# CORRECTNESS-CONDITIONED ON-POLICY DISTILLATION FOR REFLECTIVE MULTI-TURN REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reinforcement-learning-based post-training has improved the reasoning capabilities of vision-language models. Inspired by a teacher–student learning paradigm centered on questioning, receiving correctness feedback, and revising, we propose a **Reflective Multi-Turn Video Reasoning** method that enables models to verify and correct their own predictions without access to privileged feedback. First, Multi-Agent Dataset Filtering combines diverse reasoning trajectories generated by multiple agents with agent-based evaluation and human verification to construct high-quality supervision. Second, Reflective Multi-Turn Training iteratively retains mixed-outcome samples and jointly optimizes the model using Group Relative Policy Optimization (GRPO) and Correctness-Conditioned On-Policy Distillation (CCOPD), thereby improving answer verification and revision. During the second stage, CCOPD conditions the teacher model on the correctness as privileged information and provides token-level supervision along the student's on-policy trajectory. We evaluate the proposed method on 13 widely used benchmarks spanning four categories: short-video understanding, long-video understanding, video reasoning, and temporal grounding. The results show that our method achieves broadly competitive performance across all 13 benchmarks.

## 1 INTRODUCTION

Recent advances in the reasoning capabilities of large language models (LLMs) Wei et al. (2022) have stimulated growing interest in extending reinforcement-learning-based post-training to vision-language models (VLMs). Optimization methods based on proximal policy optimization (PPO) Schulman et al. (2017) and group relative policy optimization (GRPO) Shao et al. (2024) have demonstrated the potential to improve reasoning beyond supervised fine-tuning when reliable rewards are available. Extending these methods to video reasoning, however, remains challenging because video tasks often involve sparse rewards, long reasoning trajectories, unstable multi-turn optimization, and the difficulty of teaching models when and how to correct their own mistakes.

Our **motivation** draws on the educational philosophy of **Han Yu**, a prominent scholar of the Tang dynasty, who characterized the role of a teacher as transmitting principles, imparting knowledge, and resolving doubts. As illustrated in Fig. 1, learning commonly progresses through two complementary stages: guided practice and independent assessment. During practice, a student attempts a problem, receives correctness feedback from a teacher, and revises the solution accordingly. During an examination, however, such external feedback is unavailable; the student must solve the problem independently, assess the correctness of the initial response, and correct any identified errors before submitting the final answer. Through repeated practice, teacher-guided verification and correction are gradually internalized as the student's capacity for autonomous reflection and self-

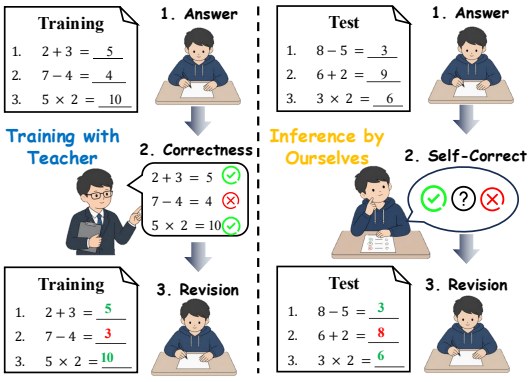



Figure 1: Motivation for reflective multi-turn reasoning, inspired by **Han Yu's** view of teaching.



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.

correction. Motivated by this transition from external guidance to autonomous reflection, we train VLMs to follow an analogous three-step process: *generating an initial answer, assessing its correctness, and correcting potential errors to produce the final answer*. This reflective capability is particularly important for VLMs, whose predictions must be grounded in visual evidence. Whereas language-model systems may verify their outputs through tool execution, code interpreters, symbolic verifiers, search engines, or interactions with other agents, VLMs generally lack comparably accessible mechanisms for determining whether their predictions are consistent with the visual input.

To address this limitation, we propose **Reflective Multi-Turn Video Reasoning**, a unified framework that enables VLMs to progressively refine their predictions through answer generation, verification, and revision. The framework spans three stages, including Multi-Agent Dataset Filtering, Reflective Multi-Turn Training, and Correctness-Based Multi-Turn Inference, which together support the construction of reliable reasoning data, the acquisition of reflective reasoning capabilities, and their autonomous use at inference. Central to the framework is **Correctness-Conditioned On-Policy Distillation (CCOPD)**, a training paradigm that transfers correctness-aware guidance from a teacher to a student without exposing such privileged information at inference. In this way, the proposed framework bridges externally guided correction during training and independent self-verification and self-correction during deployment.

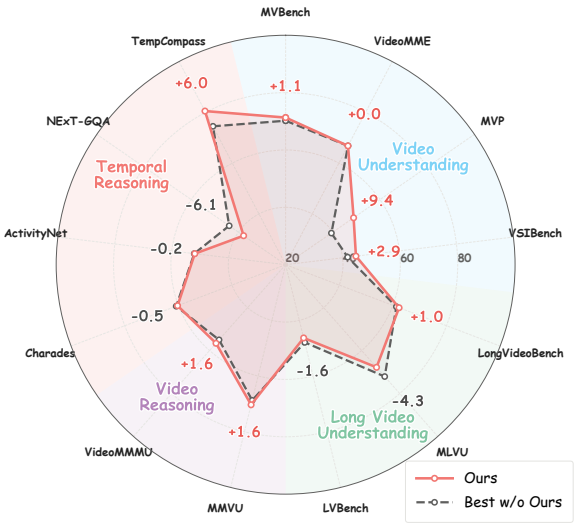



Figure 2: Performance overview across 13 benchmarks under Qwen2.5-VL-7B, which is detailed in Tab. 19.

The **main contributions** of this work are threefold. (1) We develop a **Multi-Agent Dataset Filtering** pipeline for chain-of-thought extraction and data refinement. Multiple agents independently generate diverse reasoning trajectories and candidate answers, after which a judge agent selects reliable responses based on the ground-truth annotations and human reviewers resolve ambiguous or inconsistent cases. Applied to an initial corpus of approximately 300,000 samples, the pipeline yields 200,000 high-quality chain-of-thought examples and corrects approximately 2.3K erroneous annotations. (2) We introduce an iterative sample-filtering strategy for **Reflective Multi-Turn Training**. At each training iteration, we discard samples whose rollouts are uniformly correct or uniformly incorrect and retain mixed-outcome samples, which provide informative contrasts for learning answer verification and revision. (3) We propose **Correctness-Conditioned On-Policy Distillation (CCOPD)**, in which a teacher conditioned on the correctness of the student's preceding response provides token-level supervision along the student's on-policy revision trajectory. Without access to this privileged signal, the student learns to assess the preceding response and revise it when necessary. We further formulate a joint GRPO&CCOPD objective that combines correctness-judgment, format, and final-answer rewards with reference-policy regularization and correctness-conditioned token-level on-policy distillation.

We evaluate the proposed framework on 13 benchmarks spanning four categories: short-video understanding, long-video understanding, video reasoning, and temporal reasoning. As shown in Fig. 2, Ours achieves strong and broadly competitive performance across these diverse evaluation settings. In particular, it yields clear improvements on video understanding benchmarks, *e.g.*, MVP and VSIBench, improves performance on both video reasoning benchmarks, MMVU and VideoMMMU, and delivers a substantial gain on TempCompass. These results demonstrate that reflective multi-turn training improves not only the model's initial reasoning performance but also its ability to evaluate previous responses and correct erroneous predictions. **Nevertheless, autonomous reflection is not infallible.** Analogous to students who may fail to reproduce their practice-level performance under examination conditions, the model cannot identify and correct every erroneous response and may occasionally revise an initially correct answer into an incorrect one. Importantly,

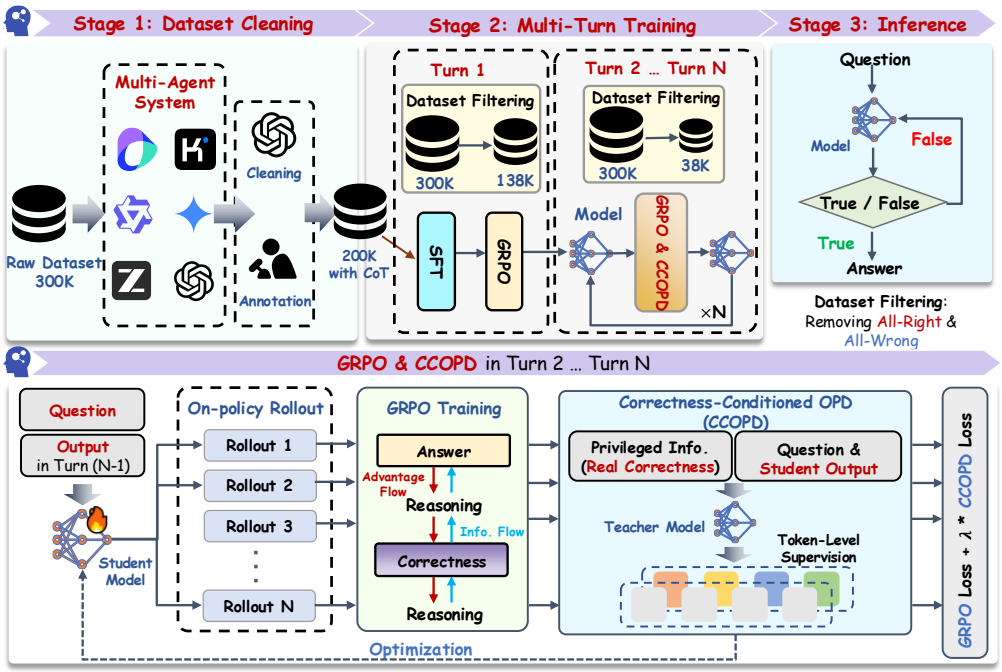



Figure 3: Overview of the proposed three-stage pipeline. Stage 1: Dataset cleaning using six agents and human verification. Stage 2: Multi-turn training, consisting of SFT & GRPO in Turn 1, followed by GRPO & CCOPD in Turns 2...$N$. Stage 3: Correctness-guided multi-turn inference.

however, across successive rounds of verification and revision, the number of errors successfully corrected consistently exceeds the number of correct answers inadvertently corrupted. This asymmetry highlights the practical value of our approach: by progressively internalizing the teacher-guided verification and correction process used during training, the model can improve and stabilize its performance at inference using only its existing knowledge, without relying on privileged correctness signals, external tools, or additional information. Overall, despite the inherent uncertainty of self-correction, a limitation shared by human, the consistent net gains across tasks with different temporal and reasoning requirements demonstrate the effectiveness and generalizability.

## 2 Methodology

### 2.1 Overview

We introduce *Reflective Multi-Turn Video Reasoning*, a unified framework that reformulates video question answering and grounding as an explicit, iterative process of answer generation, verification, and revision. Given a video and a question, the model first produces an initial response and subsequently determines whether that response should be preserved or revised. Operationally, the framework comprises three stages: multi-agent dataset construction and cleaning, reflective multi-turn training, and correctness-based multi-turn inference. Correctness-Conditioned On-Policy Distillation (CCOPD) serves as the core optimization mechanism within the second stage.

Given a video $v$, a question $q$, and a reference answer $a^\star$, we define the annotated dataset as

$$\mathcal{D}_0 = \{(v_i, q_i, a_i^\star)\}_{i=1}^{N}. \tag{1}$$

Our objective is to learn a reasoning policy $\pi_\theta$ that generates an initial response and, in subsequent reflective turns, explicitly assesses the correctness of its preceding response before preserving or revising it. In contrast to single-turn approaches that map $(v, q)$ directly to a final output, our formulation makes answer verification and revision learnable components of the reasoning process.

The proposed framework is described in three technical subsections. Sec. 2.2 outlines the multi-agent generation, task-specific verification, and judge-agent selection procedures. Sec. 2.3 then presents the split-turn reflective training procedure, whose core component, CCOPD, is introduced in Sec. 2.4. Additional details for the training and inference stages are provided in Sec. B.

| Source group | Number | | Statistics (%) | |
|---|---|---|---|---|
| | Raw | Cleaned | Retention | Share |
| LLaVA-Video-178K | 82,676 | 63,767 | 77.13 | 29.85 |
| VIRL | 23,131 | 22,308 | 96.44 | 10.44 |
| Multimath-300k | 27,000 | 13,412 | 49.67 | 6.28 |
| NeXT-QA/GQA | 10,905 | 8,418 | 77.19 | 3.94 |
| STAR | 11,455 | 6,942 | 60.60 | 3.25 |
| DAPO-Math | 6,413 | 6,127 | 95.54 | 2.87 |
| Cauldron subsets | 9,000 | 6,047 | 67.19 | 2.83 |
| CLEVRER | 8,220 | 5,556 | 67.59 | 2.60 |
| Spacellava | 10,000 | 5,491 | 54.91 | 2.57 |
| OpenSpaces | 10,284 | 4,708 | 45.78 | 2.20 |
| ThinkLite-VL-Hard | 4,336 | 4,167 | 96.10 | 1.95 |
| PerceptionTest | 6,348 | 3,857 | 60.76 | 1.81 |
| TabMWP | 4,471 | 3,799 | 84.97 | 1.78 |
| A-OKVQA | 3,999 | 3,491 | 87.30 | 1.63 |
| ScienceQA | 3,999 | 3,447 | 86.20 | 1.61 |
| PMC-VQA | 5,982 | 3,308 | 55.30 | 1.55 |
| Other sources | 100,090 | 48,766 | 48.72 | 22.83 |
| **Total** | **328,309** | **213,611** | **65.06** | **100.00** |

Table 1: Major source groups in the cleaned CoT dataset.

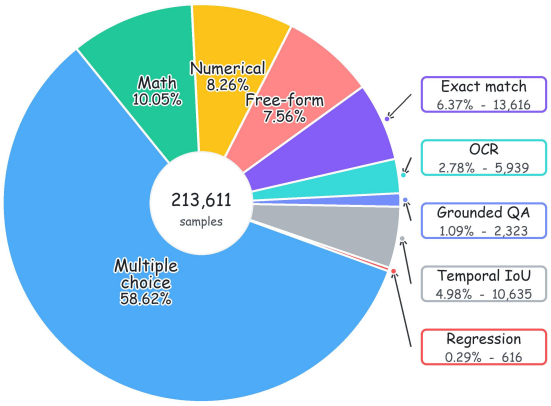



Figure 4: Task-type distribution of the cleaned CoT dataset.

### 2.2 Multi-Agent Dataset Construction and Cleaning

As illustrated in Stage 1 of Fig. 3, our data construction and cleaning pipeline transforms noisy public data into reliable reasoning supervision through three successive steps: multi-agent CoT generation, task-specific verification and judge-agent selection, and targeted human review. Starting from 328,309 samples drawn from 64 source tags spanning video, image, and text reasoning tasks, we construct a high-quality chain-of-thought (CoT) corpus before reinforcement-learning-based training. The objective is not merely to retain samples with correct final answers, but to obtain reasoning traces that are grounded in the input evidence and do not directly disclose the reference answers.

For each raw sample, six heterogeneous open- and closed-source agents independently generate a reasoning trace and a final answer from the original input, including the available video, image, or text context and the corresponding question. The reference answer is deliberately withheld during this generation process. Conditioning a model on the correct answer and asking it to generate a rationale can produce label-leaking or hallucinated justifications, *e.g*, "we already know that the ground-truth answer is ...." Although such traces may arrive at the correct final answer, they do not constitute valid supervision for reasoning. Independent generation without access to the reference answer therefore encourages each agent to derive its response from the input evidence.

An answer-extraction agent subsequently parses the final answer from each model output and normalizes it into the format required by the corresponding task. A task-specific evaluator then compares the extracted answer with the original annotation. This process converts the six independently generated traces into a set of valid candidates for subsequent selection. Formally, let $(r_{ij}, a_{ij})$ denote the reasoning trace and final answer generated by agent $j$ for sample $i$. Let $\mathcal{E}_i(a_{ij}, a_i^\star) \in \{0, 1\}$ denote the corresponding task-specific evaluator, where a value of one indicates that the generated answer satisfies the correctness criterion. We define the set of correct candidates as$\mathcal{C}_i = \{j \mid \mathcal{E}_i(a_{ij}, a_i^\star) = 1\}$. The retained output is selected according to

$$(\tilde{r}_i, \tilde{a}_i) = \begin{cases} \varnothing, & |\mathcal{C}_i| = 0, \\ (r_{ij_i^\dagger}, a_{ij_i^\dagger}), & |\mathcal{C}_i| = 1, \\ (r_{ij^\star}, a_{ij^\star}), \quad j^\star = \arg\max_{j \in \mathcal{C}_i} \mathcal{J}(r_{ij}), & |\mathcal{C}_i| > 1, \end{cases} \tag{2}$$

where $j_i^\dagger$ denotes the unique element of $\mathcal{C}_i$, and $\mathcal{J}$ denotes the judge agent's score for faithfulness to the input evidence, logical completeness, and absence of direct reference-answer leakage. equation 2 discards samples for which none of the agents produces a correct answer, directly retains the unique correct candidate when only one exists, and invokes the judge agent when multiple valid traces remain. The reference answer is used to verify and filter the generated candidates.

After automatic filtering and human review, the cleaned dataset contains 213,611 samples, corresponding to 65.06% of the original corpus. Approximately 203K of these samples contain complete and usable CoT traces, from which 200K examples are used for SFT initialization in the subsequent training stage. As shown in Tab. 1, the cleaned corpus preserves broad coverage across major video, image, and text sources. Meanwhile, Fig. 4 shows higher retention rates for multiple-choice,

mathematical reasoning, and exact-match tasks, but more aggressive filtering for temporal IoU and regression tasks. Notably, manual review of approximately 4,000 samples for which at least four agents agreed on an answer that conflicted with the original annotation corrected and retained 2,324 samples while excluding unresolved cases, demonstrating the effectiveness of multi-agent consensus in identifying annotation noise.

The resulting corpus is therefore not simply an answer-filtered dataset. Multi-agent generation provides diverse candidate reasoning processes, task-specific evaluation ensures answer validity, judge-agent selection favors evidence-grounded and logically complete traces, and human review corrects errors that cannot be resolved reliably through automatic evaluation alone. This combination provides cleaner and denser reasoning supervision for the reflective training.

### 2.3 Reflective Multi-Turn Training

As illustrated in Stage 2 of Fig. 3, reflective multi-turn training alternates between rollout-based sample filtering and policy optimization. The first reflective turn establishes the model's initial video reasoning capability, while subsequent turns teach it to evaluate and revise preceding responses. Each optimization iteration uses fresh rollouts from the current policy, allowing the training data to evolve with the model. To avoid ambiguity, we use $m$ to index optimization iterations, $\ell$ to index reflective turns, and $k$ to index multiple rollouts of the same sample.

We first initialize the Turn 1 policy through supervised fine-tuning on 200K cleaned CoT examples from Sec. 2.2. The first-turn output contains the first reasoning trace, the first answer, the second reasoning trace, and the second answer. After SFT initialization, we construct an initial reinforcement-learning candidate pool $\mathcal{U}^{(0)}$ containing approximately 300K eligible annotated samples.

Before optimization iteration $m$, the current policy generates $K$ independent rollouts for each sample $x_i = (v_i, q_i, a_i^\star)$ in $\mathcal{U}^{(0)}$. We define the correctness of rollout $k$ and the training subset as

$$c_{i,k}^{(m)} = \mathbb{I}\Big[\mathrm{Eval}\Big(a_{i,k}^{(m)}, a_i^\star\Big) = 1\Big], \qquad \mathcal{U}^{(m)} = \left\{ x_i \in \mathcal{U}^{(0)} \;\middle|\; 0 < \sum_{k=1}^{K} c_{i,k}^{(m)} < K \right\}. \tag{3}$$

This criterion excludes all-correct and all-incorrect rollout groups, which provide little within-group reward variation for group-relative optimization, while retaining mixed-outcome groups that offer informative contrasts between successful and unsuccessful reasoning trajectories.

During the first optimization iteration, the SFT-initialized policy is trained with GRPO on $\mathcal{U}^{(1)}$ to enhance its reasoning capability. For the subsequent iteration, $\mathcal{U}^{(2)}$ is independently constructed from the original dataset by applying the filtering criterion defined in Equation (3). Each example retained in $\mathcal{U}^{(2)}$ is then converted into a multi-turn interaction.Turn 2 receives the preceding response and generates a structured reflective completion. The model first reasons about whether the preceding answer is correct and emits a judgment. It then performs a second reasoning process to verify, preserve, or revise the preceding answer, followed by a final answer:

$$y_{i,k}^{(\ell=2)} = \langle \texttt{think} \rangle\, r_{i,k}^{\mathrm{jud}}\, \langle \texttt{/think} \rangle \boxed{\hat{z}_{i,k}^{(2)}} \langle \texttt{think} \rangle\, r_{i,k}^{\mathrm{ans}}\, \langle \texttt{/think} \rangle \boxed{\hat{a}_{i,k}^{(2)}}. \tag{4}$$

Here, $r_{i,k}^{\mathrm{jud}}$ denotes the reasoning process used to determine whether the preceding answer is correct, and $\hat{z}_{i,k}^{(2)} \in \{\mathrm{True}, \mathrm{False}\}$ is the correctness judgment. The second reasoning process $r_{i,k}^{\mathrm{ans}}$ verifies the preceding prediction against the input evidence and determines whether it should be preserved or revised. The final task answer after this verification process is denoted by $\hat{a}_{i,k}^{(2)}$.

The retained reflective trajectories are optimized jointly with GRPO and CCOPD. GRPO provides sequence-level rewards for correctness prediction, verification, and answer revision, whereas CCOPD provides dense token-level guidance from a teacher branch. The stage-specific objective is

$$\mathcal{L}_{\mathrm{train}}^{(m)} = \begin{cases} \mathcal{L}_{\mathrm{GRPO}}^{(1)}, & m = 1, \\ \mathcal{L}_{\mathrm{GRPO}}^{(m)} + \lambda \mathcal{L}_{\mathrm{CCOPD}}^{(m)}, & m \geq 2, \end{cases} \tag{5}$$

where $\lambda$ controls the contribution of CCOPD. The first optimization establishes the initial reasoning policy, while later iterations jointly optimize explicit verification and revision.

**This iterative design serves two complementary purposes.** The split-turn formulation separates initial answer generation from subsequent verification and correction, allowing each reflective turn to have a clearly defined learning objective. Meanwhile, repeated mixed-outcome filtering concentrates optimization on samples near the current policy's decision boundary, where additional supervision is most informative. As the policy improves, new rollouts expose its remaining weaknesses, producing an adaptive curriculum for progressively learning self-verification and self-correction.

### 2.4 CORRECTNESS-CONDITIONED ON-POLICY DISTILLATION

As illustrated by the *GRPO & CCOPD in Turn N* block of Fig. 3, each reflective turn is optimized using both sequence-level reinforcement-learning rewards and token-level correctness-conditioned distillation. The reflective completion follows the structured format defined in Equation (4): a first reasoning span followed by a correctness judgment, and a second reasoning span followed by a final answer. GRPO evaluates the format validity, correctness judgment, and final task answer of the complete reflective response, whereas CCOPD transfers dense token-level guidance from a teacher branch conditioned on the ground-truth correctness of the preceding response. We first describe the two-turn formulation and then extend it to later reflective turns.

For a two-turn interaction, the student receives the standard verification context, whereas the teacher receives the same context augmented with a privileged correctness label:

$$h_{i,k}^{S} = \left[v_i; q_i; y_{i,k}^{(1)}; p^{(2)}\right], \qquad h_{i,k}^{T} = \left[z_{i,k}^{\star}; v_i; q_i; y_{i,k}^{(1)}; p^{(2)}\right], \tag{6}$$

where $p^{(2)}$ is the second-turn verification prompt and $z_{i,k}^{\star} \in \{\mathrm{True}, \mathrm{False}\}$ indicates whether the post-reasoning answer from Turn 1 is correct according to the task-specific evaluator. The student must infer this correctness state from the video, question, and preceding response, whereas the teacher is explicitly conditioned on it. Neither branch directly observes the reference answer $a_i^{\star}$; the teacher receives only the binary correctness label derived from it.

CCOPD is on-policy because the reflective completion is sampled from the current student policy:

$$y_{i,k}^{(2)} \sim \pi_\theta\left(\cdot \mid h_{i,k}^{S}\right). \tag{7}$$

After sampling, the same student-generated token sequence is teacher-forced under both contexts. The teacher branch is evaluated without gradient tracking and provides a fixed token-level target. We use the student-to-teacher reverse KL on the student's on-policy prefixes:

$$\mathcal{L}_{\mathrm{CCOPD}} = \mathbb{E}_{i,k}\left[\frac{1}{|M_{i,k}^{(2)}|}\sum_{t \in M_{i,k}^{(2)}} \min\left(D_{\mathrm{KL}}\left(p_\theta^{S}(\cdot \mid h_{i,k}^{S}, y_{i,k,<t}) \,\|\, \mathrm{sg}\left[p_\theta^{T}(\cdot \mid h_{i,k}^{T}, y_{i,k,<t})\right]\right), c_{\mathrm{KL}}\right)\right], \tag{8}$$

where $c_{\mathrm{KL}} = 10.0$. Thus, CCOPD minimizes $D_{\mathrm{KL}}(p^S \| \,\mathrm{sg}[p^T])$; gradients flow only through the student distribution. This is the same reverse-KL objective used in Sec. B.7 and in all experiments.

Conditioning on $z_{i,k}^{\star}$ allows the teacher branch to define a differently-conditioned token distribution over the student's on-policy revision trajectory that, as we verify empirically below, exhibits more reliable preservation and revision behavior. When $z_{i,k}^{\star} = \mathrm{True}$, the privileged context encourages verification behavior that preserves a correct preceding answer. When $z_{i,k}^{\star} = \mathrm{False}$, it encourages the model to identify inconsistencies and produce an appropriate revision. The student is optimized to approximate these correctness-aware token distributions without the privileged label. For later reflective turns, the same construction is applied recursively by replacing the first-turn trajectory with the interaction history and computing the correctness for the preceding candidate answer.

**Why Is CCOPD Needed?** Unlike answer-privileged distillation, CCOPD provides the teacher only with the correctness of the preceding response, requiring it to infer whether and how the answer should be revised from the visual input and interaction history without exposing the reference answer. Compared with direct correctness-supervised SFT, a binary SFT target teaches only whether revision is needed, while trajectory-level SFT imitates fixed offline targets and does not supervise the errors encountered by the evolving student policy. CCOPD instead matches the teacher's soft token distribution along the student's on-policy trajectory across the verification, judgment, and revision spans. It therefore supplies dense and adaptive guidance at the states actually visited by the student, complementing sequence-level rewards with supervision for both when and how to revise.

Table 2: Benchmark results across video understanding and reasoning. The best score is **bold** and the second-best is underlined. * denotes running by ourselves. Other experimental results are collected from the accepted papers, *e.g.*, VideoKR Fu et al. (2026), Video-o3 Zeng et al. (2026), and so on.

| | Method | Frames | Video Understanding | | | Video Reasoning | |
|---|---|---|---|---|---|---|---|
| | | | VideoMME | MV. | LongVideo. | VideoMMMU | MMVU |
| Base Models | GPT-5.4 Singh et al. (2025) | 64 | 86.0 | 78.3 | 76.7 | 87.1 | 82.0 |
| | Gemini 3 Pro Comanici et al. (2025) | 64 | 87.7 | 74.1 | 77.4 | 87.6 | 77.5 |
| | Claude Opus 4.5 Anthropic (2025) | 64 | 81.4 | 67.2 | 67.2 | 84.4 | 77.3 |
| | Qwen3-VL-32B-Thinking Bai et al. (2025a) | 128 | 72.9 | 71.7 | 63.8 | 69.6 | 67.8 |
| | Qwen3-VL-32B-Instruct Bai et al. (2025a) | 128 | 74.4 | 72.6 | 65.5 | 72.0 | 68.2 |
| | Qwen2.5-VL-72B Bai et al. (2025b) | 128 | 72.8 | 67.4 | 63.2 | 67.0 | 65.1 |
| | InternVL3.5-8B Wang et al. (2025c) | 128 | 65.5 | 73.3 | 61.0 | 57.2 | 54.0 |
| Qwen2.5-VL-7B | Qwen2.5-VL-7B-Instruct Bai et al. (2025b) | 16 | 57.1* | 62.0* | 55.2* | 48.4* | 52.5* |
| | Video-MTR Xie et al. (2025) | 16 | 55.8* | 60.9* | 54.4* | 49.4* | 63.4* |
| | Video-R1 Feng et al. (2026a) | 16 | 59.7 | 65.5 | 55.3 | 51.1 | 53.3 |
| | Video-VER Luo et al. (2025) | 16 | 56.0 | 63.2 | - | 50.0 | 64.8 |
| | VideoRFT Wang et al. (2026b) | 16 | 57.6 | 61.7 | 53.6 | 51.1 | 53.6 |
| | VideoKR Fu et al. (2026) | 16 | 56.6 | 66.6 | **57.0** | **52.6** | 59.2 |
| | VideoKTR Wang et al. (2026e) | 16 | 57.3 | - | - | 51.3 | 65.7 |
| | **Ours** | 16 | **62.7**$_{\pm 0.6}$ | **66.8**$_{\pm 0.3}$ | 56.7$_{\pm 0.5}$ | 51.7$_{\pm 0.9}$ | **65.8**$_{\pm 0.6}$ |
| | Qwen2.5-VL-7B-Instruct Bai et al. (2025b) | 128 | 65.1* | 63.9* | 60.9* | 51.1* | 55.7* |
| | Video-MTR Xie et al. (2025) | 128 | 61.9* | 61.7* | 57.7* | 50.4* | 61.9* |
| | VideoAuto-R1 Liu et al. (2026) | 128 | **66.8** | 70.2 | 59.7 | 52.1 | 55.7 |
| | VideoKR Fu et al. (2026) | 128 | 66.4 | 68.9 | 61.3 | 52.2 | 60.5 |
| | VideoZoomer Ding et al. (2026) | 128 | 65.2 | - | 57.7 | 52.2 | 65.6 |
| | ResAdapt Liao et al. (2026) | 128 | 65.6 | - | 60.2 | 51.1 | 52.8 |
| | **Ours** | 128 | 66.2$_{\pm 0.7}$ | **71.3**$_{\pm 1.1}$ | **61.4**$_{\pm 0.5}$ | **55.8**$_{\pm 0.7}$ | **68.0**$_{\pm 0.5}$ |
| | Qwen2.5-VL-7B-Instruct Bai et al. (2025b) | 768 | 65.2* | 63.9* | 56.0* | 51.4* | 64.9* |
| | VITAL Zhang et al. (2026b) | 768 | 66.1 | - | 58.5 | 54.2 | 68.7 |
| | VideoZoomer Ding et al. (2026) | 768 | 64.6 | - | 55.9 | 51.4 | 65.6 |
| | Love-r1 Fu et al. (2025b) | 768 | 64.3 | - | 58.0 | 50.6 | 65.0 |
| | Video-o3 Zeng et al. (2026) | 768 | 66.5 | - | 60.5 | 51.7 | 67.2 |
| | **Ours** | 768 | **66.8**$_{\pm 1.0}$ | **72.1**$_{\pm 0.9}$ | **62.3**$_{\pm 0.5}$ | **56.6**$_{\pm 1.1}$ | **70.3**$_{\pm 0.9}$ |
| Qwen3-VL-8B | Qwen3-VL-8B-Instruct Bai et al. (2025a) | 128 | 68.2* | 67.9* | 61.6* | 60.3* | 59.6* |
| | OneThinker Feng et al. (2026b) | 128 | 65.8 | 69.3 | 61.4 | 62.9 | 61.6 |
| | VideoAuto-R1 Liu et al. (2026) | 128 | 68.7 | 68.8 | 58.8 | 63.1 | 59.6 |
| | Qwen3-VL-8B-Thinking Bai et al. (2025a) | 128 | 67.6 | 68.0 | 60.0 | 64.9 | 60.5 |
| | VideoKR Fu et al. (2026) | 128 | 67.8 | 67.0 | 61.5 | 63.0 | 64.8 |
| | ResAdapt Liao et al. (2026) | 128 | 67.4 | - | 61.9 | 59.6 | 56.3 |
| | **Ours** | 128 | **69.6**$_{\pm 0.9}$ | **73.9**$_{\pm 1.0}$ | **66.7**$_{\pm 0.8}$ | **67.3**$_{\pm 1.3}$ | **69.8**$_{\pm 0.9}$ |

# 3 EXPERIMENTS

## 3.1 DATASET AND IMPLEMENTATION DETAILS

The model is evaluated on 13 benchmarks spanning 4 categories: short-video understanding, long-video understanding, video reasoning, and temporal reasoning. In addition to Tabs. 2 and 3, more results are provided in Tab. 19. All experiments are conducted on 32 NVIDIA Tesla H20 GPUs.

**Statement.** Related Works (Sec. A), Dataset and Implementation Details (Sec. C), Additional Results (Sec. D.1), Analysis of Answer Transitions(Sec. D.4), Computation Comparisons(Sec. D.5), **Additional Ablation Study** (Sec. D.6), and Visualization (Sec. E) are provided in Appendix.

## 3.2 MAIN RESULTS

**Video QA Benchmarks.** As shown in Tab. 2, CCOPD provides consistent improvements across a range of video understanding and reasoning benchmarks. With Qwen2.5-VL-7B and 16 frames, it achieves 62.7 on VideoMME, 66.8 on MVBench, and 65.8 on MMVU, while remaining close to the strongest reported results on LongVideoBench and VideoMMMU. When the frame is increased to 128, CCOPD obtains 71.3 on MVBench, 61.4 on LongVideoBench, 55.8 on VideoMMMU, and 68.0 on MMVU, corresponding to improvements of 1.1, 0.1, 3.6, and 2.4 over the strongest competing

Table 3: Benchmark results on temporal grounding. The best is **bold** and the second-best is underlined. * denotes running by ourselves. Others are collected from the accepted papers.

| | Method | Temporal Grounding Benchmark | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Charades-STA | | | | ActivityNet | | | | NExT-GQA | |
| | | 0.3 | 0.5 | 0.7 | mIoU | 0.3 | 0.5 | 0.7 | mIoU | Acc | mIoU |
| SFT | TimeChat Ren et al. (2024) | 51.0 | 27.5 | 11.4 | 31.2 | 44.0 | 27.8 | 14.3 | 30.4 | 28.8 | 17.4 |
| SFT | TimeMarker Chen et al. (2024) | 73.5 | 51.9 | 26.9 | 48.4 | 67.4 | 50.7 | 33.0 | 49.5 | - | - |
| Qwen2.5-VL-7B | Qwen2.5-VL-7B Bai et al. (2025b) | 77.3* | 59.7* | 34.4* | 52.8* | 42.1* | 26.6* | 13.2* | 29.7* | 79.6* | 26.5* |
| Qwen2.5-VL-7B | Temporal-RLT Li et al. (2025) | 79.6 | 67.9 | 44.1 | 57.0 | 56.9 | 38.4 | 20.2 | 39.0 | 78.7 | 37.3 |
| Qwen2.5-VL-7B | Time-R1 Wang et al. (2026c) | 82.8 | 72.2 | **50.1** | 58.8 | **73.3** | **55.6** | 34.0 | **52.1** | - | - |
| Qwen2.5-VL-7B | VITAL Zhang et al. (2026b) | 83.1 | 72.0 | 46.7 | 59.9 | 70.9 | 50.8 | 31.6 | 49.8 | 78.7 | **43.0** |
| Qwen2.5-VL-7B | VideoChat-R1.5 Yan et al. (2026) | 82.8 | 71.6 | 48.3 | **60.6** | 52.4 | 32.3 | 16.8 | 35.3 | - | - |
| Qwen2.5-VL-7B | FlashVid (Fan et al., 2026) | 40.7 | 24.2 | 11.3 | 26.6 | 15.8 | 8.4 | 3.8 | 12.0 | 78.1 | 16.5 |
| Qwen2.5-VL-7B | ResAdapt Liao et al. (2026) | 63.5 | 43.6 | 21.3 | 42.0 | 33.1 | 19.3 | 10.2 | 24.3 | 78.1 | 27.2 |
| Qwen2.5-VL-7B | VideoAuto-R1 Liu et al. (2026) | 82.9 | 70.8 | 46.0 | 60.0 | 69.2 | 48.5 | 27.3 | 47.6 | 80.6 | 36.7 |
| Qwen2.5-VL-7B | **Ours** | **83.8** ±1.0 | **72.5** ±0.6 | 45.5 ±0.5 | 60.3 ±0.9 | 70.2 ±1.1 | 51.7 ±0.5 | 31.8 ±0.4 | 51.9 ±0.8 | **82.1** ±1.1 | 37.8 ±0.8 |
| Qwen3-VL-8B | Qwen3-VL-8B Bai et al. (2025b) | 71.5* | 45.0* | 18.9* | 44.5* | 45.1* | 30.7* | 18.7* | 33.3* | 81.3* | 37.2* |
| Qwen3-VL-8B | FlashVid (Fan et al., 2026) | 68.8 | 46.9 | 22.9 | 44.6 | 49.9 | 31.5 | 17.4 | 35.2 | 75.6 | 31.8 |
| Qwen3-VL-8B | ResAdapt Liao et al. (2026) | 64.4 | 37.0 | 15.9 | 39.8 | 40.6 | 26.7 | 15.7 | 30.0 | 76.8 | 33.3 |
| Qwen3-VL-8B | OneThinker Feng et al. (2026b) | 83.5 | 68.3 | 45.3 | 59.9 | 65.0 | 43.6 | 25.7 | 45.9 | - | - |
| Qwen3-VL-8B | VideoAuto-R1 Liu et al. (2026) | 85.1 | 74.9 | **53.7** | **63.7** | 74.1 | 54.3 | 32.4 | 51.9 | 81.1 | 44.2 |
| Qwen3-VL-8B | **Ours** | **86.4** ±1.2 | **75.2** ±0.8 | 46.8 ±0.7 | 63.5 ±1.1 | **76.7** ±1.5 | **59.9** ±1.1 | **38.6** ±1.0 | **57.1** ±1.3 | **83.8** ±1.5 | **45.6** ±1.4 |

results listed under the same setting; its VideoMME accuracy of 66.2 is also comparable to the best reported score of 66.8. With 768 frames, the model reaches 66.8, 72.1, 62.3, 56.6, and 70.3 on the five benchmarks, respectively, showing that its improvements are maintained with a larger temporal context. Similar trends are observed when applying Ours to Qwen3-VL-8B, where it achieves 69.6 on VideoMME, 73.9 on MVBench, 66.7 on LongVideoBench, 67.3 on VideoMMMU, and 69.8 on MMVU. Overall, these results suggest that CCOPD can provide reliable performance gains across different backbones and frames, particularly on benchmarks requiring temporal understanding.

**Temporal Grounding Benchmarks.** Tab. 3 summarizes the temporal grounding performance on Charades-STA, ActivityNet, and NExT-GQA. With Qwen2.5-VL-7B, CCOPD improves the mIoU of the baseline from 52.8 to 60.3 on Charades-STA and from 29.7 to 51.9 on ActivityNet. On Charades-STA, it achieves the highest reported recall at IoU thresholds of 0.3 and 0.5, while its mIoU remains close to the strongest result in the table (60.3 versus 60.6). On ActivityNet, the model obtains 51.7 and 31.8 recall at IoU thresholds of 0.5 and 0.7, respectively, together with an mIoU of 51.9, showing competitive performance relative to existing temporal reasoning methods. On NExT-GQA, CCOPD increases the base-model accuracy from 79.6 to 82.1 and the mIoU from 26.5 to 37.8. Applying the same approach to Qwen3-VL-8B leads to further gains, yielding an mIoU of 63.5 on Charades-STA, 57.1 on ActivityNet, and 45.6 on NExT-GQA. In particular, the ActivityNet mIoU exceeds the strongest comparison result by 5.2 percentage points, while the NExT-GQA accuracy reaches 83.8. Overall, these results indicate that CCOPD improves temporal localization across different backbones and datasets, with particularly clear gains on ActivityNet and NExT-GQA.

**Statement.** For consistent comparisons, results in Tabs. 2 and 3 are grouped by backbone and frame count, with comparisons made under matched evaluation protocols (*e.g.*, lmms-eval Zhang et al. (2025b)). Our experiments follow the corresponding backbone, frame-count, and evaluation settings. The reported scores for **Ours** are means across three independent training seeds, with ± corresponding to the minimum and maximum scores observed across these runs, respectively.

### 3.3 Ablation Studies

**Training Strategy.** Table 4 validates the proposed stage-wise pipeline. For Qwen2.5-VL-7B, the average score increases from 59.3 for the baseline to 59.9 after SFT, 61.6 after Turn1 GRPO, and 64.3 after Turn2 GRPO&CCOPD; Turn3 further improves the average to 64.5, although its gains are smaller and task dependent. Applying GRPO without SFT achieves a comparable Turn1 average (61.7 *vs.* 61.6), but reaches only 62.4 after Turn2, indicating that SFT provides a stronger initialization for learning subsequent reflective behavior. The Turn2 ablations further distinguish CCOPD from direct correctness supervision and conventional distillation: True/False SFT, GRPO

Table 4: Ablation of the multi-turn training strategy under 128-frame input with Qwen2.5-VL-7B and Qwen3-VL-8B. Dataset Num. denotes the number of samples, and the best score is **bold**.

| Method | Dataset Num. | Video Understanding | | | Video Reasoning | | Average |
|---|---|---|---|---|---|---|---|
| | | VideoMME | MV. | LongVideo. | VideoMMMU | MMVU | |
| *Our Strategy* | | | | | | | |
| **Qwen2.5-VL-7B** | - | 65.1 | 63.9 | 60.9 | 51.1 | 55.7 | 59.3 |
| **w/** SFT (Turn 1) | 200K | 62.2 | 65.1 | 56.9 | 52.0 | 63.4 | 59.9 |
| **w/** GRPO (Turn 1) | 138K | 63.4 | 65.3 | 60.2 | 53.2 | 66.1 | 61.6 |
| **w/** GRPO&CCOPD (Turn 2) | 38K | 65.6 | 69.1 | 61.3 | **56.9** | **68.8** | 64.3 |
| **w/** GRPO&CCOPD (Turn 3) | 18K | **66.2** | **71.3** | **61.4** | 55.8 | 68.0 | **64.5** |
| *Ablation Studies in Turn 1* | | | | | | | |
| **w/** Only GRPO (Turn 1) | 138K | 64.4 | 65.2 | 58.9 | 52.8 | 67.3 | 61.7 |
| **w/** GRPO&CCOPD (Turn 2) | 38K | 65.5 | 67.0 | 58.8 | 52.6 | 68.3 | 62.4 |
| *Ablation Studies in Turn 2* | | | | | | | |
| **SFT&GRPO** (in Turn 1) | 138K | 63.4 | 65.3 | 60.2 | 53.2 | 66.1 | 61.6 |
| **w/** Only SFT$^\star$ (Turn 2) | 38K | 62.7 | 63.5 | 58.7 | 49.9 | 60.1 | 59.0 |
| **w/** Only GRPO (Turn 2) | 38K | 64.1 | 67.8 | 60.3 | 51.0 | 67.0 | 62.0 |
| **w/** Only OPSD$^*$ (Turn 2) | 38K | 62.1 | 64.1 | 54.2 | 54.1 | 67.8 | 60.5 |
| **w/** GRPO&OPSD$^*$ (Turn 2) | 38K | 64.8 | 67.1 | 58.8 | 52.7 | 68.3 | 62.3 |
| *Ablation Studies of teacher conditioning in Turn 2* | | | | | | | |
| **w/** GRPO&CCOPD$_1^\dagger$ (Turn 2) | 38K | 63.1 | 64.4 | 56.1 | 50.5 | 64.7 | 59.8 |
| **w/** GRPO&CCOPD$_2^\dagger$ (Turn 2) | 38K | 64.2 | 65.1 | 58.7 | 53.7 | 67.1 | 61.8 |
| **w/** GRPO&CCOPD$_3^\dagger$ (Turn 2) | 38K | 64.8 | 68.3 | 59.2 | 55.1 | 67.9 | 63.1 |
| **w/** GRPO&CCOPD$_4^\dagger$ (Turn 2) | 38K | 63.5 | 64.6 | 56.2 | 52.1 | 66.5 | 60.6 |
| *Our Strategy* | | | | | | | |
| **Qwen3-VL-8B** | - | 68.2 | 67.9 | 61.6 | 60.3 | 59.6 | 63.5 |
| **w/** SFT (Turn 1) | 200K | 66.4 | 69.1 | 60.7 | 62.3 | 62.6 | 64.2 |
| **w/** GRPO (Turn 1) | 103K | 68.5 | 71.1 | 62.6 | 64.1 | 64.8 | 66.2 |
| **w/** GRPO&CCOPD (Turn 2) | 33K | 69.5 | 72.8 | 66.2 | 67.3 | 68.7 | 68.9 |
| **w/** GRPO&CCOPD (Turn 3) | 15K | **69.6** | **73.9** | **66.7** | **67.3** | **69.8** | **69.5** |

*Note:* $\star$ denotes direct supervised finetuning on the binary True/False judgment and final answer.
† denotes modified teacher conditioning. $\text{CCOPD}_1$: randomly correctness; $\text{CCOPD}_2$: removing correctness; $\text{CCOPD}_3$: correctness always sets as True; $\text{CCOPD}_4$: correctness always sets as False.
$*$ indicates using ground-truth as privileged information, while all other settings are the same as CCOPD.

alone, OPSD alone, and GRPO&OPSD obtain averages of 59.0, 62.0, 60.5, and 62.3, respectively, all below the 64.3 achieved by GRPO&CCOPD. Additional ablations on teacher conditioning confirm that meaningful correctness signals are essential: random labels (59.8), removing correctness (61.8), always-True (63.1), and always-False (60.6) all underperform the true-correctness variant, demonstrating that the gain arises from accurate correctness-conditioned distributional guidance rather than merely adding a distillation objective or label exposure. The same progression generalizes to Qwen3-VL-8B, whose average improves from 63.5 to 64.2, 66.2, 68.9, and 69.5 across the base, SFT, Turn1 GRPO, Turn2 GRPO&CCOPD, and Turn 3 GRPO&CCOPD stages. Overall, SFT establishes a stable initialization, GRPO improves sequence-level task performance, and CCOPD provides dense on-policy supervision for learning how to verify and revise preceding answers.

**More ablations** are in Sec. D.6, including CoT construction (Tab. 15), dataset filtering (Tab. 16), hyperparameters (Tab. 17), and divergence objectives and teacher updating strategies (Tab. 18).

## 4 Conclusion

In this work, we presented a reflective multi-turn framework that enables video-language models to evaluate and revise their previous predictions without privileged feedback at inference. The framework combines multi-agent dataset filtering, iterative mixed-outcome sample selection, and GRPO&CCOPD optimization, in which correctness-conditioned teacher signals provide privileged token-level supervision along the student's on-policy trajectories. Experiments on 13 benchmarks spanning short- and long-video understanding, video reasoning, and temporal grounding demonstrate consistent and broadly competitive performance across diverse tasks. These results provide evidence that learning from correctness-conditioned revision trajectories can improve verification and self-correction ability, offering a promising direction for developing reliable reasoning models.

### AI USE STATEMENT

In this work, we used generative AI tools for assisting with the design of experiments and interpreting experimental results. Additionally, we used generative AI tools for drafting and language editing of limited parts of the manuscript. All AI-assisted content was carefully reviewed, revised, and verified by the authors. In particular, all LLM-assisted code was independently reviewed and tested by at least two authors; all AI-assisted text was checked against the cited literature and revised for technical accuracy; and all experimental results were reproduced using the final implementation. We take responsibility for the final content of this work, including all text, claims, code, data, figures, and other artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We support the reproducibility of this work through detailed descriptions of the proposed method, experimental setup, and evaluation protocol in the main paper. Specifically, Sec. 2 describes the methodology and model design; Sec. 3 and Sec. C report the datasets, baselines, implementation details, and evaluation metrics; and the appendix provides hyperparameter settings, additional ablations, proofs, and expanded experimental results.

Anonymous source code, including scripts for data preprocessing, evaluation, and reproducing the reported results, is provided in the supplementary material.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Anthropic. Claude opus 4.5 system card. Technical report, Anthropic, November 2025. URL https://www.anthropic.com/claude-opus-4-5-system-card. Accessed: 2026-08-12.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024.

Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Bohan Zeng, Yang Shi, Sihan Yang, Pengfei Wan, Qiang Liu, Liang Wang, and Tieniu Tan. Versavid-r1: A versatile video understanding and reasoning model from question answering to captioning tasks. *arXiv e-prints*, pp. arXiv–2506, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 2026.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. FlashVID: Efficient video large language models via training-free tree-based spatiotemporal token merging. In *The Fourteenth International Conference on Learning Representations*, 2026.

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 *Advances in Neural Information Processing Systems*, volume 38, pp. 99114–99137, 2026a.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5432–5443, 2026b.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025a.

Lin Fu, Zheyuan Yang, Yang Wang, Tingyu Song, Arman Cohan, and Yilun Zhao. VideoKR: Towards knowledge- and reasoning-intensive video understanding. In *Forty-third International Conference on Machine Learning*, 2026.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. *arXiv preprint arXiv:2509.24786*, 2025b.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

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

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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.

Jiazhen Huang, Xiao Chen, Xiao Luo, Yong Dai, Senkang Hu, and Yuzhi Zhao. Skill-conditioned gated self-distillation for llm reasoning. *arXiv preprint arXiv:2605.28791*, 2026.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Dongwei Jiang, Jingyu Zhang, Orion Weller, Nathaniel Weir, Benjamin Van Durme, and Daniel Khashabi. Self-[in] correct: Llms struggle with discriminating self-generated responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24266–24275, 2025.

Yifan Jiang, Yueying Wang, Rui Zhao, Toufiq Parag, Zhimin Chen, Zhenyu Liao, and Jayakrishnan Unnikrishnan. Videop2r: Video understanding from perception to reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8303–8313, 2026.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*, 2025.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*, 2025.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-OPD: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. In *Forty-third International Conference on Machine Learning*, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024a.

Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *arXiv preprint arXiv:2402.12563*, 2024b.

Huanxuan Liao, Zhongtao Jiang, Yupu Hao, Yuqiao Tan, Shizhu He, Ben Wang, Jun Zhao, Kun Xu, and Kang Liu. Resadapt: Adaptive resolution for efficient multimodal reasoning. *arXiv preprint arXiv:2603.28610*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36, pp. 34892–34916, 2023.

Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, Ernie Chang, Saksham Suri, Hongyu Xu, Qi Qian, Wei Wen, Balakrishnan Varadarajan, Zhuang Liu, Hu Xu, Florian Bordes, Raghuraman Krishnamoorthi, Bernard Ghanem, Vikas Chandra, and Yunyang Xiong. Videoauto-r1: Video auto reasoning via thinking once, answering twice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32829–32839, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, August 2024.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yingzi Ma, Zichen Zhu, Ming Jiang, and Chaowei Xiao. One student, many teachers: Multi-task on-policy distillation via soft-prompt privileged context. *arXiv preprint arXiv:2607.18293*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in neural information processing systems*, volume 36, pp. 46534–46594, 2023.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41089–41099, 2026.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2024.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635, 2011.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

Ao Shen, Yongheng Zhang, Yinghui Li, Manning Wang, Di Yin, and Xing Sun. Deep thought alignment: Trajectory-level latent distillation for video reasoning. *arXiv preprint arXiv:2608.16316*, 2026.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8523–8533. IEEE, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in neural information processing systems*, volume 36, pp. 8634–8652, 2023.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Liyan Tang, Fangcong Yin, and Greg Durrett. Visually grounded self-reflection for vision-language models via reinforcement learning. *arXiv preprint arXiv:2607.02490*, 2026.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Yuanzhu Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908, 2024.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, et al. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents. *arXiv preprint arXiv:2604.10674*, 2026a.

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

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967, October 2025b.

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*, 2025c.

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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 83330–83364, 2026c.

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*, pp. 28114–28128, 2025d.

Ziyue Wang, Shiqi Huang, Weiwen Xu, Bihan Wen, and Xudong Jiang. Video-opsd: Exploiting privileged visual evidence for on-policy self-distillation in video large language models. *arXiv preprint arXiv:2608.27065*, 2026d.

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Jiang Xudong. Video-ktr: Reinforcing video reasoning via key token attribution. In *International Conference on Learning Representations*, volume 2026, pp. 71553–71570, 2026e.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*, 2022.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *arXiv preprint arXiv:2508.20478*, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 119152–119184, 2026.

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, et al. Opid: On-policy skill distillation for agentic reinforcement learning. *arXiv preprint arXiv:2606.26790*, 2026a.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026b.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, et al. Multi-rollout on-policy distillation via peer successes and failures. *arXiv preprint arXiv:2605.12652*, 2026.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, Ziang Yan, Yi Wang, Hongjie Zhang, Yali Wang, and Limin Wang. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9050–9061. IEEE, 2025a.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026b.

Haoyu Zhang, Yuwei Wu, Pengxiang Li, Xintong Zhang, Zhi Gao, Rui Gao, Mingyang Gao, Che Sun, and Yunde Jia. Mirror: Multimodal iterative reasoning via reflection on visual regions. *arXiv e-prints*, 2026c.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025b.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026.

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.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, June 2025.

# CONTENTS

**A Related work** **17**

A.1 Reinforcement Learning for Video Reasoning . . . . . 17

A.2 On-Policy Distillation . . . . . 17

A.3 Multi-Turn Self-Correction . . . . . 18

**B Model Details** **19**

B.1 Training Schedule and Data Flow . . . . . 19

B.2 CoT Construction and Supervised Initialization . . . . . 19

B.3 Task Evaluators and Output Parsing . . . . . 20

B.4 Policy-Dependent Mixed-Outcome Filtering . . . . . 20

B.5 Turn-Specific Rewards and Stage Masks . . . . . 21

B.6 Exact GRPO Objective with Turn-1 Stage Masking . . . . . 22

B.7 Correctness-Conditioned On-Policy Distillation . . . . . 23

B.8 Joint Objective and Recursive Extension . . . . . 24

B.9 Correctness-Guided Multi-Turn Inference . . . . . 25

**C Experiment Details** **26**

C.1 Implementation Details . . . . . 26

C.2 Training Dataset . . . . . 26

C.3 Evaluation Benchmarks . . . . . 27

**D Additional Results** **27**

D.1 Results on 13 VQA Benchmarks . . . . . 27

D.2 Controlled Multi-Turn Comparison . . . . . 29

D.3 Does Correctness Conditioning Create a Meaningful Information Gap? . . . . . 30

D.4 Selective Error Detection and Net Correction Utility . . . . . 31

D.5 Computational Overhead and Efficiency . . . . . 36

D.6 Ablation Studies . . . . . 37

**E Visualization** **39**

E.1 Long-Horizon Training Stability of CCOPD . . . . . 39

E.2 Experimental Results . . . . . 39

E.3 Prompts in Multi-Agent Dataset Filtering . . . . . 40

**F Limitations** **40**

# A Related work

## A.1 Reinforcement Learning for Video Reasoning

Reasoning-oriented video models can be grouped into three lines: general-purpose multimodal foundation models, verifier-based reinforcement learning for fixed video observations, and active-perception methods that learn when and where to acquire additional evidence. The first line extends vision language models to dynamic resolution, temporal alignment, and long multimodal context, while the latter two inherit chain-of-thought prompting and policy-gradient post-training from language reasoning Wei et al. (2022); Schulman et al. (2017). In particular, DeepSeek-Math Shao et al. (2024) introduced Group Relative Policy Optimization (GRPO), which replaces PPO's learned critic with group-relative advantages computed from multiple responses to the same prompt; DeepSeek-R1 Guo et al. (2025) later showed that verifiable-reward RL can elicit deliberation and self-verification without annotated reasoning traces. Video-R1 Feng et al. (2026a) was among the early video-specific extensions, introducing temporally aware GRPO to distinguish ordered from shuffled visual evidence.

General-purpose baselines include proprietary frontier models GPT-5.4 Singh et al. (2025), Gemini 3 Pro Comanici et al. (2025), and Claude Opus 4.5 Anthropic (2025), together with open models Qwen2.5-VL Bai et al. (2025b), Qwen3-VL Bai et al. (2025a), and InternVL3.5 Wang et al. (2025c). Fixed-observation video RL methods include Video-R1 Feng et al. (2026a), VideoRFT Wang et al. (2026b), Video-VER Luo et al. (2025), VideoKR Fu et al. (2026), Video-KTR Wang et al. (2026e), and OneThinker Feng et al. (2026b). Among them, Video-VER explicitly rewards visually supported reasoning, whereas Video-KTR attributes learning signals to visually and temporally important tokens. A complementary line allocates computation or retrieves evidence adaptively: Video-MTR Xie et al. (2025) performs multi-turn clip retrieval; VideoZoomer Ding et al. (2026), VITAL Zhang et al. (2026b), LOVE-R1 Fu et al. (2025b), and Video-o3 Zeng et al. (2026) learn temporal zooming or tool-mediated clue seeking; ResAdapt Liao et al. (2026) assigns query-dependent resolution budgets; and VideoAuto-R1 Liu et al. (2026) uses an answer–think–answer objective with confidence-based early exit.

These methods primarily improve how a model reasons over a single response, grounds a rationale, or acquires additional visual evidence. Our objective is different: we train a multi-turn policy to decide whether the answer should be preserved or revised. Turn 2 observes the original video and the first-turn response, while the training objective evaluates the transition itself, rewarding incorrect-to-correct changes and penalizing corruption of an already-correct answer. The proposed framework therefore complements evidence-search and adaptive-compute methods: it learns selective revision after an answer has been produced, rather than only improving the initial reasoning trajectory or expanding the observed video context.

## A.2 On-Policy Distillation

Distillation for autoregressive reasoning broadly follows three related paths: offline knowledge distillation, on-policy distillation, and privileged self-distillation. Offline methods imitate teacher-generated targets from a fixed dataset, whereas on-policy methods supervise trajectories sampled from the evolving student policy, reducing the discrepancy between training and inference distributions. Privileged self-distillation further constructs a stronger teacher from the student itself by conditioning it on information available only during training. A related line of work transforms sparse correctness feedback into dense supervision for verification and revision. These directions are particularly relevant to video reasoning, where a model must learn not only to generate an answer, but also to evaluate and revise its own response using temporally distributed visual evidence.

Classical knowledge distillation Hinton et al. (2015) and DAgger Ross et al. (2011) establish the foundations of teacher distribution matching and supervision on student-visited states, respectively. Subsequent methods construct offline reasoning traces Shi et al. (2025), distill teacher distributions along student-generated sequences Agarwal et al. (2024); Li et al. (2026), or transfer trajectory-level representations Shen et al. (2026). Privileged self-distillation instead conditions a shared self-teacher on verified solutions, soft prompts, skills, or hindsight guidance Zhao et al. (2026); Ma et al. (2026); Wang et al. (2026a); Yang et al. (2026a). In video reasoning, Video-OPSD Wang et al. (2026d) provides the teacher with annotated evidence frames to emphasize visually grounded tokens. Other

studies convert binary verifier outcomes, privileged skill labels, or cross-rollout correctness patterns into refinement supervision He et al. (2026); Huang et al. (2026); Yu et al. (2026).

The latter methods are the closest to CCOPD because they use correctness-related signals to improve a policy. However, they are primarily designed for verifier-rich settings, such as code generation and other tasks where compilers, unit tests, execution results, or rule-based evaluators can provide reliable external feedback. Video-language reasoning differs fundamentally because no such verifier is generally available at deployment: the model cannot access the reference answer, temporal annotation, or oracle correctness label and must determine from the video itself whether its preceding response is reliable. CCOPD therefore uses the gold-derived correctness state only as a training-time scaffold. A shared-backbone teacher observes this state and provides dense token-level supervision along the student's on-policy reflective trajectory, while the student receives only the video, question, and interaction history. Joint optimization with GRPO consequently internalizes a verification-and-revision policy that operates without an external verifier, separate teacher, reference rationale, or step-level human annotation at inference time.

CCOPD differs from prior approaches along three fundamental dimensions:

**(1) Multimodal Video Context.** Unlike primarily text-based refinement methods, CCOPD operates on video-language inputs whose supporting evidence may be distributed across long temporal contexts. The correctness-conditioned teacher must demonstrate how the visual evidence should be re-examined, while the student must reproduce this behavior without receiving the privileged correctness state at inference time.

**(2) Multi-Turn Reflective Structure.** CCOPD separates initial answer generation from subsequent verification and revision. Turn 1 establishes the initial reasoning state, whereas Turn 2 and later turns require the model to assess its preceding response and determine whether it should be retained or revised. This structure explicitly models inference without external correctness feedback.

**(3) Answer-Preserving vs. Answer-Correcting Conditioning.** When $z^\star = \mathrm{True}$, the teacher guides the model to verify the preceding response and generate a consistent final answer; when $z^\star = \mathrm{False}$, it guides error identification and answer correction. This dual-mode conditioning trains both conservative verification and effective revision, rather than treating correctness as only a generic outcome signal.

### A.3 Multi-Turn Self-Correction

Self-correction methods fall into inference-time self-feedback, critic- or tool-guided revision, and parameter-trained correction policies. An early training-based formulation learned a separate corrector that maps low-value sequences to improved ones Welleck et al. (2022). Self-Refine Madaan et al. (2023) and Reflexion Shinn et al. (2023) subsequently popularized inference-time iteration: Self-Refine reuses a frozen model to generate feedback and revisions, whereas Reflexion stores verbal feedback in episodic memory for later attempts. These approaches established the value of iterative refinement, but they do not by themselves ensure that a model can detect errors in its own outputs.

Empirical studies show that intrinsic self-correction is often unreliable and may replace correct answers with incorrect ones Huang et al. (2024); Jiang et al. (2025); Kamoi et al. (2024); more specifically, models are better at correcting an error once its location is given than at finding that error unaided Tyen et al. (2024). Confidence Matters therefore uses confidence-guided prompting to reduce unnecessary revision Li et al. (2024b), while CRITIC augments revision with external tools Gou et al. (2024). Among trained methods, SCoRe is the closest text-only precedent: it applies multi-turn online RL on self-generated trajectories and shapes rewards around improvement from the first answer to the second Kumar et al. (2025). Multimodal extensions instead use an external visual critic, region-level verification, or recovery from intermediate visual states, as in Critic-V, MIRROR, and VRRL Zhang et al. (2025a; 2026c); Tang et al. (2026).

Unlike confidence-guided correction, our method uses an explicit *True/False* judgment to control multi-turn reflection. A *True* judgment indicates verification-oriented behavior and prevents an additional reflective turn after the current response has been fully generated, whereas a *False* judgment indicates correction-oriented behavior and triggers another turn when the maximum horizon has

not been reached. The correctness label is available only to the privileged teacher during training, requiring no oracle or separate critic at inference time.

# B Model Details

This section provides the complete optimization and inference specification of our reflective multi-turn framework. We use $i$ to index examples, $m$ to index policy-optimization iterations, $\ell$ to index conversational turns, $k \in \{1, \ldots, K\}$ to index rollouts of the same example, $s \in \{1, 2\}$ to index the two intra-response stages of Turn 1, and $j$ to index completion tokens. In our implementation, optimization iteration $m$ trains conversational Turn $\ell = m$: the first iteration establishes the initial two-stage reasoning policy, whereas iterations $m \geq 2$ train verification and revision. All policy-gradient and distillation losses are computed only on valid assistant-response tokens; prompt, padding, and externally masked tokens are excluded. During the first GRPO iteration, two disjoint stage masks assign separate answer-conditioned advantages to the two Turn 1 response stages, preventing the reward of either boxed answer from updating tokens in the other stage.

## B.1 Training Schedule and Data Flow

Given an annotated example $x_i = (v_i, q_i, a_i^\star)$, where $v_i$ is the visual input, $q_i$ is the question, and $a_i^\star$ is the reference answer or grounding annotation, training proceeds in three steps. First, we construct a cleaned chain-of-thought (CoT) corpus and use 200K examples for supervised fine-tuning (SFT) under the two-stage Turn 1 output template defined in Equation (11). Second, the SFT checkpoint generates fresh rollout groups over an initial reinforcement-learning pool $\mathcal{U}^{(0)}$ of approximately 300K annotated examples. The first mixed-outcome subset is optimized with stage-masked group relative policy optimization (GRPO): the two intra-response stages receive independently normalized rewards and disjoint token-level policy-gradient masks. Third, the updated policy repeatedly generates fresh rollouts, applies the same mixed-outcome criterion, and learns explicit verification and revision with GRPO and Correctness-Conditioned On-Policy Distillation (CCOPD).

## B.2 CoT Construction and Supervised Initialization

For each raw example, six heterogeneous reasoning agents, as shown in Tab. 5, independently produce a candidate pair $(r_{i,s}, a_{i,s})$, where $r_{i,s}$ is a reasoning trace and $a_{i,s}$ is the final answer generated by agent $s$. The reference answer is withheld from all candidate-generation agents. A deterministic answer-extraction stage normalizes each candidate into the format required by the corresponding task, after which the task-specific evaluator $\mathcal{E}_i$ identifies the set of correct candidates:

$$\mathcal{C}_i = \{s \mid \mathcal{E}_i(a_{i,s}, a_i^\star) = 1\}. \tag{9}$$

Table 5: Six agents used for CoT construction.

| Model Version | Context | Max. output |
|---|---|---|
| Doubao Seed 2.1 | 256K | 256K |
| Kimi K3 | 1M | 1,048,576 |
| Qwen3.8-Max | 1M | 131,072 |
| Gemini 3.1 | 1,048,576 | 65,536 |
| GLM-5.3 | 1M | 128K |
| GPT-5.6 Luna | 1,050,000 | 128,000 |

If $|\mathcal{C}_i| = 0$, the example is rejected and no rationale is synthesized from the reference answer. If $|\mathcal{C}_i| = 1$, the unique valid trace is retained. If multiple candidates are correct, an aggregation agent selects the trace with the strongest visual faithfulness, logical completeness, and absence of direct answer leakage. The aggregation agent selects only among candidates in $\mathcal{C}_i$ and does not receive the reference answer as a rationale-generation prompt.

Samples for which at least four agents agree on an answer that conflicts with the original annotation are routed to human verification. Annotators inspect the original input, question, candidate traces, and source annotation, and then retain, relabel, or remove the sample. Approximately 4,000 conflicts are reviewed; 2,324 are corrected and retained. After automatic filtering and human review, the cleaned corpus contains 213,611 examples, approximately 203K of which have complete CoT traces. We use 200K of these examples for SFT.

Let $y_i^\star$ denote the cleaned assistant response and $M_i^{\mathrm{SFT}}$ the indices of its valid assistant tokens. The supervised objective is

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_i \left[ \frac{1}{|M_i^{\mathrm{SFT}}|} \sum_{j \in M_i^{\mathrm{SFT}}} \log \pi_\theta \left( y_{i,j}^\star \mid v_i, q_i, y_{i,<j}^\star \right) \right]. \tag{10}$$

The SFT target follows the Turn 1 response structure in Equation (11), containing two reasoning spans and two boxed task answers. The supervised loss is applied to all valid tokens in both stages. The resulting checkpoint initializes both the behavior policy for the first rollout pass and the frozen GRPO reference policy for the first optimization iteration. Stage masking is not applied to the SFT objective; it is introduced only for reward attribution during the first GRPO iteration.

### B.3 Task Evaluators and Output Parsing

All filtering and answer rewards use the same task-specific evaluator $\mathcal{E}_i$. For multiple-choice questions, the parser extracts the designated answer field and compares the normalized option label exactly. For open-ended question answering, both predictions and references are normalized by lowercasing, removing punctuation and articles, and collapsing whitespace before exact matching. For temporal grounding, the parser extracts a valid interval $[\hat{s}, \hat{e}]$ and returns one when its temporal intersection over union with the reference interval is at least $0.5$. For joint question answering and grounding, both the normalized answer criterion and the temporal-IoU criterion must be satisfied. An absent answer, an invalid option, a malformed interval, or a non-finite timestamp is scored as incorrect.

Turn 1 contains two consecutive reasoning-and-answer stages and follows

$$y_{i,k}^{(1)} = \langle \texttt{think} \rangle \, r_{i,k}^{(1,1)} \, \langle \texttt{/think} \rangle \boxed{\hat{a}_{i,k}^{(1,1)}} \langle \texttt{think} \rangle \, r_{i,k}^{(1,2)} \, \langle \texttt{/think} \rangle \boxed{\hat{a}_{i,k}^{(1,2)}}. \tag{11}$$

Here, $r_{i,k}^{(1,s)}$ and $\hat{a}_{i,k}^{(1,s)}$ denote the reasoning trace and boxed task answer of intra-response stage $s$. A deterministic parser identifies the two complete `think` spans and extracts the first and second boxed answers in order. The final Turn 1 prediction used for mixed-outcome filtering, downstream evaluation, and subsequent reflective turns is $\hat{a}_{i,k}^{(1,2)}$.

A reflective turn $\ell \geq 2$ follows

$$y_{i,k}^{(\ell)} = \langle \texttt{think} \rangle \, r_{i,k}^{\mathrm{jud},(\ell)} \, \langle \texttt{/think} \rangle \boxed{\hat{z}_{i,k}^{(\ell)}} \langle \texttt{think} \rangle \, r_{i,k}^{\mathrm{ans},(\ell)} \, \langle \texttt{/think} \rangle \boxed{\hat{a}_{i,k}^{(\ell)}}. \tag{12}$$

The first reasoning span evaluates whether the preceding answer is correct, and the first boxed field must contain $\hat{z}_{i,k}^{(\ell)} \in \{\mathrm{True}, \mathrm{False}\}$. The second reasoning span verifies, preserves, or revises the preceding answer, and the second boxed field contains the updated task answer $\hat{a}_{i,k}^{(\ell)}$. The latter is the only field used as the task prediction.

For Turn 1, $r_{\mathrm{fmt},i,k}^{(1,s)}$ equals one when stage $s$ contains a complete `think` span followed by exactly one parseable boxed answer, and zero otherwise. We define the full Turn 1 format score as

$$r_{\mathrm{fmt},i,k}^{(1)} = \frac{1}{2} \sum_{s=1}^{2} r_{\mathrm{fmt},i,k}^{(1,s)}. \tag{13}$$

For a reflective turn $\ell \geq 2$, $r_{\mathrm{fmt},i,k}^{(\ell)}$ equals one only when the two reasoning spans and two boxed fields occur in the prescribed order and the first boxed field is exactly `True` or `False`; it equals zero otherwise. Answer and format components are evaluated independently, so a format failure does not convert an unparseable field into a correct answer.

### B.4 Policy-Dependent Mixed-Outcome Filtering

Before optimization iteration $m$, the behavior policy $\pi_{\theta_{\mathrm{old}}}^{(m)}$ generates $K = 8$ independent completions for every example in original dataset $\mathcal{U}^{(0)}$. Let $\widetilde{a}_{i,k}^{(m)}$ denote the answer extracted from the

second boxed field of rollout $k$:

$$\widetilde{a}_{i,k}^{(m)} = \begin{cases} \hat{a}_{i,k}^{(1,2)}, & m = 1, \\ \hat{a}_{i,k}^{(m)}, & m \geq 2. \end{cases} \tag{14}$$

For $m = 1$, this is the answer after the second Turn 1 reasoning stage. For $m \geq 2$, it is the final answer after the correctness judgment and the second reflective reasoning span. We define

$$c_{i,k}^{(m)} = \mathbb{I}\left[\mathcal{E}_i\left(\hat{a}_{i,k}^{(m)}, a_i^\star\right) = 1\right] \tag{15}$$

and retain

$$\mathcal{U}^{(m)} = \left\{ x_i \in \mathcal{U}^{(0)} \;\middle|\; 0 < \sum_{k=1}^{K} c_{i,k}^{(m)} < K \right\}. \tag{16}$$

The lower bound removes rollout groups with no successful final trajectory, and the upper bound removes groups with no final-answer reward variation. The resulting set contains at least one correct and one incorrect second boxed answer under the current policy. Fresh rollouts are generated after every policy-optimization iteration; trajectories collected by an earlier policy are not reused as if they were on-policy samples for a later iteration.

### B.5 Turn-Specific Rewards and Stage Masks

For Turn 1, each intra-response stage has its own answer reward:

$$r_{\mathrm{ans},i,k}^{(1,s)} = \mathcal{E}_i\left(\hat{a}_{i,k}^{(1,s)}, a_i^\star\right), \qquad s \in \{1, 2\}. \tag{17}$$

To preserve the original total format-reward weight of $1.0$, half of the local format reward is assigned to each stage. The stage-specific rewards are

$$R_{i,k}^{(1,s)} = r_{\mathrm{ans},i,k}^{(1,s)} + \frac{1}{2} r_{\mathrm{fmt},i,k}^{(1,s)}, \qquad s \in \{1, 2\}. \tag{18}$$

Their sum recovers the aggregate Turn 1 reward used for logging:

$$\begin{aligned} R_{i,k}^{(1)} &= \sum_{s=1}^{2} R_{i,k}^{(1,s)} \\ &= r_{\mathrm{ans},i,k}^{(1,1)} + r_{\mathrm{fmt},i,k}^{(1)} + r_{\mathrm{ans},i,k}^{(1,2)}. \end{aligned} \tag{19}$$

Thus, the aggregate component weights remain $(1.0, 1.0, 1.0)$ for the first answer, complete-format score, and second answer, respectively.

The aggregate reward in Equation (19) is used only for logging and aggregate reward statistics; it is not broadcast to every Turn 1 completion token. Instead, $R_{i,k}^{(1,1)}$ and $R_{i,k}^{(1,2)}$ are normalized separately within their rollout groups and are applied only to their corresponding token masks, as defined in Sec. B.6. Consequently, the correctness reward of the first boxed answer cannot update tokens in the second stage, and the correctness reward of the second boxed answer cannot update tokens in the first stage.

For a reflective turn $\ell \geq 2$, let

$$z_{i,k}^{\star,(\ell)} = \begin{cases} \mathrm{True}, & \mathcal{E}_i\left(\hat{a}_{i,k}^{(\ell-1)}, a_i^\star\right) = 1, \\ \mathrm{False}, & \text{otherwise}. \end{cases} \tag{20}$$

The correctness-judgment component, extracted from the first boxed field, is

$$r_{z,i,k}^{(\ell)} = \begin{cases} 0.5, & \hat{z}_{i,k}^{(\ell)} = z_{i,k}^{\star,(\ell)} = \mathrm{True}, \\ 1.0, & \hat{z}_{i,k}^{(\ell)} = z_{i,k}^{\star,(\ell)} = \mathrm{False}, \\ 0, & \hat{z}_{i,k}^{(\ell)} \neq z_{i,k}^{\star,(\ell)}, \\ -1.0, & \hat{z}_{i,k}^{(\ell)} \notin \{\mathrm{True}, \mathrm{False}\}. \end{cases} \tag{21}$$

The larger reward for a correct `False` judgment reflects the greater difficulty and importance of detecting an answer that should be revised. The final-answer component is computed from the second boxed field:

$$r^{(\ell)}_{\mathrm{ans},i,k} = \mathcal{E}_i\left(\hat{a}^{(\ell)}_{i,k}, a^\star_i\right). \tag{22}$$

The complete reflective-turn reward is

$$R^{(\ell)}_{i,k} = 1.1\, r^{(\ell)}_{z,i,k} + 0.9\, r^{(\ell)}_{\mathrm{fmt},i,k} + 1.0\, r^{(\ell)}_{\mathrm{ans},i,k}. \tag{23}$$

All components are computed from the parsed response and the task evaluator; no reward model receives the reference answer as free-form generation context.

### B.6 Exact GRPO Objective with Turn-1 Stage Masking

For each retained example, $K$ completions are sampled from the frozen behavior policy $\pi^{(m)}_{\theta_{\mathrm{old}}}$. During Turn 1, the two stage rewards in Equation (18) are normalized independently. For each $s \in \{1, 2\}$, we compute

$$\mu^{(1,s)}_i = \frac{1}{K}\sum_{k=1}^{K} R^{(1,s)}_{i,k}, \tag{24}$$

and

$$\sigma^{(1,s)}_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(R^{(1,s)}_{i,k} - \mu^{(1,s)}_i\right)^2}. \tag{25}$$

The corresponding stage-specific group-relative advantage is

$$\widehat{A}^{(1,s)}_{i,k} = \begin{cases} \dfrac{R^{(1,s)}_{i,k} - \mu^{(1,s)}_i}{\sigma^{(1,s)}_i}, & \sigma^{(1,s)}_i > 0, \\ 0, & \sigma^{(1,s)}_i = 0. \end{cases} \tag{26}$$

For an optimization iteration $m \geq 2$, the reflective sequence reward in Equation (23) is normalized in the standard way. We compute

$$\mu^{(m)}_i = \frac{1}{K}\sum_{k=1}^{K} R^{(m)}_{i,k}, \qquad \sigma^{(m)}_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(R^{(m)}_{i,k} - \mu^{(m)}_i\right)^2}, \tag{27}$$

with

$$\widehat{A}^{(m)}_{i,k} = \begin{cases} \dfrac{R^{(m)}_{i,k} - \mu^{(m)}_i}{\sigma^{(m)}_i}, & \sigma^{(m)}_i > 0, \\ 0, & \sigma^{(m)}_i = 0. \end{cases} \tag{28}$$

For Turn 1, let $M^{(1,1)}_{i,k}$ contain tokens from the first opening `think` tag through the end of the first boxed answer. Similarly, let $M^{(1,2)}_{i,k}$ contain tokens from the second opening `think` tag through the end of the second boxed answer. The two stage masks are disjoint and cover all valid Turn 1 completion tokens:

$$M^{(1,1)}_{i,k} \cap M^{(1,2)}_{i,k} = \varnothing, \qquad M^{(1,1)}_{i,k} \cup M^{(1,2)}_{i,k} = M^{(1)}_{i,k}. \tag{29}$$

We define the binary stage indicator

$$\gamma^{(1,s)}_{i,k,j} = \mathbb{I}\left[j \in M^{(1,s)}_{i,k}\right]. \tag{30}$$

The token-level advantage used by GRPO is

$$\widetilde{A}^{(m)}_{i,k,j} = \begin{cases} \displaystyle\sum_{s=1}^{2} \gamma^{(1,s)}_{i,k,j}\, \widehat{A}^{(1,s)}_{i,k}, & m = 1, \\ \widehat{A}^{(m)}_{i,k}, & m \geq 2. \end{cases} \tag{31}$$

Because the stage masks are disjoint, every valid Turn 1 token receives exactly one of the two stage-specific advantages. Tokens in the first reasoning-and-answer stage receive only $\widehat{A}_{i,k}^{(1,1)}$, whereas tokens in the second reasoning-and-answer stage receive only $\widehat{A}_{i,k}^{(1,2)}$.

This stage mask performs policy-gradient credit assignment rather than differentiating through the task evaluator. The evaluator and the resulting rewards are non-differentiable rollout metadata. Gradient isolation is achieved by multiplying each stage-specific advantage by its corresponding token mask before evaluating the policy-gradient objective.

For token $j$ in the valid-token set $M_{i,k}^{(m)}$, the importance ratio is

$$\rho_{i,k,j}^{(m)}(\theta) = \frac{\pi_\theta\left(y_{i,k,j}^{(m)} \mid h_{i,k}^{(m)}, y_{i,k,<j}^{(m)}\right)}{\pi_{\theta_{\text{old}}}^{(m)}\left(y_{i,k,j}^{(m)} \mid h_{i,k}^{(m)}, y_{i,k,<j}^{(m)}\right)}. \tag{32}$$

We use asymmetric probability-ratio clipping with $\epsilon_{\text{low}} = 0.20$ and $\epsilon_{\text{high}} = 0.28$:

$$\bar{\rho}_{i,k,j}^{(m)} = \text{clip}\left(\rho_{i,k,j}^{(m)}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right). \tag{33}$$

Let $\pi_{\text{ref}}^{(m)}$ be a frozen copy of the checkpoint that initializes optimization iteration $m$. For each sampled token, we use the standard non-negative sampled GRPO estimator

$$d_{i,k,j}^{\text{ref},(m)} = \frac{\pi_{\text{ref}}^{(m)}\left(y_{i,k,j}^{(m)} \mid h_{i,k}^{(m)}, y_{i,k,<j}^{(m)}\right)}{\pi_\theta\left(y_{i,k,j}^{(m)} \mid h_{i,k}^{(m)}, y_{i,k,<j}^{(m)}\right)} - \log \frac{\pi_{\text{ref}}^{(m)}\left(y_{i,k,j}^{(m)} \mid h_{i,k}^{(m)}, y_{i,k,<j}^{(m)}\right)}{\pi_\theta\left(y_{i,k,j}^{(m)} \mid h_{i,k}^{(m)}, y_{i,k,<j}^{(m)}\right)} - 1. \tag{34}$$

The exact GRPO objective is

$$\mathcal{L}_{\text{GRPO}}^{(m)} = -\mathbb{E}_{i,k}\left[\frac{1}{|M_{i,k}^{(m)}|}\sum_{j\in M_{i,k}^{(m)}}\left(\min\left(\rho_{i,k,j}^{(m)}\widetilde{A}_{i,k,j}^{(m)}, \bar{\rho}_{i,k,j}^{(m)}\widetilde{A}_{i,k,j}^{(m)}\right) - \beta d_{i,k,j}^{\text{ref},(m)}\right)\right], \beta = 0.04. \tag{35}$$

We first average over valid tokens within each completion, and then over completions and examples. The stage masks affect only the policy-gradient advantage term during Turn 1. Reference-policy regularization is computed on all valid completion tokens. The behavior-policy log probabilities, reference-policy log probabilities, rewards, masks, and advantages are treated as fixed rollout data during each policy update.

### B.7 Correctness-Conditioned On-Policy Distillation

For a reflective Turn $\ell \geq 2$, the student receives

$$h_{i,k}^{S,(\ell)} = \left[v_i; q_i; \mathcal{H}_{i,k}^{(\ell)}; p^{(\ell)}\right], \tag{36}$$

where $\mathcal{H}_{i,k}^{(\ell)}$ contains the preceding interaction history and $p^{(\ell)}$ is the reflection prompt. The prompt instructs the model to follow the output format in Equation (12): a reasoning span, a boxed `True`/`False` judgment, a second reasoning span, and a boxed final answer.

The shared-parameter teacher receives the same information together with the binary correctness state of the immediately preceding answer:

$$h_{i,k}^{T,(\ell)} = \left[z_{i,k}^{\star,(\ell)}; v_i; q_i; \mathcal{H}_{i,k}^{(\ell)}; p^{(\ell)}\right]. \tag{37}$$

The teacher does not receive $a_i^\star$, the reference grounding interval, the raw IoU value, or the task-evaluator score. It receives only the binary state defined in Equation (20).

The reflective completion is sampled from the student-side behavior policy:

$$y_{i,k}^{(\ell)} \sim \pi_{\theta_{\text{old}}}^{(m)}\left(\cdot \mid h_{i,k}^{S,(\ell)}\right), \qquad \ell = m \geq 2. \tag{38}$$

The sampled sequence is then teacher-forced under both the student and teacher contexts. Let $g^T_{i,k,j}$ and $g^S_{i,k,j}$ denote the teacher and student logits, respectively, at valid completion token $j$. With temperature $T = 1.0$, the corresponding next-token distributions are

$$p^T_{i,k,j} = \mathrm{softmax}\left(g^T_{i,k,j}/T\right), \qquad p^S_{i,k,j} = \mathrm{softmax}\left(g^S_{i,k,j}/T\right). \tag{39}$$

The teacher forward pass is executed without gradient tracking, and its distribution is treated as a fixed reference for the current update:

$$\nabla_\theta \,\mathrm{sg}\left[p^T_{i,k,j}\right] = 0. \tag{40}$$

We compute the student-to-teacher reverse KL over the full next-token vocabulary and clip each token-level divergence before averaging over the completion length:

$$\mathcal{L}^{(m)}_{\mathrm{CCOPD}} = \mathbb{E}_{i,k}\left[\frac{1}{|M^{(m)}_{i,k}|}\sum_{j\in M^{(m)}_{i,k}} \min\left\{D_{\mathrm{KL}}\left(p^S_{i,k,j}\,\middle\|\,\mathrm{sg}\left[p^T_{i,k,j}\right]\right), c_{\mathrm{KL}}\right\}\right], \qquad c_{\mathrm{KL}} = 10.0. \tag{41}$$

Accordingly, the divergence in Equation (41) is the reverse KL

$$D_{\mathrm{KL}}\left(p^S \,\middle\|\, \mathrm{sg}[p^T]\right) = \sum_{w\in\mathcal{V}} p^S(w)\left(\log p^S(w) - \log \mathrm{sg}[p^T(w)]\right). \tag{42}$$

Although the two branches share parameters, gradients from CCOPD flow only through the student distribution. Teacher dropout and other stochastic training-time layers are disabled during target computation, making the teacher distribution deterministic for a fixed context and checkpoint.

CCOPD is computed over all valid tokens in the reflective completion, including both reasoning spans, the boxed correctness judgment, and the boxed final answer. When conditioned on $z^\star = \mathrm{True}$, the teacher defines a reference distribution that supports justifying a `True` judgment and subsequently verifying and preserving the preceding answer. Minimizing the reverse KL encourages the student to concentrate its probability mass on this teacher-supported verification behavior. When conditioned on $z^\star = \mathrm{False}$, the teacher instead defines a reference distribution that supports identifying the error in the preceding response and deriving a revised final answer. The reverse-KL objective correspondingly suppresses student predictions that are inconsistent with this correction behavior.

CCOPD is on-policy with respect to the student-side rollout distribution: the divergence is evaluated along prefixes induced by sequences sampled from the current student behavior policy rather than along teacher-generated trajectories. Here, *on-policy* characterizes the trajectory distribution on which distillation is performed, whereas *reverse KL* specifies the direction of the token-level distribution matching objective. Because the binary correctness state is derived from the reference annotation, it remains privileged training information. This restriction prevents direct exposure of the reference answer but does not eliminate the context mismatch between the teacher and student branches; the effect of this mismatch is evaluated separately through matched ablations.

### B.8 Joint Objective and Recursive Extension

The optimization objective for iteration $m$ is

$$\mathcal{L}^{(m)}_{\mathrm{train}} = \begin{cases} \mathcal{L}^{(1)}_{\mathrm{GRPO}}, & m = 1, \\ \mathcal{L}^{(m)}_{\mathrm{GRPO}} + \lambda\mathcal{L}^{(m)}_{\mathrm{CCOPD}}, & m \geq 2, \end{cases} \qquad \lambda = 0.05. \tag{43}$$

No CCOPD term is applied in the first optimization iteration because Turn 1 has no preceding answer to verify. The first-iteration GRPO term is the stage-masked objective in Equation (35): its two stage rewards are normalized separately and assigned through the disjoint masks in Equation (31). For $m \geq 2$, GRPO and CCOPD are computed on the same student-generated reflective completion and the same valid-token set. GRPO supplies sequence-level correctness-judgment, format, and final-answer feedback, whereas CCOPD supplies token-level distributional guidance under the binary correctness condition.

To extend the framework to Turn $\ell 2$, the student context includes the complete interaction history up to Turn $\ell - 1$, and $z^{\star,(\ell)}$ is recomputed from the answer in the second boxed field of Turn $\ell - 1$. The teacher receives only this new binary state in addition to the student-visible history. A fresh rollout-and-filtering pass precedes every new optimization iteration; thus, the subset used for Turn $\ell$ is not assumed to be distributionally equivalent to the subset used for Turn $\ell - 1$.

### B.9 Correctness-Guided Multi-Turn Inference

At inference time, only the student branch is used, and no reference answer, task evaluator, or privileged correctness label is available. The model performs reflective inference with a maximum horizon of $T_{\max}$ turns. A generated correctness judgment does not interrupt the response being generated in the current turn; instead, it determines whether another reflective turn is invoked after the current response has been completed. This distinction separates *within-turn generation* from *between-turn continuation*.

Turn 1 generates two consecutive reasoning-and-answer stages according to Equation (11). The answer extracted from the second boxed field initializes the candidate prediction:

$$\tilde{a}^{(1)} = \hat{a}^{(1,2)}. \tag{44}$$

Because Turn 1 does not contain an explicit correctness judgment, every example proceeds to Turn 2, where the model verifies and, if necessary, revises this initial candidate.

For each reflective turn $\ell \geq 2$, the model follows the structured output format defined in Equation (12):

$$y^{(\ell)} = \langle \text{think} \rangle r_{\text{jud}}^{(\ell)} \langle /\text{think} \rangle \hat{z}^{(\ell)} \langle \text{think} \rangle r_{\text{ans}}^{(\ell)} \langle /\text{think} \rangle \hat{a}^{(\ell)}. \tag{45}$$

Here, $r_{\text{jud}}^{(\ell)}$ denotes the reasoning used to assess the candidate from the preceding turn, and the first boxed field contains the model-predicted correctness judgment $\hat{z}^{(\ell)} \in \{\text{True}, \text{False}\}$. The second reasoning span $r_{\text{ans}}^{(\ell)}$ then re-examines the video evidence and produces the current-turn task answer $\hat{a}^{(\ell)}$ in the second boxed field.

The correctness judgment controls the semantic role of the second reasoning stage. When $\hat{z}^{(\ell)} = \text{True}$, the model is expected to verify the preceding solution and generate an answer consistent with the available evidence. When $\hat{z}^{(\ell)} = \text{False}$, it is expected to identify the potential error, reconsider the relevant evidence, and generate a corrected answer. Importantly, the judgment does not act as an intra-turn generation boundary. Even when the model generates True, it must still complete the second reasoning span and produce the current-turn final answer. Thus, both judgment outcomes result in a complete reflective response with the same output structure.

After the reflective response has been fully generated, the candidate is unconditionally updated using the answer in its second boxed field:

$$\tilde{a}^{(\ell)} = \hat{a}^{(\ell)}, \qquad \ell \geq 2. \tag{46}$$

Accordingly, the current-turn answer replaces the preceding candidate regardless of whether the preceding answer was judged True or False. A True judgment therefore does not literally return or copy the preceding answer; rather, the model is trained to verify it and generate an evidence-consistent answer. A False judgment indicates that the preceding candidate should be reconsidered, after which the newly generated answer becomes the updated candidate.

Once the current turn is complete, the generated judgment determines whether another reflective turn is required. We define the continuation indicator as

$$\text{continue}^{(\ell)} = \mathbb{I}\left[\hat{z}^{(\ell)} \neq \text{True} \ \wedge \ \ell < T_{\max}\right], \qquad \ell \geq 2. \tag{47}$$

An explicitly parsed True judgment prevents any additional reflective turn from being invoked, but only after the current response, including $\hat{a}^{(\ell)}$, has been fully generated. A False or unparseable judgment invokes the next reflective turn whenever $\ell < T_{\max}$. At $\ell = T_{\max}$, inference terminates regardless of the predicted judgment.

If another turn is required, the complete response from the current turn is appended to the interaction history:

$$H^{(\ell+1)} = H^{(\ell)} \oplus y^{(\ell)}, \qquad \text{continue}^{(\ell)} = 1, \tag{48}$$

where $\oplus$ denotes concatenation using the model's native conversation template. The next turn therefore observes the video, question, complete preceding interaction history, and the most recently generated candidate answer.

Equivalently, the stopping condition after reflective turn $\ell$ is

$$\mathrm{stop}^{(\ell)} = \mathbb{I}\left[\hat{z}^{(\ell)} = \mathrm{True} \ \vee \ \ell = T_{\max}\right], \qquad \ell \geq 2. \tag{49}$$

This stopping condition is evaluated only after the complete current-turn response has been generated. It controls whether the model proceeds to turn $\ell + 1$, rather than truncating generation within turn $\ell$.

Let $\ell_{\mathrm{stop}}$ denote the first reflective turn after which no additional turn is invoked:

$$\ell_{\mathrm{stop}} = \min\left\{\ell \in \{2, \ldots, T_{\max}\} \,\middle|\, \hat{z}^{(\ell)} = \mathrm{True} \ \vee \ \ell = T_{\max}\right\}. \tag{50}$$

The final task prediction is the answer generated in the second boxed field of the stopping turn:

$$a_{\mathrm{out}} = \hat{a}^{(\ell_{\mathrm{stop}})} = \tilde{a}^{(\ell_{\mathrm{stop}})}. \tag{51}$$

Therefore, if the model predicts $\mathrm{True}$ at Turn 2, it first completes the entire Turn 2 response and returns $\hat{a}^{(2)}$. If it predicts $\mathrm{False}$ or produces an unparseable judgment, it proceeds to Turn 3 when permitted, and the answer generated in the eventual stopping turn is returned.

This inference procedure is aligned with the reflective-turn training objective. During training, the task-answer reward is computed from the second boxed answer $\hat{a}^{(\ell)}$, and the same answer is used to update the candidate and determine the final task prediction during inference. Consequently, the model is not optimized for an answer that would later be discarded by the inference procedure. The correctness judgment has two complementary roles: within the current turn, it distinguishes verification-oriented behavior from correction-oriented behavior; after the current turn is complete, it determines whether additional reflection is required. We use $T_{\max} = 3$ in all experiments.

# C Experiment Details

## C.1 Implementation Details

The stage-wise training configuration is summarized in Tab. 7. We fine-tune models initialized from Qwen2.5-VL-7B-Instruct Bai et al. (2025b) and Qwen3-VL-8B-Instruct Bai et al. (2025a). In addition to the hyperparameters listed in the table, we apply a weight decay of $0.01$ and clip the gradient norm to $1.0$. During rollout generation, we use a sampling temperature of $1.0$ to encourage exploration. All training experiments are conducted on 32 NVIDIA H20 GPUs.

For evaluation, we use `lmms-eval` Zhang et al. (2025b) with greedy decoding (*i.e.*, a temperature of $0$). The maximum response length is set to 4,096 tokens to minimize output truncation.

For test-time visual input scaling, we follow the evaluation protocol associated with each backbone model. For Qwen2.5-VL, we allow up to 16K video tokens and evaluate the model with three maximum frame counts: $\{16, 128, 768\}$. For Qwen3-VL, we allow up to 128K video tokens and use a fixed maximum frame count of $128$. Accordingly, we report the performance of Qwen2.5-VL under all three frame-count settings, while reporting the performance of Qwen3-VL using the 128-frame setting.

## C.2 Training Dataset

The final training corpus contains 328,309 samples in total, drawn from both video-centric and vision-language datasets. By modality, it includes 160,569 video samples, 161,327 image samples, and 6,413 text-only mathematical reasoning samples. The video data cover general video understanding, temporal grounding, and video question answering, while the image data span knowledge-intensive VQA, mathematical reasoning, chart understanding, spatial reasoning, OCR, and general visual instruction following. The detailed composition of all datasets, including their names, sample counts, and task types, is summarized in Tab. 8. The data cleaning procedure used to construct the final training set has been described in Sec. 2.2. After cleaning, the source distribution and task distribution of the retained data are further reported in Tab. 1 and Fig. 4, respectively.

### C.3 Evaluation Benchmarks

To comprehensively evaluate our method, we conduct experiments on 13 video benchmarks covering general, long-form, knowledge-intensive, and temporal video understanding. Specifically, MVBench Li et al. (2024a) contains 4,000 multiple-choice questions across 20 video tasks, while Video-MME Fu et al. (2025a) includes 2,700 questions from 900 videos of varying durations. MVP-mini Krojer et al. (2025) contains approximately 9,000 paired examples for evaluating fine-grained physical and temporal understanding, and VSI-Bench Yang et al. (2025) provides over 5,000 questions on visual-spatial reasoning. For long-video understanding, LongVideoBench Wu et al. (2024) contains 1,344 multiple-choice questions from 3,763 videos, MLVU Zhou et al. (2025) covers nine multiple-choice and generative tasks on videos ranging from 3 minutes to 2 hours, and LVBench Wang et al. (2025b) comprises 1,549 questions from 103 long videos. For knowledge-intensive reasoning, MMVU Zhao et al. (2025) provides 1,000 expert-annotated questions from 1,529 specialized videos, whereas Video-MMMU Hu et al. (2025) contains 900 questions based on 300 educational videos. We further evaluate temporal understanding and grounding on Charades-STA Gao et al. (2017) and ActivityNet Caba Heilbron et al. (2015), which require localizing natural-language queries in untrimmed videos; NExT-GQA Xiao et al. (2024), which provides approximately 10,500 temporal annotations for grounded video question answering; and TempCompass Liu et al. (2024), which contains 7,540 instructions spanning multiple-choice QA, yes/no QA, caption matching, and caption generation. For compact presentation in the comprehensive result tables, several benchmark names are abbreviated in the column headers; Table 6 provides the correspondence between these table abbreviations and the standard benchmark names used in the text.

Table 6: Abbreviations used for benchmark names in the result tables.

| **Abbrev.** | **Benchmark** | **Abbrev.** | **Benchmark** |
|---|---|---|---|
| MV. | MVBench | VideoMME | Video-MME |
| MVP | MVP-mini | VSI. | VSI-Bench |
| LongVideo. | LongVideoBench | MLVU | MLVU |
| LV. | LVBench | MMVU | MMVU |
| VideoMMMU | Video-MMMU | Charades-STA | Charades-STA |
| ActivityNet | ActivityNet | NExT-GQA | NExT-GQA |
| TempCompass | TempCompass | | |

## D Additional Results

### D.1 Results on 13 VQA Benchmarks

In addition to the five benchmarks reported in Tab. 2 and the three grounding benchmarks in Tab. 3, we evaluate our method on five additional test sets, resulting in a comprehensive evaluation over 13 benchmarks. As summarized in Tab. 19, these benchmarks cover four complementary capability dimensions: general video understanding, long-video understanding, video reasoning, and temporal reasoning. We conduct experiments with both Qwen2.5-VL-7B and Qwen3-VL-8B to assess the generalizability of our reflective multi-turn training framework across model backbones. Compared with their respective base models, our Qwen2.5-VL-7B and Qwen3-VL-8B variants improve the average score over the 13 benchmarks by 8.53 and 8.52 points, respectively. With Qwen2.5-VL-7B, our method achieves the best performance on six benchmarks, *i.e.*, MVBench, MVP, VSI-Bench, LongVideoBench, MMVU, and TempCompass, and ranks second on VideoMME, VideoMMMU, ActivityNet, and NExT-GQA. Notable improvements over the base model include 12.9 points on MVP, 13.8 points on MLVU, 5.4 points on MMVU, 22.2 points on ActivityNet, and 11.3 points on NExT-GQA. With the stronger Qwen3-VL-8B backbone, our method achieves the best result on nine benchmarks, including MVBench, MVP, VSI-Bench, MLVU, LVBench, VideoMMMU, ActivityNet, NExT-GQA, and TempCompass, and ranks second on VideoMME, LongVideoBench, and Charades-STA, trailing the corresponding best results by only 2.1, 0.7, and 0.2 points, respectively. It additionally ranks third on MMVU, with a score 1.3 points below the best-performing method. The largest individual gains over the Qwen3-VL-8B base model are observed on ActivityNet ($+23.8$), Charades-STA ($+19.0$), MLVU ($+14.4$), and MMVU ($+10.2$).

Overall, the results demonstrate that the proposed framework provides broad and backbone-agnostic improvements rather than benefiting only a particular model or task category. When grouped by capability, temporal reasoning exhibits the largest average gains, reaching 12.23 points with Qwen2.5-

Table 7: Implementation details of our three-stage multi-turn training. Hyperparameters highlighted in **bold** are varied in the ablation studies.

| Category | Hyperparameter | Turn 1 | Turn 2 | Turn 3 |
|---|---|---|---|---|
| **Data configuration** | Video input token budget | min $= 16$, max $= 768$, total $= 4096$ | | |
| | Image input token budget | min $= 4$, max $= 768$ | | |
| | Max video frames | 64 | | |
| **Data augmentation** | Visual preprocessing | Qwen2.5-VL smart resize and uniform frame sampling | | |
| | Frame sampling | Deterministic, no random frame jitter | | |
| | Image/video enforcement | enforce_image_video | | |
| | Random augmentation | Not used | | |
| **Training schedule** | Number of GPUs | 32 | | |
| | Per-device train batch size | 2 | | |
| | Number of generations | 8 | | |
| | Steps per generation | 8 | | |
| | Epochs | 1 | | |
| | Optimizer | AdamW | | |
| | LR scheduler | constant with warmup | | |
| | Optimization target | Answer& Format& Answer | Correctness& Format& Answer | |
| | Learning rate | $1 \times 10^{-6}$ | $5 \times 10^{-7}$ | $2 \times 10^{-7}$ |
| **Model configuration** | Base model | Qwen2.5-VL | | |
| | Precision | bfloat16 | | |
| | DeepSpeed | ZeRO-3 | | |
| | Gradient checkpointing | Enabled | | |
| | Tuned modules | LLM and multimodal MLP | | |
| | Frozen modules | Vision encoder | | |
| | Max prompt length | 8192 | | |
| | Max completion length | 2048 | | |
| | Turn-2 max completion length | 2048 | | |
| | vLLM generation | enabled, colocate mode, tensor parallel size $= 4$ | | |
| **Loss params** | **Reward weights** | 1.0, 1.0, 1.0 | 1.1, 0.9, 1.0 | |
| | RL algorithm | GRPO | GRPO & CCOPD | |
| | CCOPD enabled | False | True | |
| | **CCOPD weight $\lambda$** | - | 0.05 | |
| | CCOPD temperature | - | 1.0 | |
| | CCOPD KL type | - | reverse KL | |
| | CCOPD token KL clip | - | 10.0 | |
| | Turn-2 IoU threshold | - | 0.5 | |
| | GRPO KL coefficient $\beta$ | 0.04 | | |
| | **Judge rewards** | True-correct $= 0.5$, False-correct $= 1.0$, wrong $= 0.0$, missing $= -1.0$ | | |

VL-7B and 14.52 points with Qwen3-VL-8B. This finding is consistent with the design of our framework: iterative verification and revision encourage the model to identify relevant temporal evidence, verify event order and temporal boundaries, and correct initially plausible but insufficiently supported predictions. Meanwhile, the consistent gains in general video understanding, long-video understanding, and video reasoning indicate that the benefits extend beyond temporal localization. Taken together, the evaluation across all 13 benchmarks confirms that reflective multi-turn training substantially strengthens video understanding and reasoning, with particularly pronounced advantages on reasoning-intensive and temporally demanding tasks.

Table 8: Training Dataset Summary

| Name | Samples | Type | Name | Samples | Type |
|---|---|---|---|---|---|
| LLaVA-Video-178K | 82676 | video; multiple-choice | General | 14987 | image; multiple-choice |
| Knowledge | 37214 | image; multiple-choice | Charades-STA(Train Set) | 12398 | video; IoU |
| Math | 36924 | image; multiple-choice | STAR | 11455 | video; multiple-choice |
| VIRL | 23131 | image; math | ActivityNet(Train Set) | 10000 | video; IoU |
| Chart | 21528 | image; multiple-choice | CLEVRER | 8220 | video; multiple-choice |
| Spatial | 20284 | image; multiple-choice | NeXT-QA | 7549 | video; multiple-choice |
| OCR | 15886 | image; OCR | DAPO-Math | 6413 | text; math |
| PerceptionTest | 6348 | video; multiple-choice | ThinkLite-VL-Hard | 4336 | image; math |
| NeXT-GQA(Train Set) | 3356 | video; GQA | TimeR1 | 2497 | video; IoU |
| STI-Bench | 1524 | video; exact match | TVBench | 1373 | video; exact match |
| MMR-VBench | 210 | video; exact match | **Total** | **328309** | – |

Table 9: Comparison of multi-turn inference performance under matched settings. All methods use Qwen2.5-VL-7B as the base model, and take 128 video frames as input.

| Method | Turn # | Max. Output Tokens | Video Understanding | | | Video Reasoning | | Average |
|---|---|---|---|---|---|---|---|---|
| | | | VideoMME | MV. | LongVideo. | VideoMMMU | MMVU | |
| Qwen2.5-VL-7B† | 1 | 256 | 65.1 | 63.9 | 60.9 | 51.1 | 55.7 | 59.3 |
| | 2 | 256 | 64.7 (−0.4) | 64.1 (+0.2) | 60.9 (0.0) | 50.6 (−0.5) | 55.1 (−0.6) | 59.1 (−0.3) |
| VideoAuto-R1† | 1 | 256 | 66.8 | 70.2 | 59.7 | 52.1 | 55.7 | 60.9 |
| | 2 | 256 | 66.5 (−0.3) | 69.1 (−1.1) | 59.0 (−0.7) | 51.7 (−0.4) | 54.2 (−1.5) | 60.1 (−0.8) |
| VideoKR† | 1 | 256 | 66.4 | 68.9 | 61.3 | 52.2 | 60.5 | 61.9 |
| | 2 | 256 | 66.6 (+0.2) | 67.3 (−1.6) | 60.1 (−1.2) | 52.4 (+0.2) | 59.1 (−1.4) | 61.1 (−0.8) |
| Video-MTR⋆ | 1 | 256 | – | – | – | – | – | – |
| | 2 | 256 | 61.9 | 61.7 | 57.7 | 50.4 | 61.9 | 58.7 |
| | 3 | 256 | 61.9 (0.0) | 61.7 (0.0) | 57.7 (0.0) | 50.4 (0.0) | 61.9 (0.0) | 58.7 (0.0) |
| **Ours** | 1 | 256 | 63.0 | 64.8 | 59.4 | 53.4 | 60.2 | 60.2 |
| | 2 | 256 | 65.0 (+2.0) | 69.5 (+4.7) | 61.8 (+2.4) | 56.3 (+2.9) | 68.5 (+8.3) | 64.2 (+4.1) |
| | 3 | 256 | 66.2 (+1.2) | 71.3 (+1.8) | 61.4 (−0.4) | 55.8 (−0.5) | 68.0 (−0.5) | 64.5 (+0.3) |

*Notes:*
1. † For Qwen2.5-VL-7B, VideoAuto-R1, and VideoKR, multi-turn inference is implemented by appending the answer generated in the previous turn to the input of the next turn and running the model again.
2. ⋆ Video-MTR does not report Turn-1 results because its inference procedure does not allow the model to directly produce a final answer in the first turn. Its Turn-2 and Turn-3 results are identical because only a small number of samples proceed from the second turn to the third turn.
3. The Max. Output Tokens denotes the number of newly generated tokens per turn.

### D.2 Controlled Multi-Turn Comparison

To examine how performance changes with additional inference turns, we compare our method with representative baselines under matched backbone, frame-count, and per-turn generation-limit settings in Tab. 9. All methods use Qwen2.5-VL-7B, receive 128 uniformly sampled video frames, and generate at most 256 new tokens per turn. For Qwen2.5-VL-7B, VideoAuto-R1, and VideoKR, we implement additional inference turns by appending the preceding answer to the conversation history and providing a correctness-neutral instruction to verify and, if necessary, revise it. Video-

Table 10: Inference-time intervention on correctness conditioning with Qwen2.5-VL-7B using 128 input frames. The two Turn-2 experiments use the identical checkpoint and differ only in whether the ground-truth correctness label $z^\star$ of the preceding answer is included in the system prompt.

| Checkpoint | Inference Condition | Video Understanding | | | Video Reasoning | | Average |
|---|---|---|---|---|---|---|---|
| | | VideoMME | MV. | LongVideo. | VideoMMMU | MMVU | |
| Turn 1 | Standard prompt | 63.4 | 65.3 | 60.2 | 53.2 | 66.1 | 61.6 |
| Turn 2 | Standard prompt (without $z^\star$) | 65.6 | 69.1 | 61.3 | 56.9 | 68.8 | 64.3 |
| | Oracle correctness (with gold $z^\star$) | **81.8** | **80.1** | **73.9** | **68.3** | **80.8** | **77.0** |
| | *Gain from $z^\star$* | *+16.2* | *+11.0* | *+12.6* | *+11.4* | *+12.0* | *+12.7* |

MTR follows its own multi-turn inference procedure and does not produce a final task answer in Turn 1. No reference answer, oracle correctness label, or external correctness feedback is provided during inference. These settings align the backbone and per-turn input and output limits, although they do not guarantee identical cumulative token usage or inference latency across methods.

Additional inference turns do not consistently improve the baseline results. From Turn 1 to Turn 2, the average scores of Qwen2.5-VL-7B, VideoAuto-R1, and VideoKR decrease from 59.3, 60.9, and 61.9 to 59.1, 60.1, and 61.1, respectively, corresponding to decreases of 0.3, 0.8, and 0.8 points when calculated before rounding. Video-MTR remains at 58.7 across its reported second- and third-turn settings. In contrast, our method improves on all five benchmarks from Turn 1 to Turn 2, increasing its average score from 60.2 to 64.2, a gain of 4.1 points before rounding. The individual gains are 2.0 points on VideoMME, 4.7 on MVBench, 2.4 on LongVideoBench, 2.9 on VideoMMMU, and 8.3 on MMVU. At Turn 2, our method exceeds the strongest competing average, achieved by VideoKR, by 3.1 points. Extending our method to Turn 3 increases the average to 64.5, an additional gain of 0.3 points: VideoMME and MVBench improve by 1.2 and 1.8 points, whereas LongVideoBench, VideoMMMU, and MMVU decrease by 0.4, 0.5, and 0.5 points, respectively. Thus, the largest improvement occurs between the first and second turns, while the third turn provides a smaller, task-dependent benefit. These results are consistent with the utility of the proposed reflective training framework and show that simply repeating inference with the evaluated baselines does not yield comparable gains; they also motivate more selective stopping criteria for tasks on which further reflection reduces accuracy.

### D.3 Does Correctness Conditioning Create a Meaningful Information Gap?

To determine whether the correctness condition provides the teacher with a meaningful information advantage over the student, we conduct a controlled inference-time intervention using the same checkpoint obtained after Turn-2 training. Each evaluation example is tested under two conditions. In the standard condition, the model receives the original system prompt without the correctness label, matching the information available to the student. In the oracle-conditioned condition, we append the ground-truth correctness label $z^\star$ of the preceding answer to the system prompt, matching the additional information available to the teacher. The model parameters, conversation histories, visual inputs, decoding configuration, and evaluation protocol are kept identical across the two conditions. Therefore, the resulting performance difference isolates the effect of exposing the model to $z^\star$, rather than differences in model capacity or training stage.

As shown in Tab. 10, the Turn-2 checkpoint achieves an average score of 64.3 without the correctness condition, whereas providing the ground-truth correctness label increases the average score to 77.0, corresponding to a substantial gain of 12.7 percentage points. The improvement is consistent across all five benchmarks, ranging from 11.0 to 16.2 points. In comparison, Turn-2 training alone improves the average score from 61.6 to 64.3 when $z^\star$ is unavailable at inference time. These results demonstrate that the model has learned to make effective use of the correctness condition and that the single-bit correctness signal creates a substantial information gap between the teacher and student distributions. Thus, the benefit of the correctness-conditioned teacher cannot be explained solely by the shared model parameters: its access to $z^\star$ materially changes the model's predictive behavior.

Table 11: Inference-time efficiency on the benchmarks reported in Tabs. 2 and 3. All methods are evaluated using the same hardware, batch size, numerical precision, inference engine, frame-processing pipeline, and decoding configuration.

| Dataset | Model | Avg. Turns | Input Tokens | Generated Tokens | E2E Latency (s) | | Peak Mem. (GB/GPU) | Throughput (samples/s) |
|---|---|---|---|---|---|---|---|---|
| | | | | | Mean | P95 | | |
| **Part I: Video Understanding and Video Reasoning** (Table 2) | | | | | | | | |
| **VideoMME** | Qwen2.5-VL-7B | 1.00 | 16,110 | 1 | 10.044 | 14.000 | ∼20.0 | 0.760 |
| | Video-MTR | 2.00 | 17,017 | 54 | 27.980 | 32.000 | ∼81.6 | 1.135 |
| | Ours | 2.19 | 18,547 | 62 | 20.501 | 28.000 | ∼21.0 | 0.390 |
| **MVBench** | Qwen2.5-VL-7B | 1.00 | 6,140 | 6 | 2.636 | 9.000 | ∼20.0 | 2.898 |
| | Video-MTR | 2.00 | 16,124 | 48 | 48.912 | 71.000 | ∼81.6 | 0.649 |
| | Ours | 2.18 | 12,579 | 63 | 5.852 | 15.000 | ∼21.0 | 1.316 |
| **LongVideoBench** | Qwen2.5-VL-7B | 1.00 | 14,935 | 1 | 9.446 | 13.000 | ∼20.0 | 0.823 |
| | Video-MTR | 2.00 | 16,642 | 59 | 58.880 | 69.000 | ∼81.6 | 0.543 |
| | Ours | 2.23 | 21,305 | 59 | 20.104 | 28.000 | ∼21.0 | 0.195 |
| **Video-MMMU** | Qwen2.5-VL-7B | 1.00 | 16,215 | 2 | 15.404 | 21.000 | ∼20.0 | 0.432 |
| | Video-MTR | 2.00 | 17,503 | 69 | 33.643 | 58.000 | ∼81.6 | 0.933 |
| | Ours | 2.25 | 21,378 | 66 | 20.662 | 26.000 | ∼21.0 | 0.173 |
| **MMVU** | Qwen2.5-VL-7B | 1.00 | 10,233 | 1 | 3.722 | 7.000 | ∼20.0 | 1.822 |
| | Video-MTR | 2.00 | 17,196 | 62 | 58.880 | 69.000 | ∼81.6 | 0.543 |
| | Ours | 2.16 | 15,448 | 61 | 8.108 | 11.000 | ∼21.0 | 0.446 |
| *Average over the five benchmarks in Part I* | | | | | | | | |
| | Qwen2.5-VL-7B | 1.00 | 12,727 | 2.2 | 8.250 | 12.800 | ∼20.0 | 1.347 |
| | Video-MTR | 2.00 | 16,896 | 58.4 | 45.659 | 59.800 | ∼81.6 | 0.761 |
| | Ours | 2.20 | 17,851 | 62.2 | 15.045 | 21.600 | ∼21.0 | 0.504 |
| **Part II: Temporal Grounding** (Table 3) | | | | | | | | |
| **Charades-STA** | Qwen2.5-VL-7B | 1.00 | 5,389 | 13 | 2.361 | 3.000 | ∼20.0 | 3.310 |
| | Ours | 2.13 | 11,129 | 109 | 5.529 | 7.000 | ∼21.0 | 1.430 |
| **ActivityNet** | Qwen2.5-VL-7B | 1.00 | 7,560 | 12 | 3.448 | 5.000 | ∼20.0 | 2.297 |
| | Ours | 2.21 | 23,637 | 280 | 13.395 | 57.000 | ∼21.0 | 0.564 |
| **NExT-GQA** | Qwen2.5-VL-7B | 1.00 | 10,379 | 15 | 4.616 | 7.000 | ∼20.0 | 1.722 |
| | Ours | 2.16 | 13,694 | 79 | 8.040 | 11.000 | ∼21.0 | 0.995 |
| *Average over the three benchmarks in Part II* | | | | | | | | |
| | Qwen2.5-VL-7B | 1.00 | 7,776 | 13.3 | 3.475 | 5.000 | ∼20.0 | 2.443 |
| | Ours | 2.17 | 16,153 | 156.0 | 8.988 | 25.000 | ∼21.0 | 0.996 |

*Notes:*
1. We report the average number of executed inference turns, cumulative input and generated tokens across all model calls, end-to-end latency, peak GPU memory, and throughput. The corresponding task performance is reported in Tabs. 2 and 3.
2. Input Tokens denotes the cumulative number of visual and textual tokens processed across all executed model calls. Generated Tokens is the sum of all tokens generated across the executed turns, including reasoning, correctness judgments, formatting tokens, and final answers.
3. Avg. Turns is the average number of model invocations performed before termination.
4. E2E Latency is the wall-clock time required to obtain the final returned answer and includes visual encoding, multimodal prefill, and autoregressive decoding. P95 is the 95th percentile over per-example E2E latency.
5. Peak Mem. denotes peak allocated GPU memory under the reported batch size, and Throughput is measured over the complete evaluation workload.

### D.4 Selective Error Detection and Net Correction Utility

Our objective is to improve answer accuracy through selective reflection without assuming that autonomous self-evaluation can identify every error. At inference time, the model has access to

Table 12: Hard-label correctness-gate confusion matrices and class-wise metrics derived from Tab. 9. For Turn 2, the gate evaluates all preceding answers; for Turn 3, it evaluates only samples judged *False* at Turn 2.

| Dataset | Turn | Confusion Matrix | | | | Hard-label Metrics (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | TP | FP | FN | TN | $\mathbf{P}_{\mathrm{True}}$ | $\mathbf{R}_{\mathrm{True}}$ | $\mathbf{P}_{\mathrm{False}}$ | $\mathbf{R}_{\mathrm{False}}$ | BAcc. |
| **VideoMME** | Turn 2 | 1607 | 570 | 105 | 418 | 73.82 | 93.87 | 79.92 | 42.31 | 68.09 |
| | Turn 3 | 119 | 181 | 45 | 178 | 39.67 | 72.56 | 79.82 | 49.58 | 61.07 |
| **MVBench** | Turn 2 | 2468 | 813 | 144 | 575 | 75.22 | 94.49 | 79.97 | 41.43 | 67.96 |
| | Turn 3 | 235 | 181 | 61 | 242 | 56.49 | 79.39 | 79.87 | 57.21 | 68.30 |
| **LongVideoBench** | Turn 2 | 748 | 290 | 61 | 245 | 72.06 | 92.46 | 80.07 | 45.79 | 69.13 |
| | Turn 3 | 55 | 144 | 21 | 86 | 27.64 | 72.37 | 80.37 | 37.39 | 54.88 |
| **VideoMMMU** | Turn 2 | 434 | 243 | 45 | 178 | 64.11 | 90.61 | 79.82 | 42.28 | 66.44 |
| | Turn 3 | 60 | 74 | 18 | 71 | 44.78 | 76.92 | 79.78 | 48.97 | 62.94 |
| **MMVU** | Turn 2 | 629 | 213 | 32 | 126 | 74.70 | 95.16 | 79.75 | 37.17 | 66.16 |
| | Turn 3 | 50 | 64 | 9 | 35 | 43.86 | 84.75 | 79.55 | 35.35 | 60.05 |
| **Overall** | **Turn 2** | **5886** | **2129** | **387** | **1542** | **73.44** | **93.83** | **79.94** | **42.00** | **67.92** |
| | **Turn 3** | **519** | **644** | **154** | **612** | **44.63** | **77.12** | **79.90** | **48.73** | **62.92** |

*Note:*
1. TP denotes an actually correct answer judged *True*; FP denotes an actually incorrect answer judged *True*; FN denotes an actually correct answer judged *False*; and TN denotes an actually incorrect answer judged *False*.
2. $P$ and $R$ denote precision and recall, respectively.
3. BAcc. denotes balanced accuracy.

neither reference answers nor an external verifier. We examine correctness judgments using Tab. 12 and changes in answer correctness using Tabs. 13 and 14. These analyses distinguish the ability to identify an incorrect preceding answer from the ability to produce a correct subsequent answer.

The gate statistics in Tab. 12 summarize 9,944 preceding answers at Turn 2, of which 1,929 (19.40%) are judged *False*. This selected group contains 1,542 actually incorrect answers and 387 actually correct answers, yielding a false-class precision of 79.94%. False-class recall is 42.00%, true-class recall is 93.83%, and balanced accuracy is 67.92%. Thus, the reported judgments identify a subset containing a high proportion of errors, while leaving a substantial proportion of incorrect preceding answers labeled *True*.

At Turn 3, Tab. 12 reports judgments on 1,929 selected contexts, corresponding to samples judged *False* at Turn 2. Of these, 766 are again judged *False*. False-class precision and recall are 79.90% and 48.73%, respectively, while true-class recall and balanced accuracy are 77.12% and 62.92%.

For the full method, Tab. 13 directly measures the benefit of reflection with a fixed final checkpoint. Turn 2 corrects 421 previously incorrect answers and corrupts 38 previously correct answers, yielding a correction-to-corruption ratio of 11.08 and a net gain of 383 correct predictions. Relative to the 6,174 initially correct answers, the corruption rate is 0.62%. Turn 3 produces 126 corrections and 37 corruptions within the 1,929 evaluated examples, yielding a ratio of 3.41 and a net gain of 89 correct predictions. This additional benefit is benchmark-dependent: VideoMME and MVBench gain 32 and 72 correct predictions, respectively, whereas LongVideoBench, VideoMMMU, and MMVU each lose five. The third turn therefore improves pooled accuracy while reducing accuracy on some individual benchmarks.

Across both reflective turns, the final checkpoint produces 547 correction events and 75 corruption events, for a correction-to-corruption ratio of 7.29 and a cumulative net gain of 472 correct predictions. These are transition-event counts rather than counts of distinct examples, since an example may contribute an event in each turn. Pooling all 9,944 examples across the five benchmarks and retaining the Turn 2 answers of examples that do not proceed to Turn 3, accuracy increases from

Table 13: Answer transitions over the full test sets during multi-turn inference (**Part 1**: Main results and baseline ablations).

| Dataset | Turn | Total Samples # | False → True # (%) | False → False # (%) | True → False # (%) | True → True # (%) |
|---|---|---|---|---|---|---|
| Our Strategy *refer to Tab. 2* | | | | | | |
| VideoMME | Turn 2 | 2700 | 69 (2.56) | 919 (34.04) | 10 (0.37) | 1702 (63.04) |
| | Turn 3 | 2700 | 21 (0.78) | 908 (33.63) | 5 (0.19) | 1766 (65.41) |
| MVBench | Turn 2 | 4000 | 166 (4.15) | 1222 (30.55) | 14 (0.35) | 2598 (64.95) |
| | Turn 3 | 4000 | 94 (2.35) | 1142 (28.55) | 6 (0.15) | 2758 (68.95) |
| LongVideoBench | Turn 2 | 1344 | 21 (1.56) | 514 (38.24) | 6 (0.45) | 803 (59.75) |
| | Turn 3 | 1344 | 3 (0.22) | 517 (38.47) | 2 (0.15) | 822 (61.16) |
| VideoMMMU | Turn 2 | 900 | 38 (4.22) | 383 (42.56) | 5 (0.56) | 474 (52.67) |
| | Turn 3 | 900 | 0 (0.00) | 388 (43.11) | 10 (1.11) | 502 (55.78) |
| MMVU | Turn 2 | 1000 | 30 (3.00) | 309 (30.90) | 3 (0.30) | 658 (65.80) |
| | Turn 3 | 1000 | 0 (0.00) | 312 (31.20) | 8 (0.80) | 680 (68.00) |
| **w/** Only GRPO (Turn 2) *refer to Tab. 2* | | | | | | |
| VideoMME | Turn 2 | 2700 | 151 (5.59) | 837 (31.00) | 132 (4.89) | 1580 (58.52) |
| MVBench | Turn 2 | 4000 | 304 (7.60) | 1084 (27.10) | 204 (5.10) | 2408 (60.20) |
| LongVideoBench | Turn 2 | 1344 | 68 (5.06) | 467 (34.75) | 67 (4.99) | 742 (55.21) |
| VideoMMMU | Turn 2 | 900 | 46 (5.11) | 375 (41.67) | 66 (7.33) | 413 (45.89) |
| MMVU | Turn 2 | 1000 | 57 (5.70) | 282 (28.20) | 48 (4.80) | 613 (61.30) |
| **w/** Only OPSD (Turn 2) *refer to Tab. 2* | | | | | | |
| VideoMME | Turn 2 | 2700 | 145 (5.37) | 843 (31.22) | 180 (6.67) | 1532 (56.74) |
| MVBench | Turn 2 | 4000 | 188 (4.70) | 1200 (30.00) | 236 (5.90) | 2376 (59.40) |
| LongVideoBench | Turn 2 | 1344 | 72 (5.36) | 463 (34.45) | 153 (11.38) | 656 (48.81) |
| VideoMMMU | Turn 2 | 900 | 51 (5.67) | 370 (41.11) | 43 (4.78) | 436 (48.44) |
| MMVU | Turn 2 | 1000 | 70 (7.00) | 269 (26.90) | 53 (5.30) | 608 (60.80) |
| **w/** GRPO&OPSD (Turn 2) *refer to Tab. 2* | | | | | | |
| VideoMME | Turn 2 | 2700 | 164 (6.07) | 824 (30.52) | 126 (4.67) | 1586 (58.74) |
| MVBench | Turn 2 | 4000 | 288 (7.20) | 1100 (27.50) | 216 (5.40) | 2396 (59.90) |
| LongVideoBench | Turn 2 | 1344 | 62 (4.61) | 473 (35.19) | 81 (6.03) | 728 (54.17) |
| VideoMMMU | Turn 2 | 900 | 49 (5.44) | 372 (41.33) | 54 (6.00) | 425 (47.22) |
| MMVU | Turn 2 | 1000 | 68 (6.80) | 271 (27.10) | 46 (4.60) | 615 (61.50) |

*Note:*
1. **False** and **True** denote the correctness of an answer before and after the corresponding optimization turn.
2. For *Our Strategy*, Turn 2 denotes the transition from Turn 1 to Turn 2, while Turn 3 denotes the transition from Turn 2 to Turn 3.

62.09% at Turn 1 to 65.94% at Turn 2 and 66.83% after Turn 3. The overall improvement is 4.75 percentage points. These pooled accuracies weight examples equally and should be distinguished from an unweighted average of benchmark accuracies.

The baseline ablations in Tab. 13 instead characterize prediction changes between training-stage checkpoints. Only GRPO, Only OPSD, and GRPO&OPSD produce 626, 526, and 631 incorrect-to-correct transitions, together with 517, 665, and 523 correct-to-incorrect transitions, respectively. Their net changes are therefore +109, -139, and +108 correct predictions, with correction-to-corruption ratios of 1.21, 0.79, and 1.21. Relative to the 6,273 correct predictions of the preceding-stage checkpoint, the corresponding correct-to-incorrect transition rates are 8.24%, 10.60%, and 8.34%. These rates describe changes across training checkpoints.

The teacher-conditioning ablations in Tab. 14 use the same stage-wise checkpoint evaluation protocol. Random conditioning ($CCOPD_1$) yields 218 incorrect-to-correct and 356 correct-to-incorrect transitions, for a net loss of 138 correct predictions. Removing correctness conditioning ($CCOPD_2$)

Table 14: Answer transitions over the full test sets during multi-turn inference (**Part 2**: Teacher conditioning ablations).

| Dataset | Turn | Total Samples # | False → True # (%) | False → False # (%) | True → False # (%) | True → True # (%) |
|---|---|---|---|---|---|---|
| | | **w/ GRPO&CCOPD$_1^\dagger$** (Turn 2) *refer to Tab. 2* | | | | |
| VideoMME | Turn 2 | 2700 | 86 (3.19) | 902 (33.41) | 94 (3.48) | 1618 (59.93) |
| MVBench | Turn 2 | 4000 | 108 (2.70) | 1280 (32.00) | 144 (3.60) | 2468 (61.70) |
| LongVideoBench | Turn 2 | 1344 | 1 (0.07) | 534 (39.73) | 56 (4.17) | 753 (56.03) |
| VideoMMMU | Turn 2 | 900 | 1 (0.11) | 420 (46.67) | 26 (2.89) | 453 (50.33) |
| MMVU | Turn 2 | 1000 | 22 (2.20) | 317 (31.70) | 36 (3.60) | 625 (62.50) |
| | | **w/ GRPO&CCOPD$_2^\dagger$** (Turn 2) *refer to Tab. 2* | | | | |
| VideoMME | Turn 2 | 2700 | 98 (3.63) | 890 (32.96) | 77 (2.85) | 1635 (60.56) |
| MVBench | Turn 2 | 4000 | 110 (2.75) | 1278 (31.95) | 118 (2.95) | 2494 (62.35) |
| LongVideoBench | Turn 2 | 1344 | 16 (1.19) | 519 (38.62) | 36 (2.68) | 773 (57.51) |
| VideoMMMU | Turn 2 | 900 | 26 (2.89) | 395 (43.89) | 22 (2.44) | 457 (50.78) |
| MMVU | Turn 2 | 1000 | 40 (4.00) | 299 (29.90) | 30 (3.00) | 631 (63.10) |
| | | **w/ GRPO&CCOPD$_3^\dagger$** (Turn 2) *refer to Tab. 2* | | | | |
| VideoMME | Turn 2 | 2700 | 141 (5.22) | 847 (31.37) | 103 (3.81) | 1609 (59.59) |
| MVBench | Turn 2 | 4000 | 277 (6.93) | 1111 (27.77) | 157 (3.92) | 2455 (61.38) |
| LongVideoBench | Turn 2 | 1344 | 36 (2.68) | 499 (37.13) | 49 (3.65) | 760 (56.55) |
| VideoMMMU | Turn 2 | 900 | 46 (5.11) | 375 (41.67) | 29 (3.22) | 450 (50.00) |
| MMVU | Turn 2 | 1000 | 58 (5.80) | 281 (28.10) | 40 (4.00) | 621 (62.10) |
| | | **w/ GRPO&CCOPD$_4^\dagger$** (Turn 2) *refer to Tab. 2* | | | | |
| VideoMME | Turn 2 | 2700 | 88 (3.26) | 900 (33.33) | 86 (3.19) | 1626 (60.22) |
| MVBench | Turn 2 | 4000 | 103 (2.57) | 1285 (32.12) | 131 (3.28) | 2481 (62.02) |
| LongVideoBench | Turn 2 | 1344 | 1 (0.07) | 534 (39.73) | 55 (4.09) | 754 (56.10) |
| VideoMMMU | Turn 2 | 900 | 14 (1.56) | 407 (45.22) | 24 (2.67) | 455 (50.56) |
| MMVU | Turn 2 | 1000 | 37 (3.70) | 302 (30.20) | 33 (3.30) | 628 (62.80) |

*Note:*
1. **False** and **True** denote the correctness of an answer before and after the corresponding optimization turn.
2. † denotes modified teacher conditioning: CCOPD$_1$: random correctness; CCOPD$_2$: removing correctness; CCOPD$_3$: correctness always sets as True; CCOPD$_4$: correctness always sets as False.

Table 15: Ablation study on the construction of CoT supervision data. We compare Turn 2 performance without CoT-based SFT, with CoT data generated by a single agent (GPT 5.6 Luna), and with CoT data constructed and refined by multiple agents. All experiments use Qwen2.5-VL-7B with 128 input frames.

| Dataset | Video Understanding | | | Video Reasoning | | Average | Remarks |
|---|---|---|---|---|---|---|---|
| | VideoMME | MV. | LongVideo. | VideoMMMU | MMVU | | |
| | | | *Performance in Turn 2* | | | | |
| **Qwen2.5-VL-7B** | 65.1 | 63.9 | 60.9 | 51.1 | 55.7 | 59.3 | – |
| **w/o** CoT Dataset | 65.5 | 67.0 | 58.8 | 52.6 | 68.3 | 62.4 | Only GRPO in Turn 1 |
| **w/** One Agent | 66.1 | 68.8 | 57.1 | 54.3 | 68.8 | 63.0 | SFT & GRPO in Turn 1 |
| **w/** All Agents | 65.6 | 69.1 | 61.3 | 56.9 | 68.8 | 64.3 | SFT & GRPO in Turn 1 |

yields 290 and 283 such transitions, for a net gain of seven. Always conditioning on *True* (CCOPD$_3$) produces 558 corrections and 378 regressions, gaining 180 correct predictions, whereas always conditioning on *False* (CCOPD$_4$) produces 243 corrections and 329 regressions, losing 86. Their correction-to-corruption ratios are 0.61, 1.02, 1.48, and 0.74, respectively. The net changes vary

Table 16: Ablation of dataset filtering under 128-frame input with Qwen2.5-VL-7B. Dataset Num. denotes the number of samples.

| Method | Dataset Num. | Video Understanding | | | Video Reasoning | | Average |
|---|---|---|---|---|---|---|---|
| | | VideoMME | MV. | LongVideo. | VideoMMMU | MMVU | |
| *Removing All-Right & All-Wrong (Our Strategy)* | | | | | | | |
| **SFT** (Turn 1) | 200K | 62.2 | 65.1 | 56.9 | 52.0 | 63.4 | 59.9 |
| **w/** GRPO (Turn 1) | **138K** | 63.4 | 65.3 | 60.2 | 53.2 | 66.1 | 61.6 |
| **w/** GRPO & CCOPD (Turn 2) | **38K** | 65.6 | 69.1 | 61.3 | 56.9 | 68.8 | 64.3 |
| *Retaining All Dataset* | | | | | | | |
| **SFT** (Turn 1) | 200K | 62.2 | 65.1 | 56.9 | 52.0 | 63.4 | 59.9 |
| **w/** GRPO (Turn 1) | **300K** | 62.7 | 64.1 | 58.5 | 51.9 | 64.3 | 60.3 |
| **w/** GRPO & CCOPD (Turn 2) | **300K** | 65.1 | 67.6 | 58.9 | 53.3 | 66.7 | 62.3 |
| *Only Removing All-Right* | | | | | | | |
| **SFT** (Turn 1) | 200K | 62.2 | 65.1 | 56.9 | 52.0 | 63.4 | 59.9 |
| **w/** GRPO (Turn 1) | **245K** | 63.1 | 65.4 | 58.9 | 51.5 | 65.5 | 60.9 |
| **w/** GRPO & CCOPD (Turn 2) | **143K** | 65.0 | 68.9 | 60.1 | 52.9 | 68.3 | 63.0 |
| *Only Removing All-Wrong* | | | | | | | |
| **SFT** (Turn 1) | 200K | 62.2 | 65.1 | 56.9 | 52.0 | 63.4 | 59.9 |
| **w/** GRPO (Turn 1) | **193K** | 62.3 | 63.9 | 57.1 | 51.1 | 63.9 | 59.7 |
| **w/** GRPO & CCOPD (Turn 2) | **190K** | 63.6 | 66.0 | 58.1 | 52.6 | 66.1 | 61.3 |

Table 17: Ablation study on the CCOPD coefficient and reward configurations. We evaluate the weight $\lambda$ of the CCOPD objective, the reward weights used in Turn 1 and Turn 2, and two correctness-judgment reward schemes. Reported results are obtained on five representative benchmarks with 128 input frames.

| Params | Video Understanding | | | Video Reasoning | | Average |
|---|---|---|---|---|---|---|
| | VideoMME | MV. | LongVideo. | VideoMMMU | MMVU | |
| $\lambda$ | $\mathcal{L}_{\text{CCOPD}}$ | | | | | |
| 0.01 | 64.7 | 67.5 | 59.8 | 56.1 | 68.1 | 63.2 |
| 0.05 | 65.6 | 69.1 | 61.3 | 56.9 | 68.8 | 64.3 |
| 0.08 | 65.1 | 68.6 | 60.6 | 57.2 | 66.5 | 63.6 |
| 0.1 | 62.5 | 66.3 | 59.2 | 54.1 | 63.4 | 61.1 |
| 0.2 | 59.9 | 61.9 | 54.9 | 54.5 | 62.1 | 58.7 |
| Answer& Format& Answer | *Reward Weights in Turn 1* | | | | | |
| 1.0, 1.0, 1.0 | 63.4 | 65.3 | 60.2 | 53.2 | 66.1 | 61.6 |
| 0.9, 1.0, 1.1 | 63.6 | 64.7 | 58.6 | 53.7 | 65.9 | 61.3 |
| 1.1, 0.9, 1.0 | 63.1 | 65.3 | 59.1 | 53.7 | 66.3 | 61.5 |
| Correctness& Format& Answer | *Reward Weights in Turn 2* | | | | | |
| 1.0, 1.0, 1.0 | 65.8 | 68.7 | 60.9 | 57.1 | 67.9 | 64.1 |
| 0.9, 1.0, 1.1 | 64.9 | 68.5 | 61.0 | 56.2 | 68.3 | 63.8 |
| 1.1, 0.9, 1.0 | 65.6 | 69.1 | 61.3 | 56.9 | 68.8 | 64.3 |
| | *Judge Rewards in Turn 2* | | | | | |
| Correct/Wrong Reward† | 65.3 | 68.1 | 61.1 | 56.3 | 67.3 | 63.6 |
| Asymmetric Reward‡ | 65.6 | 69.1 | 61.3 | 56.9 | 68.8 | 64.3 |

*Note:*
1. When testing one hyperparameter, all other hyperparameters are set to their optimal values;
2. † Correct/Wrong Reward uses Correct $= 1.0$, Wrong $= 0.0$;
3. ‡ Asymmetric Reward uses True-correct $= 0.5$, False-correct $= 1.0$, Wrong $= 0.0$, Missing $= -1.0$.

across the conditioning variants, showing that training-stage improvements and regressions depend on the conditioning configuration in the reported experiments.

Taken together, the fixed-checkpoint transitions demonstrate a positive net benefit from multi-turn reflection in the evaluated final model, while the ablations characterize how predictions change across training stages.

Table 18: Ablation of token-level divergence objectives and teacher update strategies for CCOPD. Results are reported on five representative benchmarks using 128 input frames.

| Strategy | Video Understanding | | | Video Reasoning | | Average | Remarks |
|---|---|---|---|---|---|---|---|
| | VideoMME | MV. | LongVideo. | VideoMMMU | MMVU | | |
| *KL Divergence Objective* | | | | | | | |
| Forward KL | 64.9 | 66.7 | 60.6 | 54.7 | 66.4 | 62.7 | $\mathcal{L}_{\text{dist}} = D_{\text{KL}}(p^T \| p^S)$ |
| Reverse KL | 65.6 | 69.1 | 61.3 | 56.9 | 68.8 | 64.3 | $\mathcal{L}_{\text{dist}} = D_{\text{KL}}(p^S \| p^T)$ |
| Symmetric KL | 65.3 | 66.8 | 60.2 | 52.7 | 67.9 | 62.6 | $\mathcal{L}_{\text{dist}} = 0.5[D_{\text{KL}}(p^S \| p^T) + D_{\text{KL}}(p^T \| p^S)]$ |
| JS divergence | 64.1 | 67.5 | 58.5 | 50.5 | 66.1 | 61.3 | $\mathcal{L}_{\text{dist}} = 0.5[D_{\text{KL}}(p^S \| m) + D_{\text{KL}}(p^T \| m)]$ |
| No distillation | 64.1 | 67.8 | 60.3 | 51.0 | 67.0 | 62.0 | $\mathcal{L}_{\text{dist}} = 0$; GRPO objective only. |
| *Teacher Update Strategy* | | | | | | | |
| EMA | 59.6 | 66.8 | 58.9 | 49.7 | 66.1 | 60.2 | See *Note* 3. |
| Frozen | 58.1 | 66.3 | 53.7 | 50.4 | 65.6 | 58.8 | See *Note* 3. |
| Shared | 65.6 | 69.1 | 61.3 | 56.9 | 68.8 | 64.3 | See *Note* 3. |

*Notes:*
1. $p^S$ and $p^T$ denote the student and correctness-conditioned teacher token distributions, respectively; the teacher distribution is treated as a stop-gradient target in all distillation objectives.
2. For the Jensen–Shannon divergence, $m = \frac{1}{2}(p^S + p^T)$.
3. *Frozen* keeps the teacher unchanged during optimization, *Shared* evaluates the detached teacher branch using the current student parameters, and *EMA* updates the teacher from an exponential moving average of the student parameters.
4. The *Average* column is the arithmetic mean over the five benchmarks. Shaded rows indicate our selected setting within each ablation block.

### D.5 Computational Overhead and Efficiency

Tab. 11 reports inference-time computational costs using deployment-oriented metrics: the average number of executed turns, cumulative input and generated tokens, mean and P95 end-to-end latency, peak GPU memory, and throughput. All measurements use NVIDIA H20 GPUs with 96 GB of HBM3 memory under the same hardware, batch size, numerical precision, inference engine, frame-processing pipeline, and decoding configuration. End-to-end latency includes visual encoding, multimodal prefill, autoregressive decoding, and all model calls required to produce the returned answer. Token counts accumulate the inputs and outputs across executed turns. The summary rows report unweighted averages across benchmarks; in particular, the summarized P95 values are averages of the per-benchmark P95 latencies, rather than percentiles computed over pooled requests. Throughput is measured over the complete evaluation workload and is reported separately from per-request latency.

Our method uses adaptive multi-turn inference with a maximum of three turns. Every example executes an initial answer-generation turn and a verification-and-revision turn, while examples assigned a False judgment at Turn 2 proceed to Turn 3. Accordingly, the average number of turns is $2+N_3/N$, where $N$ is the number of evaluated examples and $N_3$ is the number routed to Turn 3. On VideoMME, MVBench, LongVideoBench, VideoMMMU, and MMVU, the average turn counts are 2.19, 2.18, 2.23, 2.25, and 2.16, respectively, with a macro-average of 2.20. On Charades-STA, ActivityNet, and NExT-GQA, the corresponding values are 2.13, 2.21, and 2.16, with a macro-average of 2.17. These results indicate that most examples terminate after Turn 2, while an additional reflective turn is applied to a subset selected by the model's correctness judgment.

Across the five video-understanding and video-reasoning benchmarks in Tab. 2, the single-turn Qwen2.5-VL-7B baseline processes an average of 12,727 input tokens and generates 2.2 tokens per example. Its macro-averaged mean and P95 latencies are 8.250 s and 12.800 s, respectively. Our method processes 17,851 input tokens and generates 62.2 tokens, with mean and P95 latencies of 15.045 s and 21.600 s. Macro-averaged throughput decreases from 1.347 to 0.504 samples/s, while peak memory increases from approximately 20.0 to 21.0 GB per GPU. Under this setup, reflective inference therefore incurs additional latency and token processing with a relatively small increase in peak allocated memory. Compared with Video-MTR, our method uses approximately 5.7% more input tokens and 6.5% more generated tokens, while reducing macro-averaged mean latency from

45.659 to 15.045 s and the average per-benchmark P95 latency from 59.800 to 21.600 s. Peak memory is also lower, at approximately 21.0 versus 81.6 GB per GPU. However, Video-MTR has higher macro-averaged throughput, at 0.761 versus 0.504 samples/s, showing that the latency and memory advantages do not imply a throughput advantage.

Across the three temporal-grounding benchmarks in Tab. 3, our method processes an average of 16,153 input tokens and generates 156.0 tokens per example, compared with 7,776 and 13.3 tokens for the single-turn baseline. Macro-averaged mean latency increases from 3.475 to 8.988 s, the average per-benchmark P95 latency increases from 5.000 to 25.000 s, and throughput decreases from 2.443 to 0.996 samples/s. ActivityNet exhibits the largest overhead among these benchmarks: our method averages 2.21 turns, processes 23,637 input tokens, generates 280 tokens, and has mean and P95 latencies of 13.395 s and 57.000 s, respectively. Charades-STA and NExT-GQA have lower P95 latencies of 7.000 s and 11.000 s. The variation in latency and token usage is substantially larger than the variation in average turn counts, indicating that turn count alone is insufficient to characterize inference cost across tasks.

These measurements characterize an accuracy–compute trade-off that should be assessed jointly with the task results in Tabs. 2 and 3. Adaptive routing limits the third turn to a selected subset, but every example still incurs the cost of a verification-and-revision turn. Under the reported configurations, our method has lower latency and peak GPU memory than Video-MTR on their shared benchmarks, while remaining more computationally expensive than direct single-turn inference. Its deployment value therefore depends on the task-performance gains achievable within the application's latency, memory, and throughput requirements.

### D.6 Ablation Studies

#### D.6.1 Effect of CoT Data Construction.

Tab. 15 investigates how the construction of the CoT supervision data affects the subsequent reflective reasoning performance. Applying GRPO without CoT-based SFT already improves the average score over the original Qwen2.5-VL-7B from 59.82 to 62.44, indicating that reward-based optimization alone can substantially enhance the model, particularly on MMVU. Introducing CoT data generated by a single agent further increases the average score to 63.02 and improves MVBench and VideoMMMU to 68.8 and 54.3, respectively. However, the single-agent construction decreases LongVideoBench performance to 57.1, suggesting that reasoning traces produced from a single perspective may contain systematic biases or fail to capture sufficient evidence in long videos. In contrast, the multi-agent construction achieves the highest average score of 64.34, outperforming the setting without CoT data by 1.90 points and the single-agent variant by 1.32 points. It obtains the best performance on MVBench, LongVideoBench, and VideoMMMU, while tying the best result on MMVU. Although its VideoMME score is slightly lower than that of the single-agent variant, the multi-agent setting provides the strongest and most consistent performance across the five benchmarks. These results validate the importance of using diverse agents to generate complementary reasoning candidates and an aggregation procedure to remove unsupported reasoning and consolidate reliable evidence. Overall, high-quality multi-agent CoT supervision provides a stronger initialization for GRPO and improves the generalization of reflective reasoning across both general video understanding and knowledge-intensive video reasoning tasks.

#### D.6.2 Effect of Dataset Filtering.

Table 16 evaluates four dataset-filtering strategies under the 128-frame setting. At Turn 1, retaining all 300K rollout groups yields an average score of 60.3, whereas removing only all-correct or only all-incorrect groups yields 60.9 and 59.7, respectively. In contrast, retaining only the 138K mixed-outcome groups achieves the best average score of 61.6, improving the five benchmarks by 0.7, 1.2, 1.7, 1.3, and 1.8 points over the unfiltered setting. The advantage becomes more pronounced at Turn 2: joint GRPO and CCOPD training on 38K mixed-outcome groups achieves an average score of 64.3, outperforming retaining all 300K groups, removing only all-correct groups, and removing only all-incorrect groups by 2.0, 1.3, and 3.0 points, respectively. These results show that the substantial reduction in the optimization subset does not degrade performance; instead, consistently retaining groups with heterogeneous outcomes produces stronger results than all three alternatives.

This behavior follows from the group-relative optimization mechanism of GRPO. When all rollouts of an example are correct or all are incorrect, their answer-level rewards have zero within-group variance, so the normalized relative advantages provide no contrastive policy-gradient signal for distinguishing successful trajectories from unsuccessful ones. Retaining these homogeneous groups nevertheless consumes rollout and optimization capacity, reducing the proportion of computation allocated to informative updates. Our filtering strategy therefore retains examples for which the current policy generates at least one correct and one incorrect response, concentrating optimization near the policy's current decision boundary. The resulting 138K and 38K subsets should be interpreted as policy-dependent optimization subsets rather than datasets obtained by removing low-quality examples: all-correct groups represent consistently solved cases, while all-incorrect groups may contain zero-success cases that cannot be improved through relative outcome comparison alone.

#### D.6.3 Effect of Training Hyperparameters.

Tab. 17 examines the CCOPD coefficient, the reward weights used in the two training turns, and the reward scheme for correctness judgment. **(1) For the CCOPD coefficient** $\lambda$, the moderate value of $0.05$ achieves the best average score of $64.3$. Decreasing $\lambda$ to $0.01$ weakens the token-level distillation signal and reduces the average score to $63.2$, whereas increasing it to $0.08$ yields a competitive but lower score of $63.6$. Although $\lambda = 0.08$ achieves the highest individual result on VideoMMMU, further increasing $\lambda$ to $0.1$ and $0.2$ substantially decreases the average performance to $61.1$ and $58.7$, respectively. These results indicate that excessive distillation can dominate the GRPO objective and over-constrain the policy toward the correctness-conditioned teacher distribution. **(2) For the Turn 1 reward weights**, assigning equal importance to the answer- and format-related objectives, *i.e.*, $(1.0, 1.0, 1.0)$, produces the highest average score of $61.6$, suggesting that the initial policy benefits from balanced optimization of response quality and structural validity. **(3) In Turn 2, the correctness-oriented configuration** $(1.1, 0.9, 1.0)$ achieves the best average score of $64.3$, outperforming both uniform weighting $(64.1)$ and the answer-emphasized configuration $(63.8)$. This result highlights the importance of accurately assessing the preceding answer before preserving or revising it during reflective optimization. **(4) The design of the judgment reward** is particularly important because the correctness labels are highly imbalanced during training: preceding answers labeled *True* are substantially more frequent than those labeled *False*. Under the conventional correct/wrong reward, correctly predicting either class receives the same reward. Consequently, predicting *True* for most or even all preceding answers can become a reward-maximizing shortcut, since such a policy is correct on the majority class while largely avoiding the more difficult task of detecting erroneous answers. This degenerate behavior weakens the error-detection capability required for effective reflection and allows incorrect answers to bypass revision. Our asymmetric reward counteracts this bias by assigning a larger reward to a correct *False* judgment than to a correct *True* judgment ($1.0$ *vs.* $0.5$), while assigning zero reward to an incorrect judgment and a negative reward to a missing judgment. The larger reward for correctly identifying errors increases their optimization advantage despite their lower frequency, discouraging majority-class collapse and encouraging the model to distinguish genuine errors from already-correct answers. Empirically, this design improves the average score from $63.6$ to $64.3$ and yields consistent gains across all five benchmarks. Based on these results, we use $\lambda = 0.05$, uniform reward weights in Turn 1, correctness-oriented reward weights in Turn 2, and the asymmetric judgment reward in the final configuration.

#### D.6.4 Effect of Token-level Divergence Objectives and Teacher Update Strategies.

Tab. 18 investigates the effects of token-level divergence objectives and teacher update strategies in CCOPD. **(1) Among the divergence objectives**, reverse KL achieves the best performance on all five benchmarks, obtaining an average score of 64.3. It surpasses forward KL, symmetric KL, Jensen–Shannon divergence, and the no-distillation baseline by 1.6, 1.7, 3.0, and 2.3 points on average, respectively. In particular, the substantial improvement over the no-distillation baseline confirms the effectiveness of transferring token-level distributional knowledge from the correctness-conditioned teacher. The consistent advantage of reverse KL further indicates that encouraging the student to concentrate its probability mass on the teacher-preferred outputs is more effective than the alternative distribution-matching objectives considered here. **(2) Regarding the teacher update strategy**, the Shared teacher achieves the highest average score of 64.3, outperforming the EMA and Frozen variants by 4.1 and 5.5 points, respectively. A frozen teacher cannot adapt to the evolving

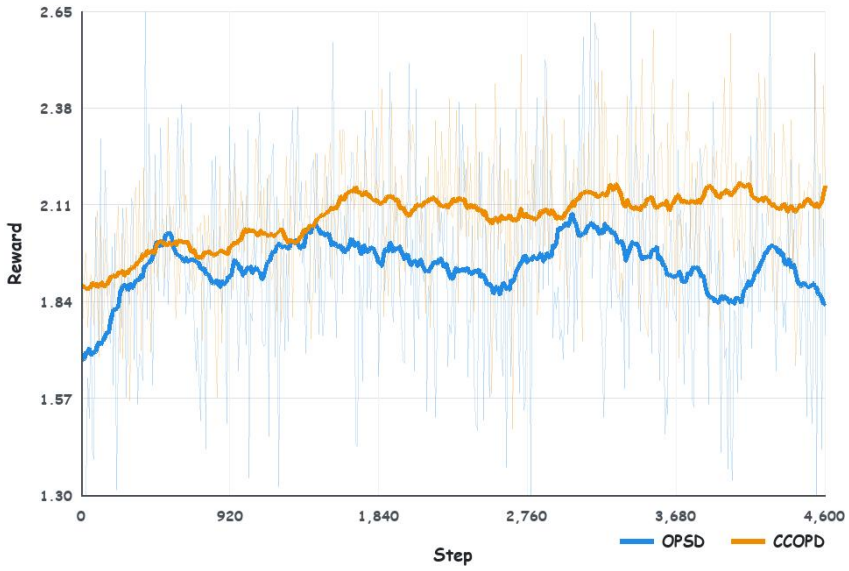



Figure 5: Reward trajectories of OPSD and the proposed CCOPD. The light curves show the raw rewards, while the solid curves represent their smoothed trends.

student policy and may therefore become increasingly misaligned as optimization proceeds, while an EMA teacher can introduce a lag between the teacher and the current policy. In contrast, the Shared strategy constructs the detached teacher distribution from the current model parameters, allowing the teacher target to remain synchronized with the student while preventing gradients from propagating through the teacher branch. Based on these results, we adopt reverse KL with the Shared teacher as the default configuration of CCOPD.

# E VISUALIZATION

## E.1 LONG-HORIZON TRAINING STABILITY OF CCOPD

Fig. 5 compares the training reward trajectories of GRPO&OPSD and GRPO&CCOPD. CCOPD maintains a higher smoothed reward during the later stages of training, whereas OPSD declines after an early peak. These trajectories provide a diagnostic view of optimization behavior. Because rollout distributions and policy-dependent training subsets may differ between the methods, the reward gap alone does not establish its underlying cause or demonstrate reduced shortcut learning or improved transfer.

## E.2 EXPERIMENTAL RESULTS

To complement the quantitative results, we provide qualitative visualizations of two-turn question answering on all 13 benchmarks across the four evaluation categories. Specifically, Fig. 6, Fig. 7, Fig. 8, Fig. 9 present examples from the video understanding benchmarks, including Video-MME, MVBench, MVP, and VSI-Bench; Fig. 10, Fig. 11, Fig. 12, Fig. 13 cover the temporal reasoning benchmarks, including Charades-STA, NExT-GQA, ActivityNet Captions, and TempCompass; Fig. 14, Fig. 15 show examples from the video reasoning benchmarks, including MMVU and Video-MMMU; and Fig. 16, Fig. 17, Fig. 18 visualize results on the long-video understanding benchmarks, including LongVideoBench, MLVU, and LVBench. Together, these examples demonstrate that our evaluation covers all four task categories and the complete set of 13 benchmarks, ranging from general and long-form video understanding to knowledge-intensive reasoning and temporal grounding.

### E.3 Prompts in Multi-Agent Dataset Filtering

The prompts used to construct and consolidate the multi-agent CoT annotations are provided in Prompts F.1 and F.2. The agent-voting prompt in Prompt F.1 instructs each of the six answering agents to independently analyze the question and the associated visual input without access to the reference answer. Each agent identifies the question's target, extracts relevant visual and temporal evidence, and produces a rationale and a final answer in a standardized JSON format.

The candidate answers are then checked against the verified annotation using the task-specific evaluator. For example $i$, let $(r_{i,j}, a_{i,j})$ denote the rationale–answer pair produced by agent $j$, and define the set of answer-correct candidates as

$$\mathcal{C}_i = \{j \in \{1, \dots, 6\} \mid E_i(a_{i,j}, a_i^\star) = 1\}. \tag{52}$$

If $\mathcal{C}_i$ is empty, the sample is discarded and excluded from SFT; no rationale is synthesized from the reference answer to recover the sample. Otherwise, the summary prompt in Prompt F.2 constructs the final rationale–answer pair from the candidates in $\mathcal{C}_i$. When multiple correct candidates are available, it selects the rationale with the strongest support from the input evidence and cleans it to remove references to voting, agreement, correctness labels, and other annotation-specific information without introducing new supporting evidence. We denote the resulting final pair by $(\tilde{r}_i, \tilde{a}_i)$.

In addition to adapting the model to the task distribution, SFT initializes the structured output format required by the subsequent training stages. We use a common response structure consisting of two reasoning spans and two boxed fields. In Turn 1, both boxed fields contain task answers; in later reflective turns, the first contains a correctness judgment and the second contains the final task answer.

To construct each Turn 1 SFT target, we uniformly sample one answer-correct candidate for the first reasoning–answer stage, $j_i \sim \mathrm{Uniform}(\mathcal{C}_i)$, and use the final pair produced by the summary step for the second stage. The target is serialized as

$$y_i^\star = \texttt{<think>}\, r_{i,j_i}\, \texttt{</think>}\boxed{a_{i,j_i}}\texttt{<think>}\, \tilde{r}_i\, \texttt{</think>}\boxed{\tilde{a}_i}. \tag{53}$$

The first-stage candidate is sampled exclusively from $\mathcal{C}_i$, while the second stage incorporates the final rationale selection and cleaning. The two stages need not contain different answers or rationales, particularly when only one correct candidate is available.

## F Limitations

Our method remains limited by imperfect autonomous verification. A mistaken *True* judgment terminates reflection and can preserve an incorrect answer, while additional turns may corrupt a previously correct response. The method also increases inference cost because every example invokes at least one reflective turn, and its benefit depends on the frame budget and the reliability of structured-output parsing. During training, CCOPD requires correctness labels derived from reference annotations, and policy-dependent filtering excludes all-correct and all-incorrect rollout groups; the resulting curriculum may underrepresent trivial examples and failure modes that the current policy cannot solve.

Table 19: **Video QA results** across two backbones (Qwen2.5-VL-7B and Qwen3-VL-8B) and 13 benchmarks in 4 categories. The best is **bold** and the second-best is underlined. * denotes running by ourselves. † denotes collecting from the accepted papers.

| | Method | Video Understanding | | | | Long Video Understanding | | | Video Reasoning | | Temporal Reasoning | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MV. | VideoMME | MVP | VSI. | LongVideo. | MLVU | LV. | MMVU | VideoMMMU | Charades-STA | ActivityNet | NExT-GQA | TempCompass |
| | *Single-turn Video MLLMs* | | | | | | | | | | | | | |
| Qwen2.5-VL-7B | Qwen2.5-VL Bai et al. (2025b) | 63.9* | 65.2* | 35.9* | 39.3* | 56.0* | 54.0* | 43.1* | 64.9* | 51.4* | 52.8* | 29.7* | 26.5* | 72.6* |
| | LLaVA-Video Liu et al. (2023) | - | 63.3 | - | - | 58.2 | 70.8 | - | - | - | - | - | - | - |
| | Video-R1 Feng et al. (2026a) | 63.9† | 59.3 | 33.0 | 35.8† | - | 45.4 | 35.9 | 63.8 | 52.3 | - | - | - | 73.2† |
| | VideoChat-R1 Yan et al. (2026) | 67.9† | 63.6 | - | 33.9† | 58.2 | 68.7 | - | 63.0 | 54.0 | **60.8**† | 36.6† | 32.4† | 72.5† |
| | Rewatch-R1 Zhang et al. (2026a) | - | 65.6 | - | - | - | - | 43.3 | - | 51.9 | - | - | - | - |
| | Video-Thinker Wang et al. (2025a) | - | - | - | - | - | - | 37.0 | - | - | - | - | - | - |
| | Time-R1 Wang et al. (2026c) | 63.1† | 59.3 | - | 29.0† | - | - | - | 62.9 | 51.0 | 58.8 | **52.1** | - | 73.7† |
| | VersaVid-R1 Chen et al. (2025) | 62.9† | 58.8 | - | 33.7† | - | - | - | 64.3 | 51.9 | - | - | - | 74.0† |
| | VideoP2R Jiang et al. (2026) | 68.1† | 60.0 | - | 36.8† | - | - | - | 65.4 | 55.0 | - | - | - | 74.5† |
| | VideoRFT Wang et al. (2026b) | 62.1 | 59.8 | - | 36.8† | - | - | - | 68.5 | 51.1 | - | - | - | 73.7† |
| | VITAL-7B Zhang et al. (2026b) | - | 66.1 | - | 41.8† | 58.5 | - | 46.2 | 68.7 | 54.2 | 59.9† | 49.8† | **43.0**† | - |
| | VideoAuto-R1 Liu et al. (2026) | 71.0† | **67.3**† | 39.4† | - | 60.5† | - | - | 69.7† | **58.6**† | 60.0† | 47.6† | 36.7† | - |
| | VideoKR Fu et al. (2026) | 68.9 | 66.4 | - | - | 61.3 | - | - | 60.5 | 52.2 | - | - | - | - |
| | *Multi-turn Video MLLMs* | | | | | | | | | | | | | |
| | Video-RTS Wang et al. (2025d) | - | 63.0 | - | - | 56.6 | - | - | 66.4 | 52.7 | - | - | - | - |
| | Video-MTR Xie et al. (2025) | 61.7* | 61.9* | - | - | 57.7* | 50.4† | 42.3† | 61.9* | 50.4* | - | - | - | 61.7* |
| | LOVE-R1 Fu et al. (2025b) | - | 64.3 | - | - | 58.0 | 69.9 | 46.7 | 65.0 | 50.6 | - | - | - | - |
| | Conan Ouyang et al. (2026) | - | 60.5 | - | - | 56.6 | 63.4 | 39.2 | - | - | - | - | - | - |
| | LongVT Yang et al. (2026b) | - | 64.3 | - | - | - | - | 41.3 | - | 45.4 | 27.2 | - | - | - |
| | Video-Zoomer Ding et al. (2026) | - | 64.6 | - | - | 55.9 | 69.9 | 44.0 | 65.6 | 51.4 | - | - | - | - |
| | Video-o3 Zeng et al. (2026) | - | 66.5 | - | - | 60.5 | **72.1** | **47.6** | 67.2 | 51.7 | 60.7 | - | - | - |
| | ResAdapt Liao et al. (2026) | - | 66.2† | - | - | 60.2† | 66.0† | 41.8† | 53.5† | 52.6† | 49.1† | 44.7† | 35.3† | - |
| | **Ours** | **72.1** | 66.8 | **48.8** | **44.7** | **62.3** | 67.8 | 46.3 | **70.3** | 56.6 | 60.3 | 51.9 | 37.8 | **80.5** |
| Qwen3-VL-8B | Qwen3-VL-8B Bai et al. (2025a) | 67.9* | 68.2* | 40.6* | 58.6* | 61.6* | 58.7* | 46.3* | 59.6* | 60.3* | 44.5* | 33.3* | 37.2* | 76.7* |
| | OneThinker Feng et al. (2026b) | - | 66.5† | - | - | 61.7† | - | - | 70.5† | 66.2† | 59.9† | 45.9† | - | - |
| | VideoAuto-R1 Liu et al. (2026) | 72.0† | **71.7**† | 43.0† | - | **67.4**† | - | - | **71.1**† | 65.0† | **63.7**† | 51.9† | 44.2† | - |
| | Qwen3-VL-8B-Thinking | 68.0 | 67.6 | - | - | 60.0 | - | - | 60.5 | 64.9 | - | - | - | - |
| | VideoKR Fu et al. (2026) | 67.0 | 67.8 | - | - | 61.5 | - | - | 64.8 | 63.0 | - | - | - | - |
| | ResAdapt Liao et al. (2026) | - | 67.4† | - | - | 61.9† | 70.8† | 43.3† | 56.3† | 59.6† | 39.8† | 30.0† | 33.3† | - |
| | **Ours** | **73.9** | 69.6 | **44.5** | **61.3** | **66.7** | **73.1** | **48.2** | 69.8 | **67.3** | 63.5 | **57.1** | **45.6** | **83.6** |

**Prompt F.1: Agent Vote**

**Role**
You are an expert multimodal reasoning assistant. Your task is to analyze the user's question together with the provided raw visual data, which may be a single image or sampled video frames arranged in temporal order.

---

**Requirement**
Carefully process the visual material and solve the question step by step. Your reasoning must be evidence-based and grounded in the visual content. If the visual material is a video, analyze the frames in temporal order and consider object changes, actions, spatial relations, scene transitions, and relevant temporal cues. If the visual material is an image, analyze the visible objects, text, people, actions, layout, and contextual clues.
In the reasoning process, explicitly include the following four parts:

1. **Question analysis:** Parse what the question is asking, identify the target information, and clarify what kind of answer is required.
2. **Visual analysis:** Analyze the image or video frames, describing only relevant visible evidence such as objects, actions, attributes, spatial relationships, temporal changes, and any readable text.
3. **Information summary:** Summarize the key facts needed to answer the question, combining the question requirements with the visual evidence.
4. **Answer derivation:** Reason over the summarized information step by step and explain how it leads to the final answer.

Do not invent details that are not supported by the visual content. If the evidence is insufficient or ambiguous, state the uncertainty clearly in the reasoning and provide the most likely answer if possible.

---

**Question**
`{question}`

---

**Output**
Output MUST be a single valid JSON object only, with no markdown, no code fences, and no extra text. The JSON object must contain exactly two keys: `reasoning` and `final answer`.
The value of `reasoning` must be a concise but complete evidence-based reasoning string that includes the four required parts: Question analysis, Visual analysis, Information summary, and Answer derivation. The value of `final answer` must be a concise answer to the question.
Required output format:

```
{
"reasoning": "Question analysis: ... Visual analysis:
... Information summary: ... Answer derivation:
...",
"final answer": "your concise final answer"
}
```

**Prompt F.2: Summary**

**Role**
You are an expert annotator with access to the verified ground-truth answer and multiple independent solution attempts for the same question.

---

**Task**
Your task is to produce a final training target by selecting or cleaning an existing rationale whenever possible.

---

**Inputs**

- {`question`}
- {`verified ground truth`}
- candidate responses, where each response contains:
  - {`final answer`}
  - {`reasoning`}

---

**Rules**
Follow these rules exactly:

1. Compare each candidate final answer with the verified ground truth.
2. If one or more candidate {`final answer`} values match the {`verified ground truth`}:
   - Do NOT write a new rationale from scratch.
   - Select reasoning only from the candidate responses whose {`final answer`} matches the {`verified ground truth`}.
   - If exactly one such response exists:
     - set difficulty = "medium"
     - use that response's reasoning, but clean it minimally if needed
   - If two or more such responses exist:
     - set difficulty = "easy"
     - choose the single best reasoning among them, and clean it minimally if needed
   - "Clean minimally" means:
     - remove references to correctness, comparison, voting, agreement, disagreement, or answer selection
     - remove unsupported, speculative, circular, or generic filler statements
     - keep the original wording and content as much as possible
     - do not add new supporting details that were not already present in the chosen reasoning
3. If no candidate {`final answer`} matches the {`verified ground truth`}:
   - set difficulty = "hard"
   - do not generate or rewrite any rationale

---

**Output**
Requirements for all outputs:

- Do NOT mention agents, models, candidates, responses, comparisons, voting, agreement, disagreement, or correctness counts.
- Do NOT use phrases such as:
  - ``the ground truth indicates''
  - ``the correct answer is''
  - ``this matches the correct answer''
  - ``one response says''

```
        - ``two answers agree''
```

- Prefer deleting content over rewriting.
- Prefer a generic safe rationale over a more specific but potentially unsupported one.

Output valid JSON only. Avoid LaTeX. Minimize escaping; use plain text for math:

```
{
"difficulty": ["hard", "medium", "easy"],
"reasoning": "a cleaned, verifiable rationale",
"final answer": "the {verified ground truth}"
}
```



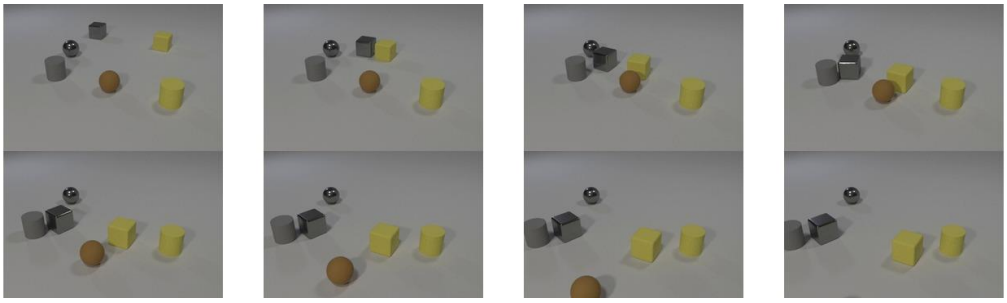

**Turn 1 Input:**
Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option.
What color of clothing is the painter wearing?
A. Blue. B. Yellow. C. Black. D. White.
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Turn 1 Output:**
**Think:** <think>Focusing on the painter's clothing and comparing its visible color with the four options, my initial observation is that it appears black. This corresponds to option C.</think>\boxed{C}<think>After reviewing the painter's clothing again under the visible lighting, I interpret its dominant color as white rather than black. Therefore, I revise the selection to option D.</think>\boxed{D}
**Answer:** D

**Turn 2 Input:**
**Question:**
Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option.
What color of clothing is the painter wearing?
A. Blue. B. Yellow. C. Black. D. White.
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
**Assistance:**
<think>Focusing on the painter's clothing and comparing its visible color with the four options, my initial observation is that it appears black. This corresponds to option C.</think>\boxed{C}<think>After reviewing the painter's clothing again under the visible lighting, I interpret its dominant color as white rather than black. Therefore, I revise the selection to option D.</think>\boxed{D}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
**Think:** The previous final answer identified the painter's clothing as white, corresponding to option D. On closer review, the clothing does not appear predominantly white, so the previous answer is incorrect.
**Judge:** False
**Think:** Re-examining the painter's clothing and comparing its dominant visible color with the available choices, blue is the best match. Therefore, the revised answer is option A.
**Answer:** A
**Ground-Truth:** A

Figure 6: Visualization of Two-turn Question-Answers of VideoMME Benchmark (Video Understanding).



**Turn 1 Input:**
Question: Are there any rubber spheres that exit the scene?
Options: (A) not sure (B) No (C) yes
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Turn 1 Output:**
Think: <think>I focus on whether any object identifiable as a rubber sphere crosses the boundary of the visible scene. In my initial review, no such exit is observed, which corresponds to option B.</think>\boxed{B}<think>After checking the relevant object movements again, I do not identify a rubber sphere leaving the frame. I therefore retain option B, "no," as the final answer.</think>\boxed{B}
Answer: B

**Turn 2 Input:**
**Question:**
Question: Are there any rubber spheres that exit the scene?
Options: (A) not sure (B) No (C) yes
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
**Assistance:**
<think>I focus on whether any object identifiable as a rubber sphere crosses the boundary of the visible scene. In my initial review, no such exit is observed, which corresponds to option B.</think>\boxed{B}<think>After checking the relevant object movements again, I do not identify a rubber sphere leaving the frame. I therefore retain option B, "no," as the final answer.</think>\boxed{B}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
Think: The previous final answer was B, meaning that no rubber sphere exited the scene. On closer review of the object trajectories,at least one rubber sphere appears to cross out of the visible area, so the previous answer is not correct.
Judge: False
Think: Re-examining the event specifically for rubber spheres leaving the frame, I identify an exit from the scene. Therefore, the revised answer is option C, "yes."
Answer: C
Ground-Truth: C

Figure 7: Visualization of Two-turn Question-Answers of MVBench Benchmark (Video Understanding).



**Turn 1 Input:**
Question: How many objects were put in the box?
Options: A. 6 B. 7
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
Put your final answer in \boxed{} with a single letter from A or B, format exactly like: '\boxed{Answer: A}' or '\boxed{Answer: B}'.

**Turn 1 Output:**
**Think:** <think>I count each completed placement into the box once, excluding any object that is only moved nearby or handled more than once. My initial count is seven objects, which corresponds to option B.</think>\boxed{B}<think>Reviewing the placements in chronological order and counting only objects that enter the box, the total remains seven. Therefore, I retain option B.</think>\boxed{B}
**Answer:** B

**Turn 2 Input:**
**Question:**
Question: How many objects were put in the box?
Options: A. 6 B. 7
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
Put your final answer in \boxed{} with a single letter from A or B, format exactly like: '\boxed{Answer: A}' or '\boxed{Answer: B}'.
**Assistance:**
<think>I count each completed placement into the box once, excluding any object that is only moved nearby or handled more than once. My initial count is seven objects, which corresponds to option B.</think>\boxed{B}<think>Reviewing the placements in chronological order and counting only objects that enter the box, the total remains seven. Therefore, I retain option B.</think>\boxed{B}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
**Think:** The previous final answer was B, corresponding to seven objects. After recounting the completed placements and checking for adduplicated or invalid count, the total is six rather than seven. The previous answer is therefore incorrect.
**Judge:** False
**Think:** Counting each object placed in the box exactly once gives a total of six objects. Since option A represents six, the revisedanswer is A.
**Answer:** A
**Ground-Truth:** A

Figure 8: Visualization of Two-turn Question-Answers of MVP Benchmark (Video Understanding).



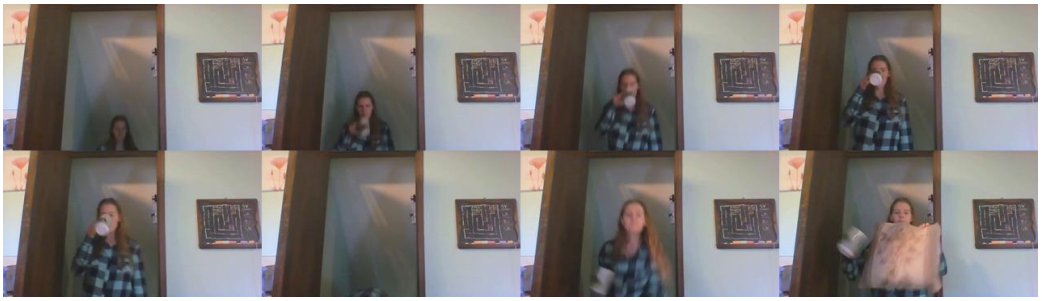
**Turn 1 Input:**
Measuring from the closest point of each object, which of these objects (window, door, trash bin, table) is the closest to the radiator?
A. Window B. door C. trash bin D. table
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Turn 1 Output:**
Think: <think>I compare the shortest visible gap between the radiator and each candidate object. In the initial view, the table appears to have the smallest separation from the radiator, corresponding to option D.</think>\boxed{D}<think>After checking the nearest edges of the objects more carefully, the window appears closer to the radiator than the table, door, or trash bin. I therefore revise the answer to option A.</think>\boxed{A}
Answer: A

**Turn 2 Input:**
**Question:**
Measuring from the closest point of each object, which of these objects (window, door, trash bin, table) is the closest to the radiator?
A. Window B. door C. trash bin D. table
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
**Assistance:**
<think>I compare the shortest visible gap between the radiator and each candidate object. In the initial view, the table appears to have the smallest separation from the radiator, corresponding to option D.</think>\boxed{D}<think>After checking the nearest edges of the objects more carefully, the window appears closer to the radiator than the table, door, or trash bin. I therefore revise the answer to option A.</think>\boxed{A}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
Think: The previous final answer was A, indicating that the window is closest to the radiator. A closer comparison of the minimum visible gaps shows that the door is nearer, so the previous answer is incorrect.
Judge: False
Think: Measuring conceptually from the nearest point of each object to the radiator, the door has the smallest apparent separation among the four choices. Therefore, the revised answer is option B.
Answer: B
Ground-Truth: B

Figure 9: Visualization of Two-turn Question-Answers of VSIBench Benchmark (Video Understanding).



**Turn 1 Input:**
Locate the start and end timestamps of the video segment corresponding to the description: a person walks in a doorway.
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question. Please provide the start and end timestamps (in seconds, precise to one decimal places) in the format \boxed{[start, end]}.

**Turn 1 Output:**
Think: <think>The target action is the interval in which a person is walking within or through a doorway. The selected segment from 0.0 to 10.0 seconds is intended to include this complete doorway-related action.</think>\boxed{[0.0, 10.0]}<think>Reviewing the selected temporal boundaries, the relevant action is localized to the opening ten seconds of the video. I therefore retain 0.0 seconds as the start and 10.0 seconds as the end.</think>\boxed{[0.0, 10.0]}
Answer: [0.0, 10.0]

**Turn 2 Input:**
**Question:**
Locate the start and end timestamps of the video segment corresponding to the description: a person walks in a doorway.
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question. Please provide the start and end timestamps (in seconds, precise to one decimal places) in the format \boxed{[start, end]}.
**Assistance:**
<think>The target action is the interval in which a person is walking within or through a doorway. The selected segment from 0.0 to 10.0 seconds is intended to include this complete doorway-related action.</think>\boxed{[0.0, 10.0]}<think>Reviewing the selected temporal boundaries, the relevant action is localized to the opening ten seconds of the video. I therefore retain 0.0 seconds as the start and 10.0 seconds as the end.</think>\boxed{[0.0, 10.0]}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
Think: The previous final answer identifies the segment from 0.0 to 10.0 seconds as the interval containing the person walking in the doorway. After reviewing the selected boundaries, this interval is retained as correct.
Judge: True
Think: Reconsidering the original description, the relevant segment is the portion in which the person walks within or through the doorway. The selected interval remains 0.0 to 10.0 seconds.
Answer: [0.0, 10.0]
Ground-Truth: [2.2, 7.8]

Figure 10: Visualization of Two-turn Question-Answers of Charades-STA Benchmark (Temporal Reasoning).



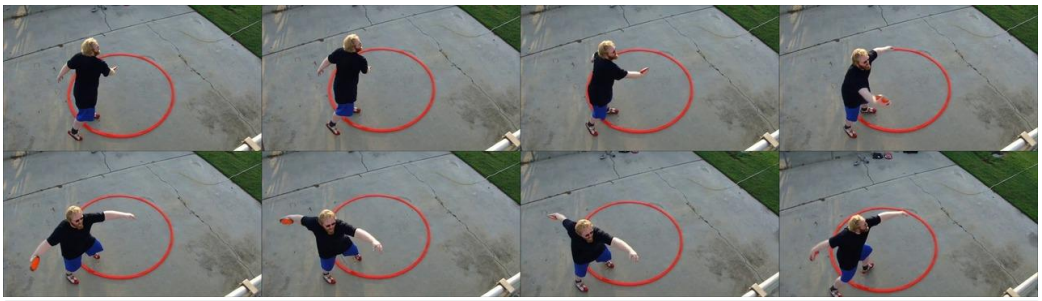
**Turn 1 Input:**
Question: why does the lady lifts the golf club and doesnt hit twice before hitting
Options: A. practice motion B. to play again C. waiting for man D. looking at boy E. talking to man beside her
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
Please answer the question, and then provide the video segment that best supports your answer. The start and end timestamps must be in seconds with exactly one decimal place.
Output format: \boxed{answer <> [start, end]}. For example: \boxed{A <> [20.3, 30.8]}.

**Turn 1 Output:**
Think: <think>The lady pauses after raising the club instead of immediately striking the ball. During this interval, her attention appears to be directed toward the man beside her, so I initially interpret the pause as part of their interaction. The relevant event occurs within the opening segment.</think>\boxed{E <> [0.0, 10.0]}<think>Reviewing the same interval, the repeated lifts without contact appear to coincide with her interaction with the nearby man. I therefore retain option E, "talking to man beside her," supported by the segment from 0.0 to 10.0 seconds.</think>\boxed{E <> [0.0, 10.0]}
Answer: E <> [0.0, 10.0]

**Turn 2 Input:**
**Question:**
Question: why does the lady lifts the golf club and doesnt hit twice before hitting
Options: A. practice motion B. to play again C. waiting for man D. looking at boy E. talking to man beside her
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
Please answer the question, and then provide the video segment that best supports your answer. The start and end timestamps must be in seconds with exactly one decimal place.
Output format: \boxed{answer <> [start, end]}. For example: \boxed{A <> [20.3, 30.8]}.
**Assistance:**
<think>The lady pauses after raising the club instead of immediately striking the ball. During this interval, her attention appears to be directed toward the man beside her, so I initially interpret the pause as part of their interaction. The relevant event occurs within the opening segment.</think>\boxed{E <> [0.0, 10.0]}<think>Reviewing the same interval, the repeated lifts without contact appear to coincide with her interaction with the nearby man. I therefore retain option E, "talking to man beside her," supported by the segment from 0.0 to 10.0 seconds.</think>\boxed{E <> [0.0, 10.0]}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
Think: The previous answer interpreted the pauses as evidence that the lady was talking to the man beside her. On closer review, however, the repeated club movements themselves are more directly relevant: she rehearses the swing without hitting the ball and then performs the actual shot. The previous answer is therefore incorrect.
Judge: False
Think: The two non-contact swings before the final hit are best understood as practice motions. The opening segment from 0.0 to 10.0 seconds contains this sequence and provides the strongest support for option A.
Answer: A <> [0.0, 10.0]
Ground-Truth: A <> [0.0, 7.7]

Figure 11: Visualization of Two-turn Question-Answers of NEXT-GQA Benchmark (Temporal Reasoning).



**Turn 1 Input:**
Locate the start and end timestamps of the video segment corresponding to the description: He throws a disc across the yard.You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question. Please provide the start and end timestamps (in seconds, precise to one decimal places) in the format\boxed{[start, end]}.

**Turn 1 Output:**
**Think:** <think>The target segment should cover the complete visible sequence associated with the man throwing a disc across the yard, including the immediate lead-up to and completion of the throw. Based on the estimated temporal boundaries, this sequence is placed between 5.0 and 35.0 seconds.</think>\boxed{[5.0, 35.0]}<think>Reviewing the selected interval, 5.0 seconds is used as the estimated beginning of the relevant sequence, while 35.0 seconds is used as its estimated endpoint. Therefore, the final timestamp range remains [5.0, 35.0].</think>\boxed{[5.0, 35.0]}
**Answer:** [5.0, 35.0]

**Turn 2 Input:**
**Question:**
Locate the start and end timestamps of the video segment corresponding to the description: He throws a disc across the yard. You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question. Please provide the start and end timestamps (in seconds, precise to one decimal places) in the format \boxed{[start, end]}.
**Assistance:**
<think>The target segment should cover the complete visible sequence associated with the man throwing a disc across the yard, including the immediate lead-up to and completion of the throw. Based on the estimated temporal boundaries, this sequence is placed between 5.0 and 35.0 seconds.</think>\boxed{[5.0, 35.0]}<think>Reviewing the selected interval, 5.0 seconds is used as the estimated beginning of therelevant sequence, while 35.0 seconds is used as its estimated endpoint. Therefore, the final timestamp range remains [5.0, 35.0].</think>\boxed{[5.0, 35.0]}
**User:**
Now verify your previous response to the original question. First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field. Next reason through the original question again and output the final answer in a second boxed field. You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
**Think:** The previous final answer, [5.0, 35.0], appears to be too broad and may not accurately isolate the relevant action. It should therefore not be accepted as correct.
**Judge:** False
**Think:** On re-evaluation, the described action is better localized near the beginning of the video. The revised segment starts at 0.0 seconds and ends at approximately 15.0 seconds.
**Answer:** [0.0, 15.0]
**Ground-Truth:** [0.0, 17.0]

Figure 12: Visualization of Two-turn Question-Answers of ActivityNet Benchmark (Temporal Reasoning).



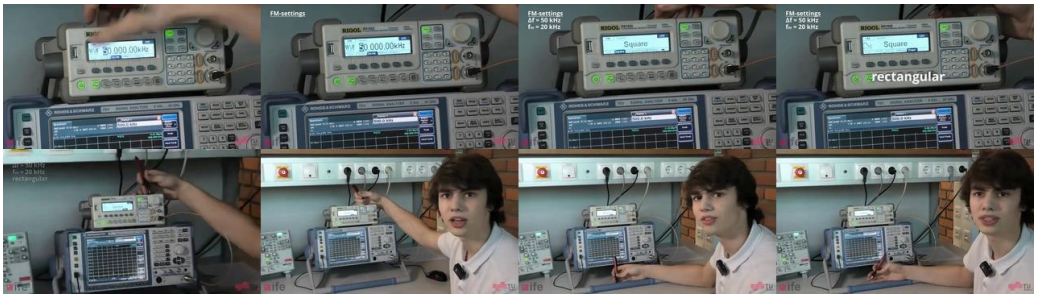
**Turn 1 Input:**
Is the tree increasing its height?
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Turn 1 Output:**
**Think:**
<think>I compare the tree's height across the observed sequence, using stable elements in the scene as reference points. I do not initially detect a clear upward change in its topmost point.</think>\boxed{No}<think>After reviewing the sequence again, the tree appears to remain approximately the same height within the observed interval. I therefore retain "No" as the final answer.</think>\boxed{No}
**Answer:** No

**Turn 2 Input:**
**Question:**
Is the tree increasing its height?
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
**Assistance:**
<think>I compare the tree's height across the observed sequence, using stable elements in the scene as reference points. I do not initially detect a clear upward change in its topmost point.</think>\boxed{No}<think>After reviewing the sequence again, the tree appears to remain approximately the same height within the observed interval. I therefore retain "No" as the final answer.</think>\boxed{No}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False ina boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
**Think:** The previous answer stated that the tree was not increasing in height. On closer comparison of its height across the sequence, an upward change is visible, so the previous answer is incorrect.
**Judge:** False
**Think:** Re-evaluating the tree relative to stable reference points in the scene, its topmost point rises over time. Therefore, the tree is increasing its height.
**Answer:** Yes
**Ground-Truth:** Yes

Figure 13: Visualization of Two-turn Question-Answers of TempCompass Benchmark (Temporal Reasoning).



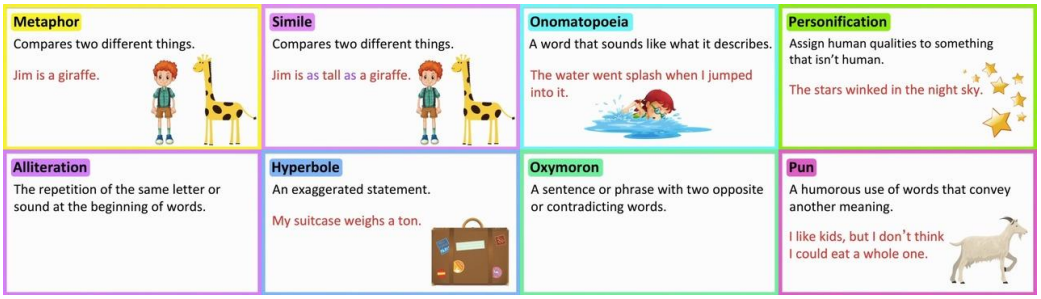

**Turn 1 Input:**
Question: According to the display on the spectrum analyzer, how is the modulated signal distributed?
Options: A. Poisson B. Exponential C. Gaussian D. Periodical E. Binomial
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Turn 1 Output:**
**Think:** <think>The question asks about the distribution indicated by the spectrum-analyzer display. Based on my initial interpretation of the displayed spectral shape, it appears to follow a bell-shaped profile, which corresponds to a Gaussian distribution.</think>\boxed{C}<think>Comparing this interpretation with the available options, Gaussian is the closest match to the apparent distribution. I therefore retain option C as the final answer.</think>\boxed{C}
**Answer:** C

**Turn 2 Input:**
**Question:**
Question: According to the display on the spectrum analyzer, how is the modulated signal distributed?
Options: A. Poisson B. Exponential C. Gaussian D. Periodical E. Binomial
You must respond in exactly this format:
<think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer}
Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
**Assistance:**
<think>The question asks about the distribution indicated by the spectrum-analyzer display. Based on my initial interpretation of the displayed spectral shape, it appears to follow a bell-shaped profile, which corresponds to a Gaussian distribution.</think>\boxed{C}<think>Comparing this interpretation with the available options, Gaussian is the closest match to the apparent distribution. I therefore retain option C as the final answer.</think>\boxed{C}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
**Think:** The previous final answer was C, Gaussian. On closer inspection, the display is better characterized by a recurring spectral pattern than by a bell-shaped distribution, so the previous answer should be rejected.
**Judge:** False
**Think:** Re-evaluating the displayed pattern and comparing it with the choices, the modulated signal appears to be distributed periodically. This corresponds to option D, Periodical.
**Answer:** D
**Ground-Truth:** D

Figure 14: Visualization of Two-turn Question-Answers of MMVU Benchmark (Video Reasoning).




**Turn 1 Input:**

Read the following four statements and decide which one is correct based on the video content. 1. The video explains the concept of "metaphor" using the example of a giraffe. 2. The video explains the concept of "simile" using the example of a rollercoaster. 3. It can be inferred from the video that a simile compares two things using "like" or "as." 4. The video implies that metaphor compares two things without using "like" or "as."
Options: A. Statement 1 is correct B. Statement 1 and 2 are correct C. Statement 3 is correct D. Statement 1, 2, and 3 are correct E. All statements are correct F. Statement 1 and 3 are correct G. Statement 1, 2, and 4 are correct H. Statement 1 and 4 are correct I. Statement 1, 3, and 4 are correct J. All statements are false
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Turn 1 Output:**

Think: <think>On an initial review, the video appears to use a giraffe when explaining metaphor, supporting statement 1. It also indcates that a simile uses "like" or "as," supporting statement 3. The rollercoaster example does not clearly support statement 2, and statement 4 is not sufficiently established in this initial assessment. Therefore, the matching option is F.</think>\boxed{F}<think>Comparing the statements with the answer choices, the initially supported combination is statements 1 and 3. This combination corresponds to option F, so I retain F as the final answer.</think>\boxed{F}
Answer: F

**Turn 2 Input:**

**Question:**

Read the following four statements and decide which one is correct based on the video content. 1. The video explains the concept of "metaphor" using the example of a giraffe. 2. The video explains the concept of "simile" using the example of a rollercoaster. 3. It can be inferred from the video that a simile compares two things using "like" or "as." 4. The video implies that metaphor compares two things without using "like" or "as."
Options: A. Statement 1 is correct B. Statement 1 and 2 are correct C. Statement 3 is correct D. Statement 1, 2, and 3 are correct E. All statements are correct F. Statement 1 and 3 are correct G. Statement 1, 2, and 4 are correct H. Statement 1 and 4 are correct I. Statement 1, 3, and 4 are correct J. All statements are false
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Assistance:**

<think>On an initial review, the video appears to use a giraffe when explaining metaphor, supporting statement 1. It also indicates that a simile uses "like" or "as," supporting statement 3. The rollercoaster example does not clearly support statement 2, and statement 4 is not sufficiently established in this initial assessment. Therefore, the matching option is F.</think>\boxed{F}<think>Comparing the statements with the answer choices, the initially supported combination is statements 1 and 3. This combination corresponds to option F, so I retain F as the final answer.</think>\boxed{F}

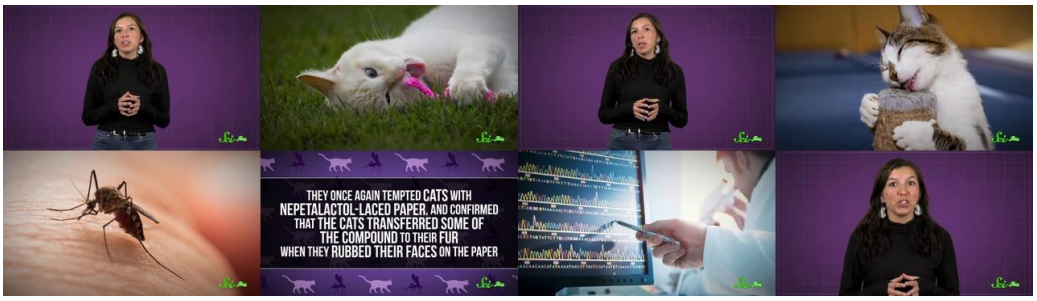
**User:**

Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**

Think: The previous final answer, F, includes only statements 1 and 3. On closer review, statement 4 is also supported because the video distinguishes metaphor from simile by indicating that metaphor makes a comparison without "like" or "as." The previous answer is therefore incomplete and incorrect.
Judge: False
Think: Re-evaluating all four statements, statements 1, 3, and 4 are supported, while statement 2 is not. The option containing exactlystatements 1, 3, and 4 is I.
Answer: I
Ground-Truth: I

Figure 15: Visualization of Two-turn Question-Answers of VideoMMMU Benchmark (Video Reasoning).




**Turn 1 Input:**

There is a lioness inside an iron mesh enclosure. Outside the iron mesh, someone is extending a pair of tongs through the mesh towards the lioness's nose. What is being clamped by the tongs that are reaching towards the lioness's nose?
Options: A. A flower B. Catnip C. A piece of meat D. A leaf E. A chicken
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Turn 1 Output:**

**Think:** <think>On an initial inspection, the object held by the tongs appears more consistent with a plant item than with an entire chicken or another large object. Among the listed options, I initially identify it as a leaf.</think>\boxed{D}<think>After reconsidering the available choices, I interpret the object as a small piece of food being presented to the lioness. Of the options given, a piece of meat is the closest match.</think>\boxed{C}
**Answer:** C

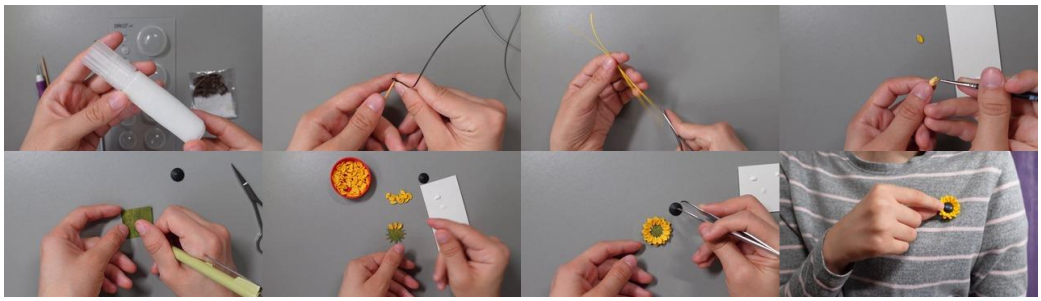
**Turn 2 Input:**

**Question:**
There is a lioness inside an iron mesh enclosure. Outside the iron mesh, someone is extending a pair of tongs through the mesh towards the lioness's nose. What is being clamped by the tongs that are reaching towards the lioness's nose?
Options: A. A flower B. Catnip C. A piece of meat D. A leaf E. A chicken
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
**Assistance:**
<think>On an initial inspection, the object held by the tongs appears more consistent with a plant item than with an entire chicken or another large object. Among the listed options, I initially identify it as a leaf.</think>\boxed{D}<think>After reconsidering the available choices, I interpret the object as a small piece of food being presented to the lioness. Of the options given, a piece of meat is the closest match.</think>\boxed{C}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**

**Think:** The previous final answer was C, a piece of meat. On closer review, that identification is not reliable: the object should not be classified merely from the assumption that meat is commonly fed to lions. The previous answer is therefore rejected.
**Judge:** False
**Think:** Re-evaluating the object itself and comparing it with the listed choices, the item held by the tongs is best identified as catnip. Therefore, the revised answer is B.
**Answer:** B
**Ground-Truth:** B

Figure 16: Visualization of Two-turn Question-Answers of LongVideoBench Benchmark (Long Video Understanding).



**Turn 1 Input:**
The sequence of handling the colors of the paper in the video
(A) Black, Yellow, Yellow, Black, Green (B) Black, Green, Yellow, Yellow, Black (C) Black, Yellow, Green, Yellow, Black (D) Yellow, Green, Black, Yellow, Black (E) Black, Yellow, Yellow, Green, Black (F) Yellow, Black, Green, Yellow, Black
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Turn 1 Output:**
**Think:** <think>I track each paper at the moment it is handled and record the colors in chronological order. My initial sequence is Black, Green, Yellow, Yellow, Black, which corresponds to option B.</think>\boxed{B}<think>Reviewing the five recorded positions in order, I retain the sequence Black, Green, Yellow, Yellow, Black. Therefore, the final selection remains option B.</think>\boxed{B}
**Answer:** B

**Turn 2 Input:**
**Question:**
The sequence of handling the colors of the paper in the video
(A) Black, Yellow, Yellow, Black, Green (B) Black, Green, Yellow, Yellow, Black (C) Black, Yellow, Green, Yellow, Black (D) Yellow, Green, Black, Yellow, Black (E) Black, Yellow, Yellow, Green, Black (F) Yellow, Black, Green, Yellow, Black
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
**Assistance:**
<think>I track each paper at the moment it is handled and record the colors in chronological order. My initial sequence is Black, Green, Yellow, Yellow, Black, which corresponds to option B.</think>\boxed{B}<think>Reviewing the five recorded positions in order, I retain the sequence Black, Green, Yellow, Yellow, Black. Therefore, the final selection remains option B.</think>\boxed{B}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
**Think:** The previous final answer was option B: Black, Green, Yellow, Yellow, Black. After reviewing the order again, the second paper was misidentified, so the previous sequence is not correct.
**Judge:** False
**Think:** Rechecking the papers chronologically gives Black first, followed by Yellow, Green, Yellow, and Black. This sequence matchesoption C.
**Answer:** C
**Ground-Truth:** C

Figure 17: Visualization of Two-turn Question-Answers of MLVU Benchmark (Long Video Understanding).



**Turn 1 Input:**
The vlogger spend ___ yen to buy water from the vending machine for the second time.
(A) 150 (B) 130 (C) 110 (D) 100
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.

**Turn 1 Output:**
**Think:** <think>The question asks specifically about the second purchase, so the relevant amount must be taken from that transaction rather than inferred from typical vending-machine prices. My initial reading suggests option D, corresponding to 100 yen.</think>\boxed{D}<think>After distinguishing the second purchase from the other transaction shown, I revise the selection to option A. This option corresponds to 150 yen.</think>\boxed{A}
**Answer:** A

**Turn 2 Input:**
**Question:**
The vlogger spend ___ yen to buy water from the vending machine for the second time.
(A) 150 (B) 130 (C) 110 (D) 100
You must respond in exactly this format: <think>your first reasoning</think>\boxed{your first answer}<think>your second reasoning</think>\boxed{your final answer} Both reasoning stages must be non-empty. The final boxed answer must be the final answer to the original question.
**Assistance:**
<think>The question asks specifically about the second purchase, so the relevant amount must be taken from that transaction rather than inferred from typical vending-machine prices. My initial reading suggests option D, corresponding to 100 yen.</think>\boxed{D}<think>After distinguishing the second purchase from the other transaction shown, I revise the selection to option A. This option corresponds to 150 yen.</think>\boxed{A}
**User:**
Now verify your previous response to the original question.
First reason about whether the second boxed answer in your previous response is correct, then output exactly True or False in a boxed field.
Next reason through the original question again and output the final answer in a second boxed field.
You must respond in exactly this format:
<think>verification_reasoning</think>\boxed{True or False}<think>reasoning</think>\boxed{final answer}
Both reasoning stages must be non-empty. The first boxed field must contain only True or False. The final boxed field must answer the original question.

**Turn 2 Output:**
**Think:** The previous final answer was option A, which corresponds to 150 yen. After rechecking the amount associated specifically with the second purchase, that selection does not appear correct.
**Judge:** False
**Think:** Re-evaluating the second vending-machine transaction and matching its amount to the choices, the appropriate selection is option C, corresponding to 110 yen.
**Answer:** C
**Ground-Truth:** C

Figure 18: Visualization of Two-turn Question-Answers of LVBench Benchmark (Long Video Understanding).