THROUGH THE LENS OF LEARNABILITY: PRIVILEGED SELF-DISTILLATION FOR VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) has emerged as an effective post-training paradigm for video large language models (Video-LLMs), where a privileged self-teacher provides token-level guidance along student-generated trajectories. For video understanding, however, privileged information can take diverse forms, including answer hints, textual descriptions, and additional temporal observations, whose learning value is not captured by teacher accuracy alone. Through systematic analysis, we find that stronger teachers do not necessarily produce better students; whether the privileged evidence is actually learnable by the student is the key factor. We therefore revisit video OPSD through the lens of *privilege learnability*, governed by two factors: *teacher reliability* and *student visibility*. Based on this analysis, we propose **LENS**, a learnable privileged evidence framework for video OPSD. LENS constructs privileged teachers from densely grounded temporal evidence, estimates teacher reliability at the sample level, filters supervision beyond the student's observable scope, and routes each sample to distillation or reinforcement supervision based on teacher reliability. Extensive experiments across four challenging video reasoning benchmarks demonstrate that LENS consistently outperforms existing approaches, with the distilled student even surpassing its privileged teacher on temporal grounding tasks, validating the importance of learnability-aware privileged distillation.

1 INTRODUCTION

Video large language models (Video-LLMs) have made rapid progress on video understanding (Cheng et al., 2024; Zhang et al., 2025; Wang et al., 2024). Unlike text reasoning (Guo et al., 2025; Jaech et al., 2024), which primarily derives correct reasoning processes through semantic inference, video reasoning requires models to jointly perform visual perception, temporal understanding, and multi-step reasoning (Feng et al., 2025; Sun et al., 2025). The complexity of these intertwined capabilities makes post-training particularly challenging, motivating increasing interest in on-policy self-distillation (OPSD), where a privileged self-teacher provides token-level guidance along student-generated trajectories using training-only privileged information (Zhao et al., 2026; Agarwal et al., 2024).

When extended to video reasoning, OPSD methods (Wang et al., 2026b;a) exploit privileged visual evidence to strengthen teacher supervision, conditioning self-teachers on additional temporal observations, visual clues, or densely grounded evidence unavailable to the student. Existing work has primarily focused on constructing stronger teachers (Yuan et al., 2026; Li et al., 2026b) or refining token-level distillation signals (Bi et al., 2026; Agarwal et al., 2024), under the implicit assumption that stronger supervision translates into better learning. A fundamental question remains underexplored: *Is the privileged supervision actually learnable by the student?*

Through systematic analysis, we find that this assumption does not hold. Stronger teachers do not necessarily produce better students, and whether the privileged evidence is learnable is the decisive factor. We identify two sources of failure. First, teacher guidance is not uniformly *reliable*: dense temporal evidence can introduce perceptual confusion, causing the teacher to err on samples the student answers correctly. Second, teacher guidance may lack *student visibility*: the teacher's reasoning can depend on observations entirely absent from the student's sparse input, making the

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

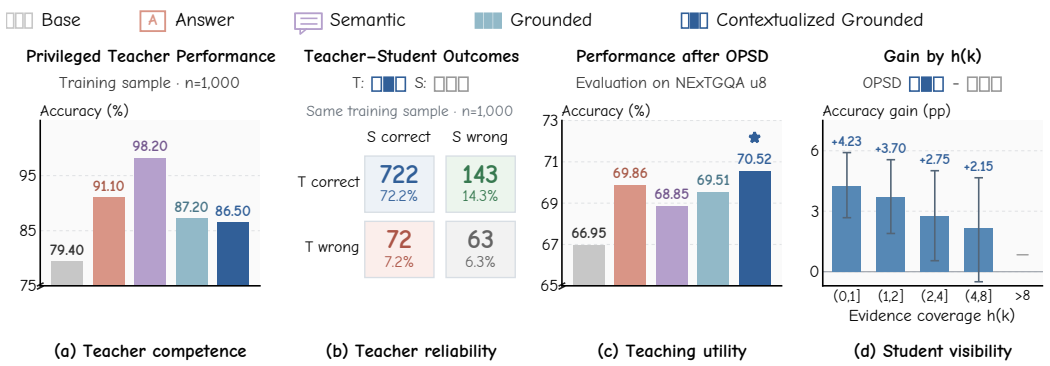

Figure 1: Preliminary analysis of privileged teachers. (a) Teacher accuracy under four privilege forms. (b) Teacher–student prediction outcomes. (c) Student accuracy after OPSD with each teacher. (d) OPSD gain grouped by student evidence coverage $h(k)$.

supervision unlearnable in principle. These findings motivate us to revisit video OPSD through the lens of *privilege learnability*, governed by two factors: *teacher reliability* and *student visibility*.

Based on this analysis, we propose **LENS** (**L**earning from Privileged **E**vidence for O**n**-Policy Video **S**elf-Distillation), a learnable privileged evidence framework for video OPSD. LENS constructs privileged teachers from densely grounded temporal evidence, filters supervision whose supporting evidence falls outside the student's observable scope, and routes each sample to distillation or reinforcement supervision based on teacher reliability. Reliable teacher guidance is distilled via a student-centered objective, while samples where privilege misleads the teacher are leveraged for discriminative self-reinforcement, turning potentially harmful supervision into a source of training signal. Extensive experiments across four challenging video reasoning benchmarks demonstrate that LENS consistently outperforms existing approaches, with the distilled student even surpassing its privileged teacher on temporal grounding tasks. Our contributions are as follows:

- We revisit video OPSD from the perspective of *privilege learnability* and identify two governing factor: teacher reliability and student visibility through systematic empirical analysis.
- We propose LENS, a framework that integrates visibility-aware filtering with dual-path reliability routing, applying reliable distillation and self-trajectory reinforcement to maximize the learning value of privileged supervision.
- Experiments on four benchmarks across the Qwen3-VL series with different model scale demonstrate consistent improvements and validate that learnability-aware privileged distillation yields gains beyond what stronger teachers alone can provide.

2 Preliminary Analysis

2.1 What Form of Privilege Works?

To understand how the form of privileged information affects distillation, we construct four teachers that differ in evidence modality and scope while sharing the same student (uniformly sampled sparse frames). We consider two linguistic forms: *Answer* (the ground-truth answer) and *Semantic* (an answer-relevant caption) and two visual forms: *Grounded* (dense frames within the GT temporal interval only) and *Contextualized Grounded* (the student's global sparse frames augmented with dense frames within the GT interval). Figure 1 and Figure 2 summarize the results.

A stronger teacher is not a better teacher. Comparing teacher competence (Figure 1a) with teaching utility (Figure 1c), the ranking of teachers is not preserved after distillation. Although the *Semantic* teacher achieves the highest standalone accuracy (98.2%), it does not yield the best student after OPSD. Instead, the *Contextualized Grounded* teacher, which combines globally sampled

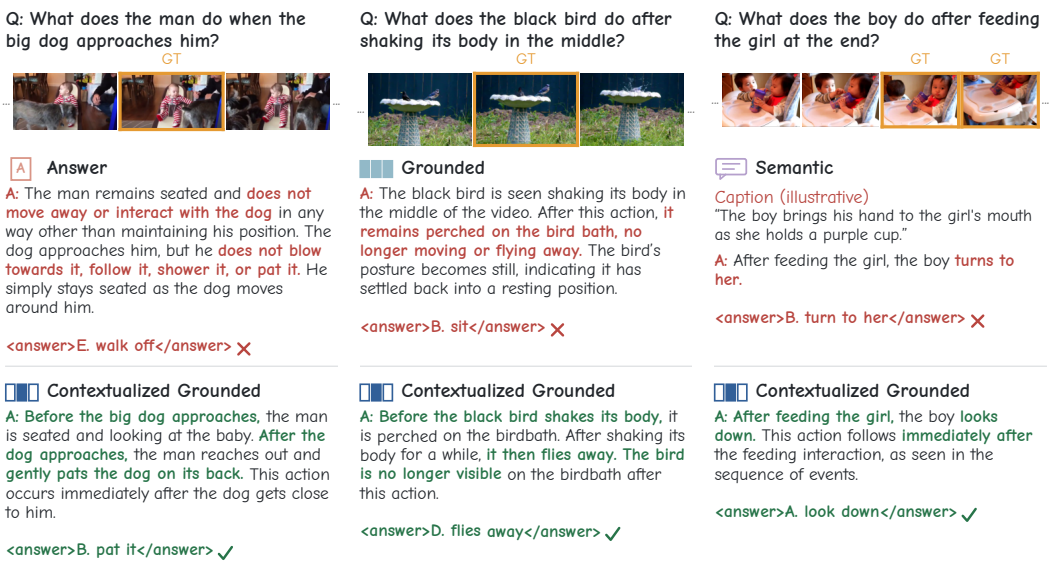

Figure 2: Case studies of different privileged teachers on three NExT-GQA examples.

frames with densely sampled grounded evidence, produces the largest improvement. This result suggests that effective supervision depends not only on teacher capability but on how transferable its guidance is: privilege that stays aligned with the student's input while adding question-relevant evidence distills best.

Different privileges induce different reasoning. Figure 2 shows how each form of privilege shapes the teacher's reasoning. *Answer* privilege invites shortcut reasoning: the teacher rationalizes the answer semantics instead of grounding its prediction in visual evidence. *Semantic* privilege supplies rich cues, but compressing the video into text blurs temporal order and leads to event-sequence confusion. *Grounded* privilege exposes the decisive evidence but, without the surrounding frames, misses state transitions beyond the GT interval. *Contextualized Grounded* privilege combines global context with localized evidence and yields grounded reasoning in all three cases. We therefore adopt it as the privileged teacher throughout the rest of the paper.

2.2 What Makes Privileged Information Learnable?

Given this teacher, we examine two factors that determine whether its guidance can be effectively learned: whether the student observes enough evidence, and whether the teacher is reliable.

Student visibility. The Contextualized Grounded teacher derives its advantage from dense frames within the GT interval, but the student's uniform sampling may cover this interval partially, minimally, or not at all. This structural gap means that even reliable teacher guidance can depend on evidence the student never observes, making it unlearnable in principle. To quantify how much of this privileged evidence is visible to the student, we define the *evidence coverage*:

$$h(k) = \left|\{f_i \in \mathbf{x}^s \mid t_{f_i} \in [t_{\text{start}}, t_{\text{end}}]\}\right|, \tag{1}$$

where $\mathbf{x}^s$ contains k uniformly sampled frames, t_{f_i} is the timestamp of frame f_i, and $[t_{\text{start}}, t_{\text{end}}]$ is the GT interval. A high $h(k)$ means the student already observes much of the decisive evidence; a low $h(k)$ indicates a **starved** sample where the student lacks the visual basis that the teacher's reasoning relies on.

Grouping OPSD gains by $h(k)$ (Figure 1d) reveals a clear pattern: gains are largest for the most starved samples ($+4.23$ points for $h(k) \in (0, 1]$) and decrease steadily as coverage grows ($+3.70$, $+2.75$, and $+2.15$ for $(1, 2]$, $(2, 4]$, and $(4, 8]$). This confirms that the teacher's value concentrates where the student's evidence is most incomplete—the privilege fills a perceptual gap the student

cannot resolve on its own. Effective distillation should therefore target the **partially starved** regime, where the student has partial evidence that the teacher's guidance can extend.

Teacher reliability. Unlike answer-based privilege, which effectively ensures teacher correctness by construction, dense sampling leaves the teacher fallible, it still reasons from visual input and can be misled by additional evidence. We partition samples by teacher correctness on 1,000 stratified examples (Figure 1b).

When the teacher is **correct** (86.5%), its guidance falls into two regimes. On corrective samples (T✓S✗, 14.3%), the teacher leverages dense evidence to answer questions that the student fails under sparse sampling; these form the core benefit region of privileged distillation. The majority are concordant samples (T✓S✓, 72.2%), where distillation acts as a stable regularizer rather than a major source of new knowledge. When the teacher is **wrong**, two cases arise. Both-wrong samples (T✗S✗, 6.3%) carry useless signal. More notable are *privilege-confounded* samples (T✗S✓, 7.2%), where the student answers correctly but dense sampling degrades the teacher's prediction. These failures concentrate on temporally adjacent events, where denser coverage introduces visually similar frames that blur event boundaries the student's sparser view preserves. Because teacher and student share the same model weights and differ only in input frames, these disagreements pinpoint decisions where the model's discrimination is most fragile.

3 LENS: Learnable Privileged Distillation

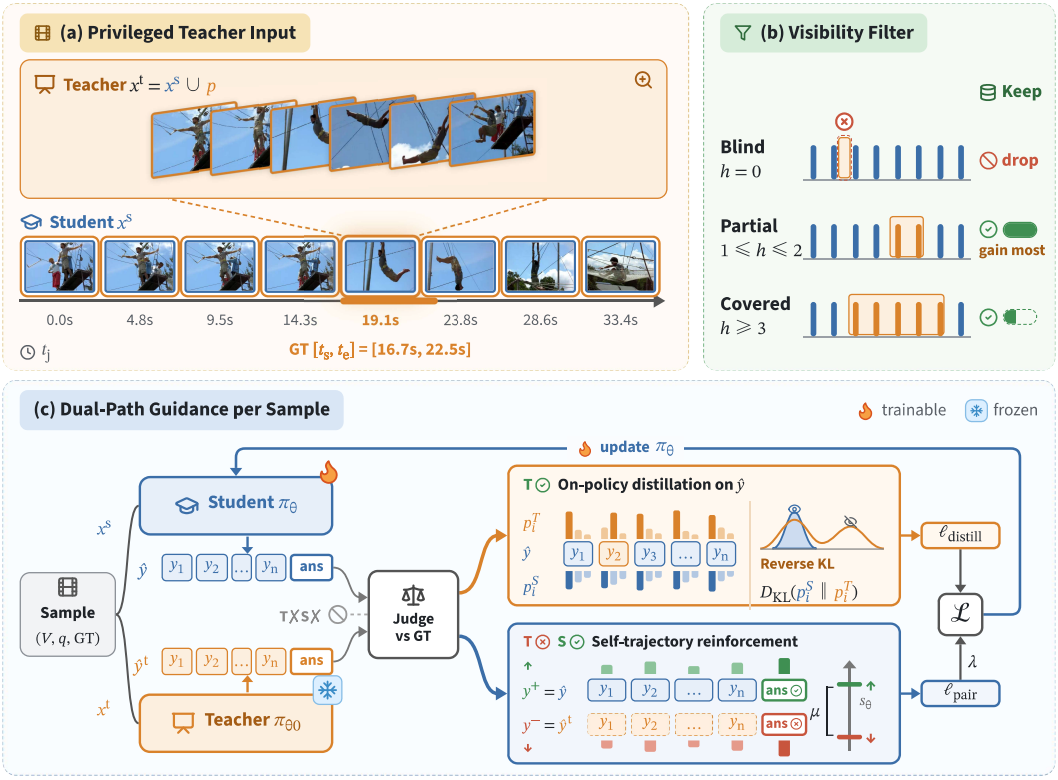

Figure 3: **Overview of our framework.** (a) The teacher sees the student's sparsely sampled frames plus frames sampled densely inside the ground-truth segment. (b) Samples are kept or dropped based on how many of the student's frames fall inside the ground-truth segment. Partially visible samples benefit most and make up most of the training data. (c) The student is optimized along two paths, chosen by whether the teacher's and the student's answers are correct.

Guided by the analysis in Section 2, we propose **LENS**, a framework that makes on-policy self-distillation aware of both *student visibility* and *teacher reliability* (Figure 3).

3.1 Problem Setup

Given a video V and a question q, the student model π_θ receives a sparsely sampled frame sequence $\mathbf{x}^s = \mathrm{Sample}(V, k)$ and generates a response trajectory $\hat{y} \sim \pi_\theta(\cdot \mid \mathbf{x}^s, q)$ via on-policy rollout.

A privileged teacher π_{θ_0} is initialized from the same checkpoint but **kept frozen** throughout training. It conditions on augmented input $\mathbf{x}^t = [\mathbf{x}^s;\ p]$, where p denotes privileged information, and evaluates the student-generated trajectory to produce token-level distributions at each position i. For notational convenience, we write p_i^S and p_i^T for the student's and teacher's next-token distributions conditioned on the shared prefix $\hat{y}_{<i}$, under their respective inputs $\mathbf{x}^s$ and $\mathbf{x}^t$. Standard OPSD minimizes the forward KL divergence $D_{\mathrm{KL}}(p_i^T \| p_i^S)$ summed over positions along the student's trajectory.

Following the findings in §2.1, we adopt the Contextualized Grounded configuration as the privileged teacher: it observes the student's k uniformly sampled frames augmented with m additional frames densely sampled within the GT temporal interval $[t_{\text{start}}, t_{\text{end}}]$. This provides fine-grained evidence in the critical region while preserving the student's global temporal context. Our method modifies *which* samples enter training and *how* they are supervised.

3.2 Learnable Privileged Distillation

At each training step, the student generates a trajectory $\hat{y}$ and the frozen teacher independently generates $\hat{y}^t$, both evaluated against the ground truth. Each sample passes through a visibility filter and a dual-path reliability router, which together determine whether and how it enters the training objective.

Visibility filtering. Distillation gains concentrate on evidence-starved samples and diminish for well-covered ones. When the student observes none of the decisive evidence, distillation may provide ungrounded supervision and encourage hallucinated reasoning. We discard samples with evidence coverage $h(k) = 0$ and filter out well-covered samples, retaining those where the student possesses at least partial grounding for the teacher's reasoning. In practice, the effective training set is dominated by partially starved samples.

Dual-path reliability routing. Samples passing the visibility filter are routed into one of two training paths by the correctness of the teacher and student trajectories:

- **Distillation path** (T✓): when the teacher is correct, its guidance is distilled into the student regardless of student correctness (§3.2.1).
- **Reinforcement path** (T✗S✓): when the teacher fails but the student succeeds, the student's own trajectory is reinforced against the teacher's incorrect one (§3.2.2).

3.2.1 Reliable Distillation

For samples where the teacher provides reliable guidance, we distill its token-level distribution into the student. Standard OPSD uses forward KL, $D_{\mathrm{KL}}(p_i^T \| p_i^S)$, which drives the student to cover all modes of the teacher—including those supported by privileged evidence that the student cannot observe. We instead adopt **reverse KL**:

$$\ell_{\text{distill}} = \sum_{i=1}^{|\hat{y}|} D_{\mathrm{KL}}\big(p_i^S \,\big\|\, p_i^T\big)\,. \tag{2}$$

Reverse KL weights the divergence by the student's own probability mass, providing a student-centered distillation signal: updates concentrate on tokens where the student already assigns non-negligible probability, rather than forcing coverage of the full teacher distribution. This preference complements the data-level visibility filtering but does not by itself guarantee that every high-probability student token is visually grounded. This objective applies to both corrective samples (T✓S✗), where it steers the student toward the teacher's correct reasoning, and concordant samples (T✓S✓), where it provides a calibrating signal on already-correct predictions.

3.2.2 SELF-TRAJECTORY REINFORCEMENT

For privilege-confounded samples, imitating the teacher would overwrite the student's correct knowledge. These samples instead offer a uniquely informative signal: the correct student trajectory $y^+ = \hat{y}$ and the incorrect teacher trajectory $y^- = \hat{y}^t$ form a natural pair of two plausible alternatives, exposing where their reasoning diverges. Contrasting them reinforces the student's discrimination at precisely the decision boundaries where it is most fragile. Both trajectories are scored under the **student's own sparse input** $\mathbf{x}^s$. We define the length-normalized trajectory score $s_\theta(y) = |y|^{-1} \sum_t \log \pi_\theta(y_t \mid \mathbf{x}^s, q, y_{<t})$ and optimize a pairwise margin objective:

$$\ell_{\text{pair}} = \tau \cdot \text{softplus}\left(\frac{\mu - s_\theta(y^+) + s_\theta(y^-)}{\tau} \right), \tag{3}$$

where μ is the target margin and τ is a temperature. The loss increases the relative likelihood of y^+ with diminishing pressure once the margin is satisfied.

3.3 TRAINING OBJECTIVE

Let $\mathcal{B}_{\text{distill}}$ denote the set of visibility-filtered, teacher-correct samples in the current batch, and $\mathcal{B}_{\text{pair}}$ the set of privilege-confounded samples. The overall objective is:

$$\mathcal{L} = \frac{1}{|\mathcal{B}_{\text{distill}}|} \sum_{i \in \mathcal{B}_{\text{distill}}} \ell^{(i)}_{\text{distill}} \; + \; \lambda \cdot \frac{1}{|\mathcal{B}_{\text{pair}}|} \sum_{j \in \mathcal{B}_{\text{pair}}} \ell^{(j)}_{\text{pair}}, \tag{4}$$

where λ balances the two terms. The procedure is summarized in Algorithm 1.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and evaluation. We train on DynFrame zhang2026dynframe, which provides temporally grounded evidence intervals for video question–answer pairs, making it suitable for visibility-aware training. With k=8 uniformly sampled student frames, training is dominated by samples with $h(8) \in \{1, 2\}$, yielding 5.8k examples in the partially starved regime. Each sample is further partitioned by teacher and student correctness for dual-path routing as described in Section 3.2. We evaluate on four benchmarks with temporal annotations: NExT-GQA xiao2024nextgqa, QA-Ego4D barmann2022qaego4d, CG-Bench chen2024cgbench, and S-EMBER wang2026sember, spanning short- to long-form videos in both offline and streaming settings.

Implementation and baselines. We experiment with Qwen3-VL-Instruct bai2025qwen3vl at 4B and 8B scales. The 2B model is reported in Appendix C.1 due to its weaker instruction-following, which use another evaluation prompt. The student observes 8 uniformly sampled frames; the teacher observes the same 8 frames plus 32 frames densely sampled within the GT interval, and is kept frozen throughout training. We compare against GRPO, SFT on ground-truth answers (SFT-GT), SFT on teacher trajectories (SFT-Teacher), and answer-privileged OPSD (OPSD-Answer) zhao2026self, all trained on the same data. Further training details are provided in Appendix B.

4.2 MAIN RESULTS

Consistent gains across benchmarks. Table 1 presents the main results. On the 8B backbone, LENS achieves the best performance on every benchmark, improving the base model by $+4.90$ points on average and outperforming the strongest baseline (OPSD-Answer) by $+2.19$. The baseline progression from GRPO through SFT variants to OPSD-Answer forms a monotonic staircase where each step introduces a richer form of supervision, and LENS extends this trend by making the privileged signal reliability- and visibility-aware. The same ranking holds on the 4B backbone ($+3.88$ on average), and LENS remains the strongest method on 2B (Appendix C.1), though with narrower absolute gains, suggesting that privileged supervision and model scaling are complementary.

Model	NExT-GQA		QA-Ego4D	CG-Bench	S-EMBER	Avg.	Δ
Frames	*8*	*16*	*64*	*64*	*64*		
Qwen3-VL-8B (Base)	66.95	68.88	45.70	32.89	21.93	47.27	—
+ *GRPO*	66.86	69.17	45.91	32.41	21.96	47.26	−0.01
+ *SFT (GT)*	67.73	69.98	45.08	32.57	22.84	47.64	+0.37
+ *SFT (Teacher)*	68.76	70.61	47.07	33.70	22.55	48.54	+1.27
+ *OPSD-Answer*	70.47	73.35	48.98	33.94	23.14	49.98	+2.71
+ *LENS (Ours)*	**73.04**	**75.40**	**51.52**	**34.70**	**26.18**	**52.17**	**+4.90**
Teacher	72.75	74.84	55.54	48.69	33.14	56.99	
Qwen3-VL-4B (Base)	65.05	66.70	44.50	29.63	20.92	45.36	—
+ *GRPO*	65.14	67.15	44.13	28.90	21.27	45.32	−0.04
+ *SFT (GT)*	67.24	68.85	43.42	30.11	22.62	46.45	+1.09
+ *SFT (Teacher)*	67.71	69.55	46.24	29.71	**23.57**	47.36	+2.00
+ *OPSD-Answer*	70.95	72.45	47.76	30.27	21.92	48.67	+3.31
+ *LENS (Ours)*	**71.31**	**73.40**	**47.99**	**30.33**	23.16	**49.24**	**+3.88**
Teacher	71.10	72.65	54.79	43.49	32.76	54.96	

Table 1: Main results on four video reasoning benchmarks. *Frames* denotes the number of evaluation frames per benchmark. The Teacher row reports the frozen base model evaluated with the same number of frames, sampled densely within the GT interval and sparsely outside it.

Surpassing the teacher and benchmark-specific gains. On NExT-GQA, the LENS student exceeds the privileged teacher with only uniformly sampled frames (73.04 vs. 72.75 at 8 frames; 75.40 vs. 74.84 at 16 frames), a result that standard distillation cannot produce, since the student's performance is nominally upper-bounded by the teacher. We attribute this to the self-trajectory reinforcement path: on privilege-confounded samples where dense observations mislead the teacher, LENS reinforces the student's own correct trajectory rather than imitating an incorrect one, converting potentially harmful samples into discriminative training signal. More broadly, the largest gains appear on benchmarks that demand precise temporal localization: NExT-GQA (+6.09) and S-EMBER (+4.25), while improvements are more moderate on CG-Bench (+1.81), which relies more on holistic comprehension. We analyze this variation through the evidence coverage distribution in §4.4.

	NExT-GQA		QA-Ego4D	CG-Bench	S-EMBER		
Training recipe	u8	u16	u64	u64	u64	Avg.	Δ
Base	66.95	68.88	45.70	32.89	21.93	47.27	–
(a) Components							
+ Distillation path (T✓)	70.52	72.50	48.65	33.15	24.62	49.89	+2.62
+ Reinforcement path (T✗S✓)	72.54	74.10	**51.60**	**35.03**	25.57	51.77	+1.88
+ Visibility filter (**LENS**)	**73.04**	**75.40**	51.52	34.70	**26.18**	**52.17**	+0.40
(b) Divergence of the distillation path (T✓ samples)							
Forward KL	70.39	72.30	46.20	31.96	**25.38**	49.25	+1.98
JSD (β=0.5)	70.43	**72.72**	47.86	**33.45**	24.64	49.82	+2.55
Reverse KL†	**70.52**	72.50	**48.65**	33.15	24.62	**49.89**	+2.62
(c) Supervision for T✗S✓ samples							
Distillation path	68.85	71.35	47.49	33.33	24.77	49.16	−0.73
Reinforcement path†	**72.54**	**74.10**	**51.60**	**35.03**	**25.57**	**51.77**	+1.88

Table 2: Ablation study on the 8B backbone. †Shared run with (a). Δ: change in Avg. from the preceding row in (a), from Base in (b), and from the first row of (a) in (c).

4.3 Ablation Study

Component contribution. Table 2 validates each component and design choice of LENS on the 8B backbone. The largest gain ($+2.62$) comes from distillation on teacher-correct samples, consistent with the premise that privileged temporal evidence provides effective guidance when the teacher is reliable. (Table 2a); The reinforcement path and visibility filtering contribute a further $+1.88$ and $+0.40$, confirming that both mechanisms improve upon standard distillation.

Design validation. Two design choices are validated by the remaining panels. The routing choice matters: privilege-confounded samples (T✗S✓) are noise for distillation but signal for reinforcement (Table 2c)—distilling them lowers accuracy by 0.73 points, whereas the reinforcement path turns the same samples into a $+1.88$ gain. For the divergence, reverse KL outperforms forward KL and JSD (Table 2b): since on-policy rollouts are sampled from the student distribution, reverse KL naturally aligns the optimization with the trajectories being generated, sharpening the student toward teacher modes it can already support.

4.4 Deeper Analysis

Gains concentrate on starved samples. Table 3 breaks down LENS's accuracy gain by evidence coverage $h(k)$. At the benchmark level, gain magnitude tracks the density of GT evidence: benchmarks where the GT interval occupies a larger fraction of the video (NExT-GQA: 8 s / 40 s; S-EMBER: 56 s / 4.3 min) yield consistently larger improvements, while CG-Bench, a long-video benchmark with sparser GT coverage (39 s / 27 min), shows more moderate gains. Within each benchmark, the main improvement concentrates in the starved regime ($h(k) \leq 2$), with peak positions varying by benchmark characteristics: short-GT benchmarks peak at $h(k) \in (0, 1]$, while benchmarks with longer evidence intervals such as S-EMBER peak at higher buckets ($h(k) \in (4, 8]$), as their evidence requires multiple frames to cover. Across all benchmarks, gains taper off beyond $h(k) > 8$, confirming the partially starved regime identified in Section 2.2 as the core benefit region.

Benchmark		GT (s)	Video (min)	Accuracy gain (pp) by $h(k)$ (0,1]	(1,2]	(2,4]	(4,8]	>8
Short videos								
NExT-GQA	k=8	7.96	0.67	**+6.91** (2200)	+5.93 (1703)	+5.32 (1091)	+4.83 (559)	— (0)
	k=16			**+8.94** (738)	+5.54 (1462)	+6.11 (1703)	+7.70 (1091)	+4.83 (559)
Long videos								
QA-Ego4D		9.35	30.40	**+6.38** (1960)	0.00 (137)	+1.72 (58)	— (33)	— (7)
CG-Bench		39.14	26.88	+1.29 (2863)	+2.31 (737)	+1.97 (964)	**+3.89** (334)	+3.57 (140)
Streaming videos								
S-EMBER		56.48	4.29	+1.27 (863)	+2.71 (813)	+5.51 (1234)	**+5.78** (1003)	+4.46 (3474)

Table 3: Accuracy gain (pp) of LENS over the base model by $h(k)$; sample counts in parentheses. GT, Video: mean duration of the GT interval (s) and of the video (min). Bold marks the largest gain per row; cells with <50 samples are omitted.

Privileged supervision repairs early reasoning. To understand where privileged trajectories affect reasoning, we analyze the divergence position between repaired and original student trajectories (Figure 4). 71.5% of repaired trajectories diverge within the first 20% of generated tokens, indicating that privileged supervision primarily redirects reasoning direction rather than correcting the final answer. This is consistent with the "forking tokens" observation in OPD opd: once an incorrect interpretation is established early, subsequent tokens reinforce that trajectory, whereas repairing the early reasoning path cascades into a correct conclusion. A complementary context-probing experiment (Appendix C.3) suggests that dense context near key events teaches the student to draw stronger inferences from sparse frames by leveraging surrounding context, rather than merely improving anchor-frame detection.

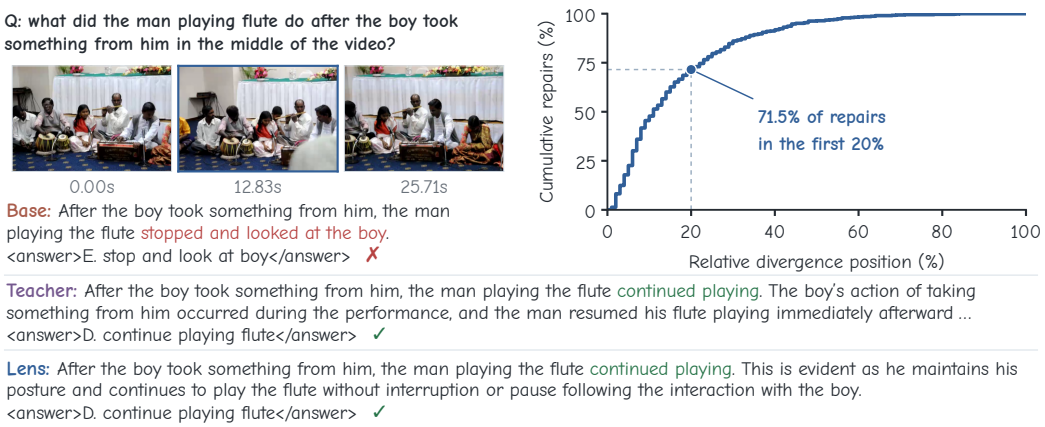

Figure 4: Divergence positions between repaired and original student trajectories. 71.5% of repairs occur within the first 20% of tokens; the example shows how correcting an early deviation cascades into a correct conclusion.

5 Related Work

On-Policy Distillation with Privileged Information Recent on-policy distillation (OPD) methods (Agarwal et al., 2024; Zhao et al., 2026; Ye et al., 2026) improve multimodal reasoning by leveraging stronger training-time supervision, including privileged information unavailable at inference. In vision-language reasoning, Vision-OPD (Yuan et al., 2026), RP-OPSD (Zhu et al., 2026), Visual-OPSD (Li et al., 2026b), and Thinking Without Images (Cai et al., 2026) construct privileged teachers from regional crops, high-resolution views, zoomed evidence, or visual thoughts to transfer stronger reasoning policies. These works demonstrate the effectiveness of privileged supervision, but largely assume that stronger observations always provide beneficial training signals. In contrast, we focus on the learnability of privileged evidence and explicitly model whether a grounded teacher is reliable and whether its privileged observations are observable from the student's sparse inputs.

Distillation for Video Reasoning Extending OPD to video understanding introduces additional challenges including long temporal dependencies, sparse evidence distribution, and spatio-temporal grounding. Video-OPD (Li et al., 2026a) and Video-OPSD (Wang et al., 2026b) adapt OPD to long-video reasoning through stronger temporal teachers and dense supervision, while VISD (Lin et al., 2026) further incorporates structured feedback on answer correctness, reasoning consistency, and grounding quality. Despite these advances, existing approaches primarily optimize reasoning outputs. Our work instead focuses on supervising evidence-selection behavior itself. By leveraging grounded spatio-temporal evidence, we teach the student not only what answer to produce, but also what evidence is worth learning from under limited observations.

6 Conclusion

This paper introduces LENS, a framework that revisits on-policy self-distillation for Video-LLMs through the lens of privilege learnability. We identify two complementary factors, teacher reliability and student visibility, that govern whether privileged guidance can be effectively absorbed by the student, and incorporate them into a unified training framework that filters samples by visibility and routes supervision by reliability. Experiments across four video reasoning benchmarks demonstrate consistent improvements over existing approaches, with the distilled student surpassing its privileged teacher on temporal grounding tasks, validating that learnability-aware design is essential for effective privileged distillation in video understanding.

AI USE STATEMENT

We used AI assistants throughout this work. Claude Code and OpenAI Codex were used for writing and debugging experiment code. GPT-6-Astra (OpenAI) and Claude Opus 5.5 (Anthropic) were the primary models used for prose drafting, proofreading, and polishing the paper writing. AI tools were also used for generating figures. All scientific contributions—including the research idea, experimental design, analysis, and interpretation of results—are entirely the work of the authors.

ETHICS STATEMENT

This work focuses on improving the training of video large language models through on-policy self-distillation, using publicly available video QA benchmarks. We do not foresee specific ethical concerns arising from the proposed method. All datasets used in our experiments are publicly released benchmarks with established usage licenses.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed training hyperparameters and implementation details in Appendix B. Our method builds on the publicly available Qwen3-VL model family and the DynFrame dataset. We will release our code and trained model checkpoints upon acceptance.

REFERENCES

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

Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, Tat-Seng Chua, et al. Opd-v: Visual on-policy self-distillation with modality balance. *arXiv preprint arXiv:2608.05131*, 2026.

Yishuo Cai, Jiahui Liu, Yuanxin Liu, Haobo Deng, Linli Yao, Yuhao Zheng, Kun Ouyang, Zhimo Li, Ziyue Wang, Xu Sun, et al. Thinking without images: Internalizing visual manipulation with on-policy self-distillation. *arXiv preprint arXiv:2606.08719*, 2026.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99114–99137, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

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

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*, 2026a.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *arXiv preprint arXiv:2606.18974*, 2026b.

Hao Lin, Kunyang Lv, Xu Jiang, Jingqi Tian, Zhongjing Du, Jiayu Ding, Qiaoman Zhang, and Hongbo Jin. Visd: Enhancing video reasoning via structured self-distillation. *arXiv preprint arXiv:2605.06094*, 2026.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN-o1: Reasoning-enhanced audio-visual large language model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57808–57827, 2025.

Kaishen Wang, Dongdi Zhao, Yijun Liang, Dingqiang Ye, Ruibo Chen, Heng Huang, and Di Fu. Where to look matters: On-policy self-distillation for long-video understanding. *arXiv preprint arXiv:2608.25356*, 2026a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2025.

Peng Zhang, Guanghao Zhang, Wanggui He, Longxiang Zhang, Mushui Liu, Yan Xia, Zhenhao Peng, Weilong Dai, Jinlong Liu, Haobing Tang, Le Zhang, Hao Jiang, and Pipei Huang. DynFrame: Adaptive reasoning-driven multimodal framework with dynamic frame augmentation for complex video understanding. *arXiv preprint arXiv:2605.26680*, 2026. URL `https://arxiv.org/abs/2605.26680`.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. Rp-opsd: Resolution-privileged on-policy self-distillation for multimodal large language models. *arXiv preprint arXiv:2607.24447*, 2026.

A Training Data

We construct the LENS training set by sampling from DynFrame (Zhang et al., 2026), a temporally grounded video QA dataset with annotated evidence intervals. All questions are answered by the same Qwen3-VL-8B checkpoint in both the student and teacher roles, differing only in the input frames. The construction proceeds in three steps.

Step 1: Candidate validation. Each candidate carries a video path, a question with multiple-choice options, the correct answer, and a temporal evidence interval $[t_{\text{start}}, t_{\text{end}}]$. We verify that all required fields are present and that the video is accessible with valid metadata (fps, duration, resolution). Candidates with missing fields or corrupted video files are discarded.

Step 2: Frame sampling. For each valid candidate, we construct paired student–teacher inputs. The student receives $k{=}8$ frames sampled uniformly across the video duration; the teacher receives the same k frames augmented with $m{=}32$ frames densely sampled within $[t_{\text{start}}, t_{\text{end}}]$. We record the actual frame indices for both views and compute the evidence coverage $h(k)$ (Eq. 1).

Step 3: Stratification and filtering. Samples with $h(8) = 0$ (no student frame falls within the GT interval) are discarded. We retain samples with $h(8) \in \{1, 2\}$—the partially starved regime where the student possesses minimal but non-zero evidence. The frozen base model independently answers each retained question via on-policy rollout in both roles; samples with unparseable responses from either role are filtered out. The correctness of each valid trajectory determines the dual-path routing (Section 3.2). The 4B and 2B models follow the same data construction pipeline. The training set has no overlap with any of the evaluation benchmarks in either videos or questions.

Dataset statistics. The resulting training set contains 5,762 samples, dominated by the $h(8) \in \{1, 2\}$ regime. Table 4a reports the partition by teacher and student correctness. The teacher is correct on 5,120 samples (88.9%), which enter the distillation path: 3,072 are concordant (T✓ S✓) and 2,048 are corrective (T✓ S✗). The remaining 642 samples are all privilege-confounded (T✗ S✓), entering the reinforcement path; no both-wrong cases (T✗ S✗) survive the unparseable-response filtering. The teacher-correct samples originate from multiple video sources aggregated in DynFrame (Table 4b). Each sample is stored as a JSON record containing the prompt, the video path with the student and teacher frame indices, the evidence interval and its coverage $h(k)$, the gold answer, and the routing label; reinforcement-path samples additionally store the paired correct and incorrect trajectories.

(a) Partition by teacher and student correctness

T	S	Case	Samples	Routing
✓	✓	Concordant	3,072	Distillation
✓	✗	Corrective	2,048	Distillation
✗	✓	Privilege-confounded	642	Reinforcement
✗	✗	Both wrong	0	—
Total			5,762	

(b) Sources of teacher-correct samples

Source dataset	Samples
NExT-QA / NExTVideo	1,824
YouTube	1,565
STAR	686
ActivityNet	429
Charades	282
YouCook2	196
PerceptionTest	110
Ego4D	15
CLEVRER	13
Total	5,120

Table 4: Training set statistics on Qwen3-VL-8B with $h(8) \in \{1, 2\}$. (a) Partition by teacher (T) and student (S) correctness and the resulting routing. (b) Source distribution of the teacher-correct samples by underlying video dataset in DynFrame.

B Training and Evaluation Setup

B.1 Training

Training procedure. Algorithm 1 summarizes one training step of LENS.

Algorithm 1 One training step of LENS

Require: student π_θ, frozen teacher π_{θ_0}, batch $\mathcal{B}$ of (V, q) with GT interval $[t_{\text{start}}, t_{\text{end}}]$
$\mathcal{B}_{\text{distill}} \leftarrow \emptyset$, $\mathcal{B}_{\text{pair}} \leftarrow \emptyset$
for each sample in $\mathcal{B}$ **do**
  $\mathbf{x}^s \leftarrow \text{Sample}(V, k)$
  if $h(k) = 0$ **then continue** ▷ visibility filtering (Eq. 1)
  end if
  $\mathbf{x}^t \leftarrow [\mathbf{x}^s; p]$, with p the m frames densely sampled in $[t_{\text{start}}, t_{\text{end}}]$
  $\hat{y} \sim \pi_\theta(\cdot \mid \mathbf{x}^s, q)$, $\hat{y}^t \sim \pi_{\theta_0}(\cdot \mid \mathbf{x}^t, q)$
  if either answer is invalid **then continue**
  else if $\hat{y}^t$ is correct **then**
    add $\hat{y}$ to $\mathcal{B}_{\text{distill}}$ ▷ reverse KL, Eq. 2
  else if $\hat{y}$ is correct **then**
    add $(y^+{=}\hat{y},\ y^-{=}\hat{y}^t)$ to $\mathcal{B}_{\text{pair}}$ ▷ pairwise margin, Eq. 3
  end if ▷ both wrong: discarded
end for
Compute $\mathcal{L}$ over $\mathcal{B}_{\text{distill}}$ and $\mathcal{B}_{\text{pair}}$ (Eq. 4) and update θ

Training infrastructure. We implement our training pipeline using verl[1] with Fully Sharded Data Parallel (FSDP). We perform full-parameter fine-tuning, including the vision encoder. All experiments are conducted on single-node eight-GPU configurations using either NVIDIA A100 (40 GB) or H100 (80 GB) GPUs for both training and evaluation. Model parameters are stored in FP32, while forward computation uses BF16 mixed precision. Gradient checkpointing and CPU offloading of parameters and optimizer states are enabled to reduce GPU memory usage. The privileged teacher is initialized from the same base checkpoint as the student and remains frozen throughout training.

Hyperparameter configuration. Table 5 summarizes the optimization and generation settings for all compared methods. All methods use AdamW with a constant learning rate of 10^{-6}, a global batch size of eight prompts, and 721 optimizer updates (one epoch). Generation settings are method-specific: GRPO samples eight responses per prompt, whereas OPSD-Answer and LENS sample

[1] https://github.com/volcengine/verl

one. For distillation, we retain the teacher's top-100 vocabulary entries and aggregate the remaining probability mass into an additional tail category. Both OPSD-Answer and LENS use reverse KL divergence; LENS additionally incorporates pairwise reinforcement.

Parameter	GRPO	SFT-GT / SFT-Teacher	OPSD-Answer	LENS
Optimization				
Fine-tuning scope	Full param.	Full param.	Full param.	Full param.
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	10^{-6}	10^{-6}	10^{-6}	10^{-6}
LR schedule	Constant	Constant	Constant	Constant
Warmup ratio	0	0	0	0
Adam (β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Weight decay	0.01	0.01	0.01	0.01
Gradient clipping	1.0	1.0	1.0	1.0
Global batch size	8	8	8	8
Optimizer updates	721	721	721	721
Rollout and sequence				
Max student prompt len.	8,192	—[1]	8,192	8,192
Max teacher prompt len.	—	—	10,240	10,240
Max response length	256	—[1]	256	256
Responses per prompt	8	—	1	1
Sampling temperature	1.0	—	1.0	1.0
Top-p	1.0	—	1.0	1.0
Top-k sampling	Off	—	Off	Off
Update epochs per rollout	1	—	1	1
Objective-specific				
Distillation divergence	—	—	Reverse KL	Reverse KL
Distillation support	—	—	Top-100 + tail	Top-100 + tail
Ref. policy KL coeff.	0.001	—	Off	Off
PPO clip ϵ	0.2	—	—	—
Pairwise loss coeff. λ	—	—	—	1.0
Pairwise margin μ	—	—	—	0.25
Pairwise temperature τ	—	—	—	0.10

Table 5: Training configurations for the 8B experiments. [1]SFT methods train on fixed target sequences without on-policy rollout.

B.2 Evaluation

Evaluation benchmarks. We evaluate on four video question-answering benchmarks covering causal and temporal reasoning, long-video understanding, and streaming episodic retrieval. The evaluation annotations provide reference answers together with temporal information identifying answer-relevant moments or segments, enabling both answer accuracy scoring and evidence-coverage analysis. Table 6 summarizes the evaluation settings and principal question types. The number of evaluation frames for each benchmark is held fixed for all models, before and after training.

Annotation usage. During standard student evaluation, models receive only the question, answer options, and uniformly sampled video frames; no temporal annotations are provided as input. Reference answers are used solely for scoring, while temporal annotations support the evidence-coverage analysis (Section 4.4) and the separately reported privileged-teacher evaluation. For S-EMBER, frames are sampled only from the video prefix preceding the question timestamp, following the streaming evaluation protocol. Temporal annotations differ in granularity across benchmarks: NExT-GQA provides question-specific grounding intervals; QA-Ego4D includes start/end timestamps with some instantaneous moments; CG-Bench annotates one or more clue intervals of

Benchmark	#Q	Frames	Main question types	Answers	Temporal evidence
NExT-GQA	5,553	8 / 16	Causal reasoning (why/how); temporal relations (before, after, during)	5-way MC	Question-specific grounding intervals
QA-Ego4D	2,409	64	Egocentric episodic recall: object identity, location, attributes, past actions	4-way MC	Start/end timestamps; some instantaneous
CG-Bench	5,038	64	Entity, event, scene understanding; temporal/spatial reasoning; hallucination detection	5–8-way MC	One or more clue intervals
S-EMBER	7,387	64	Visual-detail recall, counting, duration estimation, action sequences, temporal ordering	5-way MC	Evidence start/end; question timestamp

Table 6: Evaluation benchmarks, question types, and available annotations.

varying length; and S-EMBER supplies both evidence timestamps and the time at which each question is asked.

C Additional Results

C.1 Results on Qwen3-VL-2B

The 2B model exhibits weaker instruction-following, frequently failing to produce the evidence-first format used for 4B and 8B evaluation. We therefore evaluate it with a simpler letter-only prompt (see Appendix D for both prompts). Table 7 reports the results.

Model	NExT-GQA		QA-Ego4D	CG-Bench	S-EMBER	Avg.	Δ
Frames	*8*	*16*	*64*	*64*	*64*		
Qwen3-VL-2B (Base)	66.30	68.90	50.60	30.10	24.40	48.06	—
+ *GRPO*	66.60	68.80	50.80	**31.00**	24.10	48.26	+0.20
+ *SFT (GT)*	**66.80**	69.10	48.40	28.80	23.70	47.36	−0.70
+ *SFT (Teacher)*	66.50	69.10	**51.40**	29.40	25.90	48.46	+0.40
+ *OPSD-Answer*	66.60	69.10	50.10	30.10	24.50	48.08	+0.02
+ LENS (Ours)	**66.80**	**69.70**	51.30	29.70	**26.50**	**48.80**	**+0.74**
Teacher	72.40	74.70	60.80	43.70	36.00	57.52	

Table 7: Results on Qwen3-VL-2B, evaluated with the letter-only prompt (Appendix D). Columns and the Teacher row follow Table 1.

C.2 Different Privileged Teachers

Table 8 reports OPSD with each of the four privileged teachers in Section 2.1 on all benchmarks, complementing the NExT-GQA results in Figure 1c.

C.3 Context Probing

The analysis in Section 4.4 shows that LENS improves reasoning on samples where only one or two student frames fall in the GT interval. A natural question is whether this improvement relies solely on better use of those anchor frames, or on the model's ability to leverage surrounding context.

We select 193 samples where the student has exactly one frame in the GT interval ($h(8) = 1$) and measure how different video manipulations affect the base model's token-level support for fixed teacher claims—specifically, content and temporal tokens in the first 20 positions of the reasoning

Privileged teacher	NExT-GQA		QA-Ego4D	CG-Bench	S-EMBER	Avg.	Δ
Frames	*8*	*16*	*64*	*64*	*64*		
Qwen3-VL-8B (Base)	66.95	68.88	45.70	32.89	21.93	47.27	—
Answer	69.86	72.01	**48.68**	33.24	23.11	49.38	+2.11
Semantic	68.85	71.01	47.32	**34.21**	23.88	49.05	+1.78
Grounded	69.51	72.07	46.45	33.13	22.91	48.81	+1.54
Contextualized Grounded	**70.52**	**72.50**	48.65	33.15	**24.62**	**49.89**	**+2.62**

Table 8: OPSD with different privileged teachers on Qwen3-VL-8B (Section 2.1). All rows share the same student input; only the teacher's privileged information differs. These runs use only the distillation path (teacher-correct samples); privilege-confounded negative pairs are not included.

trace. Table 9 reports the drop in mean log-probability (nats) under four interventions relative to the original 8-frame input.

Intervention	Anchor	Context	Order	Video
Δ (nat)	+0.058	**+0.647**	+0.149	+1.388

Table 9: Support drop for fixed teacher claims under video manipulation (N=193, $h(8)$=1). Each value: decrease in mean log-prob (nats) when the anchor is mean-filled, repeated to replace all frames, non-anchor frames reversed, or video removed.

Removing the anchor frame itself causes a negligible drop ($+0.058$ nat), whereas removing all surrounding context—by repeating the anchor frame eight times—produces a far larger degradation ($+0.647$ nat). Disrupting temporal order among non-anchor frames causes a moderate effect ($+0.149$ nat), and removing the video entirely results in the largest drop ($+1.388$ nat). These results suggest that the teacher's grounded reasoning relies more on contextual frames than on the anchor frame alone. This pattern is consistent with the view that the capability LENS enhances is not solely better evidence detection at the GT frame, but also the ability to **construct coherent context from sparse observations**—using surrounding frames to situate, interpret, and reason about the limited evidence the student can directly observe.

D Evaluation Prompts

We use an evidence-first prompt for Qwen3-VL-8B and Qwen3-VL-4B, and a letter-only prompt for Qwen3-VL-2B. Each input is a single user message containing sampled video frames followed by the text below. Five options are shown for illustration; the option list and, for 2B, the allowed letters are adapted to each question. Assistant panels illustrate the requested output format and are not part of the input.

Qwen3-VL-8B / 4B: Evidence-first

User

```
[Video frames]
[Question]
Options:
A. [Option A]
B. [Option B]
C. [Option C]
D. [Option D]
E. [Option E]
Briefly verify the relevant visual evidence and its order. Use only
relative temporal words such as before, after, then, or during; do
not state seconds, timestamps, or frame numbers. Finish by writing
the full text of the correct option inside <answer>...</answer>. Do
not answer with only an option letter.
```

Assistant *Output format*

```
[Brief visual evidence and temporal order]
<answer>[Full text of the selected option]</answer>
```

Qwen3-VL-2B: Letter-only

User

```
[Video frames]
Select the best answer to the following multiple-choice question
based on the video. Respond with only the letter (A, B, C, D, or E)
of the correct option.
Question: [Question]
A. [Option A]
B. [Option B]
C. [Option C]
D. [Option D]
E. [Option E]
The best answer is:
```

Assistant *Output format*

```
[A single option letter]
```