized entropy of the answer distribution $p$:

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

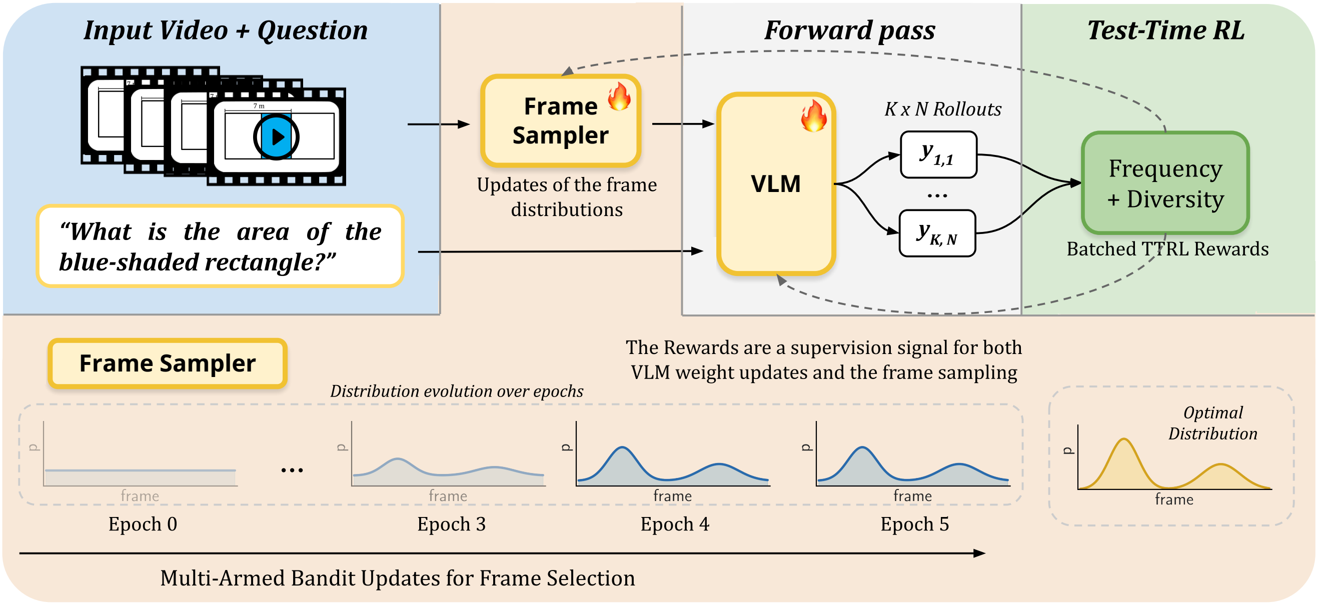

Input Video + Question
"What is the area of the blue-shaded rectangle?"
Frame Sampler
Updates of the frame distributions
Forward pass
VLM
K x N Rollouts
$y_{1,1}$
...
$y_{K,N}$
Test-Time RL
Frequency + Diversity
Batched TTRL Rewards
Frame Sampler
The Rewards are a supervision signal for both VLM weight updates and the frame sampling
Distribution evolution over epochs
p
frame
Epoch 0
Epoch 3
Epoch 4
Epoch 5
Optimal Distribution
Multi-Armed Bandit Updates for Frame Selection

Figure 2. Overview of TTA-Vid: Our method simultaneously performs test-time adaptation of model parameters through a batch-aware reinforcement learning objective and adaptively selects the most informative frames using a multi-armed bandit approach. Both components leverage a shared reward signal computed across diverse video frame subsets within a single batch, enabling the model to learn what to predict and which frames to attend to.

The final reward for an individual output $y_{k,n}$ combines the frequency score with this entropy penalty:

$$r(y_{k,n}) \;=\; p\big(h(y_{k,n})\big) - \alpha \cdot H_{\text{norm}}(p), \tag{3}$$

where $\alpha$ 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 $\theta$ are updated accordingly:

$$\theta \leftarrow \theta + \eta\, \nabla_\theta\, \mathbb{E}_{S_k}\, \mathbb{E}_{y\sim\pi_\theta(\cdot|S_k)}\big[\, r(y)\,\big], \tag{4}$$

where $\eta$ is the learning rate. Following [73], we implement this using Group Relative Policy Optimization (GRPO) [42].

### 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 [3], 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 $w_t = 1$ for all $t$. From these, we define a learnable probability distribution over frames as $\mathbf{p} = [p_1, p_2, \ldots, p_T]$, where

$$p_t = \frac{w_t}{\sum_{j=1}^{T} w_j}. \tag{5}$$

Initially, since all $w_t = 1$, the distribution is uniform ($p_t = 1/T$ for all $t$). At each epoch, the $K$ subsets $\{S_1, \ldots, S_K\}$ are sampled stochastically according to $\mathbf{p}$.

The reward signal calculated for the TTA component (see Eq. 3) 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:

$$\bar{r}_k = \frac{1}{N}\sum_{n=1}^{N} r(y_{k,n}). \tag{6}$$

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:

$$w_t^{\text{new}} = w_t \cdot \exp\left(\eta_{fs} \sum_{k=1}^{K} (\bar{r}_k - \bar{r}_{\text{baseline}}) \cdot \mathbf{1}[t \in S_k]\right), \tag{7}$$

where $\eta_{fs}$ 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:

$$p_t^{\text{new}} = \frac{w_t^{\text{new}}}{\sum_{j=1}^{T} w_j^{\text{new}}}. \tag{8}$$

**Integrated Adaptation Across Epochs.** Our two components, model parameter adaptation and frame distribution updates, are executed simultaneously within each test-time epoch. The frame distribution evolves across epochs, starting from a uniform distribution and gradually becoming more focused on discriminative frames as the model's parameters simultaneously adapt to the specific video and task. A key contribution of our approach is that the reward signal computed for model adaptation (Eq. 3) is directly repurposed to supervise frame selection without requiring additional labels. This enables us to leverage test-time RL supervision to optimize two distinct objectives simultaneously: *what* the model should predict and *where* (which frames) it should look. By grounding frame selection in the empirical reward distribution from test-time rollouts, we ensure that frames are prioritized based on their actual contribution to the model's predictions.

# 4. Experiments

## 4.1. Benchmarks

We evaluate the proposed method on two challenging and diverse instructional video question-answering benchmarks: VideoMMMU [24] and MMVU [71].

**VideoMMMU** [24] is a multimodal, multi-disciplinary benchmark that assesses LMM's ability to acquire and utilize knowledge from videos. The dataset is divided into six categories: Science, Engineering, Art, Humanities, Medicine, and Business. Furthermore, the dataset has three splits: Perception, Comprehension, and Adaptation, which assess the performance of the models at different cognitive stages. 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** [71] 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.

## 4.2. Implementation Details

For the test-time adaptation, we consider batches of 32 samples, thus video question-answer pairs. We train on each batch with 32 samples independently for 5 epochs. While training, to derive the answer, we sample four frames from the frame distribution and do a step-by-step reasoning over 8 rollouts per subset. At test time, for dataset splits - VideoMMMU (Perception and Comprehension) and MMVU - we sample the top 4 frames from the learned distribution and generate the final answer. For VideoMMMU - Adaptation split, we use the top four sampled frames and include the last frame as additional input, as this subset requires the last frame to be used as the default.

For hyperparameters, we use cosine learning rate schedule with a peak value of $5 \times 10^{-7}$ and adopt the AdamW optimizer for the policy model. In total, we use 4 subsets, consisting of 4 frames each and we sample 8 rollouts per subset using a temperature of 1.0 and keep the same number of responses for label estimation and training. We set the maximum prompt length to 7524 and the maximum response length to 1024 tokens. We set $\alpha$ in the final reward to 0.75. We set the number of epochs to 5 for all the datasets. All experiments were conducted on 4 × NVIDIA A100 40GB GPUs. For the adaptive frame selection module, we set the initialization of the weight to be 1. We set the frame selection learning rate $\eta_{fs}$ to 3 for all our experiments.

## 4.3. Comparison to State-of-the-Art

In Table 1, we present the comparison of existing works and TTA-Vid on the two instructional video benchmarks. Consider our method adapted to two baseline models, InternVL3-2B and Qwen2.5-VL-3B, and compare against three types of models: leading proprietary MLLMs [26, 50], open-source general-purpose MLLMs [6, 29], and video reasoning LLMs [43, 60]. The evaluation shows that the proposed method leads to substantial increases across all benchmarks and model backbones. Specifically, our approach achieves an accuracy of 55.44 on the perception subset of the VideoMMMU benchmark, compared to the baseline of 43.89, which is +11.55 points without any supervised finetuning. A similar boost is observed in all the splits of the VideoMMMU benchmark. On the MMVU benchmark, our method obtains 56.41 with an increase of 5.69 over the base model. Our method achieves competitive performance when compared with larger VLMs such as InternVL-2-8B and LLaVA-OneVision-7B, which demonstrates the effectiveness of the proposed method in pushing the boundaries of smaller VLMs using smart strategies.

| Model | #params | #frames | VideoMMMU |  |  |  | MMVU (mc) |
|---|---|---|---|---|---|---|---|
|  |  |  | Perception | Comprehension | Adaptation | Avg |  |
| Random | - | - | 12.00 | 14.00 | 16.00 | 14.00 | 20.00 |
| GPT-4o | - | 50 | 66.00 | 62.00 | 55.67 | 61.22 | 75.40 |
| Gemini 1.5 Flash | - | - | 57.33 | 49.00 | 43.00 | 49.78 | - |
| LLaVA-OneVision | 72B | - | 59.67 | 42.33 | 43.00 | 48.33 | - |
| Qwen-2.5-VL | 72B | - | 69.33 | 61.00 | 50.33 | 60.22 | - |
| InternVL-2 | 8B | 32 | 47.33 | 33.33 | 31.67 | 37.44 | - |
| LLaVA-OneVision | 7B | 64 | 40.00 | 31.00 | 30.67 | 33.89 | 49.20 |
| Qwen-2.5-VL | 7B | 16 | 58.33 | 44.33 | 39.67 | 47.44 | 59.20 |
| Video-RTS | 7B | 51.2 | - | - | - | 52.70 | 66.40 |
| Video-R1 | 7B | 64 | - | - | - | 52.40 | 64.20 |
| InternVL-3 | 2B | 4 | 43.89 | 32.11 | 30.00 | 35.33 | 50.72 |
| InternVL-3+TTA-Vid | 2B | 4 | **55.44** (↑ 11.55) | **35.66** (↑ 3.55) | **31.56** (↑ 1.56) | **40.89** (↑ 5.56) | **56.41** (↑ 5.69) |
| Qwen2.5-VL | 3B | 4 | 51.33 | 37.33 | 29.33 | 39.99 | 56.96 |
| Qwen2.5-VL+TTA-Vid | 3B | 4 | **60.41** (↑ 9.08) | **39.16** (↑ 1.83) | **30.93** (↑ 1.6) | **43.50** (↑ 3.51) | **58.51** (↑ 1.55) |

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**.

## 4.4. Ablation Studies

We first evaluate the contribution of both major component, the test-time adaptation via batched rewards as well as the frame sampling, in our proposed methodology. Table 2 shows the performance evolution as we ablate both components, showing 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 alone provides consistent improvements of 2 to 6% notably boosting VideoMMMU Comprehension (+2.59) and MMVU (+4.20). Frame sampling with self-supervised reward signals achieves even larger gains, particularly on VideoMMMU Perception and MMVU, indicating that these sets benefit most from identifying key frames. This consistent with the fact that those sets feature questions requiring reasoning over critical moments rather than comprehensive temporal understanding. Only for VideoMMMU Adaptation, TTRL alone decreases performance, while the combination maintains positive gains. The full method (highlighted row) achieves the best overall performance, with particularly strong improvements on VideoMMMU Perception and MMVU.

## 4.5. Generalized Test-time Analysis

**In-dataset generalization:** We further evaluate the generalization ability of TTA-Vid by testing each adapted model on the held-out test data, excluding the samples used during adaptation. The results for the generalized evaluation are shown in Table 3. Note that for the evaluation of the full method, we rely on the frame distribution that was learned for each video. While this means that the last line can be seen as an upper bound, it also allows for a direct comparison with Table 2 as the same frame distribution is used in both cases. Overall, it shows that even on samples not seen during adaptation, the adapted models consistently outperform the baseline across all benchmarks. This indicates that incorporating an RL-based reward signal together with our frame-sampling strategy during test-time adaptation improves the model's ability to generalize beyond the adapted examples. This is consistent with recent observations by the NLP community reported in works such as LIMR [32] and 1-shot RLVR [59] where models trained on drastically reduced train sets, even as small as single example maintains strong generalization capabilities. We hypothesize that determining whether similar effects emerge in multimodal contexts is an interesting direction for the future.

**Cross-dataset generalization:** To further investigate this generalization behavior across different datasets, we trained a model on a VideoMMMU-Perception 32-video subset, then tested it on the MMVU dataset. Compared to the non-adapted baseline (41.05% accuracy), adaptation on this small subset raises performance to 53.92% with random frame sampling, and further to 55.68% when using the learned frame-importance distribution from the adapted model. We do a similar analysis on VideoMMMU Perception data, where we test perception split on a model train on MMVU with 32 samples. We obtain accuracy of 47.67 with random frame sampling and 53.67 with frame distribution which improves baseline by +3.78 and +9.78 respectively.

| TTRL | Batched Rewards | Frame Sampling | VideoMMMU | | | | MMVU (mc) |
|---|---|---|---|---|---|---|---|
| | | | Perception | Comprehension | Adaptation | Avg | |
| × | × | × | 43.89 | 32.11 | 30.00 | 35.33 | 50.72 |
| ✓ | ✓ | × | 46.77 (↑ 2.88) | 36.04 (↑ 3.93) | 26.56 (↓ 3.44) | 36.45 (↑ 1.12) | 55.21 (↑ 4.49) |
| ✓ | ✓ | ✓ | 55.44 (↑ 11.55) | 35.66 (↑ 3.55) | 31.56 (↑ 1.56) | 40.89 (↑ 5.56) | 56.41 (↑ 5.69) |

Table 2. **Component ablation on InternVL3-2B.** Models are trained on 32-video subsets of each dataset for 5 epochs and evaluated on their respective test sets, with results averaged across all subsets. Models without TTRL are run one on the full test set. Arrows indicate performance changes relative to the baseline. Combining all components (highlighted) achieves the best overall performance across benchmarks.

| TTRL | Batched Rewards | Frame Sampling | VideoMMMU | | | | MMVU (mc) |
|---|---|---|---|---|---|---|---|
| | | | Perception | Comprehension | Adaptation | Avg | |
| × | × | × | 43.89 | 32.11 | 30.00 | 35.33 | 50.72 |
| ✓ | ✓ | × | 49.17 (↑ 5.28) | 36.25 (↑ 4.14) | 26.96 (↓ 3.04) | 37.46 (↑ 2.13) | 54.96 (↑ 4.24) |
| ✓ | ✓ | ✓ | 56.04 (↑ 12.15) | 35.66 (↑ 3.55) | 32.07 (↑ 2.07) | 41.26 (↑ 5.93) | 56.99 (↑ 4.27) |

Table 3. **Generalization analysis.** We train TTA-Vid on subsets of 32 samples and evaluate on the full test set excluding the training samples. Models trained on small subsets still outperform the baseline on unseen data, demonstrating that test-time adaptation with RL-based rewards and frame sampling enhances generalization capability.

## 4.6. Frame selection via frame optimization

We finally evaluate the proposed frame optimization process. Our approach selects the top-$k$ frames according to the final optimization distribution and compares against random frame selection baselines. For this evaluation, we use the models adapted on 4 frames during TTA and evaluate then for multiple frame budgets to generate the final answer: $k \in \{1, 2, 3, 4, 8, 16, 40\}$, where $k = 40$ represents the full video. As shown in Figure 3, our optimization significantly improves the final performance of the model, especially in the low-frame regime. Specifically, optimized frame selection achieves $3.68\%$ improvement when selecting $k = 4$ frames compared to random selection. Notably, selecting $4$ optimized frames outperforms selecting $4$ random frames ($55.36\%$ vs $51.68\%$), demonstrating the effectiveness of our approach. Note that while this improvement is higher for lower frame regimes, our results further shown that frame selection based on the self-supervised reward still performs better than just using the full range of frames, enabling efficient video understanding with minimal frame budgets.

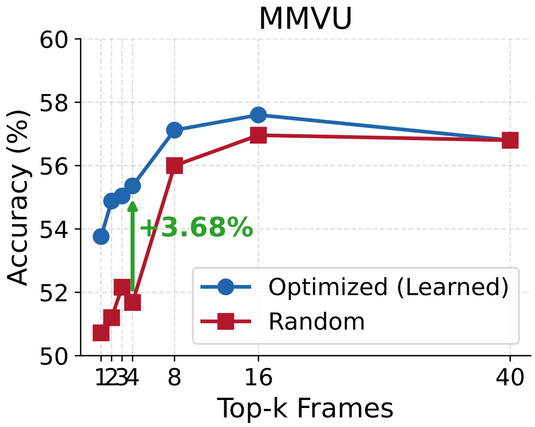


Figure 3. Comparison of optimized frame selection versus random frame selection across different numbers of frames on MMVU dataset. Our method shows substantial improvements in the low-frame regime, with 4 optimized frames outperforming 16 random frames on both datasets.

## 4.7. Qualitative Analysis

To illustrate the effectiveness of our learned frame selection, we compare the combination of the test-time adaptation and frames selection by our optimization-based approach against random sampling. Figure 4 presents two representative examples from VideoMMMU-Perception. In the first example, the question asks “What is the break-even price as shown in the final example of the video?” requiring both temporal localization (identifying the “final example”) and some numerical understanding across multiple price values. Our adapted model retrieves the relevant frame showing the break-even calculation of $6.25 (option C), whereas the baseline model based on random frame sampling leads to an incorrect prediction of $6.00 (option B). In the second example, the question requires identifying the specific textual content visible on the video: “What is $J(0.5)$ as shown in the video (highlighted in blue)?”. The adapted model successfully selects the frames that lead to the correct answer, i.e. first identifying the frames in which $J(0.5)$ is highlighted in blue color (as specified in the question), followed by continuous selection of important frames, to finally se-

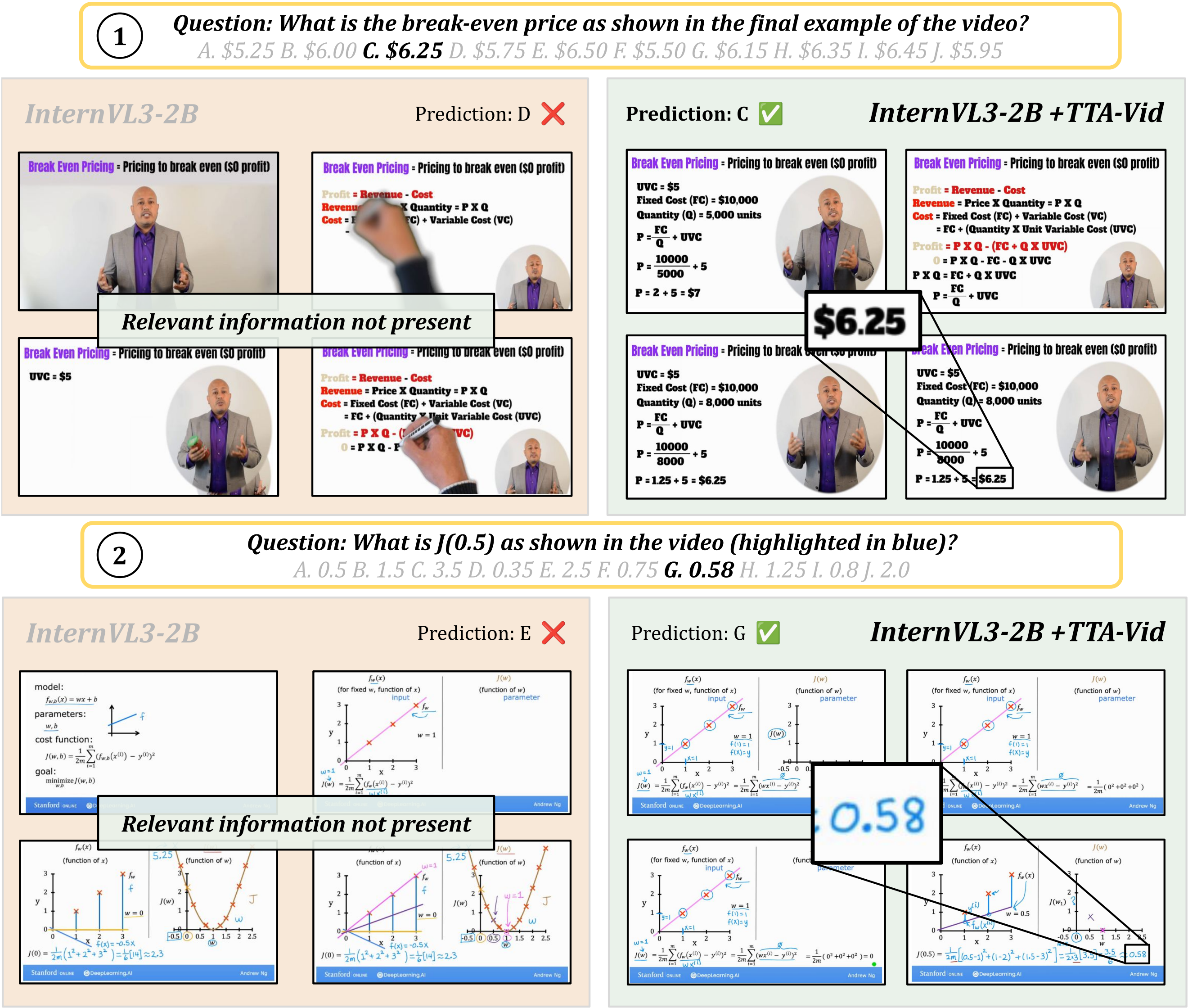

1
Question: What is the break-even price as shown in the final example of the video?
A. $5.25 B. $6.00 C. $6.25 D. $5.75 E. $6.50 F. $5.50 G. $6.15 H. $6.35 I. $6.45 J. $5.95
InternVL3-2B
Prediction: D ❌
Relevant information not present
Prediction: C ✅
InternVL3-2B +TTA-Vid
Break Even Pricing = Pricing to break even ($0 profit)
$6.25
2
Question: What is J(0.5) as shown in the video (highlighted in blue)?
A. 0.5 B. 1.5 C. 3.5 D. 0.35 E. 2.5 F. 0.75 G. 0.58 H. 1.25 I. 0.8 J. 2.0
InternVL3-2B
Prediction: E ❌
Relevant information not present
Prediction: G ✅
InternVL3-2B +TTA-Vid
0.58

Figure 4. **Qualitative comparison of frame selection strategies.** Random sampling (left) versus our learned selection (right) on two VideoMMMU (Perception) examples. **(1)**: Design question requiring identification of break-even price in the video. Our method selects the critical frame (correct answer C), while random sampling misses it (predicts D). **(2)**: Accounting question requiring localization of the value. Our method identifies the highlighted value in blue to be (0.58, option G), while random sampling fails (predicts 2.5, option E).

lect the frame that has the answer in it, which is option G same as the ground truth. In contrast, the random sampling misses these key frames, leading to an incorrect baseline prediction (J). These examples demonstrate that the proposed adaptive frame selection coupled with the test-time adaptation with reinforcement learning, learns to prioritize frames containing task-relevant information.

## 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 interpretable frame importance distributions. Extensive evaluation across VideoMMMU and MMVU demonstrates consistent improvements over strong baseline models, with gains up to 15.63% on MMVU using only 4 frames. Beyond empirical gains, TTA-Vid enhances video understanding abilities without explicit supervision, pointing to test-time optimization through reinforcement learning as a powerful paradigm for bridging pre-trained vision-language models and downstream instructional video understanding.

## References

[1] 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. *arXiv preprint arXiv:2411.07279*, 2024. 3

[2] 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. 2

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002. 4

[4] 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*, 1(2):3, 2023. 1, 2

[5] 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*, 2025.

[6] 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*, 2025. 1, 2, 5

[7] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024. 3

[8] 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. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[9] Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. Vlengagement: A dataset of scientific video lectures for evaluating population-based engagement. *arXiv preprint arXiv:2011.02273*, 2020. 3

[10] Sahan Bulathwela, Maria Perez-Ortiz, Erik Novak, Emine Yilmaz, and John Shawe-Taylor. Peek: A large dataset of learner engagement with educational videos. *arXiv preprint arXiv:2109.03154*, 2021. 3

[11] 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. *arXiv preprint arXiv:2503.18923*, 2025. 3

[12] 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. *arXiv preprint arXiv:2504.11468*, 2025. 2

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1

[14] 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. 2

[15] 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 (accessed 14 April 2023)*, 2(3):6, 2023. 2

[16] 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. *arXiv preprint arXiv:2505.24718*, 2025. 2

[17] 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. *arXiv preprint arXiv:2503.17352*, 2025. 2

[18] 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. 2023. 2

[19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala,

Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. 2

[20] Kartik Dutta, Minesh Mathew, Praveen Krishnan, and CV Jawahar. Localizing and recognizing text in lecture videos. In *2018 16th international conference on frontiers in handwriting recognition (ICFHR)*, pages 235–240. IEEE, 2018. 3

[21] 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. *arXiv preprint arXiv:2503.21776*, 2025. 1, 2

[22] 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. 2

[23] Anchit Gupta, CV Jawahar, Makarand Tapaswi, et al. Unsupervised audio-visual lecture segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5232–5241, 2023. 3

[24] 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. 2, 3, 5

[25] 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. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 2

[26] 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. 5

[27] 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. 2

[28] Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20087–20098, 2023. 3

[29] 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. 2, 5

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[31] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[32] Xuefeng Li, Haoyang Zou, and Pengfei Liu. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*, 2025. 6

[33] 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. 2

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[35] 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. *arXiv preprint arXiv:2503.20783*, 2025. 2

[36] 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 (33)*, pages 235–252. Springer, 2024. 2

[37] 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. 2

[38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24152–24161, 2023. 3

[39] 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. 2

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[41] Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Shahbaz Khan. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025. 3

[42] 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. 2, 4

[43] 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. 2, 5

[44] 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. *arXiv preprint arXiv:2510.06783*, 2025. 3

[45] Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark. *arXiv preprint arXiv:2504.14693*, 2025. 3

[46] 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. 2

[47] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A Efros, and Moritz Hardt. Test-time training for out-of-distribution generalization. 2019. 3

[48] 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 *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[49] 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. 3

[50] 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. 5

[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. 2

[52] 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*, pages 24290–24315, 2025. 2

[53] 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. *CoRR*, abs/2302.13971, 2023. 2

[54] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025. 2

[55] 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*, 2025. 1

[56] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 2

[57] 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 *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 2

[58] 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*, 2025. 2

[59] 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*, 2025. 6

[60] 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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28114–28128, 2025. 1, 2, 5

[61] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098, 2025. 2

[62] 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 *European Conference on Computer Vision*, pages 98–115. Springer, 2024. 2

[63] 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. 2

[64] 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. 2

[65] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 1

[66] 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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737, 2024. 2

[67] 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*, 2024. 2

[68] Wenqi Zhang, Hang Zhang, Xin Li, Jiashuo Sun, Yongliang Shen, Weiming Lu, Deli Zhao, Yueting Zhuang, and Lidong Bing. 2.5 years in class: A multimodal textbook for vision-language pretraining. *arXiv preprint arXiv:2501.00958*, 2025. 3

[69] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025. 2

[70] 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*, 2024. 2

[71] 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*, pages 8475–8489, 2025. 2, 3, 5

[72] 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. 1, 2

[73] 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. *arXiv preprint arXiv:2504.16084*, 2025. 3, 4, 2

# TTA-Vid: Test-Time Adaptation for Long Instructional Videos

Supplementary Material

## 6. Overview

We provide additional ablation studies, and experimental settings in this Supplementary Material. In all experiments, we follow the generalization principle established in the main paper (Section 4.5) by training only on two 32-sample subsets from each dataset. This setup is applied across all ablations and evaluation settings in the experiments presented in supplementary material. Evaluation is conducted on the full test set, and the final results are computed as the average across evaluations from both the trained models.

**Note.** In Table 1 (Performance Comparison) of the main paper, the MMVU baseline score and the reported difference between the baseline and the proposed method contain a typo. The MMVU baseline score was reported as 41.05; the correct value is 50.72 and the difference between the baseline and proposed method is 5.69.

## 7. Further Ablation Studies

To assess the effectiveness and scalability of the proposed method, we conduct a series of ablation studies. Specifically, we examine the impact of applying TTA-Vid to larger models, the effect of varying the number of training epochs, the influence of different configurations of subsets (K), and rollouts (N), and the performance sensitivity to different reward functions combinations. Together, these ablations provide a comprehensive understanding of the factors that contribute to the method's overall performance.

| Model | VideoMMMU-Perception | MMVU |
|---|---|---|
| LLaVA-OneVision-7B | 40.00 | 49.20 |
| Qwen2.5-VL-7B | 58.33 | 59.20 |
| InternVL-2-8B | 47.33 | - |
| InternVL3-8B | 63.00 | 62.08 |
| InternVL3-8B + TTA-Vid | **67.66** | **63.60** |

Table 4. **Performance analysis with InternVL3-8B as the base model.** We compare the performance of models with comparable parameter size with the proposed method: TTA-Vid with InternVL3-8B as the base model. It can be seen that, for both the datasets, the proposed method achieves the best results.

### 7.1. Impact of TTA-Vid on Bigger Models

To further evaluate the effectiveness of the proposed method, we extend the analysis beyond small-scale model and perform test-time adaptation on InternVL3-8B. We train the model using the same subset configuration of $K = 4$ and $N = 8$ for five epochs, and keep all other hyperparameters identical to those used for training InternVL3-2B variant. As shown in Table 4, applying TTA-Vid to InternVL3-8B yields the strongest results among all models of comparable size. This demonstrates that the proposed method scales effectively to larger models.

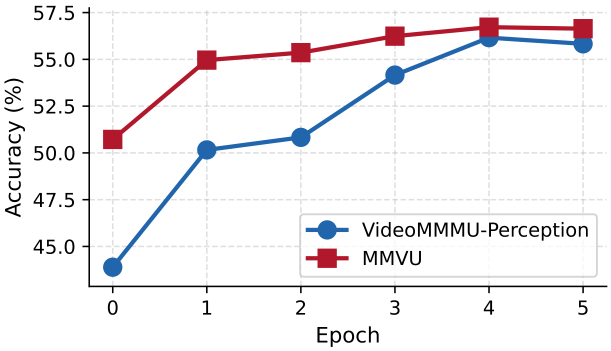


Figure 5. **Epoch ablation.** We examine the accuracy of the model per epoch. Values at epoch 0 represent the original baseline (InternVL3-2b) performance.

### 7.2. Impact of Number of Training Epochs

We ablate the impact of the number of training epochs by training InternVL3-2B on VideoMMMU-Perception and MMVU for 5 epochs. We save the model every epoch, and further test it. From Figure 5, we observe that the base model starts with relatively low accuracy, but as training with TTA-Vid progresses, the accuracy steadily improves across both datasets.

### 7.3. Impact of Subsets (K) and Rollouts (N)

To understand the effect of different number of subsets (K) and rollouts per subset (N), we ablate over different values of K and N. We experiment this on two datasets: VideoMMMU-Perception split and MMVU. We keep all remaining hyperparameters the same as for the original setting. From Table 5, it can be seen that our method generalizes to different values of K and N. For MMVU, the default configuration of 4 subsets and 8 rollouts yields the best performance, while for VideoMMMU-Perception the same default setup achieves the second-best results. Although increasing the number of subsets or rollouts can improve performance in some cases, it also increases computational cost. Therefore, these hyperparameters can be adjusted based on the task difficulty and the available computational budget.

### 7.4. Impact of Different Reward Functions

We explore several alternative configurations for the reward functions which include: (a) $R_{maj}$ is the majority reward based on TTRL [73] i.e. the rollouts receive a reward of 1 if it matches the majority answer and 0 otherwise, (b) $R_{freq}$ is based on the counts and the empirical frequency for each unique answer, defined as:

$$c(a) = \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbf{1}\big[h(y_{k,n}) = a\big], \quad R_{freq} = \frac{c(a)}{\sum_{a' \in \mathcal{A}} c(a')} \tag{9}$$

and, (c) $R_{freq} + \alpha R_{div}$, where $\alpha$ is weighting factor and $R_{div}$ is used to control diversity, which is defined as:

$$\begin{aligned} H(p) &= -\sum_{a \in \mathcal{A}} p(a) \log p(a), \\ R_{div} &= \frac{H(p)}{\log |\mathcal{A}|} \in [0, 1]. \end{aligned} \tag{10}$$

From Table 6, it can be observed that the reward configuration used in TTA-Vid outperforms other alternative design choices.

| Rewards | VideoMMMU-Perception |
|---|---|
| $R_{maj}$ | 55.16 |
| $R_{freq}$ | 55.33 |
| $R_{freq} + \alpha R_{div}$ $(\alpha = 1)$ | 55.50 |
| $R_{freq} + \alpha R_{div}$ $(\alpha = 0.75)$ | 56.04 |

Table 6. **Impact of different rewards.** In this table, we analyze the impact of different reward functions. We conduct ablations using: (a) the $R_{maj}$ reward from [73], which assigns a binary value of 0 or 1 based on majority voting, and (b) $R_{freq}$ and (c) $R_{freq} + \alpha R_{div}$, in which the relative contributions of the frequency and entropy components are controlled by a weighting factor ($\alpha$).

## 8. Comparison to Self-Consistency Baseline

Building on the success of self-consistency as a stronger alternative to greedy chain-of-thought decoding, we adopt it as a natural test-time baseline for our setting. Following [56], we evaluate InternVL3-2B with self-consistency by sampling 32 reasoning rollouts per input and returning the most common answer. It is important to note that the computational cost of self-consistency grows linearly with the size of the test set, for example, it requires 9600 forward passes for the 300 sample set in VideoMMMU Perception and $20,000$ forward passes for the 625 sample test set for MMVU. In contrast, our method requires a one-time training overhead on only 32 samples, after which inference requires essentially one forward pass per test example (e.g., 300 and 625 passes for the two datasets). Despite this large reduction in test-time cost, our method achieves higher accuracy on both datasets. Specifically, self-consistency obtains 50.00% vs. ours 56.04 on VideoMMMU-Perception, and 55.04% vs our 56.99 on MMVU.

## 9. TTA-Vid Prompt Details

In this section, we provide prompt used in our experiments. As shown in Figure 8, and following TTRL [73], the model is instructed to generate step-by-step reasoning followed by an answer in a specified format. The number of choices varies for each dataset.

## 10. Category-based Performance Analysis

We analyze the performance of the baseline model (InternVL3-2b) vs our model (InternVL3-2b+TTA-Vid). From figures, 6 and 7, it can be seen that for both datasets, the proposed method outperforms baseline model in all the categories. Specifically, in Figure 6, we show the scores for VideoMMMU Perception. The highest gains are observed in Humanities category, followed by Medicine, and the least gains are observed in Science category. In Figure 7, we show the scores for MMVU dataset. The highest gains are observed in Healthcare category.

## 11. Frame Distribution Progression Examples

This section provides a visualization of the frame distributions ($p$ from Equation 5) across different epochs during the optimization process. The heatmaps as shown in Figure 9 illustrate how the model's perception of frame importance evolves over time for one representative video in each dataset. "init" represents the initial state of the model with uniform distribution. As the optimization progresses through epochs 0 to 4, the heatmaps reveal that the model begins to focus on specific frames, indicating their higher importance for the video. This behavior suggests that the model is learning to prioritize certain frames that are more relevant for its task, effectively filtering out less significant frames. The plots highlight the model's ability to adapt and refine its frame selection strategy over time.

| Subsets (K) | Rollouts (N) | VideoMMMU-Perception (Acc. in %) | MMVU (Acc. in %) |
|---|---|---|---|
| 4 | 8 | 56.04 | 56.99 |
| 4 | 16 | 55.83 | 54.72 |
| 4 | 32 | 55.83 | 54.64 |
| 8 | 8 | 56.33 | 54.32 |
| 8 | 16 | 54.50 | 54.96 |
| 8 | 32 | 54.66 | 55.28 |

Table 5. **Component ablation in TTA-Vid.** In this table, we analyze the performance of the proposed method with different subsets (K) and rollouts (N) with InternVL3-2B as the base model. For MMVU, our default configuration of $K = 4$ and $N = 8$ yields the best performance and is the second-best setting for the VideoMMMU-Perception set. Since computational cost increases with larger values of $K$ and $N$, these hyperparameters can be adjusted according to the available computational resources.

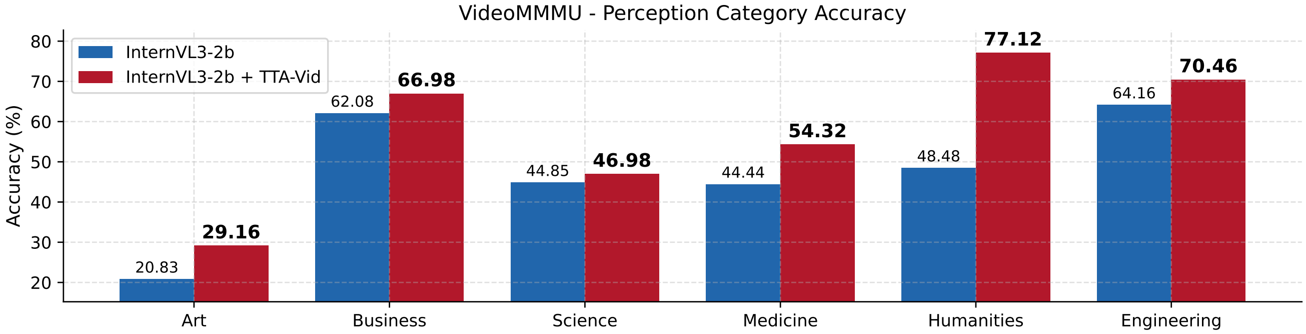


Figure 6. **Performance per category: VideoMMMU-Perception** VideoMMMU dataset has multiple categories such as Art, Business, Science, Medicine, Humanities and Engineering. It can be seen that TTA-Vid performs better than the baseline on all the different categories.

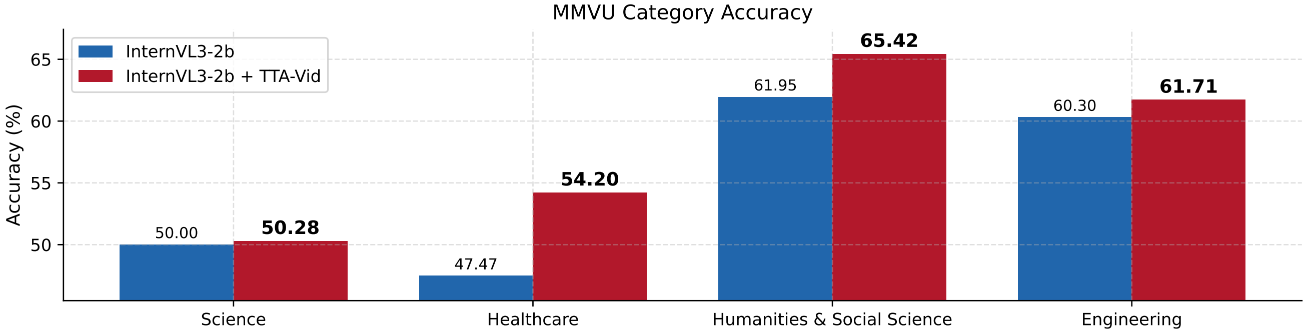
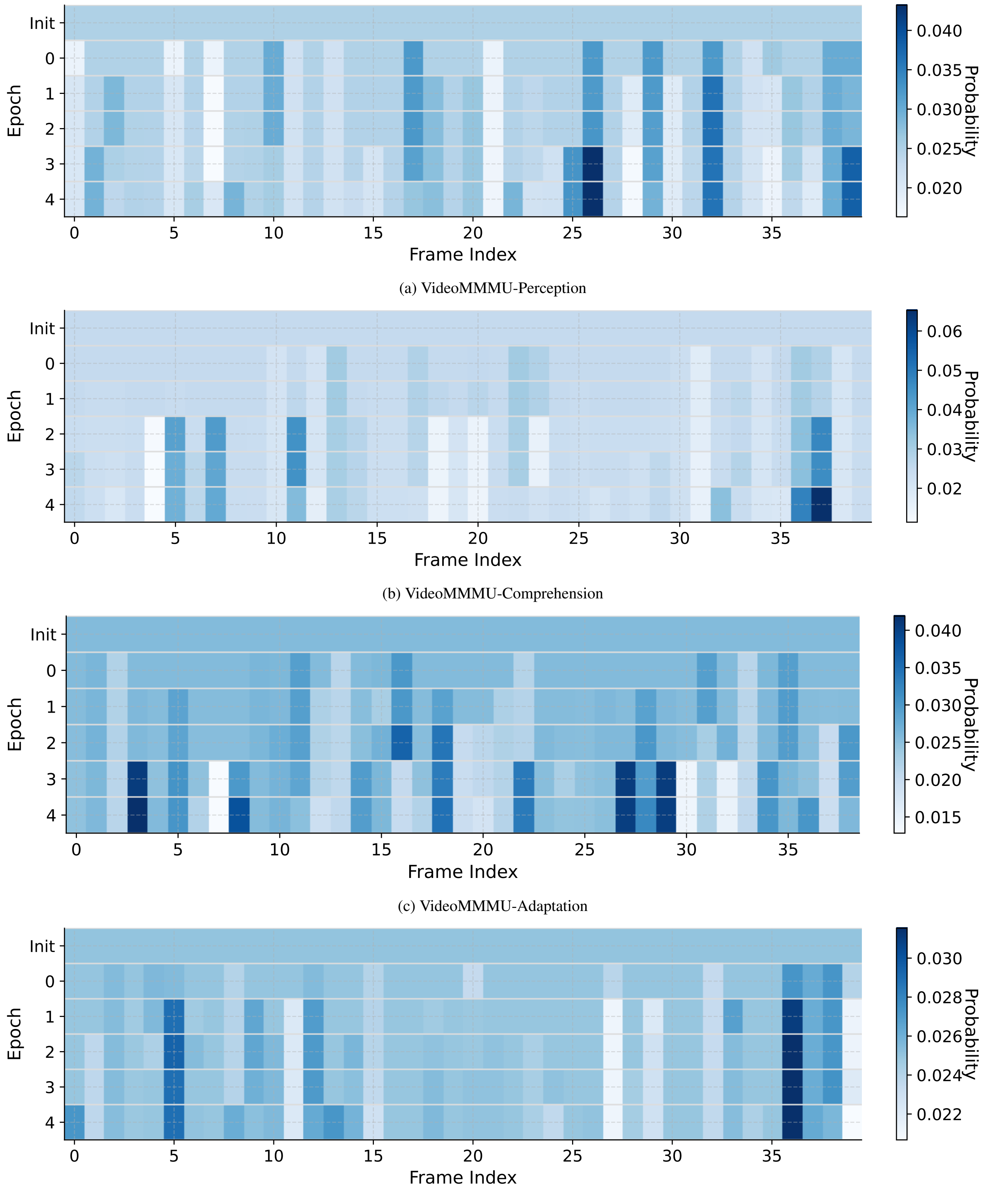


Figure 7. **Performance per category: MMVU.** MMVU dataset has four categories: Science, Healthcare, Humanities and Social Science and Engineering. TTA-Vid outperforms the baseline on all the categories.

```
"Frame-1:<image> Frame-2:<image> Frame-3:<image> Frame-4:<image>
Answer the following multiple choice question based on the video.
First, briefly summarize the content shown in each frame. Think step by step
   before answering.
Finally, the last line of your response should be of the following format: '
   Answer: $LETTER' (without quotes) where LETTER is one of ABCDEFGHIJ.
Question: <QUESTION_PLACEHOLDER>
Choices: <CHOICES_PLACEHOLDER>"
```

Figure 8. We provide the prompt used for training TTA-Vid



(a) VideoMMMU-Perception

(b) VideoMMMU-Comprehension

(c) VideoMMMU-Adaptation

(d) MMVU

Figure 9. **Heatmap examples of frame distributions across epochs:** Visualizing the evolution of frame distributions before and after frame optimization of a sample from each dataset. "Init" represents the initial state with uniform distribution, while epoch 0-4 show post optimization results.