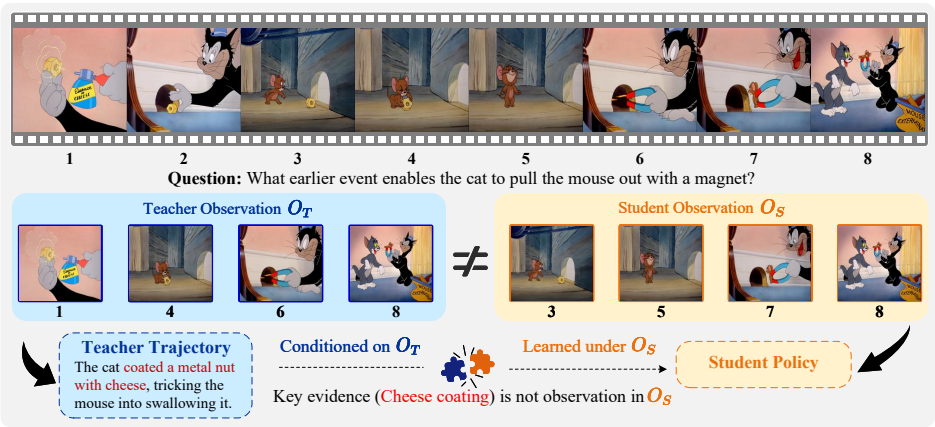
REASONING BEYOND IMITATION: OBSERVATION CONDITIONED DISTILLATION FOR VIDEO LLMS

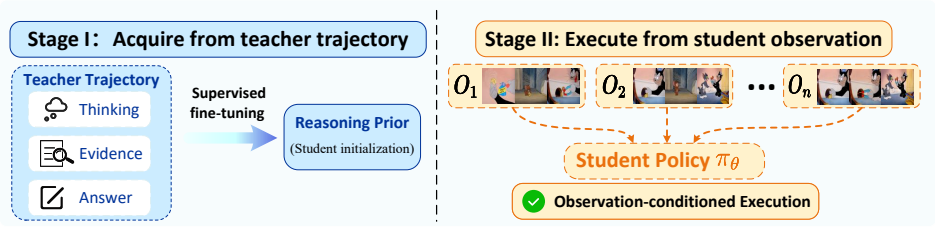
Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning distillation offers a promising approach to transfer temporal reasoning capabilities of powerful Video-LLMs to open-source student models. However, teacher and student Video-LLMs may construct different temporal observations from the same video due to the adoption of distinct video sampling pipelines. A reasoning trajectory valid under the teacher's observations may rely on evidence inaccessible to the student and thus cannot be directly executed based on the student's observations. We formalize this as a conditional target mismatch, thereby revealing the Reasoning Acquisition–Execution Gap. To bridge this gap, we propose an acquisition-to-execution framework that first distills teacher trajectories into a student reasoning prior and then further optimizes the execution capability of this prior across multiple observations. We further introduce Execution-Aware Reference Regularization, which retains the existing prior when it is executable under current observations, while allowing a greater degree of adaptive adjustment when it is not. Experiments across multiple Video-LLM backbones and six video reasoning benchmarks show that explicitly optimizing observation-conditioned execution consistently outperforms supervised distillation and standard RL-based optimization, with gains extending to held-out out-of-domain benchmarks.



(a) Teacher–student observation mismatch, making direct distillation not feasible.



(b) Overview of our cross-observation video reasoning distillation framework.

Figure 1: Motivation and framework for observation-conditioned distillation for Video-LLMs.

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.

1 INTRODUCTION

Recently, proprietary Video Large Language Models (Video-LLMs) (Anthropic, 2024; Comanici et al., 2025; Achiam et al., 2023) have demonstrated significant advantages in complex video temporal understanding and reasoning tasks (Cheng et al., 2025; Chen et al., 2025b; 2026), while open-source models still exhibit a substantial capability gap (Li et al., 2026a; Krojer et al., 2025). These strong teacher models are often accessible only through APIs, exposing only their outputs without providing model parameters, logits, or internal visual representations (Jiang et al., 2023). Therefore, black-box knowledge distillation (Hinton et al., 2015; Guo et al., 2025; Yan et al., 2025) based on generated reasoning trajectories emerges as a natural approach for capability transfer: reasoning trajectories from the teacher are used to supervise an open-source student, enabling it to acquire stronger video-reasoning capabilities. However, video scenarios introduce a challenge that differs fundamentally from text-based knowledge distillation: the teacher and the student may not be reasoning under the *same* visual conditions.

As illustrated in Fig. 1, the same video does not necessarily imply the same model observation. Video-LLMs typically process only a limited sequence of frames obtained via temporal sampling, and different models often vary in their frame budgets, sampling rates, temporal offsets, and frame locations. As a result, even when given the same video and question, the teacher and student may reason over *different temporal observations*. The key issue is that a teacher-generated trajectory is inherently conditioned on the teacher's observation, whereas rationale distillation typically uses it directly as the supervision target under the student's observation. Consequently, a reasoning path that is valid and effective given what the teacher observes may no longer remain valid or executable given what the student observes (Wang et al., 2025a; Bai et al., 2025; Achiam et al., 2023; Xu et al., 2025; Li et al., 2023). This implies that, in video reasoning distillation (Agarwal et al., 2024; Ko et al., 2024), the validity of the supervision target is itself observation-dependent, giving rise to a conditional target mismatch.

This observation-dependent mismatch further reveals a more fundamental limitation of rationale Supervised Fine-Tuning (SFT) (Fu et al., 2026; Ouyang et al., 2022). Although it can transfer teachers' reasoning trajectories to students, it does not directly address how to select and implement these behaviors based on the students' actual observations. Consequently, acquiring teacher-demonstrated reasoning patterns does not necessarily translate into reliable observation-conditioned execution. Such discrepancy can be referred to as the *Reasoning Acquisition–Execution Gap*. This perspective reframes the problem from further improving teacher rationales to a different objective: transforming the distilled reasoning prior into reasoning capability that can be reliably executed from the student's own observations.

To address the above problem, we decompose video reasoning distillation into two stages: *reasoning acquisition* and *observation-conditioned* execution. In the first stage, the student is supervised with teacher-generated trajectories through SFT, yielding a reasoning prior. As shown in our ablation, this prior remains useful for subsequent execution optimization even when SFT alone does not improve task performance. In Stage II, the student performs outcome-driven optimization over multiple temporal observations sampled from its own input stream, learning to execute reasoning behaviors conditioned on what it actually observes. Since the Stage-I prior can exhibit varying levels of executability across observations, we further introduce Execution-Aware Reference Regularization, which adaptively adjusts the reference constraint based on the prior's executability in the current observation, balancing the preservation of already executable reasoning behaviors with the discovery of observation-specific alternative reasoning strategies.

Our main contributions are as follows:

- We identify a conditional target mismatch induced by heterogeneous temporal observations and formulate the Reasoning Acquisition–Execution Gap to characterize the discrepancy between reasoning acquisition and observation-conditioned execution.
- We propose a two-stage acquisition-to-execution framework, where teacher-guided SFT is used to acquire a reasoning prior, followed by student-side Multi-Observation GRPO (MO-GRPO) to learn its observation-conditioned execution.
- We introduce Execution-Aware Reference Regularization, which adaptively adjusts the reference constraint based on the observation-wise executability of the reasoning prior, thereby balancing prior preservation and alternative reasoning discovery.

2 PROBLEM FORMULATION

Given a video-question instance (V, q, a^*), where V is the raw video, q the question, and a^* the ground-truth answer, a Video-LLM reasons over a finite temporal observation rather than the complete video. We denote the teacher and student observations as $O_T = S_T(V; \xi_T)$ and $O_S = S_S(V; \xi_S)$, where ξ specifies the sampling configuration, such as frame budget, sampling rate, temporal offset, and frame locations. Accordingly, the teacher and student may operate under different observation distributions, $p_T(O \mid V)$ and $p_S(O \mid V)$, and thus observe different evidence from the same video: $O_T \neq O_S$.

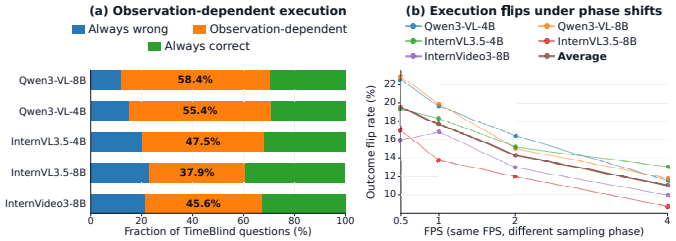


Figure 2: **Video-LLM execution varies across temporal observations on TimeBlind.** (a) Questions are categorized as always incorrect, observation-dependent, or always correct across eight sampling configurations. (b) Changing only the temporal sampling phase, while keeping the frame rate fixed, still causes substantial outcome flips across models.

Crucially, reasoning outcomes are not invariant to temporal observation. Let $c_\pi(O, q) = \mathbb{I}[\mathrm{Eval}(\hat{a}_\pi(O, q), a^*) = 1]$ indicate whether policy π answers correctly under observation O. If temporal sampling only changed input representation, we would expect $c_\pi(O_1, q) = c_\pi(O_2, q)$ for different observations of the same video. However, Fig. 2 shows that many TimeBlind (Li et al., 2026a) questions are solved under only a subset of observations; the experimental details are provided in Section D.5. Prediction flips remain frequent even when the frame rate is fixed and only the sampling phase changes, indicating that execution depends on *which* temporal evidence is observed, not merely on the number of sampled frames.

This observation dependence creates a mismatch in reasoning distillation. A teacher-generated trajectory is conditioned on the teacher observation, $y_T \sim p_T(y \mid O_T, q)$, whereas direct rationale distillation optimizes the student to reproduce that trajectory under a potentially different observation:

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(V,q,a^*)\sim\mathcal{D}}\mathbb{E}_{O_T,O_S}\left[\log \pi_\theta(y_T \mid O_S, q)\right], \qquad y_T \sim p_T(\cdot \mid O_T, q). \tag{1}$$

The teacher trajectory may be fully valid under O_T and still contain useful reasoning structure, yet its validity as a direct execution target is not guaranteed once the conditioning observation changes. We refer to this discrepancy between the context under which supervision is generated and the context under which it is learned as **conditional target mismatch**.

At deployment, however, the student should maximize task performance over observations induced by its own sampler:

$$J_{\mathrm{exec}}(\pi) = \mathbb{E}_{(V,q,a^*)\sim\mathcal{D}}\mathbb{E}_{O\sim p_S(O|V)}\mathbb{E}_{y\sim\pi(\cdot|O,q)}\left[R(y, a^*)\right], \tag{2}$$

where $R(y, a^*)$ evaluates whether the generated trajectory leads to the correct task outcome. Thus, rationale SFT transfers teacher-demonstrated reasoning behaviors, whereas successful execution requires selecting and realizing reasoning that is effective under the student's actual observation. Minimizing imitation loss therefore does not necessarily optimize student-side execution.

We call this discrepancy the **Reasoning Acquisition–Execution Gap**. Reasoning *acquisition* learns reusable reasoning patterns from teacher demonstrations, while observation-conditioned *execution* adapts those patterns to the evidence available in the current student observation. This distinction is particularly important for video reasoning, where different temporal observations may expose different events and cues and therefore support different valid reasoning paths.

Accordingly, we denote the model obtained after rationale SFT as π_{ref} and treat it as a **reasoning prior**: it captures reasoning behaviors acquired from the teacher without assuming that they are optimally executable under every student observation. Closing the acquisition–execution gap therefore requires further optimizing reasoning directly over student-side observations while preserving useful teacher-acquired structure. This motivates our acquisition-to-execution framework in Section 4.

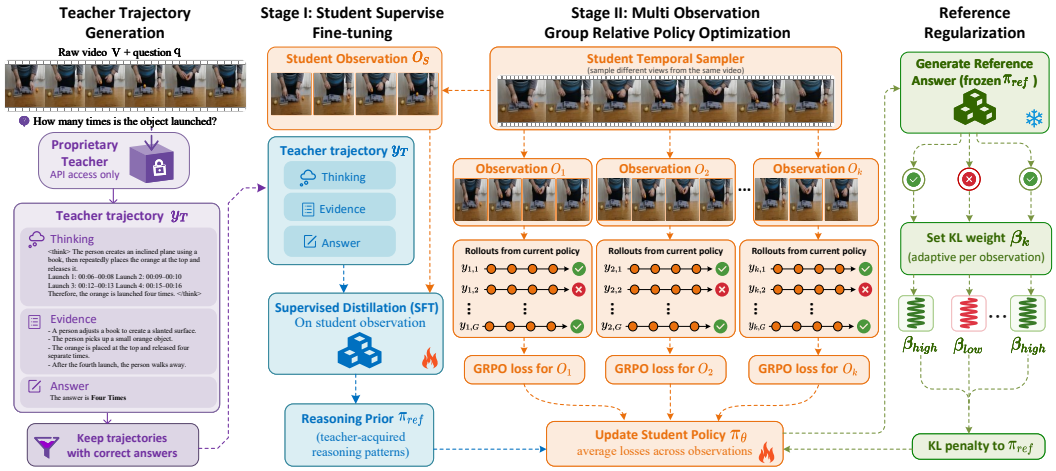


Figure 3: **Overview of our acquisition-to-execution framework.** Stage I uses SFT on student observations to distill black-box teacher trajectories with correct answers into a reasoning prior. Stage II computes GRPO losses separately for multiple temporal observations of the same video and aggregates them to update a shared student policy. Reference regularization uses a larger KL weight when the frozen Stage-I model answers correctly under the current observation and a smaller weight otherwise.

3 Related Work

Video Reasoning in Video-LLMs Recently, the ability of Video-LLMs (Team et al., 2025; Chen et al., 2024; Li et al., 2024a) to perform complex video reasoning has been significantly enhanced, particularly in determining temporal sequences, analyzing event causality, tracking changes in object states, and understanding spatial relationships (Kim et al., 2025; Cheng et al., 2025; Kim et al., 2025; Yuan et al., 2025). By breaking down video into timestamped frames and employing progressive multimodal training (An et al., 2026; Ren et al., 2024; Feng et al., 2025; Wang et al., 2025a;b), the model is better equipped to handle video-related tasks involving localization and event sequencing. As video complexity increases, Chain-of-Thought (Fei et al., 2024; Zhang et al., 2025; Han et al., 2025; Chen & Lu, 2024) breaks down complex video tasks into multiple perceptual and cognitive steps, enabling the model to arrive at the final answer progressively through an intermediate reasoning process. Meanwhile, multiple video reasoning benchmarks Li et al. (2024b); Shangguan et al. (2025); Cheng et al. (2026); Qi et al. (2025); Upadhyay et al. (2026); Zhu et al. (2026) indicate that high accuracy on traditional VideoQA tasks does not equate to reliable temporal reasoning capabilities. These studies demonstrate that, although current Video-LLMs possess strong video understanding capabilities, a significant gap in reasoning performance remains between proprietary and open-source models. Consequently, leveraging knowledge distillation to transfer the video reasoning capabilities of a powerful teacher model to a lightweight student model has emerged as an attractive research direction (Guo et al., 2025; Shi et al., 2025; Lu et al., 2025; Panigrahi et al., 2026).

Knowledge Distillation Knowledge distillation (Gou et al., 2021; Gu et al., 2024; Yang et al., 2025; Zhang et al., 2026b) primarily transfers knowledge from large models to smaller ones via prediction distributions or intermediate representations. Recently, Gradually extended the scope of distillation to the model's explicit reasoning process (Li et al., 2023; Lu et al., 2025). Reasoning distillation (Hsieh et al., 2023; Wang et al., 2023; Zhang et al., 2026b; Shridhar et al., 2023; Wang et al., 2026) leverages rationales or CoT sequences generated by a teacher model to provide richer intermediate supervision, demonstrating the significant impact of explicit reasoning supervision on the performance of smaller models. Offline teacher trajectories can lead to distribution discrepancies between the teacher and the student. Research (Gu et al., 2024; Ko et al., 2024; Agarwal et al., 2024; Ding et al., 2026) has shifted the distillation process toward student-generated responses and integrated it with reinforcement learning. This concept has been extended to multimodal scenarios, incorporating on-policy supervision and reinforcement learning into multimodal knowledge distil-

lation to enhance visual grounding and reasoning capabilities (Xu et al., 2025; Shi et al., 2025; Luo et al., 2026). However, we observe an *observation mismatch* between teacher supervision and student learning. A correct reasoning trajectory from the teacher may not be directly grounded in the evidence visible to the student, resulting in suboptimal knowledge distillation performance.

4 METHOD

To address the mentioned phenomenon described in Section 2, we propose a video reasoning distillation framework as illustrated in Fig. 3. First, we set a proprietary video-LLM as a black-box teacher to generate video reasoning trajectories. In stage I (Section 4.1), the student model is supervised fine-turning on these trajectories to acquire the teacher's reasoning patterns. Stage II (Section 4.2), we construct multiple observation states and perform group relative policy optimization across these observations. Moreover, since the Stage I prior can be executable under some observations but ineffective under others, we further introduce Execution-Aware Reference Regularization (Section 4.3), which adjusts the strength of the reference constraint separately for each observation.

4.1 STAGE I: TEACHER TRAJECTORY DISTILLATION

Black-box teacher trajectory generation. Given a training sample (V_i, q_i, a_i^*), where V_i denotes the raw video, q_i the question, and a_i^* corresponding the ground truth answer. The teacher obtains temporal observations using its own video sampler: $O_i^T = S_T(V_i, \xi_i^T)$. ξ_i^T is the corresponding sampling configuration. Given O_i^T and q_i, the teacher generates a structured reasoning trajectory

$$y_i^T = \left(T_i^T, E_i^T, \hat{a}_i^T\right) \sim p_T\left(\cdot \mid O_i^T, q_i\right). \tag{3}$$

T_i^T denotes the thinking process generated by the teacher, E_i^T represents the intermediate visual evidence referenced during reasoning, and $\hat{a}_i^T$ is the final predicted answer. To avoid misleading students with erroneous trajectories, we retain only those trajectories that yield valid answers—specifically, those satisfying $\mathrm{Eval}\left(\hat{a}_i^T, a_i^*\right) = 1$.

Student Supervised Fine-tuning. Given the same video, the student uses its own video sampler to obtain observation $O_i^S = S_S(V_i, \xi_i^S)$. Since the teacher and student may use different sampling strategies, O_i^S is not equal to O_i^T. We optimize the rationale-distillation objective in Eq. (1) over the retained correct teacher trajectories, instantiated with standard autoregressive token-level supervision. Importantly, filtering by answer correctness only ensures that a teacher trajectory is valid under O_i^T; it does not guarantee that the same trajectory is directly executable from O_i^S. Stage I therefore deliberately uses the teacher trajectory as an *acquisition signal*, despite the observation mismatch formalized in Section 2. The objective of this stage is therefore not to establish y_i^T as the unique reasoning path that should be reproduced under every student observation. Instead, teacher trajectories expose the student to structured problem decomposition, temporal evidence organization, and multi-step inference patterns that may be reusable beyond the specific observation from which they were generated. After Stage I, we use the resulting model to initialize Stage II and freeze an identical copy as the reference policy, denoted as π_{ref}.

4.2 STAGE II: MULTI-OBSERVATION GROUP RELATIVE POLICY OPTIMIZATION

Stage II further optimizes the student under its own video sampling distribution. For each training example (V_i, q_i, a_i^*), we construct K temporal observations using the student sampler, denoted as $O_{i,k} = S_S(V_i; \xi_{i,k})$, where $k = 1, \ldots, K$. Different $\xi_{i,k}$ correspond to different video sampling configurations, such as different sampling phases or frame locations.

These observations correspond to the same video-question pair, but may expose different visual evidence to the model due to their different video sampling positions. For each $O_{i,k}$, the behavior policy $\pi_{\theta_{\mathrm{old}}}$ independently samples G reasoning trajectories $y_{i,k,g}$. We do not require consistent reasoning trajectories across different observations, since different temporal views may support distinct yet equally valid reasoning paths. For each rollout $y_{i,k,g} = (R_{i,k,g}, E_{i,k,g}, \hat{a}_{i,k,g})$, we compute the reward based on the final answer and output format:

$$r_{i,k,g} = \lambda_{\mathrm{ans}} R_{\mathrm{ans}}(\hat{a}_{i,k,g}, a_i^*) + \lambda_{\mathrm{fmt}} R_{\mathrm{fmt}}(y_{i,k,g}). \tag{4}$$

R_{ans} provides the primary optimization signal based on the correctness of the final answer, while R_{fmt} maintains the structured reasoning–evidence–answer output format used in Stage I.

Observation-Local Relative Optimization. For each temporal observation $O_{i,k}$, we perform relative comparison only among the G trajectories generated under the same observation. Specifically, we compute the reward mean and standard deviation within each group: $\mu_{i,k} = \frac{1}{G}\sum_{g=1}^{G} r_{i,k,g}$, then $\sigma_{i,k} = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(r_{i,k,g} - \mu_{i,k}\right)^2}$, and define the corresponding group-relative advantage as $A_{i,k,g} = \frac{r_{i,k,g} - \mu_{i,k}}{\sigma_{i,k} + \epsilon}$. Unlike normalizing all rollouts from the same video-question pair together, we keep the rollout groups of different temporal observations independent. Since different observations may contain different task-relevant visual evidence, comparing rewards across observations would conflate the variation caused by the observation itself with the quality differences among reasoning trajectories. As a result, a successful trajectory from $O_{i,k'}$ does not reduce the relative advantage of trajectories generated from $O_{i,k}$. When all G trajectories under the same observation receive identical rewards, we set the advantages of that group to zero.

We then define the token-level importance ratio as

$$\rho_{i,k,g,t}(\theta) = \frac{\pi_\theta\left(y_{i,k,g,t} \mid O_{i,k}, q_i, y_{i,k,g,<t}\right)}{\pi_{\theta_{\text{old}}}\left(y_{i,k,g,t} \mid O_{i,k}, q_i, y_{i,k,g,<t}\right)}. \tag{5}$$

For each observation group, we independently apply the clipped Group Relative Policy Optimization (GRPO) (Guo et al., 2025) objective:

$$\mathcal{L}_{i,k}^{\text{GRPO}} = -\mathbb{E}_{g,t}\left[\min\left(\rho_{i,k,g,t}A_{i,k,g}, \text{clip}\left(\rho_{i,k,g,t}, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}}\right)A_{i,k,g}\right)\right]. \tag{6}$$

The GRPO losses from all K temporal observations are jointly used to update the same student model. In this way, multiple video views of the same video-question pair can all provide training signals for policy optimization, while each reasoning trajectory is compared only against alternatives generated from the same visual observation.

4.3 Execution-Aware Reference Regularization

Stage II uses the Stage-I model as a frozen reference policy π_θ to prevent the student from drifting excessively from the reasoning behaviors learned during supervised fine-tuning. However, different temporal observations may expose different visual evidence, and the same reference policy can therefore perform differently across observations. It may already answer the question correctly under one observation while failing under another. Using the same KL coefficient for all observations is thus suboptimal. When the reference policy succeeds under the current observation, we would like to better preserve its behavior; when it fails, the constraint should be relaxed to allow the student to explore alternative reasoning trajectories.

For each temporal observation $O_{i,k}$, we first obtain a deterministic response from the frozen reference policy. We then determine whether the reference policy correctly answers the question under this observation: $e_{i,k} = \mathbb{I}\left[\text{Eval}\left(\hat{a}_{i,k}^{\text{ref}}, a_i^*\right) = 1\right]$. Based on this result, we assign an observation-specific KL coefficient: $\beta_{i,k} = \beta_{\text{low}} + e_{i,k}\left(\beta_{\text{high}} - \beta_{\text{low}}\right)$, where $0 \le \beta_{\text{low}} < \beta_{\text{high}}$. When $e_{i,k} = 1$, we use the larger coefficient β_{high} to better preserve the reference behavior that already succeeds under $O_{i,k}$. When $e_{i,k} = 0$, we instead use β_{low}, allowing the policy to deviate further from the reference and explore alternative reasoning trajectories under the current observation.

Let $\mathcal{L}_{i,k}^{\text{KL}}(\pi_\theta, \pi_{\text{ref}})$ denote the reference regularization term for observation $O_{i,k}$. The complete Stage-II objective is

$$\mathcal{L}_{\text{StageII}} = \mathbb{E}_i\left[\frac{1}{K}\sum_{k=1}^{K}\left(\mathcal{L}_{i,k}^{\text{GRPO}} + \beta_{i,k}\mathcal{L}_{i,k}^{\text{KL}}\right)\right]. \tag{7}$$

Stage II performs GRPO optimization independently for each temporal observation, while adapting the reference constraint based on the reference policy's performance at that observation.



Figure 4: **Main comparison across two student reasoning regimes and six video reasoning benchmarks.** We compare SFT, GRPO, LUFFY, T-GRPO, and our method on Qwen3-VL-4B-Thinking and InternVL3.5-4B. Each point shows the score change from the pretrained baseline, where positive values indicate improvement. Shaded regions denote the best method for each backbone. Our method achieves the largest gain in 11 of 12 backbone-benchmark settings.

5 Experiments

5.1 Experiment Setup

Training and Evaluation Benchmarks. To comprehensively evaluate the effectiveness and generalization of our method, we conduct experiments on six video reasoning benchmarks, including three *in-domain* benchmarks: TimeBlind (Li et al., 2026a), MVBench (Li et al., 2024b), and TempCompass (Liu et al., 2024); and three *out-of-domain* benchmarks: TOMATO (Shangguan et al., 2025), Video-MME (Fu et al., 2025), and VideoReasonBench (Liu et al., 2026). More details are provided in Section C.

Implementation Details. We use Gemini 2.5 Pro (Comanici et al., 2025) as the black-box teacher and Qwen3VL-4B, 8B (Bai et al., 2025) and InternVL-3.5-4B, 8B (Wang et al., 2025a) as student models. All students are fine-tuned with LoRA ($r = 32$, $\alpha = 64$) in BF16. Stage I and Stage II are trained for 2 and 1 epochs at learning rates of 1×10^{-5} and 1×10^{-6}, respectively. Full training hyperparameters and observation sampling configurations are in Section D. All RL-based methods are evaluated under the same compute budget. We compare against standard SFT/GRPO pipelines and representative reasoning-distillation baselines; details are provided in Section D.3.

5.2 Main Result

Overall Comparison. Fig. 4 compares our method with direct SFT, standard GRPO (Guo et al., 2025), LUFFY (Yan et al., 2025), and T-GRPO (Feng et al., 2025) on Qwen3-VL-4B-Thinking (Bai et al., 2025) and InternVL3.5-4B (Wang et al., 2025a) across six video reasoning benchmarks. Our method achieves the strongest post-training performance across all twelve backbone–benchmark settings, and surpasses the corresponding pretrained baseline in eleven of them. In contrast, the competing methods show substantially less consistent behavior across backbones and tasks: improvements on one setting often do not transfer to another. This consistency is particularly notable given the two students' different reasoning behaviors, suggesting that the benefit of our method is not tied to a specific model family or reasoning initialization. The same trend is reproduced with the open-source Qwen3-VL-235B-A22B-Thinking teacher as shown in Section E.3.

Direct Trajectory Distillation. Direct teacher-trajectory SFT exhibits markedly different behavior across the two students. On InternVL3.5, SFT improves performance on several benchmarks, suggesting that teacher trajectories can provide useful structured reasoning supervision when the student does not rely on an explicit native thinking process. In contrast, SFT consistently degrades

Table 1: **Scaling and generalization across student model families and sizes.** We report results for 4B and 8B variants of InternVL3.5 and Qwen3-VL-Thinking on three in-domain and three out-of-domain video reasoning benchmarks.

Params	Method	*In-Domain*			*Out-of-Domain*		
		TimeBlind	MVBench	TempCompass	TOMATO	Video-MME	VideoReason
Proprietary Teacher Model: Gemini 2.5 Pro							
-	-	81.13	72.25	83.85	46.91	90.56	56.00
Student Model: InternVL 3.5							
4B	Baseline	47.88	57.00	67.84	21.29	67.00	0.97
	CoT SFT	55.50	54.88	66.15	28.23	68.88	1.78
	SFT + GRPO	58.50	58.33	72.34	29.92	69.44	2.88
	Ours	**60.13**	**59.50**	**77.67**	**30.80**	**70.44**	**3.47**
8B	Baseline	64.38	62.50	77.47	26.62	73.43	1.32
	CoT SFT	63.38	58.63	76.37	31.27	68.11	1.39
	SFT + GRPO	65.25	60.62	76.95	33.62	72.44	2.43
	Ours	**66.50**	**62.88**	**79.10**	**34.23**	**75.45**	**4.12**
Student Model: Qwen3-VL-Thinking							
4B	Baseline	64.50	66.63	79.69	32.01	**76.67**	5.21
	SFT	60.12	62.00	73.37	24.30	68.33	1.46
	SFT + GRPO	65.87	66.00	79.30	33.62	75.32	5.82
	Ours	**66.50**	**67.47**	**81.32**	**34.97**	76.00	**9.51**
8B	Baseline	67.00	65.38	80.34	34.57	78.56	6.11
	SFT	64.72	64.28	75.55	31.66	71.56	3.54
	SFT + GRPO	67.88	67.75	79.86	35.37	**81.04**	7.36
	Ours	**69.11**	**68.62**	**81.34**	**36.74**	80.22	**11.04**

Qwen3-VL-Thinking across all six benchmarks, where the student already possesses its own explicit reasoning behavior. This contrast suggests that direct reasoning distillation is not uniformly beneficial: teacher trajectories may introduce useful reasoning structure for one student, while requiring another to reshape an already established reasoning policy.

RL-Based Baselines. RL-based optimization consistently improves upon direct trajectory SFT, with standard GRPO recovering much of the performance lost through direct imitation. More specialized approaches provide further gains in some settings: LUFFY (Yan et al., 2025) incorporates off-policy teacher guidance, while T-GRPO (Feng et al., 2025) explicitly encourages sensitivity to temporal ordering. However, their improvements vary substantially across students and benchmarks, and neither consistently dominates standard GRPO. In contrast, our method remains the strongest post-training approach across all settings in Fig. 4 by optimizing reasoning under multiple student-side temporal observations rather than treating execution as observation-invariant.

5.3 Scaling and Generalization

Scaling Across Model Sizes. Table 1 extends the main comparison to both 4B and 8B students. The gains from observation-conditioned execution persist as the student scales up: our method improves over the corresponding SFT+GRPO pipeline in all 4B settings and in all but one 8B setting. The improvement therefore does not disappear as the underlying student becomes stronger. Notably, post-training also closes a substantial part of the capacity gap: on several benchmarks, the 4B model trained with our method matches or exceeds the corresponding pretrained 8B model, showing that stronger reasoning execution can compensate for part of the difference in model scale.

Out-of-Domain Generalization. The improvement also transfers beyond the benchmarks used for training. While Stage I and Stage II are trained on TimeBlind (Li et al., 2026a), MVBench (Li et al., 2024b), and TempCompass (Liu et al., 2024), the same advantage is retained on TOMATO (Shangguan et al., 2025), Video-MME (Short) (Fu et al., 2025), and VideoReasonBench (Liu et al., 2026), which are held out entirely from training. Across the four student backbones, our method improves over standard SFT+GRPO in all in-domain settings and in all but one out-of-domain setting. This consistency across benchmarks with different task formats and reasoning requirements suggests that the learned execution behavior is not restricted to the training tasks.

5.4 Component Ablation

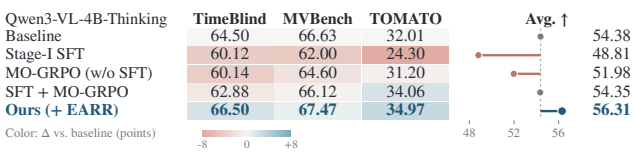


Figure 5: **Ablation study of the proposed framework.** We evaluate the contributions of Stage-I SFT, multi-observation GRPO, and execution-aware reference regularization.

Complementary Acquisition and Execution. Fig. 5 isolates the roles of teacher-guided reasoning acquisition and student-side execution optimization. Neither Stage-I SFT nor multi-observation GRPO is sufficient on its own: direct trajectory distillation degrades the pre-trained model, while applying multi-observation optimization without the teacher-acquired prior only partially recovers its performance. Combining the two substantially improves all evaluated benchmarks, indicating that the two stages serve complementary roles rather than interchangeable ones. Stage I provides structured reasoning patterns acquired from the teacher, whereas Stage II adapts these patterns to reasoning trajectories that are effective under the student's own temporal observations.

Execution-Aware Regularization. Adding execution-aware reference regularization further improves the combined model across all evaluated benchmarks. This result shows that preserving the Stage-I reasoning prior with a uniform constraint is still suboptimal: the prior may already be effective under one temporal observation but fail under another. EARR explicitly accounts for this observation-wise difference, preserving successful reasoning behavior while allowing larger policy updates when the prior is not executable. The improvement is most pronounced on TimeBlind, which is also the benchmark where Fig. 2 shows substantial sensitivity to temporal sampling, making the result consistent with the motivation for observation-dependent reference control.

5.5 Execution Across Temporal Observations

Section 2 shows that reasoning execution can change with temporal observation, making a teacher trajectory useful for acquisition but not necessarily executable from every student-side view. We therefore evaluate Qwen3-VL-4B-Thinking (Bai et al., 2025) on TimeBlind (Li et al., 2026a) under the five observations used in Stage II. As shown in Table 2, direct SFT remains sensitive to temporal sampling. Standard GRPO partially recovers this loss, but optimization under a single observation does not consistently transfer when the observed frames change. In contrast, our method improves execution across different observations by explicitly optimizing the student on multiple views and conditioning the reference constraint on whether the acquired prior is executable under each view. The case study in Section E further shows that this does not require reproducing the same reasoning trajectory: our model uses different observation-specific evidence while reaching the correct conclusion.

Table 2: **Multi-observation ablation.** Performance under different temporal observations in the TimeBlind benchmark, where ϕ denotes the temporal sampling phase.

Method	FPS = 1		FPS = 2		
	ϕ=0	ϕ=.5	ϕ=0	ϕ=.25	ϕ=.5
Baseline	56.75	60.75	64.50	64.25	63.88
SFT	53.75	52.50	60.12	58.62	59.00
GRPO	54.50	56.12	65.88	64.50	65.88
Ours	**62.75**	**63.38**	**66.50**	**67.88**	**67.75**

6 Conclusion

In this paper, we identify the Reasoning Acquisition–Execution Gap in video reasoning distillation, where teacher-generated reasoning may not be directly executable under the student's temporal observations. To address this issue, we propose an acquisition-to-execution framework that first distills teacher reasoning into a student prior and then optimizes its execution across multiple student-side observations with execution-aware reference regularization. Extensive experiments across multiple Video-LLM backbones and reasoning benchmarks demonstrate consistent improvements over supervised distillation and standard RL-based optimization. These results highlight that effective video reasoning distillation requires not only acquiring teacher reasoning, but also adapting its execution to the evidence actually observed by the student. We study reasoning distillation for Video-LLMs under heterogeneous temporal observations.

AI USE STATEMENT

We used generative AI tools to assist with manuscript revision, literature search, and code development. The authors take responsibility for the accuracy and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility throughout the appendix. Section C describes the dataset construction, training/evaluation splits, and teacher trajectory generation procedure; Section D provides the model configurations, training hyperparameters, temporal sampling settings, evaluation protocols, and prompts. Additional analyses and experiments, including reproduction with an open-source teacher model, are provided in Section D.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL `https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf`. Model card.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26209–26218, 2026.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Live: Learning video llm with streaming speech transcription at scale. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29083–29095. IEEE, 2025a.

Junzhou Chen and Sidi Lu. An advanced driving agent with the multimodal large language model for autonomous vehicles. In *2024 IEEE International Conference on Mobility, Operations, Services and Technologies (MOST)*, pp. 1–11. IEEE, 2024.

Junzhou Chen, Jindong Wang, and Gang Zhou. Fast enough to act: Spatio-temporal visual token merging for low-latency robotic vlms and vlas. *arXiv preprint arXiv:2606.29350*, 2026.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Zixuan Chen, Jiaxin Li, Junxuan Liang, Liming Tan, Yejie Guo, Cewu Lu, and Yong-Lu Li. Mˆ 3-vos: Multi-phase, multi-transition, and multi-scenery video object segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29193–29202, 2025b.

Tongtong Cheng, Rongzhen Li, Yixin Xiong, Tao Zhang, Jing Wang, and Kai Liu. Mcam: Multimodal causal analysis model for ego-vehicle-level driving video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5479–5489, 2025.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9155–9164, 2026.

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.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*, 2024.

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. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2025.

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

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 102430–102459, 2026.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 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. *arXiv preprint arXiv:2501.12948*, 2025.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26181–26191. IEEE, 2025.

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

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the association for computational linguistics: ACL 2023*, pp. 8003–8017, 2023.

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.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3134–3154, 2023.

Dahun Kim, AJ Piergiovanni, Ganesh Mallya, and Anelia Angelova. Videocomp: Advancing fine-grained compositional and temporal alignment in video-text models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29060–29070, 2025.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 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.

Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. Timeblind: A spatio-temporal compositionality benchmark for video llms. *arXiv preprint arXiv:2602.00288*, 2026a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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. *arXiv preprint arXiv:2602.02994*, 2026b.

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, 2024b.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2665–2679, 2023.

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, 2024.

Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Xinyu Zhou, and Xu Sun. Videoreasonbench: Can mllms perform vision-centric complex video reasoning? In *International Conference on Learning Representations*, volume 2026, pp. 3475–3502, 2026.

Yujie Lu, Yale Song, William Wang, Lorenzo Torresani, and Tushar Nagarajan. Vited: Video temporal evidence distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8501–8511. IEEE, 2025.

Yinyi Luo, Yiqiao Jin, Weichen Yu, Mengqi Zhang, Srijan Kumar, Xiaoxiao Li, Weijie Xu, Xin Chen, and Jindong Wang. Agentark: Distilling multi-agent intelligence into a single llm agent. *arXiv preprint arXiv:2602.03955*, 2026.

OpenRouter Team. Openrouter: A unified interface for large language models, 2023. URL `https://openrouter.ai/`. Accessed: September 26, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham Kakade, and Surbhi Goel. In good graces: Principled teacher selection for knowledge distillation. In *International Conference on Learning Representations*, volume 2026, pp. 54779–54822, 2026.

Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025.

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. IEEE, 2024.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 7593–7734, 2025.

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.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7059–7073, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. Time blindness: Why video-language models can't see what humans can? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30906–30918, 2026.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. Scott: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5546–5558, 2023.

Sudong Wang, Weiquan Huang, Xiaomin Yu, Zuhao Yang, Hehai Lin, Keming Wu, Chaojun Xiao, Chen Chen, Wenxuan Wang, Beier Zhu, et al. Beyond sft-to-rl: Pre-alignment via black-box on-policy distillation for multimodal rl. *arXiv preprint arXiv:2604.28123*, 2026.

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

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, 2025b.

Ziyu Wang, Yue Xu, Cewu Lu, and Yong-Lu Li. Dancing with still images: Video distillation via static-dynamic disentanglement. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6296–6304. IEEE, 2024.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*, 2025.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *Advances in Neural Information Processing Systems*, 38:117157–117186, 2025.

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, Limin Wang, Yu Qiao, and Yi Wang. Internvideo3: Agentify foundation models with multimodal contextual reasoning, 2026. URL https://arxiv.org/abs/2606.12195.

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. Survey on knowledge distillation for large language models: methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–27, 2025.

Yuyang You, Yongzhi Li, Jiahui Li, Yadong Mu, Quan Chen, and Peng Jiang. Adaptive video distillation: Mitigating oversaturation and temporal collapse in few-step generation. *arXiv preprint arXiv:2603.21864*, 2026.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18970–18980, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

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, 2026a.

Ruichen Zhang, Rana Muhammad Shahroz Khan, Zhen Tan, Dawei Li, Song Wang, and Tianlong Chen. The quest for efficient reasoning: A data-centric benchmark to cot distillation. In *International Conference on Learning Representations*, volume 2026, pp. 94555–94578, 2026b.

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

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Mmr-v: What's left unsaid? a benchmark for multimodal deep reasoning in videos. In *International Conference on Learning Representations*, volume 2026, pp. 67949–67991, 2026.

APPENDIX CONTENTS

A Outline of the appendix **16**

B Full Related Work **16**

B.1 Video Reasoning in Video-LLMs 16

B.2 Knowledge Distillation 16

B.3 Temporal Observation Construction in Video-LLMs 17

C Data Generation Methods and Pipeline **18**

C.1 Dataset Composition and Evaluation Split 18

C.2 Multi-Configuration Teacher Trajectory Generation 18

C.3 Error-Focused RL Dataset Construction 19

D Experimental Details **19**

D.1 Models and Compared Systems 19

D.2 Implementation Details for Training 20

D.3 Baselines 21

D.4 Evaluation Protocol 22

D.5 Observation-Sensitivity Analysis 22

D.6 Prompt 23

E Additional Studies **23**

E.1 Reasoning under Different Temporal Observations 23

E.2 Observation Mismatch in Practice 23

E.3 Reproducibility with an Open-Source Teacher 24

A Outline of the appendix

Below, we outline the structure of the appendix.

Section B provides additional discussion of related work on video reasoning and knowledge distillation. Section C details the data construction pipeline, including dataset composition, teacher trajectory generation, and the construction of the Stage-II RL training set. Section D provides complete experimental details, including model configurations, training hyperparameters, baselines, evaluation protocols, and prompts. Finally, Section E presents qualitative case studies on reasoning under different temporal observations and practical observation mismatch across Video-LLMs.

B Full Related Work

B.1 Video Reasoning in Video-LLMs

Recently, the ability of Video-LLMs (Team et al., 2025; Chen et al., 2024; Li et al., 2024a) to perform complex video reasoning has been significantly enhanced, particularly in determining temporal sequences, analyzing event causality, tracking changes in object states, and understanding spatial relationships (Kim et al., 2025; Cheng et al., 2025; Kim et al., 2025; Yuan et al., 2025). By breaking down video into timestamped frames and employing progressive multimodal training (An et al., 2026; Ren et al., 2024; Feng et al., 2025; Wang et al., 2025a;b), the model is better equipped to handle video-related tasks involving localization and event sequencing.

As video complexity increases, Chain-of-Thought (Fei et al., 2024; Zhang et al., 2025; Han et al., 2025; Chen & Lu, 2024) breaks down complex video tasks into multiple perceptual and cognitive steps, enabling the model to arrive at the final answer progressively through an intermediate reasoning process. Meanwhile, multiple video reasoning benchmarks (Li et al., 2024b; Shangguan et al., 2025; Cheng et al., 2026; Qi et al., 2025; Upadhyay et al., 2026; Zhu et al., 2026) indicate that high accuracy on traditional VideoQA tasks does not equate to reliable temporal reasoning capabilities. These studies demonstrate that, although current Video-LLMs possess strong video understanding capabilities, a significant gap in reasoning performance remains between proprietary and open-source models.

Many studies (Shi et al., 2025; Lu et al., 2025; Wang et al., 2024; You et al., 2026) focus directly on optimizing the reasoning capabilities of video LLMs. Video-R1 (Feng et al., 2025) introduces reinforcement learning to video reasoning, employing a temporally aware reward design to encourage the model to more fully utilize temporal information from the video during the reasoning process. Thinking with Videos (Zhang et al., 2026a) enables models to actively locate and resample relevant video segments during inference. VideoP2R (Jiang et al., 2026) explicitly distinguishes between perception and reasoning processes and introduces more fine-grained process supervision during reinforcement learning. These methods primarily investigate how a single Video-LLM can more effectively use the video information it has acquired to reason. In contrast, we focus on a distinct issue arising in knowledge transfer for cross-model reasoning: the teacher and student may obtain different temporal observations due to differing video sampling processes; consequently, a reasoning process valid for the teacher cannot necessarily be directly executed based on the video evidence actually seen by the student.

B.2 Knowledge Distillation

Knowledge distillation (Gou et al., 2021; Gu et al., 2024; Yang et al., 2025; Zhang et al., 2026b; Hinton et al., 2015) primarily transfers knowledge from large models to smaller ones via prediction distributions or intermediate representations. Recently, the scope of distillation has gradually been extended to the model's explicit reasoning process (Li et al., 2023; Lu et al., 2025). Reasoning distillation (Hsieh et al., 2023; Wang et al., 2023; Zhang et al., 2026b; Shridhar et al., 2023; Wang et al., 2026) leverages rationales or CoT sequences generated by a teacher model to provide richer intermediate supervision, demonstrating the significant impact of explicit reasoning supervision on the performance of smaller models. Offline teacher trajectories can lead to distribution discrepancies between the teacher and the student.

On the other hand, relying on fixed reasoning trajectories generated offline by the teacher introduces new challenges (Zhang et al., 2026b; Shridhar et al., 2023). During training, the student is required to fit teacher-generated sequences, whereas at inference time it generates outputs under its own policy, potentially leading to a substantial policy mismatch between the teacher and the student. MiniLLM (Gu et al., 2024) improves generative distillation for autoregressive language models through reverse KL optimization.

Knowledge distillation has also been extended to multimodal and video reasoning scenarios. Agent-of-Thoughts Distillation (AoTD) (Shi et al., 2025) first employs multiple video agents with specialized capabilities to construct verified spatio-temporal reasoning chains, and then distills these reasoning capabilities into a single Video-LLM. VOLD (Bousselham et al., 2026) combines GRPO with on-policy distillation to continuously incorporate teacher supervision during reinforcement learning for vision-language models, and highlights the importance of a reasoning-SFT cold start before online knowledge transfer. Video-OPD (Li et al., 2026b) leverages a stronger teacher to provide fine-grained token-level supervision on trajectories generated by the student itself. These studies suggest that the conditions under which reasoning supervision is generated, and the conditions under which the student learns from that supervision, are becoming increasingly important in video reasoning distillation.

Research (Gu et al., 2024; Ko et al., 2024; Agarwal et al., 2024; Ding et al., 2026) has shifted the distillation process toward student-generated responses and integrated it with reinforcement learning. This concept has been extended to multimodal scenarios, incorporating on-policy supervision and reinforcement learning into multimodal knowledge distillation to enhance visual grounding and reasoning capabilities (Xu et al., 2025; Shi et al., 2025; Luo et al., 2026). Unlike existing approaches that distill from privileged visual evidence, we focus on the observation mismatch that naturally arises from differing temporal sampling strategies between the teacher and the student, where a reasoning trajectory valid under the teacher's observations may no longer be executable given the evidence actually observed by the student.

Table 3: **Representative temporal sampling pipelines in Video-LLMs.** Different models construct temporal observations using different sampling rules and frame constraints.

Model	Reference Setting	Frame Selection	Budget / Control
Gemini 2.5 Pro	1 FPS	Fixed-rate sampling	FPS configurable
Qwen3-VL	2 FPS	FPS to frame count, then uniform indices	4–768 frames or explicit frame count
InternVL3.5	Fixed number of segments	One frame per temporal segment	Segment count configurable
LLaVA-Video	1 FPS	FPS sampling or uniform sampling	Maximum frame count configurable
VideoLLaMA3	1 FPS	FPS sampling with uniform subsampling	Maximum frame count configurable

B.3 Temporal Observation Construction in Video-LLMs

Video-LLMs (Chen et al., 2025a; Zhang et al., 2025; 2024; 2025) do not share a standardized procedure for converting a raw video into the finite sequence of frames used for inference. Table 3 summarizes representative temporal preprocessing pipelines documented in official model interfaces, model cards, and released code. We emphasize that these configurations are not immutable properties of the models: several ystems allow users to modify the sampling rate or frame budget. Rather, the comparison illustrates that the effective temporal observation is determined by a model-specific sampling pipeline, instead of a single universal FPS parameter.

As shown in Table 3, these pipelines differ in both their parameterization and frame-selection rules. Gemini's (Comanici et al., 2025) static video interface uses fixed-rate sampling by default, whereas Qwen3-VL (Bai et al., 2025) converts a requested FPS into a target number of frames and distributes the selected indices over the temporal range. InternVL3.5 (Wang et al., 2025a) instead exposes a segment-based video loader that divides the video into a configurable number of temporal intervals and samples one representative frame from each interval. Other open-source Video-LLMs (Zhang et al., 2024; 2025) combine rate-based sampling with explicit frame budgets, switching or resampling to a fixed number of frames when the budget becomes active. Thus, two systems receiving the same raw video can construct different frame sequences even when their nominal sampling rates appear similar.

This distinction motivates treating the temporal observation as $O = S(V; \xi)$, where ξ describes the complete sampling configuration rather than only the nominal frame rate. In practice, ξ may include the sampling rate, frame budget, temporal phase or starting offset, frame-index construction, and budget-induced resampling rule. Consequently, matching the nominal FPS alone does not in general guarantee identical temporal observations across Video-LLMs.

C Data Generation Methods and Pipeline

C.1 Dataset Composition and Evaluation Split

Table 4: **Benchmark card.** Summary of the video reasoning benchmarks used in our experiments. "Scale" reports the number of videos and question–answer pairs.

Benchmark	Scale	Video Duration	Main Focus	Model Output
TimeBlind (Li et al., 2026a)	600 instances / 2.4K QA	8.5s avg.	Temporal Compositionality; Event Order; Structural Logic; Event Attribute;	Two-choice answer
Video-MME (Fu et al., 2025)	900 videos / 2.7K QA	11s–60min	Temporal Reasoning; Action Reasoning; Perception; Recognition; Long-Video Understanding	Four-choice answer
TOMATO (Shangguan et al., 2025)	1.4K videos / 1.5K QA	9.2s avg.; 72.7s max.	Action Count; Direction; Rotation; Shape & Trend; Visual Cues; Velocity & Frequency;	Multiple-choice answer
VideoReasonBench (Liu et al., 2026)	240 videos / 1.4K QA	54.3s avg.; 154.8s max.	Operation Recall; State Tracking; State Comparison; State Prediction; Operation Planning	Open-ended structured state; Operation sequence
MVBench (Li et al., 2024b)	3.6K Video clips / 4.0K QA	∼16s avg.	Action Understanding; Motion Understanding; Object Interaction; State Change; Episodic Reasoning; Counterfactual Reasoning	Multiple-choice answer
TempCompass (Liu et al., 2024)	410 videos / 7.5K instructions	∼11.4s avg.	Action; Direction; Speed; Event Order; Attribute Change	Multiple-choice, yes/no,

We construct the training pool from three video reasoning benchmarks, as summarized in Table 4: TimeBlind (Li et al., 2026a), MVBench (Li et al., 2024b), and TempCompass (Liu et al., 2024). Specifically, we sample 1,600 video-question pairs from TimeBlind, 3,000 from MVBench, and 4,000 from TempCompass. For TempCompass, we exclude caption-oriented questions and retain only questions with deterministic answers. This results in a total of 8,600 video-question pairs for teacher trajectory generation and subsequent training.

The remaining samples from these three benchmarks are reserved for in-domain evaluation and are never used during training. We further ensure that the training and evaluation splits are **completely disjoint at the video level**, i.e., no video appears in both training and evaluation sets. For out-of-domain evaluation, we use TOMATO (Shangguan et al., 2025), the Short subset of Video-MME (Fu et al., 2025), and VideoReasonBench (Liu et al., 2026). No samples from these benchmarks are used for either Stage-I supervised distillation or Stage-II reinforcement learning. The dataset composition and training/evaluation usage are summarized in Table 5.

C.2 Multi-Configuration Teacher Trajectory Generation

For each training video-question pair, we generate four reasoning trajectories using the black-box teacher model. The four generations use different generation configurations to expose the student to diverse teacher reasoning behaviors for the same underlying problem. Across configurations, we keep the task prompt and required output format fixed, with only minimal benchmark-specific changes such as formatting multiple-choice options. The complete teacher response, including the reasoning trajectory, evidence, final answer, and generation metadata, is stored for every rollout.

Importantly, we distinguish between *raw trajectory generation* and the trajectories used for Stage-I supervised distillation. All four raw teacher generations are retained so that their correctness can later be used to construct the Stage-II training set. For Stage-I SFT, however, we only use teacher trajectories whose final answers are correct according to the benchmark ground truth. Incorrect

Table 5: **Dataset split and training usage.** The training pool is constructed from TimeBlind, MVBench, and TempCompass. Held-out samples from these benchmarks are used for in-domain evaluation, while TOMATO, Video-MME (Short), and VideoReasonBench are used exclusively for out-of-domain evaluation.

Benchmark	Available QA	Training QA	Training Exposure	Domain
TimeBlind	2,400	**1,600**	Yes	In-Domain
MVBench	4,000	**3,000**	Yes	In-Domain
TempCompass	5,500*	**4,000**	Yes	In-Domain
Video-MME (Short)	900	0	No	Out-of-Domain
TOMATO	1,500	0	No	Out-of-Domain
VideoReasonBench	1,400	0	No	Out-of-Domain
Total	15,600	**8,600**	–	–

*Caption-related questions are excluded from TempCompass, leaving approximately 5.5K objective QA samples.
†Only the Short subset of Video-MME is used for evaluation. N_{short} denotes the number of QA pairs in the evaluated Short subset.

teacher generations are therefore never used as supervised reasoning targets, but their correctness labels are retained for identifying examples on which the teacher reasoning process is not consistently successful.

Formally, let

$$\mathcal{Y}_i^T = \{y_{i,1}^T, \ldots, y_{i,4}^T\}$$

denote the four teacher trajectories generated for training example $x_i = (V_i, q_i, a_i^*)$, and let $\hat{a}_{i,j}^T$ denote the final answer of the j-th trajectory. The set of valid Stage-I supervision trajectories is

$$\mathcal{Y}_{i,\text{SFT}}^T = \left\{ y_{i,j}^T \in \mathcal{Y}_i^T \;\middle|\; \text{Eval}\big(\hat{a}_{i,j}^T, a_i^*\big) = 1 \right\}. \tag{8}$$

Only trajectories in $\mathcal{Y}_{i,\text{SFT}}^T$ are used as supervised targets during Stage I.

C.3 Error-Focused RL Dataset Construction

Stage II focuses reinforcement learning on training examples for which the teacher generations are not uniformly correct. After generating the four teacher trajectories for each training example, we evaluate the final answer of every trajectory against the corresponding ground-truth answer. An example is included in the RL training set if *at least one* of its four teacher generations is incorrect.

Specifically, the Stage-II dataset is defined as

$$\mathcal{D}_{\text{RL}} = \left\{ x_i \in \mathcal{D}_{\text{train}} \;\middle|\; \sum_{j=1}^{4} \mathbb{I}\left[\text{Eval}\big(\hat{a}_{i,j}^T, a_i^*\big) = 1\right] < 4 \right\}. \tag{9}$$

Equivalently, a training example is retained whenever there exists at least one teacher generation that fails to reach the correct final answer. Examples for which all four teacher generations are correct are excluded from Stage-II training.

This construction concentrates reinforcement learning on examples that remain non-trivial under the teacher-generation process, while reusing the same trajectory-generation results without requiring additional teacher inference. Note that the teacher trajectories themselves are not used as reward targets during Stage II; once the RL subset is selected, policy optimization is performed using student-generated rollouts under student-side temporal observations, as described in Section 4.2.

D Experimental Details

D.1 Models and Compared Systems

We conduct experiments on two Video-LLM families: Qwen3-VL-4B-Thinking, Qwen3-VL-8B-Thinking, InternVL3.5-4B-HF, and InternVL3.5-8B-HF. Qwen3-VL is used with its native thinking

behavior enabled, whereas InternVL3.5 is used in its standard instruction configuration without enabling an additional explicit thinking mode. This provides two complementary student reasoning regimes for evaluating reasoning distillation.

For each backbone, we evaluate the pretrained model, Stage-I SFT, standard GRPO, multi-observation GRPO (MO-GRPO), and our full model with Execution-Aware Reference Regularization (EARR). For the main comparison in Fig. 4, we additionally evaluate LUFFY (Yan et al., 2025) and T-GRPO (Feng et al., 2025) on the 4B students.

Unless otherwise specified, all reinforcement-learning methods are initialized from the same Stage-I SFT checkpoint of the corresponding backbone. They are trained on the same error-focused RL dataset defined in Section C, and use the same question set, output format, and answer-based reward signal. This ensures that differences among RL variants arise from their optimization strategies rather than different initialization or training data.

All student models are fine-tuned with LoRA for parameter-efficient adaptation. We set the LoRA rank to 32, scaling factor to 64, and dropout rate to 0. LoRA is applied to the `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` modules of the language model. Training is performed in BF16 precision with SDPA attention.

D.2 Implementation Details for Training

Stage-I Supervised Fine-Tuning. We train the Stage-I models for 2 epochs with a per-GPU batch size of 1 and 8 gradient accumulation steps. We use AdamW with a learning rate of 1×10^{-5}, zero weight decay, a cosine learning-rate scheduler, and a warmup ratio of 0.03. The gradient norm is clipped to 1.0, and the maximum sequence length is set to 32,768 tokens.

For Qwen3-VL, videos are processed at 2 FPS with the maximum number of pixels per frame set to 400,704. For InternVL3.5, we uniformly sample at 2 FPS from each video to control the number of visual tokens.

Stage I optimizes the standard autoregressive cross-entropy objective over the retained correct teacher trajectories, as defined in Section 4.1. The resulting model is used both to initialize Stage II and as the frozen reasoning reference for EARR.

Stage-II Reinforcement Learning. All GRPO-based variants are trained for 1 epoch with a learning rate of 1×10^{-6} and zero weight decay. To control for training resources, all RL-based experiments use the same hardware and wall-clock budget: four NVIDIA H100 GPUs for approximately 48 hours per run, corresponding to approximately 192 H100 GPU-hours. The RL dataset is described in Section C. We use a per-GPU batch size of 1 and accumulate gradients over four original video-question examples. The maximum input sequence length and maximum generation length are set to 40,960 and 2,048 tokens, respectively. The GRPO clipping range is 0.2, and the gradient norm is clipped to 1.0. Online student rollouts are sampled with temperature 1.0 and top-p 0.95.

All Stage-II methods are trained on the same error-focused RL subset introduced in Section C. Teacher correctness is used only to determine whether a video-question pair is included in this subset. Once an example enters Stage II, optimization uses student-generated rollouts and ground-truth task outcomes; teacher trajectories are not used as reward targets.

All RL variants use the same answer- and format-based reward function. The primary reward is determined by final-answer correctness. Producing a valid final answer and following the required output format each provide an auxiliary reward of 0.05. Outputs that reach the maximum generation length without producing a valid answer receive a truncation penalty of 0.5, while excessive repeated 4-grams receive a repetition penalty of 0.2. No teacher correctness signal or cross-observation difficulty signal is directly included in the rollout reward.

Standard GRPO. Standard GRPO uses a single student-side temporal observation for each video-question pair and generates four student rollouts under that observation. Rewards are normalized within the four-rollout group. The KL coefficient is fixed to $\beta = 0.003$ throughout training. This configuration serves as the single-observation RL baseline for evaluating the contribution of multi-observation execution.

Multi-Observation Configuration. MO-GRPO constructs $K = 5$ temporal observations for each video-question pair:

$$(1\ \text{FPS}, 0), \qquad (1\ \text{FPS}, 0.5), \qquad (2\ \text{FPS}, 0), \qquad (2\ \text{FPS}, 0.25), \qquad (2\ \text{FPS}, 0.5), \tag{10}$$

where the second term denotes the sampling phase ratio ϕ. For a sampling rate f and phase ratio ϕ, sampling timestamps are given by $t_k = \frac{k+\phi}{f}$.

Accordingly, the two 1-FPS observations use absolute offsets of 0 and 0.5 seconds, while the three 2-FPS observations use offsets of 0, 0.125, and 0.25 seconds. All observations are generated from the same raw video. Qwen3-VL and InternVL3.5 use the same timestamp-based frame-selection logic so that the observation configurations have consistent temporal semantics across model families.

Observation-Local Optimization. For MO-GRPO without EARR, each of the five temporal observations independently generates four student rollouts. Reward statistics and group-relative advantages are computed only among rollouts generated from the same observation; rewards are never normalized across different temporal observations.

With four original video-question pairs accumulated per optimizer update, This variant uses a fixed KL coefficient of $\beta = 0.003$.

Execution-Aware Reference Regularization. For the full model, we preserve the same total generation budget of four generations per observation. Each observation produces four stochastic student rollouts together with one deterministic response from the frozen Stage-I reference policy.

The reference response uses the same temporal observation, question, and maximum generation length as the student rollouts, but is decoded deterministically with temperature 0 and top-p 1.0. Its final-answer correctness is used only to determine the observation-specific reference coefficient:

$$\beta_{i,k} = \begin{cases} \beta_{\text{high}} = 0.005, & \text{if the reference answer is correct,} \\ \beta_{\text{low}} = 0.001, & \text{otherwise.} \end{cases} \tag{11}$$

The reference response is not included in the GRPO rollout group and does not participate in the computation of the reward mean, standard deviation, group-relative advantage, or policy loss. These quantities are computed exclusively from the four student rollouts generated under the same observation.

D.3 Baselines

We compare against supervised reasoning distillation and representative RL-based reasoning optimization methods. All RL baselines use the same Stage-I initialization and Stage-II training subset whenever their original formulation permits, and we keep the student backbone, training questions, answer format, and evaluation protocol fixed.

SFT. SFT (Ouyang et al., 2022) directly trains the student on correct teacher-generated reasoning trajectories using the Stage-I supervised objective described above. The same LoRA configuration is used as in our method.

GRPO. GRPO (Guo et al., 2025) performs on-policy reinforcement learning with group-relative reward normalization over multiple student rollouts. In our implementation, it operates on a single temporal observation per video-question pair and uses the same Stage-I SFT initialization and Stage-II RL dataset as our method.

SFT $\rightarrow$ GRPO. This baseline follows the same two-stage initialization used by our framework but applies standard single-observation GRPO in Stage II. Importantly, it uses the same Stage-I SFT checkpoint and the same error-focused RL training set as our method; no separate data split is introduced between the SFT and RL stages.

LUFFY. LUFFY (Yan et al., 2025) is an off-policy reasoning optimization method that combines external teacher trajectories with on-policy student rollouts. We adapt LUFFY to the same video reasoning training pool and use the same teacher trajectories available to the other distillation methods. The student initialization, task data, output format, and evaluation protocol are kept consistent with the other RL baselines.

T-GRPO. T-GRPO (Feng et al., 2025) introduces temporal awareness into GRPO by contrasting model behavior under temporally ordered and shuffled video frames. We implement its temporal reward on the same Stage-II examples and student backbone used by the other RL methods while keeping the remaining training and evaluation settings unchanged.

D.4 Evaluation Protocol

All pretrained and post-trained models are evaluated using the same unified evaluation pipeline. At inference time, the model receives only the raw video and benchmark question; ground-truth answers are used solely for post-generation evaluation. Both Qwen3-VL and InternVL3.5 are evaluated using a fixed 2-FPS temporal sampling configuration with deterministic decoding: temperature 0, top-p 1.0, and a maximum of 2,048 newly generated tokens.

For TimeBlind (Li et al., 2026a), MVBench (Li et al., 2024b), TempCompass (Liu et al., 2024), TOMATO (Shangguan et al., 2025), and the Short subset of Video-MME (Fu et al., 2025), we extract the final answer from the required `Answer:` field and evaluate it against the benchmark ground truth using the corresponding deterministic answer-normalization rule.

VideoReasonBench (Liu et al., 2026) contains open-ended reasoning and structured state-prediction questions. We follow a fixed evaluation pipeline for all candidate models and use the same judge configuration and prompts throughout evaluation. For structured state-prediction questions, the generated response is first converted into the required structured representation before the benchmark's state-transition verification is applied. No candidate model receives access to the evaluation reference answer during generation.

D.5 Observation-Sensitivity Analysis

The analysis in Fig. 2 is an inference-only diagnostic experiment on the held-out TimeBlind (Li et al., 2026a) evaluation set and does not involve any additional training or post-training. We evaluate Qwen3-VL-4B, Qwen3-VL-8B (Bai et al., 2025), InternVL3.5-4B, InternVL3.5-8B (Wang et al., 2025a), and InternVideo3-8B (Yan et al., 2026) on the same 800 held-out video-question pairs. Within each model, the raw video, question, prompt, answer extraction, spatial preprocessing, and decoding configuration are kept fixed; only the temporal sampling configuration is changed. All observations are evaluated independently using deterministic decoding with temperature 0 and top-$p = 1$. We use the same TimeBlind prompt and deterministic answer evaluation described in Sections D.4 and D.6.

Observation-dependent execution. For Fig. 2(a), we construct five temporal observations for every video-question pair:

$$(1, 0),\ (1, 0.5),\ (2, 0),\ (2, 0.25),\ (2, 0.5),$$

where each tuple denotes (f, ρ), with sampling rate f in FPS and sampling phase ρ. For a fixed model and question, we record binary correctness under each of the five observations. A question is categorized as *always incorrect* if all five predictions are incorrect, *always correct* if all five are correct, and *observation-dependent* otherwise. We report the fraction of held-out TimeBlind questions in each category.

Outcome flips under phase shifts. For Fig. 2(b), we vary the sampling rate over $f \in \{0.5, 1, 2, 4\}$ FPS while comparing two sampling phases, $\rho \in \{0, 0.5\}$. For a given (f, ρ), sampling timestamps follow$t_k = \frac{k+\rho}{f}$. At each fixed sampling rate, an outcome flip occurs when the same model changes from correct to incorrect, or vice versa, between the two phase settings. For model m, the flip rate at sampling rate f is

$$\mathrm{FlipRate}_m(f) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[c_{m,i}^{(f,0)} \neq c_{m,i}^{(f,0.5)}\right], \tag{12}$$

where $N = 800$ and $c_{m,i}^{(f,\rho)} \in \{0, 1\}$ denotes answercorrectness. The *Average* curve in Fig. 2(b) reports the macro-average of the flip rates over the five evaluated models.

D.6 Prompt

We provide the complete benchmark-specific input prompts used in our experiments. The same task prompt and output format are used during knowledge distillation and evaluation for each benchmark.

We do not explicitly request a thinking trace in the user prompt. Instead, the user prompt specifies only the task, the evidence field, and the final-answer field. For models with a native reasoning mechanism, such as Qwen3-VL-Thinking, intermediate reasoning is produced through the model's native thinking behavior. InternVL3.5 is used in its standard instruction configuration.

TimeBlind uses the prompt in Fig. 9, TempCompass uses Fig. 10, MVBench uses Fig. 11, TOMATO and Video-MME (Short) share the prompt in Fig. 12, and VideoReasonBench uses Fig. 13.

The answer-normalization prompt shown in Fig. 14 is used only to map ambiguous free-form model outputs to the benchmark's allowed answer labels. It is not provided with the video or reference answer and is not used to solve the underlying video question.

E Additional Studies

E.1 Reasoning under Different Temporal Observations

To qualitatively complement the observation-sensitivity analysis in Section 2, we examine a TimeBlind example that requires comparing the rubbing speed of a hand over time. Fig. 7 and Fig. 8 show responses from Qwen3-VL-4B-Thinking (Bai et al., 2025) and our model under different temporal observations obtained by varying the sampling FPS and phase. The underlying video and question remain unchanged; only the temporal evidence exposed to the model differs.

The baseline exhibits observation-dependent execution: its prediction changes across sampling configurations, and under several observations it reverses the temporal trend from fast-to-slow" to slow-to-fast." In contrast, our model consistently predicts the correct transition while grounding its reasoning in the evidence available under each observation. Importantly, the evidence itself is not identical across observations: the model refers to different timestamps and motion cues depending on the sampled frames. This illustrates the intended behavior of observation-conditioned execution—rather than reproducing a fixed reasoning trajectory, the model adapts its reasoning path to the current temporal evidence.

E.2 Observation Mismatch in Practice

Fig. 6 provides a controlled illustration of the observation mismatch discussed in Section 2. We construct a five-second synthetic video containing 150 consecutive frames, where frame t displays the integer t. Because the underlying visual sequence is deterministic and contains no semantic ambiguity, differences in model responses can be directly related to the temporal evidence exposed to each model.

As shown in Fig. 6, standard Video-LLMs do not necessarily recover the complete temporal sequence from the same raw video. For example, different models report sparse or shifted subsequences rather than the full progression from 1 to 150. These responses reflect the frames made available by their internal video sampling procedures: changing the sampling rate, phase, or frame locations produces a different effective observation even though the underlying video remains unchanged. In contrast, the agentic systems shown in the figure explicitly inspect the video frames and recover the complete sequence. We emphasize that this comparison is not intended as a direct capability ranking between standard and agentic models, since the latter have access to additional

Table 6: **Results with an open-source teacher model.** Distillation performance using Qwen3-VL-235B-A22B-Thinking as the teacher and InternVL3.5-4B and Qwen3-VL-Thinking-4B as student models across three in-domain and three out-of-domain benchmarks. Consistent improvements over the corresponding pretrained and SFT baselines indicate that the proposed framework remains effective with an open-source teacher, supporting the reproducibility of the main findings.

Method	*In-Domain*			*Out-of-Domain*		
	TimeBlind	MVBench	TempCompass	TOMATO	Video-MME	VideoReason
Teacher Model: Qwen3-VL-235B-A22B-Thinking						
-	69.75	71.38	84.77	44.85	88.50	27.48
Student Model: InternVL 3.5 4B						
Baseline	47.88	57.00	67.84	21.29	67.00	0.97
CoT SFT	50.14	58.98	65.34	23.70	68.24	1.12
Ours	**60.13**	**59.50**	**77.67**	**30.80**	**70.44**	**2.76**
Student Model: Qwen3-VL-Thinking 4B						
Baseline	64.50	66.63	79.69	32.01	76.67	5.21
SFT	63.45	65.42	77.92	24.30	74.26	4.88
Ours	**66.75**	**67.20**	**80.96**	**34.97**	**78.91**	**7.82**

frame-inspection tools. Rather, the example isolates the central issue considered in this work: the raw video V is shared, while the effective model observation $O = S(V; \xi)$ can differ substantially across systems.

This distinction becomes particularly important for reasoning distillation. A teacher trajectory is generated from the evidence contained in its observation O_T, while the student is trained to reproduce that trajectory from a potentially different observation O_S. If the teacher reasoning explicitly relies on an event or frame that is absent from O_S, the trajectory can remain logically correct under O_T while being difficult or impossible to execute from the evidence available to the student. The issue is therefore not necessarily an incorrect teacher rationale, but a mismatch between the observation under which the rationale was generated and the observation under which it is subsequently learned. This is the conditional target mismatch introduced in Section 2.

E.3 Reproducibility with an Open-Source Teacher

To further examine whether the main findings depend on a specific proprietary teacher, we replace Gemini 2.5 Pro with the open-source Qwen3-VL-235B-A22B-Thinking model (Bai et al., 2025). Teacher inference is performed through the OpenRouter API (OpenRouter Team, 2023), with videos processed at 2 FPS to generate reasoning trajectories for distillation. The resulting trajectories are then used to train InternVL3.5-4B (Wang et al., 2025a) and Qwen3-VL-4B-Thinking (Bai et al., 2025) following the same acquisition-to-execution pipeline as in the main experiments. All remaining dataset splits, training procedures, and evaluation protocols are kept unchanged, with performance evaluated on the same three in-domain and three out-of-domain video reasoning benchmarks.

As shown in Table 6, the overall behavior observed in the main experiments persists after replacing the proprietary teacher with an open-source model. Direct trajectory SFT remains strongly dependent on the student reasoning regime: teacher trajectories provide useful supervision for InternVL3.5 on several tasks, while direct imitation can still degrade the pretrained Qwen3-VL-Thinking model. In contrast, the proposed acquisition-to-execution framework consistently outperforms both the corresponding pretrained and SFT models across all evaluated in-domain and out-of-domain settings. These results indicate that the benefit of observation-conditioned execution is not specific to Gemini-generated trajectories, while providing an open-source teacher setting for reproducing the main experimental findings.

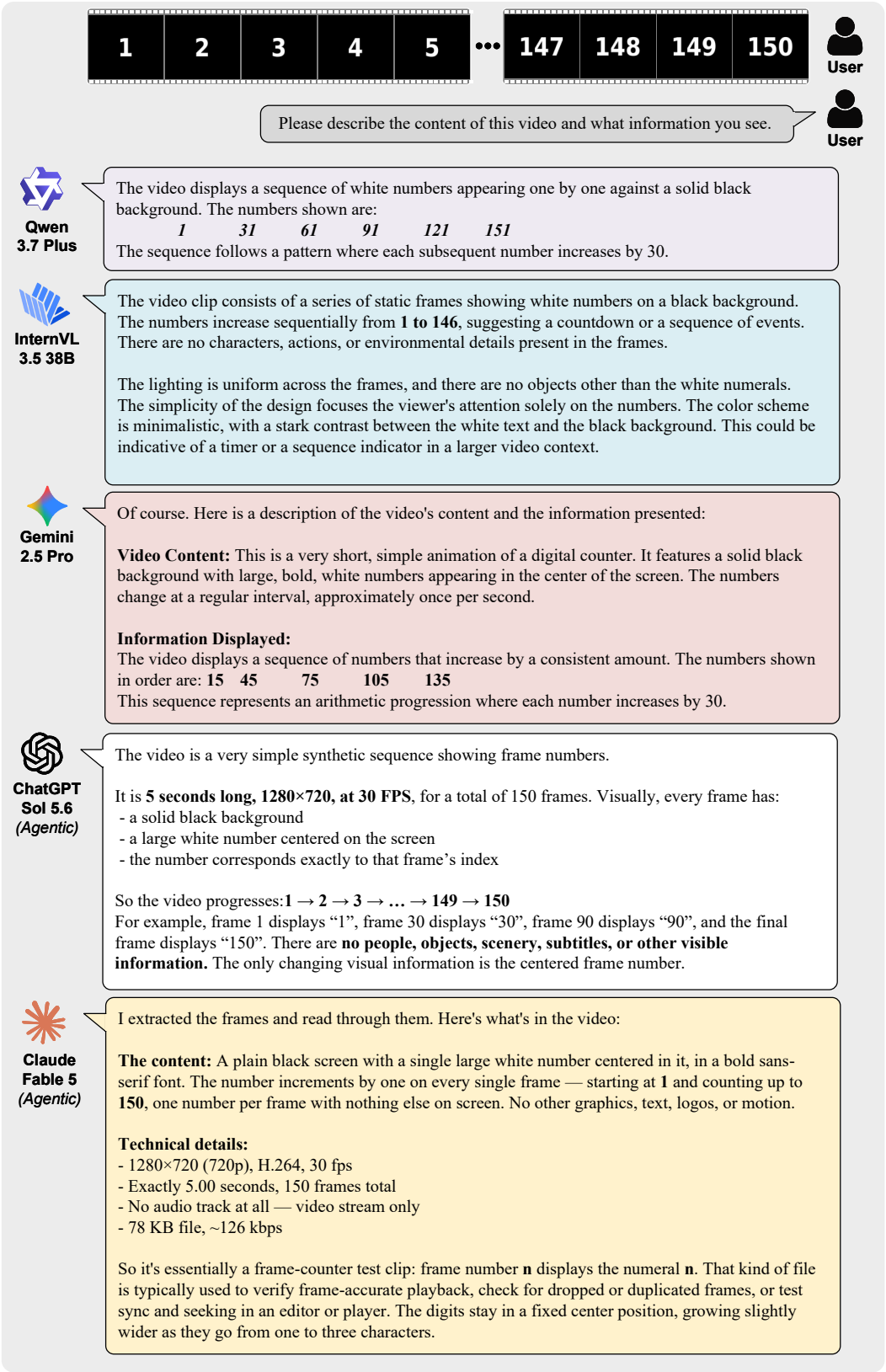


Figure 6: **Case study of frame-level temporal perception.** Standard Video-LLMs misinterpret sparsely sampled frames, while agentic models recover the full frame sequence.

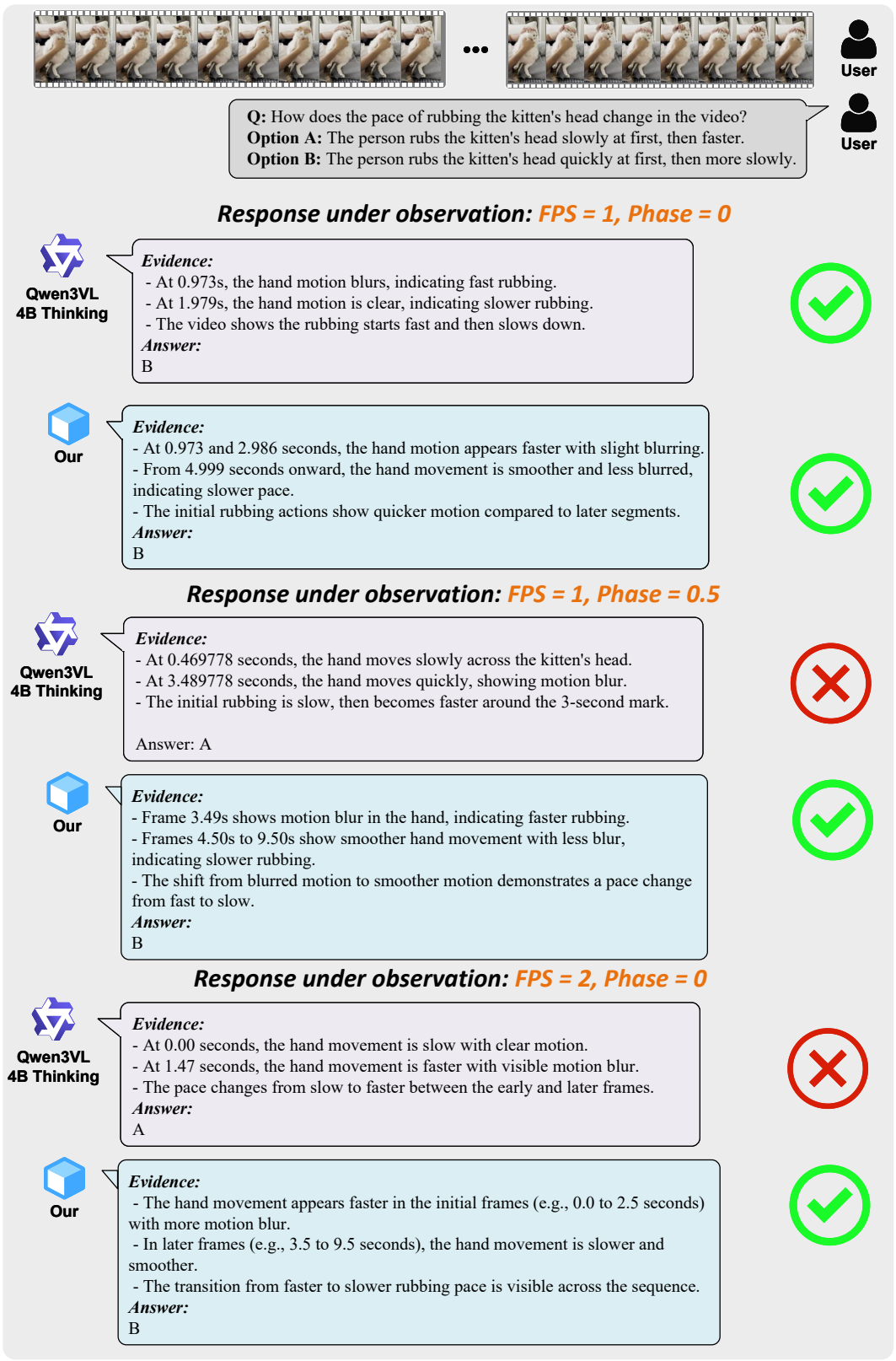


Figure 7: **Case Study 1 (Part I)** Comparison between Qwen3-VL-4B-Thinking and our method under different temporal observations of the same video-question pair.

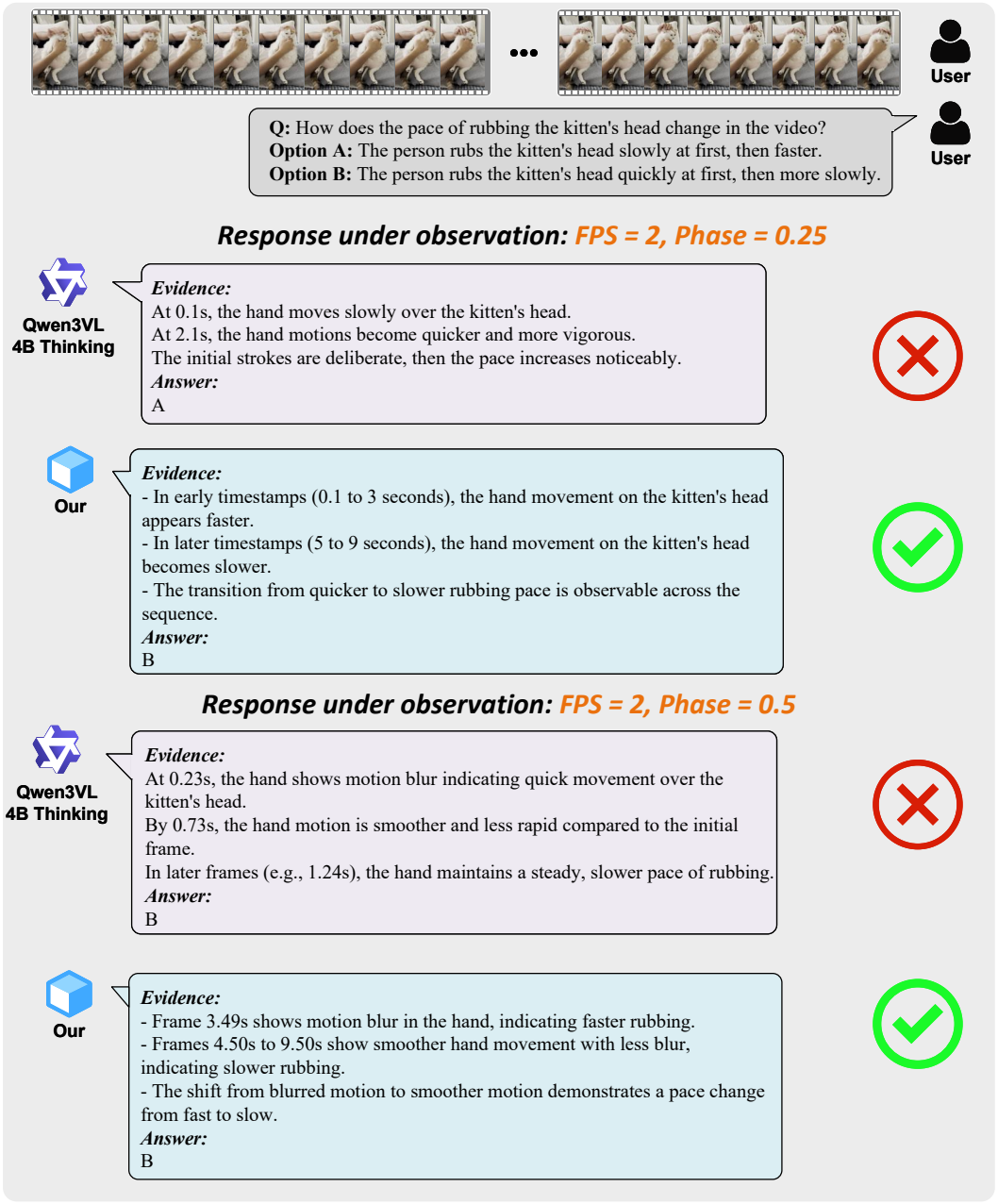


Figure 8: **Case Study 1 (Part II)** Additional comparison between Qwen3-VL-4B-Thinking and our method under different temporal observations of the same video-question pair.

Input Prompt for TimeBlind Benchmark

```
Answer the question using the video.

Do not repeat the question. Do not include extra explanation after the final answer.

Question:
{{question}}

Output format:
Evidence:
Use as many short evidence items as needed. For simple questions, 2-3 items are enough.
Each item should describe a relevant temporal event, spatial relation, or comparison.
Moreover, One short sentence connecting the evidence to the answer.

Answer:
State the final answer as yes/no or A/B without any other words.
```

Figure 9: Prompt Template for TimeBlind Benchmark.

Input Prompt for TempCompass Benchmark

```
Answer the question using the video.

Use temporal evidence from the video. Do not repeat the question and do not add
anything after the final answer label.

Task type: {{task}}

Allowed answers: {{allowed_answers}}

Question:
{{question}}

Output format:
Evidence:
- Give 2-4 short observations about the relevant action, direction, speed,
  event order, or attribute change.
- End with one short sentence connecting the observations to the answer.

Answer:
Write exactly one label from the Allowed answers list. Do not copy the choice
text and do not add punctuation, explanation, or any text after the label.
```

Figure 10: Prompt Template for TempCompass Benchmark.

Input Prompt for MVBench

```
Answer the multiple-choice question using the video.

Use temporal evidence from the video. Do not repeat the question and do not add
anything after the final answer label.

Task type: {{task}}

Allowed answers: {{allowed_answers}}

Question:
{{question}}

Output format:
Evidence:
- Give 2-4 short observations about the relevant actions, objects, event order,
  state changes, or causal relation.
- End with one short sentence connecting the observations to the answer.

Answer:
Write exactly one letter from the Allowed answers list. Do not copy the choice
text and do not add punctuation, explanation, or any text after the letter.
```

Figure 11: Prompt Template for MVBench Benchmark.

Input Prompt for TOMATO and Video MME Benchmark

```
Answer the multiple-choice question using the video.

Do not repeat the question. Do not include extra explanation after the final answer.

Question:
{{question}}

Options:
{{options}}

Output format:
Evidence:
Use as many short evidence items as needed. Each item should describe a relevant
temporal event, spatial relation, or comparison. Include one short sentence connecting
the evidence to the answer.

Answer:
State only the option letter without any other words.
```

Figure 12: Prompt Template for TOMATO and Video-MME Benchmark.

Input Prompt for VideoReason Benchmark

Answer the question using the video.

Question:
{{question}}

Output format:
Evidence:
Use concise evidence items describing the relevant observations and state transitions.
Include one short sentence connecting the evidence to the answer.

Answer:
Provide the complete final answer requested by the question, without any explanation after it.

Figure 13: Prompt Template for VideoReason Benchmark.

Input Prompt for LLM Judgement

You are an answer-normalization judge, not a question-answering model.
For each case, infer which allowed answer the MODEL RESPONSE intended to submit.
Do not solve the video question yourself. No video and no reference answer are provided.

Rules:
1. Return an allowed label when the response explicitly names that label, copies its choice text, or gives a final natural-language conclusion that unambiguously matches it.
2. Use UNVERIFIABLE if the response is empty, conflicting, truncated, or does not unambiguously select one allowed answer.
3. Do not prefer an answer based on your own world knowledge.
4. Preserve each allowed label exactly as written.
5. Return only a compact JSON array in the same order.

Figure 14: Prompt Template for LLM Judgment.