OWLET: ON-POLICY WEIGHTED LEARNING VIA EVIDENCE TRANSFER FOR VIDEO LARGE LANGUAGE MODEL POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training has become a key stage for eliciting reasoning capabilities in Video Large Language Models (Video-LLMs). Existing post-training methods mainly follow two paradigms: reinforcement learning (RL) and on-policy distillation (OPD). However, they do not fully exploit visual evidence for effective training. RL supervises the policy with a single trajectory-level reward, which is too coarse-grained for video reasoning: the same response-level advantage is applied to every generated token, disregarding that the tokens of a video response are grounded in the visual stream to very different degrees. OPD provides token-level supervision through output-distribution matching, but lacks direct supervision of the spatiotemporal representations that produce those outputs. We observe across extensive experiments that the same model reasons more accurately when its video view combines global temporal context with question-relevant details. Comparing the same student-generated response under this view and uniform sampling reveals changes in token likelihoods and internal representations, providing a training signal for evidence use. Building on this observation, we propose OWLET, an evidence-privileged RL method that derives training signals from a richer teacher view. This view couples uniform global context with question-conditioned local windows, providing visual evidence without requiring a larger model. To better exploit this privileged visual evidence, OWLET introduces (i) **evidence-aware credit redistribution**, which compares token likelihoods across the two views and redistributes positive response-level advantages according to their relative evidence support; and (ii) **spatiotemporal state alignment**, which aligns the two views' hidden states on verifier-accepted responses. To facilitate training at scale, we release Video-Evidence-96K, containing 95,705 automatically curated question-conditioned temporal-evidence records over 21,832 videos. Experiments on the Qwen3.5 and Qwen3-VL families, ranging from 0.8B to 27B parameters, demonstrate the effectiveness of OWLET, yielding up to 10.9 point accuracy improvement over the vanilla model.

1 INTRODUCTION

Recent advances in video large language models (Video-LLMs) (Team, 2026; Google, 2026; OpenAI, 2025a;b) have extended large language models to the video domain, achieving strong downstream performance (Tang et al., 2025; Zhang et al., 2023). However, answering a video question often requires more than recognizing what a clip contains. The evidence that determines the answer may be concentrated in a few key frames scattered among thousands (Wang et al., 2026b), or span the video as a whole, requiring an understanding of event order, duration, and changes over time (Xiao et al., 2021). The central post-training question is how to make the model use visual input more effectively.

To this end, two on-policy paradigms have been widely adopted for reasoning-oriented post-training (Song & Zheng, 2026). Reinforcement learning (RL) samples responses from the current policy and reinforces those accepted by a verifier, keeping optimization on the distribution the model uses at test time (Shao et al., 2024). On-policy distillation (OPD) preserves this sampling scheme but replaces the scalar reward with a teacher that re-scores the student's responses (Li et al., 2026b). How-

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: Illustration of how OWLET improves video reasoning through evidence-guided post-training. Uniform frame sampling misses the temporal transitions needed to answer the question, whereas OWLET learns from a correct evidence-privileged teacher.

ever, these paradigms do not fully exploit visual evidence for effective training, leaving two aspects of visual supervision underexplored: **(i)** RL supervises the policy with a single trajectory-level reward, which is too coarse-grained for video reasoning. The same response-level advantage is applied to every generated token, disregarding that the tokens are grounded in the visual stream to very different degrees. **(ii)** OPD provides token-level supervision through output-distribution matching, but lacks direct supervision of the spatiotemporal representations that produce those outputs. This omission is particularly vital for video reasoning, where cues from distant frames need to be integrated into an internal account of the temporal structure: matching the output distributions alone does not ensure that the student acquires representations formed by the teacher (please refer to Section 3 for details).

Addressing these limitations calls for supervision that reveals both which tokens benefit from visual evidence and how that evidence is organized into spatiotemporal representations. We therefore construct an evidence-privileged *teacher view* that couples a uniform global context with question-conditioned local windows, preserving the video's overall structure while resolving relevant moments in greater detail as shown in Figure 1. Our motivated experiments in Section 3 show that the same model reasons more accurately when it combines global temporal context with question-relevant details. Comparing the same student-generated response under this richer view and uniform sampling reveals changes in visual token support and internal representations. These differences arise from the contribution of visual evidence and provide supervision without requiring a larger model.

Building on this observation, we propose OWLET, an evidence-privileged on-policy RL framework to enhance video-LLM reasoning capabilities. The student samples responses from uniformly sampled frames, and the teacher provides supervision through two complementary components. **(i) Evidence-aware credit redistribution** compares token likelihoods under the two views and redistributes positive response-level advantages into token-level advantages according to their relative evidence support. The verifier thus determines the direction of learning, while the teacher view guides the allocation of positive credit within each response. **(ii) Spatiotemporal state alignment** transfers the spatiotemporal representations formed under the teacher view and aligns the response-end hidden states of the two views on verifier-accepted responses. Both components operate on responses sampled by the student itself, keeping supervision tied to its own reasoning contexts. To facilitate training at scale, we release **Video-Evidence-96K**, containing 95,705 automatically curated question-conditioned temporal-evidence records over 21,832 videos. Our contributions are as follows:

- **A large-scale temporal-evidence resource.** We release Video-Evidence-96K, 95,705 automatically curated question-conditioned temporal-evidence records over 21,832 videos, which provides the global tracks and question-conditioned windows needed to construct the teacher view at scale.
- **Evidence-privileged on-policy learning.** We propose OWLET, which turns a temporally richer teacher view into fine-grained supervision for video-LLM post-training, through evidence-aware credit redistribution at the token level and spatiotemporal state alignment at the response level.
- **Consistent and versatile gains.** Extensive experiments demonstrate the effectiveness of OWLET, yielding up to 10.9 point accuracy improvement over the vanilla model.

2 PRELIMINARY

Video-LLM Post-Training. A video-LLM answers a question q about a video $\mathcal{V}$. Since $\mathcal{V}$ contains more frames than the model can process, it is reduced to a *view*: a sequence of frames that is encoded into visual tokens and prepended to q. We use v for the view the student receives and $\tilde{v}$ for the teacher. Conditioned on a view, the policy π_θ generates a response $y = (y_1, \ldots, y_{|y|})$ autoregressively.

Reinforcement Learning with Verifiable Rewards. RL has become the dominant route to reasoning ability, from preference optimization with policy-gradient algorithms (Ouyang et al., 2022; Schulman et al., 2017) to verifiable rewards, where a rule checks the final answer and no learned critic is required (Shao et al., 2024). For each question, a group of G responses $\{y^i\}_{i=1}^G$ is sampled from the current policy under the student view, each receives a scalar reward $R(q, y^i) \in \{0, 1\}$ from the verifier, and the advantage is estimated from the relative rewards within the group:

$$\mathcal{J}_{\mathrm{RL}}(\theta) = \mathbb{E}_{\substack{q\sim P(Q),\\ \{y^i\}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q,v)}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|y^i|}\sum_{t=1}^{|y^i|} \min\Big(\rho_t^i(\theta)\, A^i,\ \mathrm{clip}\big(\rho_t^i(\theta), 1-\epsilon, 1+\epsilon\big)\, A^i\Big) \right], \quad (1)$$

where $\rho_t^i(\theta) = \pi_\theta(y_t^i \mid q, v, y_{<t}^i) \,/\, \pi_{\theta_{\mathrm{old}}}(y_t^i \mid q, v, y_{<t}^i)$ is the importance ratio and $A^i = \big(R(q, y^i) - \mathrm{mean}(\{R(q, y^j)\})\big)/\mathrm{std}(\{R(q, y^j)\})$. Since sampling follows the evolving policy, training stays on the distribution the model uses at test time. However, the supervision is a single scalar per trajectory: every token is reinforced by the same reward regardless of how much it depends on the video.

On-Policy Self-Distillation. On-policy distillation replaces the scalar reward with a teacher that re-scores the student's own responses (Agarwal et al., 2024; Lu & Lab, 2025). OPSD removes the external teacher: the two policies are initialized from the same weights, and what separates them is privileged information withheld from the student, so that any advantage the teacher holds cannot be attributed to extra capacity (Hübotter et al., 2026; Shenfeld et al., 2026; Zhang et al., 2026a). The response is still sampled from the student, so the supervision is applied exactly on the states the student visits, but the objective becomes a token-level divergence to the teacher rather than a reward:

$$\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_{\substack{q\sim P(Q),\\ y\sim\pi_{\mathrm{stu}}(\cdot|q,v)}} \left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\mathrm{KL}}(\pi_{\mathrm{stu}}(\cdot \mid q, v, y_{<t}) \,\|\, \pi_{\mathrm{tea}}(\cdot \mid q, \tilde{v}, y_{<t})) \right]. \quad (2)$$

This yields a dense signal at every position, yet it remains an output-space objective: Equation 2 constrains only the next-token distributions, and weights all positions equally through the uniform factor $1/|y|$. Section 3.2 shows that both properties are consequential for video, and OWLET addresses them by reshaping the per-token weight in Equation 1 and by adding a term that acts upstream of the distributions compared in Equation 2.

3 MOTIVATED EXPERIMENTS

3.1 CONSTRUCTING THE EVIDENCE-PRIVILEGED TEACHER VIEW

Our design of OWLET is motivated by a simple question: *how should a visually privileged view be constructed so that it actually improves reasoning?* A naive answer is a view assembled purely from key frames. However, as shown in Figure 2, a view that contains only the annotated key frames may focus better on the local event that answers the question, but it also loses the surrounding context needed to understand the video as a whole. We therefore compare four input views on 1,000 held-out samples from Video-Evidence-96K, each assembled from exactly 64 frames so that the views differ only in which moments are shown: the *vanilla* samples 64 frames uniformly over the whole video; *key frames* takes the 64 frames ranked highest by question relevance; *key windows* takes the 16 highest-ranked frames and expands each into a disjoint four-frame window; and *global context with local windows* combines 32 uniform frames with disjoint four-frame windows around the 8 highest-ranked frames. We evaluate both overall accuracy and a subset of *global-temporal questions*, which require information spread across the video. These are questions about the overall purpose of a video, or about the causal relation between events (e.g., how a state evolves from beginning to end). Detailed categorization and composition of Video-Evidence-96K are in Appendix E.1.

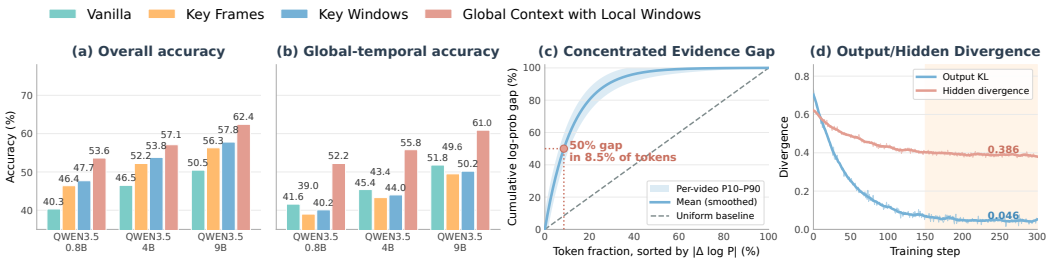


Figure 2: **(a, b)** Evidence-only views improve overall accuracy but lose global-temporal accuracy; only pairing global context with local windows improves both. **(c)** Half of the teacher–student log-probability gap concentrates on 8.5% of response tokens, so spreading training credit uniformly dilutes the evidence signal. **(d)** Output-only distillation collapses the output KL, yet the hidden divergence plateaus: output supervision alone does not effectively align spatiotemporal representations.

Figure 2(a,b) illustrates the trend across Qwen3.5-0.8B, 4B, and 9B. Using only key frames or windows improves the overall score over the vanilla, suggesting that focused visual evidence is generally beneficial. However, these focused views perform worse on global-temporal questions. The global context with local windows view performs best on both metrics: global frames preserve the broad temporal structure, while local windows provide details around question-relevant moments.

3.2 Supervision Gaps Unique to Video Post-Training

Constructing the teacher view settles what the teacher should see, and we further need to design how the resulting advantage is transferred to the student. Since the teacher and the student are initialized with the same weight, differences in token log-probabilities for the same response under the two views reflect differences in visual input. Using Qwen3.5-9B with a 64-frame student view as an example, we identify two supervision gaps that existing methods for video reasoning leave underexplored. Results for other models and frame budgets are in Appendix E.6.

(i) Token credit is uniformly distributed, yet video grounding is not. Compared with text-only reasoning, video reasoning can exhibit greater variation in how strongly response tokens depend on input evidence. Tokens naming a temporal order, an event location, or a cross-frame causal relation often depend strongly on visual evidence, while connectives and narrative scaffolding tend to depend less on it (Liu et al., 2026a). Figure 2(c) quantifies the resulting structure. For student responses collected on 1,000 held-out samples from Video-Evidence-96K, we re-score each one under the teacher view and rank its tokens by $|\Delta \log P|$ between the two views. The two passes share the same question, and the same generated prefix, and differ only in the frames shown; a large gap at a token therefore signals that predicting it relies on the additional visual evidence. Half of the total shift falls on **8.5%** of the tokens, far from the uniform baseline, and the 10th–90th percentile band shows the concentration is consistent across videos. Existing post-training does not effectively exploit this gap: RL copies one scalar reward to every token, and OPD matches the two output distributions position by position without asking which positions the video decides. Both dilute supervision exactly where evidence matters most. This motivates **evidence-aware credit redistribution**, which reads the per-token gap as a diagnostic of evidence-dependence and redistributes training credit accordingly.

(ii) Output-level alignment leaves spatiotemporal representations unsupervised. Before its first answer token, a video-LLM must compress thousands of frames into hidden states carrying its internal account of the video's temporal structure. That integration happens during visual encoding and cross-modal attention, upstream of any token generation and therefore beyond the reach of an objective defined on the token distribution. Figure 2(d) confirms the gap: once on-policy distillation converges, the output KL reduces to 0.046, yet the response-end hidden states of the two views stay far apart, with $1 - \mathrm{CKA}$ (Kornblith et al., 2019), a divergence index invariant to rotation and scaling of the representations, plateauing at 0.386 (please refer to Appendix D.6 for details). The student produces similar tokens without having internalized the spatiotemporal context the teacher view assembles. This motivates **spatiotemporal state alignment**, which aligns response-end hidden states at matched depths to transfer the representations output-level supervision cannot reach.

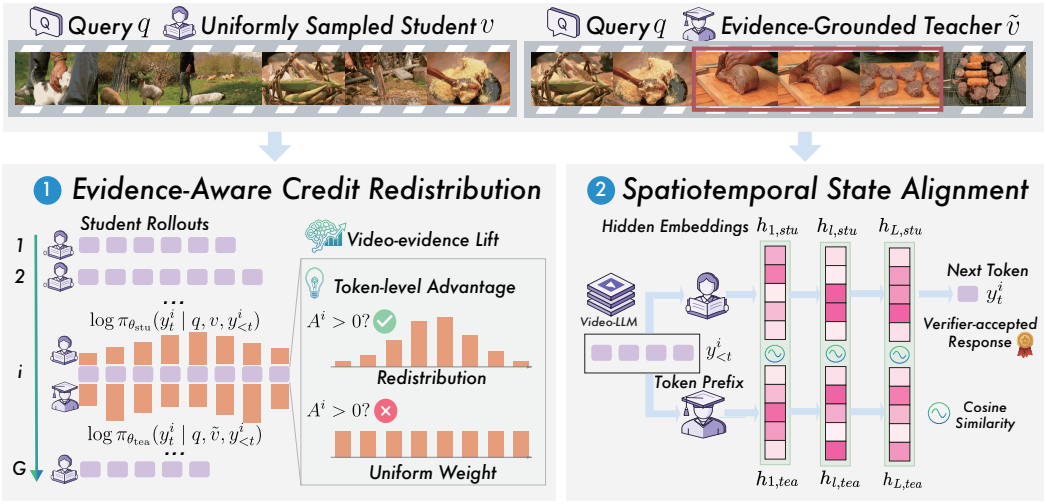


Figure 3: **Overview of OWLET.** The student samples responses from uniformly sampled frames, while the teacher scores the same responses using a view that combines global temporal coverage with question-conditioned local windows (red box). OWLET consists of two components: **(1)** Evidence-aware credit redistribution uses teacher–student token log-probability contrasts to redistribute positive advantages within each response; non-positive advantages remain unchanged; **(2)** Spatiotemporal state alignment aligns hidden states through cosine similarity on verifier-accepted responses.

4 Method

4.1 Question-Conditioned Video-Evidence Data

We construct Video-Evidence-96K from 104,257 video–question–answer records drawn from LLaVA-Video-178K (Zhang et al., 2024), NExT-QA (Xiao et al., 2021), Perception (Patraucean et al., 2023), and STAR (Wu et al., 2024a). We validate non-empty questions and answers, well-formed multiple-choice options with unique labels, valid video paths, and sufficient temporal candidates for evidence sampling. An automatic question-conditioned frame-relevance annotator (Wang et al., 2026b) identifies temporally localized evidence; we form local windows around the selected frames and combine them with global frames. The resulting training set contains 95,705 records over 21,832 videos. We checked for video-level overlap between the Video-Evidence-96K and evaluation benchmarks in Section 5, ensuring that no evaluation videos were included in our training data. The complete construction pipeline and dataset composition are in Appendix E.1.

4.2 Evidence-Aware Credit Redistribution

Video-evidence lift. Different parts of a video response refer to different temporal evidence. For example, a token describing a state change may depend on a short local transition, whereas a token used mainly to connect statements may depend little on any particular frame. A trajectory-level reward cannot express this distinction. ECR converts the teacher–student contrast into token-level weights that emphasize relative support within each response. Let $y^i = (y_1^i, \ldots, y_{T_i}^i)$ be the i-th response in a group of G student rollouts, sampled from $\pi_{\theta_{\text{old}}}(\cdot \mid q, v)$. Here T_i is the number of valid response tokens and $y_{<t}^i$ is the prefix before position t. We run the same response through both views and define the *token-level evidence contrast* as

$$\ell_t^i = \log \pi_{\bar{\theta}}(y_t^i \mid q, \tilde{v}, y_{<t}^i) - \log \pi_{\theta_{\text{old}}}(y_t^i \mid q, v, y_{<t}^i), \qquad t = 1, \ldots, T_i, \tag{3}$$

where $\bar{\theta}$ denotes the teacher parameters. A positive ℓ_t^i indicates that the global-plus-local view assigns more support to the observed token than the uniform view under the same text prefix. Tokens that express an evidence-sensitive transition or cross-frame relation can therefore exhibit a larger positive contrast, while purely discourse-oriented tokens can exhibit a smaller contrast. The quantity is used as a view-conditioned training signal; it does not require a teacher-generated trajectory. Appendix F.3 reports the proportions of rollouts with positive, negative, and zero advantages across training stages.

Within-response evidence profile. The evidence can vary across questions and responses, so ECR redistributes credit according to the relative evidence profile of each response. We first bound extreme contrasts and center them by their response mean:

$$z_t^i = \mathrm{clip}\big(\ell_t^i, -\log c, \log c\big), \qquad \mu^i = \frac{1}{T_i}\sum_{s=1}^{T_i} z_s^i,$$
$$\widetilde{A}_t^i = A_-^i + A_+^i \exp\big(z_t^i - \mu^i\big), \quad A_+^i = \max(A^i, 0), \quad A_-^i = \min(A^i, 0). \tag{4}$$

The clipping threshold c bounds the teacher-to-student likelihood ratio before centering; we use $c = 5$. The residual $z_t^i - \mu^i$ compares position t with the response's own evidence baseline. Equation 4 converts a single response-level advantage into a sequence of token-level advantages: each position t receives its own $\widetilde{A}_t^i$ according to the evidence contrast at that position. For $A^i > 0$, positions with above-baseline evidence receive larger advantages and positions with below-baseline evidence receive smaller ones. For $A^i \le 0$, $A_+^i = 0$ and $\widetilde{A}_t^i = A_-^i$, preserving the ordinary RL penalty. This design effectively lets the privileged view redistribute positive trajectory credit.

Evidence-weighted policy update. Let $R(q, y^i)$ be the verifier reward and A^i be the normalized $R(q, y^i)$ within G responses. The current policy's importance ratio for each sampled token is

$$\rho_t^i(\theta) = \frac{\pi_\theta(y_t^i \mid q, v, y_{<t}^i)}{\pi_{\theta_{\mathrm{old}}}(y_t^i \mid q, v, y_{<t}^i)}.$$

ECR then optimizes the following clipped objective, where ϵ controls the clipping range:

$$\mathcal{J}_{\mathrm{ECR}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\min\Big\{\rho_t^i(\theta)\widetilde{A}_t^i,\ \mathrm{clip}\big(\rho_t^i(\theta), 1-\epsilon, 1+\epsilon\big)\,\widetilde{A}_t^i\Big\}\right]. \tag{5}$$

4.3 Spatiotemporal State Alignment

Token-level weighting does not explicitly constrain the hidden representations produced by the visual computation. SSA addresses this gap by aligning hidden states produced from the student and teacher views for the same sampled response. Let $h_{\ell,v}^i$ and $h_{\ell,\tilde{v}}^i$ denote the response-end hidden states for response y^i after decoder layer ℓ under views v and $\tilde{v}$, respectively. The response-end state is the hidden state at the final valid response token after layer ℓ. Let $\mathcal{S}$ denote the decoder layers, and let $C = \{i : R(q, y^i) = 1\}$ be the verifier-accepted response indices, with cardinality $|C|$. We define

$$\mathcal{L}_{\mathrm{SSA}} = \frac{1}{\max(1, |C|)\,|\mathcal{S}|}\sum_{i\in C}\sum_{\ell\in\mathcal{S}}\left(1 - \frac{\left\langle h_{\ell,v}^i, h_{\ell,\tilde{v}}^i\right\rangle}{\left\|h_{\ell,v}^i\right\|_2\left\|h_{\ell,\tilde{v}}^i\right\|_2}\right). \tag{6}$$

When C is empty, we set $\mathcal{L}_{\mathrm{SSA}} = 0$. SSA provides a shared anchor after both models have processed the question and the same generated text. They can incorporate local appearance changes together with relations among distant events. Because the targets are computed on student-generated responses, SSA acts on the student's own reasoning contexts. It encourages the student to integrate uniformly sampled observations consistently with the local evidence and global temporal structure.

4.4 Joint Training and Inference

The two components are trained jointly by minimizing

$$\mathcal{L}_{\mathrm{OWLET}}(\theta) = -\mathcal{J}_{\mathrm{ECR}}(\theta) + \lambda_{\mathrm{SSA}}\mathcal{L}_{\mathrm{SSA}}(\theta), \tag{7}$$

where the negative sign converts policy maximization into loss minimization and λ_{SSA} controls the strength of state alignment. We set λ_{SSA} as 0.05. Both terms use responses sampled on-policy from the student view, while the teacher view supplies only token scores and representation targets.

Table 1: Accuracy (%) results on five video-understanding benchmarks for Qwen3.5 at 64 input frames. We **bold** the best result for each backbone and benchmark.

Backbone	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2	MLVU	Avg.
Qwen3.5-0.8B	Vanilla	49.4	51.8	48.9	15.8	41.9	41.6
	SFT	49.5	52.2	49.5	15.7	42.1	41.8
	GRPO	50.8	53.9	50.4	18.5	43.3	43.4
	OPSD	50.9	53.9	51.4	17.2	44.1	43.5
	OWLET	**54.9**	**58.6**	**55.9**	**19.9**	**45.2**	**46.9**
Qwen3.5-4B	Vanilla	64.7	67.4	58.0	24.7	50.3	53.0
	SFT	65.1	68.0	58.4	25.6	50.7	53.6
	GRPO	65.2	68.3	59.8	26.2	50.8	54.1
	OPSD	65.5	68.9	58.3	25.6	51.1	53.9
	OWLET	**68.3**	**72.2**	**63.4**	**28.5**	**58.8**	**58.2**
Qwen3.5-9B	Vanilla	65.0	67.1	61.9	25.6	52.7	54.4
	SFT	65.9	67.2	62.5	26.5	53.7	55.2
	GRPO	66.3	67.7	63.4	26.8	53.5	55.5
	OPSD	66.1	68.7	62.9	26.7	53.3	55.5
	OWLET	**70.7**	**72.8**	**65.2**	**28.9**	**61.6**	**59.8**
Qwen3.5-27B	Vanilla	71.4	70.2	64.2	30.4	61.5	59.5
	SFT	71.8	70.5	64.2	30.6	61.8	59.8
	GRPO	72.0	71.1	64.5	32.1	61.1	60.2
	OPSD	71.4	70.9	65.1	30.7	62.9	60.2
	OWLET	**76.1**	**76.2**	**68.8**	**34.6**	**66.7**	**64.5**

5 EXPERIMENTS

We evaluate OWLET on: (1) main comparisons with post-training baselines across four Qwen3.5 scales and five video benchmarks; (2) robustness across student frame budgets; (3) ablations of the teacher view, ECR, and SSA; (4) training dynamics case studies; and (5) generalization to Qwen3-VL in non-thinking and thinking modes (Appendix E.5). Appendices F.1–F.9 additionally include: comparison with on-policy distillation variants, different teacher parameterization settings, advantage and token-weight statistics, SSA-weight sensitivity, response examples of different post-trained models, training-time comparisons, response-length statistics, global-temporal accuracy after post-training, and extended training schedules.

5.1 EXPERIMENT SETUPS

Implementations We post-train Qwen3.5-0.8B/4B/9B/27B (Team, 2026) in non-thinking mode, each with a student frame budget $N \in \{16, 32, 64\}$. We also evaluate Qwen3-VL 2B/4B/8B in both non-thinking and thinking modes in Appendix E.5 to demonstrate the generalization of OWLET. We report accuracy on Video-MME (Fu et al., 2025), MVBench (Li et al., 2024a), LongVideoBench (Wu et al., 2024b), Video-MME-2 (Fu et al., 2026a), and MLVU (Zhou et al., 2025). We use eight NVIDIA H200 (140GB) GPUs and a learning rate of 1×10^{-6}. Details are in Appendices E.2 and E.3.

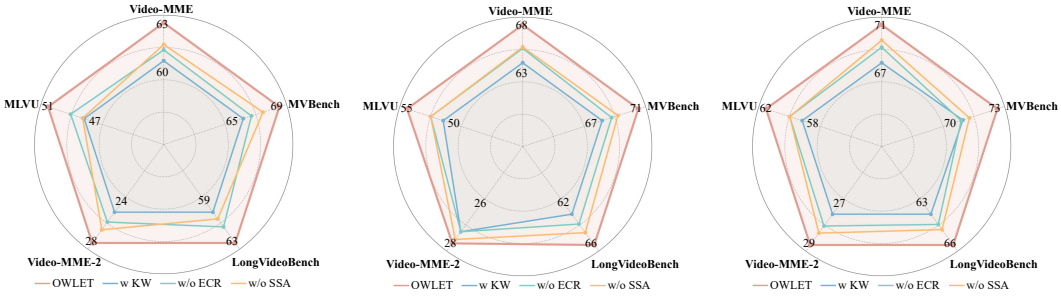
5.2 MAIN RESULTS

As shown in Table 1, OWLET outperforms SFT, GRPO, and OPSD on all five benchmarks for the four measured Qwen3.5 scales. We also observe that the largest per-benchmark gains typically appear on MVBench and MLVU, which stress fine-grained temporal understanding and long-video reasoning. This pattern matches the design of the teacher view: local windows supply the temporal detail that evidence-sensitive questions require, while the global track preserves the long-range structure that long-video benchmarks demand in Section 3.1.

Table 2 reports the macro-average at 16, 32, and 64 input frames. OWLET achieves the best result in all twelve combinations of model size and frame budget, demonstrating that its advantage is robust to the amount of visual evidence available to the student. Qwen3.5-4B with OWLET also exceeds the

Table 2: Macro-average accuracy (%) over the five benchmarks for Qwen3.5, at 16, 32, and 64 input frames. We **bold** the best result for each backbone and frame budget.

Method	Qwen3.5-0.8B			Qwen3.5-4B			Qwen3.5-9B			Qwen3.5-27B		
	16	32	64	16	32	64	16	32	64	16	32	64
Vanilla	37.8	40.1	41.6	48.0	49.8	53.0	49.0	51.3	54.4	54.3	56.8	59.5
SFT	38.6	40.9	41.8	48.5	50.7	53.6	49.8	52.2	55.2	54.8	56.1	59.8
GRPO	39.6	41.9	43.4	49.8	51.8	54.1	50.5	52.9	55.5	55.7	57.9	60.2
OPSD	39.1	41.5	43.5	49.1	51.0	53.9	50.3	52.6	55.5	55.7	58.1	60.2
OWLET	**42.9**	**45.2**	**46.9**	**53.3**	**54.6**	**58.2**	**54.2**	**57.1**	**59.8**	**57.9**	**60.4**	**64.5**



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 4: Ablation results for Qwen3.5-9B across different frame budgets.

strongest 9B baseline at the same budget, indicating that the proposed evidence-guided supervision can improve the effectiveness of a smaller model. Appendix E.4 provides per-benchmark results at 16 and 32 frames, detailed question-group results for Video-MME 2, and the Video-MME duration splits. Beyond the Qwen3.5 family, we also post-train Qwen3-VL-2B/4B/8B in both non-thinking and thinking modes; OWLET again attains the highest macro-average in all settings. The detailed results and analyses are provided in Appendix E.5.

5.3 Ablation Study

To investigate the contribution of the proposed components, we conduct an ablation study on Qwen3.5-9B across 16, 32, and 64 input frames. We consider three variants: OWLET with a key-window-only teacher view (w/ Key Windows), OWLET without evidence-aware credit redistribution (w/o ECR), and OWLET without spatiotemporal state alignment (w/o SSA). All variants share the training and evaluation protocol with the full model. As shown in Figure 4, removing any component degrades performance across the five benchmarks and frame budgets. Replacing the global-plus-local teacher view with key windows produces the largest average decrease, indicating that global temporal coverage is complementary to localized evidence. The full OWLET remains strongest at every budget, while the key-window variant remains the weakest on average, showing that additional local frames do not replace the benefit of combining global coverage with local evidence. Additional ablations across the remaining Qwen3.5 scales and Qwen3-VL modes are provided in Appendix E.7.

5.4 Case Study

To understand where the improvements of OWLET come from, we conduct case studies on Qwen3.5-9B with 64 frames, tracking how the supervision gaps identified in Section 3.2 evolve over training.

Exploration of Representation Dynamics. Figure 5(a–c) tracks the response-end hidden divergence ($1 - \mathrm{CKA}$), the output KL between the two views, and the rollout accuracy on the training set for OPSD and OWLET. We observe a clear dissociation between the two interfaces. OPSD rapidly minimizes its own objective—the output KL falls below 0.05—yet its hidden divergence plateaus near 0.39, the level of the diagnostic run in Figure 2(d), and its accuracy saturates near 61%. In

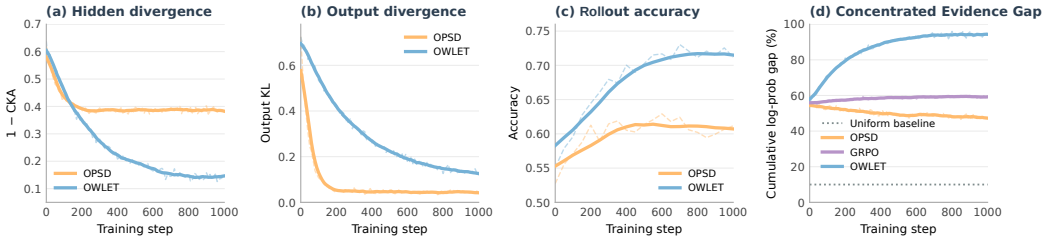


Figure 5: **Training dynamics.** Qwen3.5-9B with 64 student frames. **(a)** Hidden divergence, measured by response-end $1 - \mathrm{CKA}$, for OPSD and OWLET. **(b)** Output KL between student and teacher views for OPSD and OWLET. **(c)** Rollout accuracy on the training set for OPSD and OWLET. **(d)** Cumulative absolute log-probability gap at the top 10% of response tokens for OPSD, GRPO, and OWLET; the dotted line marks the uniform baseline. Diagnostic samples are spaced 10 steps apart and accuracy samples 50 steps apart. Thin dashed lines connect sampled trajectories; thick lines show rolling means.

contrast, OWLET steadily drives the hidden divergence from the shared starting point of 0.62 down to about 0.13 while its accuracy climbs from 54% to 73%; notably, its output KL also decreases to about 0.11 without ever being optimized directly. This dissociation confirms the gap identified in Section 3.2: matching output distributions alone leaves the upstream spatiotemporal representation misaligned, whereas explicitly aligning that representation through SSA closes both divergences and accompanies substantially higher rollout accuracy.

Exploration of Evidence-Gap Concentration. Figure 5(d) examines how each objective reshapes the structure of the evidence gap. We report the share of the total per-response gap $\sum_t |\Delta \log p_t|$ carried by its top 10% of tokens for post-trained methods; at initialization every method inherits the base model's concentration of about 56%. The three objectives then diverge. Under OWLET, the share climbs steeply over 90%: as ECR keeps routing positive credit toward evidence-dependent positions, the two views come to agree on scaffolding tokens, and nearly the entire residual gap collapses onto the small evidence-bearing set. GRPO, which never sees the teacher view during training, rises mildly to about 60%: reward-driven training strengthens visual grounding and thereby makes evidence tokens somewhat more view-sensitive, but nothing in a scalar reward concentrates credit further. OPSD moves in the opposite direction, easing down to about 48%: its KL gradient is largest precisely at the high-gap positions, so it preferentially erodes the concentrated head—the flattening that Section 3.2 shows discards the most informative part of the signal—while the erosion remains mild because the objective converges (panel b) well before the relative structure is gone. Together with panel (c), these trends indicate that OWLET converts the concentrated evidence signal into higher rollout accuracy rather than averaging it away.

6 Conclusion

In this paper, we introduce OWLET, a novel evidence-privileged reinforcement-learning framework that performs **o**n-policy **w**eighted **l**earning via **e**vidence **t**ransfer for video-LLM post-training. OWLET supervises each student rollout with a teacher view of the same video that pairs global temporal context with question-conditioned local windows, and closes the two supervision gaps identified by our motivated experiments through two key innovations: (i) evidence-aware credit redistribution, which converts the teacher–student token contrast into **token-level credit** while the verifier retains control of the learning direction; and (ii) spatiotemporal state alignment, which transfers the **spatiotemporal representations** that output-level supervision cannot reach. Trained on the proposed Video-Evidence-96K corpus, OWLET adds no model capacity and no test-time cost: the student answers from uniformly sampled frames in a single pass, with no evidence annotation or teacher computation. Experiments on Qwen3.5 and Qwen3-VL backbones from 0.8B to 27B parameters across five benchmarks demonstrate the effectiveness of OWLET, yielding up to 10.9 accuracy improvement over the vanilla model.

AI USE STATEMENT

In this work, we used LLM-based assistants to polish the wording of text. We have not used generative AI tools to produce research ideas or experimental designs. We also used generative AIs to assign global-temporal and local-evidence question labels, assess the quality of question-conditioned evidence annotations, and to generate the chain-of-thought traces for the SFT baseline. No generative AI tool had access to non-public data. We have reviewed all AI-assisted work: all AI-assisted text was checked, corrected, and approved by the authors, and all AI-assisted code was verified for correctness, with every reported result produced by the authors' own runs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work post-trains and evaluates video language models exclusively on publicly available academic datasets and benchmarks, used under their respective licenses; no private or personally collected data is involved, and no human-subject study was conducted beyond the authors' own verification of automatic annotations. Video-Evidence-96K is derived from existing public corpora, and we will document its provenance and filtering rules upon release so that upstream restrictions remain traceable. The proposed method improves how models use visual evidence and, to our knowledge, does not introduce risks beyond those already present in the underlying video-LLMs, such as biases inherited from public video corpora. As with any advance in video understanding, stronger temporal reasoning could in principle be misused in surveillance settings; our training recipe neither targets nor specifically facilitates such uses.

REPRODUCIBILITY STATEMENT

We document everything needed to reproduce our results in the paper and its appendix. The construction pipeline of Video-Evidence-96K, including source datasets, validation rules, evidence annotation, and splits, is described in Appendix E.1; the protocols behind the motivated experiments are specified in Appendices D.1–D.6; baselines, benchmarks, and the complete set of training and evaluation hyperparameters are listed in Appendices E.2 and E.3; the method itself is fully specified by Equations 3–7. We will release the training code, the dataset construction scripts, and the evaluation process upon publication.

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 *Proc. ICLR*, Vienna, 2024.

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

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: Reasoning transfer from LLMs to vision-language models via on-policy distillation. In *Proc. CVPR*, Denver, 2026.

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *Proc. CVPR*, Nashville, 2025.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. LLaVA-KD: A framework of distilling multimodal large language models. In *Proc. ICCV*, Honolulu, 2025.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. CG-Bench: Clue-grounded question answering benchmark for long video understanding. In *Proc. ICLR*, Singapore, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proc. AAAI*, Singapore, 2026.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. In *Proc. NeurIPS*, San Diego, 2025a.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025b.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025a.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025b.

Henghui Du, Chang Zhou, Xi Chen, and Di Hu. APPO: Attention-guided perception policy optimization for video reasoning. *arXiv preprint arXiv:2602.23823*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLM. In *Proc. NeurIPS*, San Diego, 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 LLM in video analysis. In *Proc. CVPR*, Nashville, 2025.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026a.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026b.

Google. Gemini 3.5 flash. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *Proc. ICLR*, Vienna, 2024.

Maxime Guigon, Lucas Dixon, and Michaël E Sander. A study on hidden layer distillation for large language model pre-training. *arXiv preprint arXiv:2605.11513*, 2026.

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.

Dongze Hao, Zhiwei Jin, Chen Chen, and Haonan Lu. Stabilizing on-policy distillation for MLLM reasoning with global normalization. *arXiv preprint arXiv:2606.09091*, 2026.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

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

ZhiYan Hou, Xinyu Tang, Hongyan An, Jianjin Zhang, Weizhen Wang, Yunyun Han, Gengsheng Li, Xiangzhao Hao, Haiyun Guo, Wenbin Hu, et al. DASH: Divergence-adaptive supervision horizons for on-policy self-distillation of reasoning models. *arXiv preprint arXiv:2608.06243*, 2026.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-LLM based video frame selection for efficient video understanding. In *Proc. CVPR*, Nashville, 2025.

Shiqi Huang, Ziyue Wang, Zhongrong Zuo, Han Qiu, Qi She, and Bihan Wen. EvoVid: Temporal-centric self-evolution for video large language models. *arXiv preprint arXiv:2605.21931*, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proc. EMNLP*, Austin, 2016.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proc. ICML*, Long Beach, 2019.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *Proc. ICLR*, Rio de Janeiro, 2026a.

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 *Proc. CVPR*, Seattle, 2024a.

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

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025a.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. VideoChat-Flash: Hierarchical compression for long-context video modeling. In *Proc. ICLR*, Rio de Janeiro, 2026d.

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

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026e.

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, et al. TreePO: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *arXiv preprint arXiv:2508.17445*, 2025b.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. *arXiv preprint arXiv:2607.21556*, 2026.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, et al. Unleashing hour-scale video training for long video-language understanding. In *Proc. NeurIPS*, San Diego, 2025.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026a.

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, et al. VideoTemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. *arXiv preprint arXiv:2602.07801*, 2026b.

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 Proc. ACL*, pp. 8731–8772, Bangkok, 2024.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Yujie Lu, Yale Song, William Wang, Lorenzo Torresani, and Tushar Nagarajan. VITED: Video temporal evidence distillation. In *Proc. CVPR*, Nashville, 2025.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Qitan Lv, Tianyu Liu, Wen Wu, Xuenan Xu, Bowen Zhou, Feng Wu, and Chao Zhang. Hippo: Accelerating video large language models inference via holistic-aware parallel speculative decoding. *arXiv preprint arXiv:2601.08273*, 2026.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3-Video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

OpenAI. GPT-5.1. https://openai.com/index/gpt-5-1/, 2025a.

OpenAI. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025b.

OpenAI. GPT-5.6 Preview System Card, 2026. URL https://deploymentsafety.openai.com/gpt-5-6-preview.

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. In *Proc. NeurIPS*, New Orleans, 2022.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. TimeSearch-R: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. In *Proc. ICLR*, Rio de Janeiro, 2026.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. DeepVideo-R1: Video reinforcement fine-tuning via difficulty-aware regressive GRPO. *arXiv preprint arXiv:2506.07464*, 2025.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. In *Proc. NeurIPS*, New Orleans, 2023.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael Ryoo. Understanding long videos with multimodal language models. In *Proc. ICLR*, Singapore, 2025.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. CRISP: Compressed reasoning via iterative self-policy distillation. *arXiv preprint arXiv:2603.05433*, 2026.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. LongVU: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-LLM reasoning via agent-of-thoughts distillation. In *Proc. CVPR*, Nashville, 2025.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-XL: Extra-long vision-language model for hour-scale video understanding. In *Proc. CVPR*, Nashville, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. MovieChat: From dense token to sparse memory for long video understanding. In *Proc. CVPR*, Seattle, 2024.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP 3: A training-free approach for list-wise frame selection in video-LLM. In *Proc. ICCV*, Honolulu, 2025.

Zhiquan Tan and Yinrong Hong. Self-supervised on-policy distillation for reasoning language models. *arXiv preprint arXiv:2605.17497*, 2026.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025. URL https://doi.org/10.48550/arXiv.2506.15220.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. ViCuR: Visual cues as recoverable privilege for multimodal on-policy distillation. *arXiv preprint arXiv:2606.05718*, 2026.

Chen Wang, Zhaochun Li, Jionghao Bai, Yining Zhang, Hexuan Deng, Ge Lan, and Yue Wang. Distilled reinforcement learning for LLM post-training. *arXiv preprint arXiv:2607.17247*, 2026a.

Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Liu, Jan Kautz, Jose M Alvarez, Lei Zhang, and Zhiding Yu. VideoITG: Multimodal video understanding with instructed temporal grounding. In *Proc. CVPR*, Denver, 2026b.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. LVBench: An extreme long video understanding benchmark. In *Proc. ICCV*, Honolulu, 2025a.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. ReTaKe: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*, 2024a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Proc. ECCV*, Milan, 2024b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-R1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proc. CVPR*, Nashville, 2025c.

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Jiang Xudong. Video-KTR: Reinforcing video reasoning via key token attribution. In *Proc. ICLR*, Rio de Janeiro, 2026c.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024a.

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

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking Kullback-Leibler divergence in knowledge distillation for large language models. In *Proc. COLING*, Abu Dhabi, 2025.

Yecheng Wu, Song Han, and Hai Cai. Lightning OPD: Efficient post-training for large reasoning models with offline on-policy distillation. *arXiv preprint arXiv:2604.13010*, 2026.

Haotian Xia, Zilin Xiao, Junbo Zou, Vicente Ordonez, and Hanjie Chen. Perception before reasoning: Dynamic latent reasoning for video understanding and question answering. *arXiv preprint arXiv:2608.04124*, 2026.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proc. CVPR*, Conference held virtually, 2021.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? Visually grounded video question answering. In *Proc. CVPR*, Seattle, 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.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. LLAVADI: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024.

Weiwen Xu, Jia Liu, Hou Pong Chan, Long Li, Deng Cai, Min Chen, and Hao Zhang. Beyond entropy: Correctness-aware advantage shaping via contrastive policy optimization. *arXiv preprint arXiv:2607.14614*, 2026a.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026b.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026a.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, et al. OPRD: On-policy representation distillation. *arXiv preprint arXiv:2606.06021*, 2026b.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. VCA: Video curious agent for long video understanding. In *Proc. ICCV*, Honolulu, 2025.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. LongVT: Incentivizing "thinking with long videos" via native tool calling. In *Proc. CVPR*, Denver, 2026c.

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

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. *arXiv preprint arXiv:2606.00564*, 2026.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025a.

Fangxu Yu, Weijia Xu, Michael Xu, Tianyi Zhou, and Zinan Lin. Weak-to-strong on-policy distillation. *arXiv preprint arXiv:2607.26246*, 2026.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-Voyager: Learning to query frames for video large language models. In *Proc. ICLR*, Singapore, 2025b.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. VideoExplorer: Think with videos for agentic long-video understanding. *arXiv preprint arXiv:2506.10821*, 2025.

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.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Temporal sentence grounding in videos: A survey and future directions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (8):10443–10465, 2023.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *Proc. ICCV*, Honolulu, 2025a.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *Proc. NeurIPS*, San Diego, 2025b.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. OPSDL: On-policy self-distillation for long-context language models. *arXiv preprint arXiv:2604.17535*, 2026a.

Yongheng Zhang, Zhipeng Xu, Hao Wu, Yinghui Li, Di Yin, Xing Sun, and Philip S Yu. Latent visual cache for video reasoning. *arXiv preprint arXiv:2607.02607*, 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.

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.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *Proc. CVPR*, Nashville, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Heqing Zou, Tianze Luo, Guiyang Xie, Victor, Zhang, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, and Huaijian Zhang. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. *arXiv preprint arXiv:2409.18938*, 2024.

Appendix Table of Contents

A Limitations **19**

B Formalized Algorithms of OWLET **19**

C Related Work **21**
C.1 Long-Video Understanding 21
C.2 Reinforcement Learning for Video-LLMs 21
C.3 On-Policy Distillation and Self-Distillation 22

D Details of Experiment Settings **23**
D.1 Evaluation Data and Decoding 23
D.2 Construction of the Four Input Views 23
D.3 Global-Temporal and Local-Evidence Questions 24
D.4 Teacher-View Comparison on Qwen3-VL Model Family 25
D.5 Measuring the Concentration of the Evidence Gap 27
D.6 Measuring Output and Hidden-State Alignment 27

E Experimental Data and Supplementary Experiments **28**
E.1 Construction of Video-Evidence-96K 28
E.2 Baselines and Benchmarks 30
E.3 Training Hyper-Parameters 31
E.4 Per-Benchmark Results and Detailed Analysis 32
E.5 Results on the Qwen3-VL Family 33
E.6 Additional Motivation Evidence Across Models and Frame Budgets 41
E.7 Ablation Across Models and Frame Budgets 43

F Additional Analyses **44**
F.1 Comparison with On-Policy Distillation Variants 45
F.2 Teacher Parameterization 47
F.3 Statistics of Advantage Signs and Highly Weighted Tokens 48
F.4 Sensitivity to the SSA Weight 50
F.5 Qualitative Comparison Across Post-Training Methods 51
F.6 Training-Time Comparison 53
F.7 Response Length during On-Policy Training 54
F.8 Global-Temporal Accuracy after Post-Training 55
F.9 Extended Training Schedules 56

A LIMITATIONS

We consider several limitations and directions for future work. **(i) Reliance on automatically localized evidence.** Video-Evidence-96K is built with an automatic question-conditioned frame-relevance annotator, which the privileged teacher view uses to construct local windows. Although the global track provides broad temporal coverage, inaccurate or noisy relevance estimates can still expose the teacher to misleading evidence and affect both ECR and SSA. Future work could incorporate confidence-aware or jointly learned evidence selection, and multiple evidence proposals. **(ii) Dependence on verifiable outcome supervision.** Our training objective uses binary answer verification, and our experiments focus on multiple-choice video benchmarks. ECR preserves the verifier's response-level learning direction, while SSA is applied to verifier-accepted responses; for open-ended, ambiguous, or subjective questions, an equally reliable verifier may be unavailable. Extending the framework to free-form video QA and richer temporal reasoning will require semantic or process-level verifiers, preference signals, or other forms of outcome supervision. **(iii) Training cost and fixed design choices.** Inference of OWLET has no additional teacher computation, but the training process requires an extra forward pass through a frozen teacher under the global-plus-local view. Scaling OWLET to much longer videos, larger model scales, or different non-Qwen model families remains to be studied. Adaptive teacher views, competence-aware teacher weighting, and layer- or token-level state alignment are promising directions. **(iv) Indirect evidence attribution.** ECR uses the teacher–student likelihood contrast as an indirect measure of how strongly each response token is supported by visual evidence; it does not provide an explicit causal attribution of the underlying visual cues. This lightweight signal nevertheless remains effective across model scales, frame budgets, and backbone families in our experiments, while future work could explore its interpretability and combine it with explicit grounding or process-level supervision.

B FORMALIZED ALGORITHMS OF OWLET

Algorithm 1 summarizes the training procedure of OWLET. For each training instance, the student first samples a group of on-policy responses from a uniformly sampled video view v. The same student-generated responses and prefixes are then evaluated by the rollout policy under v and by a fixed teacher initialized at θ_0 under a global-local view $\tilde{v}$. A verifier assigns binary rewards, from which group-normalized advantages are computed. Evidence-aware credit redistribution (ECR) contrasts the teacher and rollout token log-likelihoods, clips and centers the contrasts within each response, and redistributes positive advantage toward tokens better supported by the privileged view. In parallel, spatiotemporal state alignment (SSA) aligns response-end hidden states from the student and teacher on verifier-accepted responses, with gradients propagated only through the student. The student is optimized using the joint objective $(-\mathcal{J}_{\mathrm{ECR}} + \lambda_{\mathrm{SSA}}\mathcal{L}_{\mathrm{SSA}})$, while the teacher remains frozen throughout training. At inference time, only the trained student, the uniform video view, and the question are required.

Algorithm 1 Training Procedure of OWLET

Require: Training records $\mathcal{D} = \{(\mathcal{V}, q, a, \mathcal{W})\}$; initial checkpoint θ_0; frame budget N; group size G; training steps K; clipping range ϵ; ECR threshold c; SSA weight λ_{SSA}; decoder layers $\mathcal{S}$
Ensure: Trained student parameters θ

$\theta \leftarrow \theta_0, \bar{\theta} \leftarrow \theta_0$ ▷ the teacher $\pi_{\bar{\theta}}$ stays fixed
for $k = 1, \ldots, K$ **do**
  $\theta_{\mathrm{old}} \leftarrow \theta$
  Sample a batch $\mathcal{B} \subset \mathcal{D}$
  for each $(\mathcal{V}, q, a, \mathcal{W}) \in \mathcal{B}$ **do**
    $v \leftarrow \mathrm{UniformSample}(\mathcal{V}, N)$
    $\tilde{v} \leftarrow \mathrm{GlobalLocalView}(\mathcal{V}, \mathcal{W}; N)$
    Sample $\{y^i\}_{i=1}^{G} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid q, v)$; let T_i be the valid length of y^i
    $R(q, y^i) \leftarrow \mathrm{Verify}(q, y^i; a) \in \{0, 1\}$ for $i = 1, \ldots, G$
    $A^i \leftarrow \dfrac{R(q, y^i) - \mathrm{mean}_{j=1}^{G} R(q, y^j)}{\mathrm{std}_{j=1}^{G} R(q, y^j)}$ for $i = 1, \ldots, G$
    $C \leftarrow \{i : R(q, y^i) = 1\}$
    for $i = 1, \ldots, G$ **do**
      for $t = 1, \ldots, T_i$ **do**
        $\ell_t^i \leftarrow \log \pi_{\bar{\theta}}(y_t^i \mid q, \tilde{v}, y_{<t}^i) - \log \pi_{\theta_{\mathrm{old}}}(y_t^i \mid q, v, y_{<t}^i)$
        $z_t^i \leftarrow \mathrm{clip}(\ell_t^i, -\log c, \log c)$
      end for
      $\mu^i \leftarrow T_i^{-1} \sum_{s=1}^{T_i} z_s^i, \quad A_+^i \leftarrow \max(A^i, 0), \quad A_-^i \leftarrow \min(A^i, 0)$
      for $t = 1, \ldots, T_i$ **do**
        $\widetilde{A}_t^i \leftarrow A_-^i + A_+^i \exp(z_t^i - \mu^i)$
        $\rho_t^i(\theta) \leftarrow \dfrac{\pi_\theta(y_t^i \mid q, v, y_{<t}^i)}{\pi_{\theta_{\mathrm{old}}}(y_t^i \mid q, v, y_{<t}^i)}$
      end for
      Compute $\{h_{\ell,v}^i\}_{\ell \in \mathcal{S}}$ with π_θ on (q, v, y^i)
      Compute $\{h_{\ell,\tilde{v}}^i\}_{\ell \in \mathcal{S}}$ with frozen $\pi_{\bar{\theta}}$ on $(q, \tilde{v}, y^i)$, without gradients
    end for
    $\mathcal{J}_{\mathrm{ECR}}(q) \leftarrow \dfrac{1}{G} \sum_{i=1}^{G} \dfrac{1}{T_i} \sum_{t=1}^{T_i} \min\Big(\rho_t^i(\theta)\widetilde{A}_t^i, \mathrm{clip}(\rho_t^i(\theta), 1-\epsilon, 1+\epsilon)\widetilde{A}_t^i\Big)$
    $\mathcal{L}_{\mathrm{SSA}}(q) \leftarrow \dfrac{1}{\max(1, |C|)|\mathcal{S}|} \sum_{i \in C} \sum_{\ell \in \mathcal{S}} \Big[1 - \cos\Big(h_{\ell,v}^i, \mathrm{sg}(h_{\ell,\tilde{v}}^i)\Big)\Big]$
    $\mathcal{L}_{\mathrm{SSA}}(q) \leftarrow 0$ when $C = \varnothing$
  end for
  $\mathcal{J}_{\mathrm{ECR}} \leftarrow |\mathcal{B}|^{-1} \sum_{(\mathcal{V},q,a,\mathcal{W}) \in \mathcal{B}} \mathcal{J}_{\mathrm{ECR}}(q)$
  $\mathcal{L}_{\mathrm{SSA}} \leftarrow |\mathcal{B}|^{-1} \sum_{(\mathcal{V},q,a,\mathcal{W}) \in \mathcal{B}} \mathcal{L}_{\mathrm{SSA}}(q)$
  $\mathcal{L}_{\mathrm{OWLET}}(\theta) \leftarrow -\mathcal{J}_{\mathrm{ECR}}(\theta) + \lambda_{\mathrm{SSA}} \mathcal{L}_{\mathrm{SSA}}(\theta)$
  Update only θ by back-propagating $\mathcal{L}_{\mathrm{OWLET}}$; keep $\bar{\theta} = \theta_0$ fixed
end for
return θ

C Related Work

C.1 Long-Video Understanding

Video-LLMs extend image-language models to the temporal domain. A fixed number of frames is sampled from the video, turned into visual tokens, and placed before the question in a single input sequence (Bai et al., 2025; Team, 2026; Zhu et al., 2025; Chen et al., 2024; Tang et al., 2025). This is the standard pipeline for video question answering, and benchmarks now test it on longer videos and harder question types (Fu et al., 2025; Wu et al., 2024b; Wang et al., 2025a; Zhou et al., 2025; Chen et al., 2025). Its central constraint is that context length and compute cap how many frames the model can read, while the videos themselves keep growing. An hour of footage reaches the model as a few dozen frames, so the frame budget decides how much of the video is actually seen.

One line of work enlarges this budget by making frames cheaper. Token compression reduces the tokens spent per frame (Li et al., 2026d; Shen et al., 2024; Li et al., 2024b; Lv et al., 2026), memory mechanisms summarize past content into a compact state that is carried forward (Song et al., 2024; Zhang et al., 2025a; Shu et al., 2025), and redundancy-aware encoders drop content that repeats across neighboring frames (Wang et al., 2024a). Training recipes for hour-scale material push in the same direction (Ranasinghe et al., 2025; Lin et al., 2025). These methods change how much video reaches the model, but the supervision that teaches it what to do with the video is still the answer alone.

A second line asks which frames to keep. The evidence that decides an answer usually sits in a short span, and the rest of the video contributes little or actively distracts (Zou et al., 2024; Xiao et al., 2024). Query-agnostic selectors score frames by coverage or diversity (Sun et al., 2025; Wang et al., 2025c), while question-conditioned selectors and grounding models locate the span the question refers to (Wang et al., 2026b; Yu et al., 2025b; Hu et al., 2025; Buch et al., 2025; Zhang et al., 2023). Some datasets annotate this span directly, giving clue intervals or evidence frames alongside each question (Chen et al., 2025; Wu et al., 2024a; Cheng et al., 2025b; Lu et al., 2025). Selection alone is not enough, however: questions about order, duration, or change are answered from what happens *between* the selected frames (Xiao et al., 2021; Liu et al., 2024), so keeping only the evidence frames removes the context that makes them readable.

More recently, *thinking-with-videos* methods move evidence localization into inference. The model searches for relevant segments, inspects them closely, and feeds the result back into its reasoning (Wang et al., 2024b; Yang et al., 2025; Yuan et al., 2025; Liu et al., 2026b; Yang et al., 2026c; Zhang et al., 2025b; Pan et al., 2026). Localization becomes explicit, but each question now costs several rounds of search or tool calls. OWLET pays this cost during training instead. The teacher view, uniform global coverage together with question-conditioned local windows, is built only at training time, and the student is measured against it. At inference the student runs the ordinary single-pass pipeline: no search, no grounding module, and no extra forward passes.

C.2 Reinforcement Learning for Video-LLMs

Reinforcement learning is now the standard route to reasoning ability in large models. Early alignment methods optimize a learned preference model with policy-gradient algorithms (Ouyang et al., 2022; Schulman et al., 2017). Reinforcement learning with verifiable rewards instead checks the final answer with a rule, and estimates the advantage by comparing responses sampled for the same prompt, so no critic is needed (Shao et al., 2024; Guo et al., 2025). Reinforcement learning samples from the current policy, so training stays on the distribution the model uses at test time.

Adapting the recipe to video has produced a large body of work. Video-R1 adds temporal rewards that discourage answering from a single frame (Feng et al., 2025), and related methods target temporal consistency and self-evolution (Huang et al., 2026; Yu et al., 2025a), spatiotemporal perception (Li et al., 2025a; Wang et al., 2025b), grounded reasoning with explicit evidence (Meng et al., 2025), difficulty-aware optimization (Park et al., 2025), and adaptive temporal search (Pan et al., 2026). These works change what the reward measures or which trajectories are sampled, but the reward stays a single scalar attached to the whole trajectory.

This creates a credit-assignment problem that video makes worse. The same reward is copied to every token, yet the tokens of a video response rely on the visual stream to very different degrees: a token

that names an observed action and a token that closes a sentence receive equal credit. In the text domain, this has been addressed with process reward models and step-level credit (Luo et al., 2024; Cui et al., 2025a; Cheng et al., 2025a; Li et al., 2025b), and with token-level signals read off the policy's own uncertainty (Cui et al., 2025b; Cheng et al., 2026; Xu et al., 2026a; Wang et al., 2026a). Both define importance over the reasoning text and are blind to the video. We also draw inspiration from these text domain token-level RL methods. Closer to our setting, some video methods score tokens by how much the model attends to visual tokens (Du et al., 2026; Wang et al., 2026c). These scores are read from a single forward pass and can be driven by language priors: they say where the model looked, not whether what it saw supports the token. OWLET obtains the weight by comparing the token likelihood assigned by a fixed teacher under the privileged view with that assigned by the student rollout policy under the uniform view, using the same question and response prefix.

C.3 On-Policy Distillation and Self-Distillation

Knowledge distillation transfers capability by matching a student to a teacher's predictive distribution (Hinton et al., 2015; Kim & Rush, 2016). Run on a fixed corpus of teacher-generated sequences, it supervises the student on states that may differ from those visited during generation. On-policy distillation (OPD) removes this mismatch: the response is sampled from the student, and a token-level divergence to the teacher is minimized on exactly the states the student reaches (Gu et al., 2024; Agarwal et al., 2024; Lu & Lab, 2025). This gives supervision at every position for roughly one extra teacher forward pass per rollout. Later work studies the choice of divergence, the weighting of tokens, the interaction with RL, and the failure modes of the objective (Wu et al., 2025; Xu et al., 2025; Jin et al., 2026; Xu et al., 2026b; Fu et al., 2026b; Li et al., 2026e; Wu et al., 2026).

On-policy self-distillation (OPSD) derives the teacher from the same underlying model and supplies privileged information that the student does not receive (Hübotter et al., 2026; Zhao et al., 2026; Shenfeld et al., 2026). This provides supervision without requiring a larger external teacher; the choice of whether to freeze or update the teacher is a separate design decision. In our implementation, teacher and student share their initial checkpoint, and the teacher remains fixed as the student is trained. So far the privileged information has been almost always textual: a reference rationale (Yang et al., 2026a), a longer context (Ye et al., 2026; Zhang et al., 2026a), or feedback from the environment (Shenfeld et al., 2026). Related variants study weak-to-strong pairings, self-supervised teachers, reasoning compression, and adaptive supervision horizons (Yu et al., 2026; Tan & Hong, 2026; Sang et al., 2026; Hou et al., 2026).

In the multimodal setting the privileged information becomes visual. Vision-OPD gives the teacher evidence-centered crops while the student sees the full image (Yuan et al., 2026). Later methods reweight vision-critical tokens (Liu et al., 2026a), transfer reasoning from a text teacher (Bousselham et al., 2026), steer gradients toward visual grounding (Yoon et al., 2026; Tian et al., 2026), stabilize the objective for multimodal reasoners (Hao et al., 2026), and study contrastive forms of asymmetric visual supervision (Li et al., 2026c; Liang et al., 2026). In the temporal domain, Video-OPD applies OPD to video grounding (Li et al., 2026b), and rationale- or agent-based methods transfer temporal reasoning through generated text (Shi et al., 2025; Lu et al., 2025). All of them work in output space: what is matched is a distribution over the next token.

That interface loses information. A logit is the last step of a computation that has already merged the visual stream, and a growing literature argues that much of what matters for reasoning lives in the intermediate states rather than in the tokens they produce (Hao et al., 2024; Li et al., 2026a; Xia et al., 2026; Zhang et al., 2026b). Hidden-state distillation has accordingly been studied for multimodal models and for pre-training (Xu et al., 2024; Cai et al., 2025; Guigon et al., 2026), and OPRD extends OPD into representation space for text LLMs (Yang et al., 2026b). Video sharpens the point: cues from distant frames often need to be merged into one spatiotemporal representation well before the first answer token, so matching output distributions leaves that merging step unsupervised.

OWLET differs from this literature in three ways. First, the privileged information is temporal rather than textual: another view of the same video, combining uniform global coverage with question-conditioned local windows, which does not require a stronger model. Second, the teacher view does not replace the reward with an imitation objective; the verifier still sets the direction of learning, and the two views are compared only to redistribute credit inside a response. Third, the transfer acts at

two levels, on token weights and on response-end spatiotemporal representations, so that both what the student reinforces and what it internally computes are shaped by the richer view.

D Details of Experiment Settings

This section provides the detailed experimental setup for Sections 3 and 5: the data and decoding protocol (Appendix D.1), the construction of the four input views and the question taxonomy used by the teacher-view comparison (Appendices D.2 and D.3), the teacher-view results on a second model family (Appendix D.4), and the two measurements behind Figure 2(c,d) (Appendices D.5 and D.6). We use the notation from Section 2: v denotes the student view, $\tilde{v}$ the teacher view, and the student model and the teacher model are both initialized from the same checkpoint.

D.1 Evaluation Data and Decoding

The view comparison and the two motivation diagnostics use $1{,}000$ samples held out from Video-Evidence-96K and disjoint from the training split. Questions are in multiple-choice form, and accuracy is the exact match of the predicted option letter; for the credit measurement we cap generation at 2048 new tokens.

For a controlled comparison, every view in Appendix D.2 is assembled from exactly the same number of frames.

D.2 Construction of the Four Input Views

Every record in Video-Evidence-96K stores a question-conditioned relevance score over a densely sampled pool of candidate frames, from which we take the ranking used below. We construct the four views at total frame budgets of $B \in \{16, 32, 64\}$, reusing the same candidate pool and relevance ranking across budgets. Due to computational resource constraints, we do not train OWLET with frame budgets larger than 64. A set of windows is *disjoint* when no two of them share a frame, which is what keeps the total frame count exact. Window centers are selected in decreasing relevance order, skipping any candidate whose window overlaps an earlier one. For all sampling methods, the selected frames are arranged in chronological order before being fed to the model. Table 3 gives the composition at each budget.

Table 3: Composition of the four input views defined in Section 3.1 at total budgets of 16, 32, and 64 frames. Ranked frames are selected by question-conditioned relevance and used directly in Key frames or as centers of disjoint windows in Key windows and Global with local. All four views use the same total number of frames within each budget.

View	Uniform frames	Ranked frames	Window size	Total
16-frame budget				
Vanilla	16	—	—	16
Key frames	—	16	—	16
Key windows	—	4	4	16
Global with local	8	2	4	16
32-frame budget				
Vanilla	32	—	—	32
Key frames	—	32	—	32
Key windows	—	8	4	32
Global with local	16	4	4	32
64-frame budget				
Vanilla	64	—	—	64
Key frames	—	64	—	64
Key windows	—	16	4	64
Global with local	32	8	4	64

The trained student uses the *vanilla* view at inference: B frames sampled uniformly over the whole video, with no access to the relevance ranking. *Key frames* spends the entire budget on the B highest-ranked frames, and *key windows* spends it on $B/4$ disjoint four-frame windows selected by the ranking; the two views use the same ranking but allocate the budget between evidence breadth and local temporal context differently. *Global with local* splits the budget equally, combining $B/2$ uniform frames with $B/8$ disjoint four-frame windows, so that it retains a coarse view of the whole video while still resolving the question-relevant moments more finely than uniform sampling would. At 32 frames this gives 16 uniform frames and four local windows; at 16 frames it gives eight uniform frames and two local windows. The window size stays fixed across budgets, so the smaller views reduce the number of selected windows while preserving the same allocation between global and local coverage.

Key frames and key windows use the same relevance ranking but allocate the frame budget differently, whereas global with local combines ranked windows with uniform coverage. All four views use the same frame budget, allowing us to compare their allocation of global and local coverage.

D.3 Global-Temporal and Local-Evidence Questions

Section 3.1 splits the evaluation into *global-temporal* and *local-evidence* questions. A question is global-temporal when answering it requires integrating information across separated parts of the video: the annotated evidence identifies relevant moments, but their interpretation depends on broader context, such as how an activity develops, whether a behavior recurs, or how a scene changes from beginning to end. A question is local-evidence when the answer can be determined from the annotated span and its immediate temporal neighborhood. Local-evidence questions may still require several consecutive frames, for example to distinguish the steps of a short action. The distinction concerns the temporal scope of the information needed for the answer.

We use GPT-5.6 Sol (OpenAI, 2026) to assign the labels from the question, the answer options, and the annotated visual evidence with its temporal locations. The annotation protocol has two steps:

1. **Identify the required visual facts.** Specify the observable facts needed to distinguish the correct answer from the alternatives, and record the temporal location of each fact. These may include an action, an object state, an interaction, or a relation between events. A plausible answer based on language alone does not count as visual support.
2. **Determine the temporal scope.** Assign *local-evidence* when all required facts are available within one coherent local event and its immediate temporal neighborhood. Assign *global-temporal* when answering requires combining facts from separated episodes or making a judgment about the video as a whole. The local event must be a coherent activity; an interval drawn broadly enough to include unrelated episodes does not make the question local.

Question labels are assigned once, before model evaluation. We then assess the quality of the key-frame annotations separately from the local/global labels. An annotation passes this check when the selected frames support the claimed question-relevant visual facts at their recorded temporal locations and no unresolved annotation error, unclear visual evidence, or question ambiguity prevents applying the labeling rule. On random sample of 300 records, GPT-5.6 Sol judges 94.3% of the annotations acceptable, and manual inspection accepts 275 of 300 annotations (91.67%). On these 300 records, the GPT and human assessments agree on 96.67% of the pass/fail judgments, and Cohen's κ (Cohen, 1960) is approximately 0.74.

The four categories below account for the global-temporal side of the split. For each we give the reason a view restricted to the annotated evidence is insufficient, together with two examples beyond those quoted in the main text.

- **Theme and purpose.** The answer summarizes the video as a whole rather than any of its parts, so evidence frames may identify one activity without establishing that it is the dominant one. *"What is the main topic of this video?"*; *"Which word best describes the atmosphere of the event?"*
- **Event progression and temporal structure.** The answer summarizes how a multi-stage activity unfolds across the video, so an evidence-only view may show individual stages without revealing their transitions or overall organization. *"What are the main stages of the experiment, from setting*

up the apparatus to reporting the result?"; *"How does the cooking demonstration progress from preparation to serving?"*

- **Recurring patterns and routines.** The answer identifies a behavior that recurs across multiple episodes, so one annotated window may show an instance but cannot establish the pattern or its consistency throughout the video. *"What does the instructor consistently do before introducing each new exercise?"*; *"Which actions does the player repeat after each failed attempt throughout the video?"*
- **State evolution.** The answer is a change between two points in time, whereas a view that keeps only the most relevant window keeps only one endpoint of the comparison. *"How does the colour of the liquid change during the video?"*; *"What is different about the room at the end compared with the beginning?"*

D.4 Teacher-View Comparison on Qwen3-VL Model Family

Figure 2(a,b) compares the four input views on Qwen3.5-0.8B, 4B, and 9B. We extend the teacher-view comparison to Qwen3-VL-2B, 4B, and 8B, considering both non-thinking and thinking variants. The views follow the definitions in Appendix D.2, and the global-temporal subset follows the question taxonomy in Appendix D.3. Figure 6 reports the non-thinking results in the top row and the thinking results in the bottom row, with overall accuracy on the left and global-temporal accuracy on the right. Overall accuracy uses all 1,000 held-out questions, and global-temporal accuracy uses the 500-question subset. All differences discussed below are absolute percentage-point differences in accuracy; model sizes are listed in the order 2B, 4B, and 8B, respectively.

Non-thinking models: overall improvements can conceal a loss of global context. Key frames improve overall accuracy over vanilla by 4.0, 5.0, and 3.6 points, while key windows improve it by 5.4, 4.6, and 5.5 points. These gains show that focusing the input on question-relevant evidence helps when performance is aggregated over all questions. Neither evidence-only construction improves the global-temporal subset: key frames reduce accuracy by 1.4, 3.8, and 3.0 points, and key windows reduce it by 1.6, 3.0, and 1.6 points. Expanding selected frames into local windows therefore does not consistently recover the broader context needed by these questions. The ordering of key frames and key windows varies across models and question subsets.

Combining global coverage with local evidence improves both metrics. Global context with local windows reaches overall accuracies of 51.1%, 54.3%, and 58.9%, exceeding vanilla by 13.2, 12.8, and 13.2 points. Its global-temporal accuracies are 50.2%, 53.2%, and 56.6%, corresponding to gains of 13.0, 13.0, and 10.2 points. The advantage remains substantial when the reference is the stronger of the two evidence-only views for each model and metric: the margins are 7.8, 7.8, and 7.7 points overall, and 14.4, 16.0, and 11.8 points on global-temporal questions. Thus, the combined view retains the overall benefit of focused evidence while avoiding the global-temporal degradation observed with either evidence-only view.

Thinking models show the same trade-off across views. For the thinking variants, global context with local windows obtains overall accuracies of 55.8%, 59.3%, and 63.5%, compared with vanilla accuracies of 42.9%, 47.1%, and 50.2%. The gains are 12.9, 12.2, and 13.3 points. Key frames reach 44.3%, 52.8%, and 54.6%, and key windows reach 44.8%, 53.5%, and 54.8%. Although key windows are the stronger evidence-only view on overall accuracy for all three thinking models, the combined view still improves over them by 11.0, 5.8, and 8.7 points. The overall advantage of combining broad temporal coverage with focused local evidence therefore persists across both evaluated reasoning modes.

The global-temporal results reveal the same limitation of evidence-only inputs in thinking mode. Relative to vanilla accuracies of 42.2%, 46.4%, and 50.6%, key frames reduce accuracy by 1.4, 1.8, and 1.2 points, and key windows reduce it by 0.4, 1.6, and 1.4 points. Global context with local windows instead reaches 54.4%, 59.6%, and 62.0%, improving over vanilla by 12.2, 13.2, and 11.4 points. Its margins over the stronger evidence-only view are 12.6, 14.8, and 12.6 points. Thus, the overall gains of evidence-only views continue to coexist with a loss on global-temporal questions in the thinking variants, whereas the combined view improves both metrics at every evaluated scale.

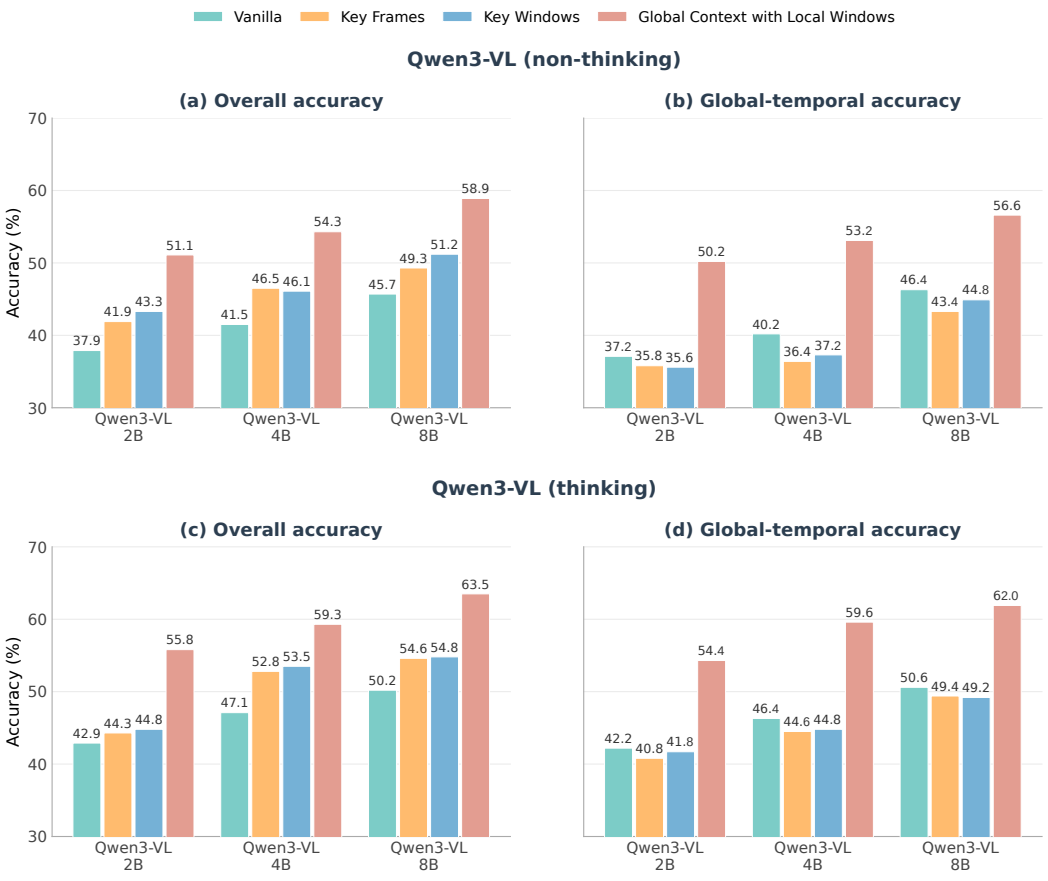


Figure 6: **Teacher-view comparison on the Qwen3-VL model family. (a, b)** Non-thinking variants. **(c, d)** Thinking variants. The left column reports accuracy on all questions; the right column reports accuracy on the global-temporal subset. Each panel compares the four input views defined in Table 3 on the 2B, 4B, and 8B models. Bars are labeled with accuracy in percent, and all four panels use the same vertical scale. In both modes and at all three model sizes, evidence-only views improve overall accuracy while reducing global-temporal accuracy relative to vanilla; global context with local windows improves both metrics.

Implications for teacher-view design. Both Qwen3-VL modes reproduce the qualitative trade-off observed for Qwen3.5: selecting relevant evidence can improve the aggregate score while reducing accuracy on questions about the video as a whole. Across all six Qwen3-VL model–mode configurations, global context with local windows is the only tested alternative to vanilla that improves both overall and global-temporal accuracy, and it achieves the highest accuracy on both metrics. The deployed OWLET student continues to use uniformly sampled frames without evidence annotations.

D.5 Measuring the Concentration of the Evidence Gap

For each question we sample one response from the student under its own view v, keep the responses the verifier accepts, and re-score exactly those token sequences under the teacher view $\tilde{v}$ in a single forward pass. We restrict the diagnostic to verifier-accepted responses and compare their token likelihoods under the two views. This yields a per-token log-probability difference

$$\Delta \log p_t = \log \pi_{\text{teacher}}(y_t \mid q, \tilde{v}, y_{<t}) - \log \pi_{\text{student}}(y_t \mid q, v, y_{<t}), \tag{8}$$

for the same question, response token, and prefix. In the static diagnostic of Figure 2(c), both policies use the common initial checkpoint θ_0, so $\Delta \log p_t$ isolates the effect of changing the visual input at position t. This equality of weights applies to the initial diagnostic; during training, the student evolves while the teacher remains fixed.

Within each response we sort tokens by $|\Delta \log p_t|$ in decreasing order and accumulate their share of $\sum_t |\Delta \log p_t|$. We normalize each response separately before averaging the concentration curves. Figure 2 (c) in Section 3 plots the mean curve over the evaluation set against the *uniform baseline*, the diagonal that a response would trace if every one of its tokens contributed equally to the gap. The shaded band spans the 10th to 90th percentile of the individual per-video curves at each token fraction, so its width reports how much the concentration varies from video to video rather than any uncertainty in the mean. The reported statistic is the token fraction at which the mean curve reaches 50%.

Section 5.4 tracks a scalar variant of this measurement during training: the share of $\sum_t |\Delta \log p_t|$ carried by the top decile of tokens within each response. Here the student distribution comes from the evaluated training checkpoint under v, and the teacher distribution comes from the fixed initial checkpoint θ_0 under $\tilde{v}$.

D.6 Measuring Output and Hidden-State Alignment

Figure 2(d) measures whether output-level alignment also brings the student and teacher hidden states closer. We run on-policy self-distillation from Equation 2: teacher and student are initialized from the same checkpoint θ_0, the teacher remains fixed, and responses are sampled from the evolving student under v. Only the student is trained, using the token-level KL divergence to the fixed teacher under $\tilde{v}$. There is no reward term or representation-level loss in this diagnostic. We train for 300 steps and evaluate both quantities at fixed intervals on the held-out split described in Appendix D.1.

For the output measurement, we score each sampled response with the current student under v and the fixed teacher under $\tilde{v}$, holding the question, generated tokens, and prefix at each response position identical. We report the KL divergence from the student distribution to the teacher distribution, averaged over valid response positions and then over the evaluation batch. A smaller value means that the student more closely matches the fixed teacher's next-token distributions on its own responses.

To measure the corresponding hidden-state alignment, we use centered kernel alignment (CKA) (Kornblith et al., 2019). CKA compares the similarity structure of response-end representations across matched responses: it asks whether pairs of responses that are similar under one view are also similar under the other.

For decoder layer ℓ, let $X_\ell, Y_\ell \in \mathbb{R}^{m \times d_\ell}$ be the response-end hidden states for m matched responses, computed by the current student under v and the fixed teacher under $\tilde{v}$, respectively. As in Section 4.3, we take the final valid response token after layer ℓ, so both states refer to the same generated text position. After centering the columns of X_ℓ and Y_ℓ, we form $K_\ell = X_\ell^c (X_\ell^c)^\top$ and $L_\ell = Y_\ell^c (Y_\ell^c)^\top$. Linear CKA is the normalized similarity between these two Gram matrices:

$$\mathrm{CKA}(X_\ell, Y_\ell) = \frac{\langle K_\ell, L_\ell \rangle_F}{\|K_\ell\|_F \, \|L_\ell\|_F}, \qquad \langle K_\ell, L_\ell \rangle_F = \sum_{i=1}^{m} \sum_{j=1}^{m} (K_\ell)_{ij} (L_\ell)_{ij}. \tag{9}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. CKA is bounded between 0 and 1 for the centered Gram matrices used here. A value of 1 means that the two views induce the same pairwise similarity structure up to scale; a value near 0 means that the structures are weakly aligned. The measure is invariant to rotations and isotropic rescaling of the hidden coordinates, so it captures the organization of the representations rather than their arbitrary basis or norm.

We plot $1 - \mathrm{CKA}$, averaged over the evaluated decoder layers:

$$D_{\text{hidden}} = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \left(1 - \mathrm{CKA}(X_\ell, Y_\ell)\right). \tag{10}$$

Each layer's CKA is computed across the batch before the layer average. The SSA loss in Equation 6 instead uses cosine similarity between paired response-end states.

The output KL and $1 - \mathrm{CKA}$ use the same matched responses but measure different quantities and are not on the same scale. In Figure 2(d), output KL decreases to 0.046, while $1 - \mathrm{CKA}$ remains at 0.386. Thus, output-only matching reduces the next-token divergence while leaving a distinct hidden-state similarity structure; SSA aligns the response-end states at matched depths.

E Experimental Data and Supplementary Experiments

This section details the construction and composition of Video-Evidence-96K, including the global-temporal and local-evidence questions defined in Appendix D.3. We then describe the baselines, benchmarks, and training settings, followed by the full results across model families and frame budgets and the corresponding component ablations.

E.1 Construction of Video-Evidence-96K

We construct Video-Evidence-96K in four stages: source aggregation, record validation, question-conditioned evidence annotation, and view assembly. It retains 96,705 question–video records, of which 1,000 are held out for diagnostic evaluation and 95,705 are used for training. Table 4 tracks these counts through validation and the held-out split, Table 5 reports the training sources, and Table 6 reports the temporal scope of the questions in each split.

Source aggregation. We start from 104,257 video–question–answer records collected from four public sources: LLaVA-Video-178K (Zhang et al., 2024), NExT-QA (Xiao et al., 2021), Perception (Patraucean et al., 2023), and STAR (Wu et al., 2024a). The sources cover complementary aspects of video understanding: LLaVA-Video-178K contributes broad-domain footage with diverse question types, NExT-QA emphasizes causal and temporal questions, Perception stresses fine-grained perceptual skills, and STAR provides situated-reasoning questions over short action-centric clips. Every record is normalized into a shared schema consisting of the video path, the question, lettered multiple-choice options, and a single correct option letter, so that the rule-based verifier of Section 2 applies uniformly across sources. The unit of data is a question–video pair: questions referring to the same video remain separate records because they can require different evidence and different temporal scopes.

Record validation. We then apply four checks in order. (i) The question and answer fields must be non-empty after whitespace normalization. (ii) The options must be well formed: unique option labels, non-duplicate option texts, and exactly one option matching the reference answer. (iii) The video file must exist and decode end to end. (iv) The video must supply sufficient temporal candidates for evidence sampling: after candidate-pool extraction (below), a record is kept only when its pool can support the disjoint local windows and the global track of the largest teacher configuration. These checks remove 7,552 records in total, leaving 96,705 valid records. The subsequent held-out split reserves 1,000 of these records for evaluation; these are valid examples withheld from training.

Question-conditioned evidence annotation. For each retained record we extract a candidate pool of up to 512 frames sampled uniformly over the full video; shorter videos contribute every decodable frame. An automatic question-conditioned frame-relevance annotator (Wang et al., 2026b) scores each candidate frame against the question, and we store the full relevance ranking with the record

Table 4: Construction pipeline of Video-Evidence-96K. Validation rows report excluded records and the number remaining after each check. The final row reserves 1,000 valid records for diagnostic evaluation, leaving 95,705 training records.

Stage	Criterion	Excluded / reserved	Remaining
Source aggregation	four public sources	—	104,257
Field validation	non-empty question and answer	812	103,445
Option validation	unique labels, single correct option	1,134	102,311
Video validation	file exists and decodes end to end	1,313	100,998
Temporal validation	sufficient evidence candidates	4,293	96,705
Held-out diagnostics	evaluation split of Section 3	1,000	95,705

rather than a hard top-k cut. Storing the ranking allows the four diagnostic views of Appendix D.2 and the training views at every student budget $N \in \{16, 32, 64\}$ to use the same underlying evidence annotation. Since the scores are conditioned on the question, two records from the same video can select different moments. The ranking identifies where relevant evidence occurs; the local windows and global track determine how that evidence is presented to the teacher.

Evidence quality. We use the annotation audit described in Appendix D.3. On the same 300 randomly sampled records, GPT-5.6 Sol accepts 94.3% of the evidence annotations and human inspection accepts 275 of 300 (91.67%). An annotation is accepted when the selected frames support the question-relevant visual facts at the recorded temporal locations and no unresolved ambiguity prevents applying the labeling rule.

View assembly. At training time, the teacher view combines a global track of $N/2$ uniformly sampled frames with $N/8$ disjoint local windows of four frames centered on the top-ranked candidates, following the window construction of Appendix D.2. Windows are expanded in rank order, a window that would overlap an earlier local window is skipped, and the selected frames are arranged in chronological order before encoding. The local windows contribute $N/2$ frames, giving the teacher the same total frame budget N as the student: 16, 32, or 64 frames. The global track provides coverage of the video as a whole, while the local windows resolve the moments most relevant to the question. The student receives only N uniformly sampled frames and does not access the relevance ranking at training or inference.

Dataset splits and source composition. From the 96,705 validated records we hold out 1,000 for the diagnostic experiments of Section 3, leaving 95,705 training records, with approximately 4.4 questions per video. Table 5 counts training records after the held-out examples have been removed. Its source percentages use the 95,705-record training split as the denominator; they describe the distribution of question–video pairs rather than the distribution of unique videos.

Table 5: Composition of the Video-Evidence-96K training split.

Source	Records	Share
LLaVA-Video-178K	70,468	73.63%
NExT-QA	7,542	7.88%
Perception	6,345	6.63%
STAR	11,350	11.86%
Total	95,705	100.00%

Global-temporal and local-evidence composition. We use the question taxonomy in Appendix D.3 to describe the temporal scope of the dataset. GPT-5.6 Sol assigns the labels from the question, answer options, and annotated visual evidence with its temporal locations, first identifying the facts needed to distinguish the answer and then determining how widely those facts are distributed in time. A local-evidence question can be answered within one coherent event and its immediate temporal neighborhood, even when several consecutive frames are needed. A global-temporal question requires

combining separated episodes or making a judgment about the video as a whole. Video duration and source-dataset identity do not determine the label: different questions about the same video can fall into different subsets.

The global-temporal subset includes questions about theme and purpose, event progression, recurring patterns, and state evolution. The local-evidence subset covers questions whose relevant action, object state, or interaction can be resolved within one selected event and its surroundings.

Table 6 separates the composition of the training set from that of the held-out diagnostic set. Each count refers to question–video records, with the share computed within the corresponding split. Thus, the global-temporal share in training is the number of global-temporal training records divided by 95,705, whereas its share in the diagnostic set uses 1,000 as the denominator. The complementary local-evidence share is computed over the same records. Reporting both splits specifies how much global-temporal supervision is available during training and how much weight that subset carries in the overall diagnostic accuracy.

Table 6: Composition of Video-Evidence-96K by question type, following Appendix D.3. Counts are question–video records, and each percentage uses the total of its own split as the denominator.

		Global-temporal		Local-evidence	
Split	Total	Records	Percentage	Records	Percentage
Training	95,705	20,137	21.04%	75,568	78.96%
Held-out diagnostics	1,000	500	50%	500	50%

The same held-out question labels define the subsets throughout the teacher-view comparisons and the post-training analysis in Appendix F.8. Overall accuracy is weighted by the number of questions in each subset, while global-temporal and local-evidence accuracies are calculated within their respective subsets. This is why an evidence-only view can improve overall accuracy while reducing global-temporal accuracy: gains on local-evidence questions can outweigh losses on the global-temporal subset. Reporting the subset sizes alongside both accuracies makes the contribution of each question type explicit.

E.2 Baselines and Benchmarks

E.2.1 Baselines

Within each model and mode, all post-trained methods start from the same released checkpoint and use Video-Evidence-96K as the common source of question–video pairs. GRPO, OPSD, and OWLET use the same training split, optimizer settings, rollout budget, and number of updates. SFT instead uses traces generated for these question–video pairs, as described below.

- **Vanilla** is the released checkpoint with no post-training. It fixes the starting point every other row is measured against.
- **SFT** is supervised fine-tuning on chain-of-thought traces generated by Qwen3.5-397B-A17B for the training question–video pairs. It is the off-policy counterpart of the methods below: the supervision is dense and comes from a far larger model, but it is collected once rather than sampled from the evolving student policy.
- **GRPO** is the verifier-driven RL objective of Equation 1, sampled from the student view alone.
- **OPSD** is on-policy self-distillation as described in Equation 2. It keeps the student's own rollouts and the same fixed teacher checkpoint and teacher view as OWLET, but replaces the reward with the token-level divergence between the student and teacher next-token distributions.

E.2.2 Benchmarks

We evaluate on five multiple-choice video benchmarks, all read with uniform frame sampling at the configuration's budget N and with no evidence annotation. These benchmarks assess transfer beyond the held-out Video-Evidence-96K diagnostics. A prediction is correct when the predicted option letter matches the reference, and **Avg.** is the unweighted macro-average over the five benchmark scores.

The global-temporal/local-evidence labels describe the Video-Evidence-96K diagnostic subsets; they are distinct from benchmark-specific divisions such as Video-MME duration and Video-MME 2 question groups.

- **Video-MME** (Fu et al., 2025) provides a broad evaluation of video understanding across diverse content domains and temporal spans. Its questions cover perception and reasoning, while its short, medium, and long duration splits examine how well models handle increasingly extended video context. This combination makes it useful for assessing whether post-training gains extend beyond a particular scene type or temporal scale. Under our fixed input-frame budgets, the duration breakdown also shows how the gains vary as the sampled frames must cover longer videos. We report overall accuracy and the three duration splits, with the Qwen3.5 breakdown in Tables 11–14.
- **MVBench** (Li et al., 2024a) targets temporal understanding through multiple-choice tasks that require information from more than one frame, spanning temporal perception and reasoning about actions, objects, and their changes. It complements the broader domain and duration coverage of Video-MME by testing how models use sequences of observations.
- **LongVideoBench** (Wu et al., 2024b) was introduced to evaluate referring reasoning in long, interleaved video–language contexts. A question refers to a particular part of the video, and the model must locate the relevant content within a much longer context before answering. The tasks range from perceiving details at the referred moment to reasoning about relations between events, making both evidence localization and contextual interpretation important. Its value for our evaluation is that a model must retain question-relevant detail despite substantial surrounding content; a coarse account of the video's overall topic is insufficient. We evaluate on the validation split using the uniform frame-sampling protocol described above.
- **Video-MME 2** (Fu et al., 2026a) extends video evaluation to visual information aggregation, temporal dynamics, and complex reasoning. Its grouped-question design examines whether a model's understanding remains coherent across related questions, adding a dimension that isolated question answering does not capture. The *logic* groups probe successive steps in a reasoning chain, while the *relevance* groups probe different aspects or levels of detail of the same capability. These demands complement the duration-based analysis of Video-MME: the model must combine observations and reason about their relationships, in addition to recognizing individual events. We report the two subsets separately, using the scoring convention specified below.
- **MLVU** (Zhou et al., 2025) was introduced to evaluate long-video understanding across diverse video genres and complementary task types. Its task design distinguishes holistic understanding, retrieval of a single relevant detail, and integration of multiple details. The multiple-choice tasks include topic reasoning, retrieval of a brief event from a long video, and reasoning about action order or repetition. These tasks expose different requirements: global understanding depends on broad temporal coverage, detail retrieval depends on preserving a localized observation, and action ordering or counting requires combining observations over time. We evaluate the multiple-choice portion, which tests both global context use and access to local evidence within a common answer format.

For Video-MME 2, we report per-question accuracy in the *logic* and *relevance* subsets. The combined score weights the two subset accuracies by their question counts, and Tables 8–10 report the Qwen3.5 subset results separately. For consistency and reproducibility, we use the official evaluation prompts for each benchmark, respectively.

E.3 Training Hyper-Parameters

Table 7 reports the training and evaluation settings for GRPO, OPSD, and OWLET. Following Section 5.1, each run uses 1,000 training steps, with the optimizer settings, prompt and rollout counts, and rollout decoding held fixed within a model–mode–budget configuration. The two OWLET-specific settings are the evidence-contrast clipping threshold c of Equation 4 and the alignment weight λ_{SSA} of Equation 7. Verifier rewards and group-normalized advantages apply to GRPO and OWLET; OPSD optimizes the token-level distillation objective without a reward term.

Table 7: Training and evaluation hyper-parameters.

Setting	Value
Optimization (shared by GRPO, OPSD, and OWLET*)*	
Learning rate	1×10^{-6}, constant, no warm-up
Prompts per step / rollouts per prompt G	16 / 8
Mini-batch size	16
Training steps	$1,000$
Clipping range ϵ	0.2
KL penalty to a reference policy	none
Advantage estimator (GRPO and OWLET)	group-normalized (GRPO)
Reward (GRPO and OWLET)	rule-based match on the option letter
Devices	8 NVIDIA H200 GPUs (140GB each)
Rollout decoding	
Temperature / top-p / top-k	1.0 / 0.99 / disabled
Max response length	2048 tokens
Views	
Student frames N	16, 32, or 64 (uniform)
Teacher global track	$N/2$ uniform frames
Teacher local windows	$N/8$ disjoint windows of 4 frames
Teacher parameters (OPSD and OWLET*)*	
Default teacher	fixed at the common initial checkpoint θ_0
OWLET-*specific*	
ECR clipping threshold c	5
ECR applied to	positively-advantaged responses
SSA weight λ_{SSA}	0.05
SSA responses	verifier-accepted only
Evaluation decoding	
Temperature	0.0
Max new tokens	2048
Frames	N uniform, no evidence annotation

E.4 Per-Benchmark Results and Detailed Analysis

Tables 8–10 give the per-benchmark results for Qwen3.5-0.8B, 4B, 9B, and 27B in non-thinking mode at student frame budgets of 16, 32, and 64. They extend Table 1 to the smaller frame budgets and report the two Video-MME 2 question groups separately. Tables 11–14 further break the Video-MME results down into short, medium, and long videos, with one table per model. Together, these results show how the comparison changes across benchmark tasks, temporal spans, input budgets, and model scales.

Table configuration. The per-benchmark tables group rows by backbone and compare the base model, SFT, GRPO, OPSD, and OWLET at a fixed frame budget. The duration tables instead fix the backbone and group rows by frame budget. All scores are accuracies in percent. Within Video-MME 2, *Logic* and *Rel.* denote the logic and relevance question subsets, and *All* combines their accuracies using the respective question counts. In the Video-MME duration tables, *All* averages the short, medium, and long splits. The **Avg.** column weights the five benchmarks equally, with Video-MME 2 contributing its combined score once. Boldface identifies the best score in each metric column within a model and frame budget. Scores are displayed to one decimal place, and improvements below are expressed in percentage points.

Comparison with post-training baselines. Across the twelve model–budget configurations, OWLET achieves the highest accuracy on all five benchmarks and the highest macro-average. The strongest baseline by macro-average varies between GRPO and OPSD, while OWLET improves on it by 2.2–4.3 points. At 64 frames, the margins are 3.4, 4.1, 4.3, and 4.3 points for the 0.8B, 4B, 9B, and 27B models, respectively. Thus, the advantage persists across model scales and is not limited to the smallest backbones. The comparison with GRPO supports the value of supplementing verifiable

rewards with supervision from the evidence-privileged view. The comparison with OPSD, which uses the same teacher view, further supports transferring that supervision through evidence-aware credit and hidden-state alignment. The separate contributions of ECR and SSA are examined in Section 5.3 and Appendix E.7.

Performance across video durations. The Video-MME duration tables show improvements over the baselines on short, medium, and long videos at every evaluated model size and frame budget. At 27B with 64 frames, OWLET reaches 87.8, 74.8, and 65.8 on the three splits, exceeding the strongest baseline in each split by 2.7, 4.0, and 5.2 points, respectively. The larger margin on long videos in this configuration is consistent with the role of the global track: the teacher retains temporal context around the question-relevant evidence even when the video spans many events. Improvements on the short split also show that the benefit extends to videos with a smaller temporal span, where local transitions can still determine the answer. Across the duration tables, the size of the gain varies with the backbone and frame budget, while the improvement over the corresponding baselines remains consistent.

Temporal understanding and long-video reasoning. The gains also extend across MVBench, LongVideoBench, and MLVU, which place different demands on the use of temporal evidence. At 4B with 64 frames, OWLET scores 72.2 on MVBench, 63.4 on LongVideoBench, and 58.8 on MLVU, improving over the strongest baseline for each benchmark by 3.3, 3.6, and 7.7 points. MLVU also shows a substantial gain at 9B with 64 frames: OWLET reaches 61.6, compared with 53.7 for the strongest baseline. These results cover reasoning about temporal changes, locating and interpreting referred content in a long video, and combining global understanding with access to local details. The gains across these complementary tasks are consistent with the teacher view's combination of global coverage and question-conditioned local windows.

Logic and relevance in Video-MME 2. OWLET improves on the strongest baseline in both question subsets for all twelve model–budget configurations. At 27B with 64 frames, logic accuracy increases to 37.6 from a strongest-baseline score of 35.2, while relevance accuracy increases to 32.9 from 30.4. The corresponding improvements are 2.4 and 2.5 points, so the increase in the combined score is supported by both subsets. The margins need not be equally large: at 9B with 64 frames, the gains are 0.5 points on logic and 2.5 points on relevance.

Robustness across frame budgets. Table 2 summarizes macro-average accuracy across input budgets. For each model size, the macro-average of OWLET increases as the student budget grows from 16 to 32 and then to 64 frames, and it leads the baselines at every budget. The 27B model reaches 57.9, 60.4, and 64.5, exceeding the strongest baseline at the corresponding budget by 2.2, 2.3, and 4.3 points. More input frames and evidence-guided post-training therefore provide complementary gains in these configurations. Across the four model sizes, OWLET also exceeds the strongest baseline using twice as many frames in seven of eight adjacent-budget comparisons: 16 against 32 frames and 32 against 64 frames for each model. The exception is 27B at 16 frames, which scores 57.9 against 58.1 for the strongest 32-frame baseline. At 32 frames, the same model exceeds the strongest 64-frame baseline, 60.4 versus 60.2. These comparisons under uniform sampling suggest that OWLET learns to make more effective use of the student's available observations.

Comparison across model scales. The benefits also affect comparisons between backbones. Qwen3.5-4B with OWLET scores 53.3, 54.6, and 58.2 at 16, 32, and 64 frames, respectively. These scores exceed the strongest Qwen3.5-9B baseline at the same budgets, which reaches 50.5, 52.9, and 55.5, giving margins of 2.8, 1.7, and 2.7 points. At the same time, OWLET continues to improve the 9B and 27B models over their own baselines. The results thus support evidence-guided post-training both as a way to strengthen smaller models and as a complementary improvement for larger backbones.

E.5 Results on the Qwen3-VL Family

Section 5.2 reports the Qwen3.5 family in non-thinking mode. We extend the evaluation to Qwen3-VL (Bai et al., 2025) at 2B, 4B, and 8B, considering both non-thinking and thinking modes at student frame budgets of 16, 32, and 64. The resulting eighteen model–mode–budget configurations test

Table 8: Accuracy (%) for Qwen3.5 in non-thinking mode at 16 input frames, with the two Video-MME 2 question groups reported separately. *Logic* and *Rel.* are the group types and *All* combines them by question count. We **bold** the best result for each backbone and benchmark.

Backbone Model	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2			MLVU	Avg.
					Logic	Rel.	All		
Qwen3.5-0.8B	Vanilla	45.1	49.4	44.9	17.8	14.4	15.6	34.2	37.8
	SFT	45.5	50.8	45.9	18.5	14.8	16.1	34.7	38.6
	GRPO	46.8	50.7	47.0	19.6	16.3	17.5	35.8	39.6
	OPSD	46.2	50.4	47.5	18.2	14.9	16.1	35.1	39.1
	OWLET	**50.3**	**56.4**	**50.6**	**20.5**	**17.6**	**18.6**	**38.9**	**42.9**
Qwen3.5-4B	Vanilla	57.2	64.8	53.7	24.3	19.9	21.5	42.7	48.0
	SFT	58.3	64.9	54.3	24.3	20.8	22.0	43.0	48.5
	GRPO	58.8	67.6	56.3	25.7	20.8	22.5	43.7	49.8
	OPSD	59.0	65.5	54.2	25.4	21.1	22.6	44.1	49.1
	OWLET	**62.8**	**69.2**	**59.5**	**27.3**	**22.5**	**24.2**	**50.8**	**53.3**
Qwen3.5-9B	Vanilla	58.2	63.4	56.7	26.7	20.0	22.3	44.3	49.0
	SFT	58.9	64.6	57.6	27.4	20.8	23.1	44.6	49.8
	GRPO	60.2	64.2	58.5	29.0	21.0	23.8	45.6	50.5
	OPSD	60.0	64.7	57.7	28.9	21.2	23.9	45.4	50.3
	OWLET	**62.2**	**68.2**	**62.4**	**32.9**	**25.1**	**27.8**	**50.6**	**54.2**
Qwen3.5-27B	Vanilla	64.8	66.4	60.7	30.5	25.6	27.3	52.5	54.3
	SFT	64.7	66.5	61.6	31.3	26.2	28.0	53.3	54.8
	GRPO	66.0	68.9	62.2	30.9	26.0	27.7	53.8	55.7
	OPSD	65.8	68.8	61.6	32.6	26.7	28.8	53.3	55.7
	OWLET	**68.2**	**73.6**	**63.2**	**32.9**	**27.7**	**29.6**	**55.0**	**57.9**

Table 9: Accuracy (%) for Qwen3.5 in non-thinking mode at 32 input frames, with the two Video-MME 2 question groups reported separately. *Logic* and *Rel.* are the group types and *All* combines them by question count. We **bold** the best result for each backbone and benchmark.

Backbone Model	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2			MLVU	Avg.
					Logic	Rel.	All		
Qwen3.5-0.8B	Vanilla	48.2	51.8	46.4	17.7	14.7	15.8	38.1	40.1
	SFT	48.8	52.4	46.9	19.5	15.5	16.9	39.4	40.9
	GRPO	50.1	54.8	48.2	19.5	16.4	17.5	38.9	41.9
	OPSD	49.6	53.4	48.0	19.5	15.5	16.9	39.7	41.5
	OWLET	**53.9**	**57.8**	**53.0**	**22.2**	**18.0**	**19.4**	**42.0**	**45.2**
Qwen3.5-4B	Vanilla	61.4	65.4	55.1	24.9	20.1	21.8	45.3	49.8
	SFT	62.6	66.7	56.5	25.9	20.2	22.2	45.5	50.7
	GRPO	63.2	67.6	58.0	26.0	21.8	23.2	46.9	51.8
	OPSD	62.9	66.6	56.1	25.7	20.5	22.3	47.0	51.0
	OWLET	**67.0**	**68.7**	**61.4**	**28.7**	**22.7**	**24.8**	**51.2**	**54.6**
Qwen3.5-9B	Vanilla	62.0	64.4	58.7	30.6	22.0	25.0	46.5	51.3
	SFT	62.6	65.6	60.0	32.3	23.2	26.4	46.6	52.2
	GRPO	63.2	65.5	60.9	32.0	24.0	26.8	48.0	52.9
	OPSD	63.2	65.2	60.3	32.4	23.2	26.4	47.9	52.6
	OWLET	**67.7**	**70.0**	**65.8**	**32.8**	**24.7**	**27.5**	**54.2**	**57.1**
Qwen3.5-27B	Vanilla	68.9	68.5	63.9	32.8	26.5	28.7	53.9	56.8
	SFT	69.0	67.2	61.9	33.6	27.8	29.8	52.4	56.1
	GRPO	70.4	69.4	64.5	35.5	27.4	30.2	54.9	57.9
	OPSD	70.7	69.7	64.9	35.5	28.1	30.7	54.6	58.1
	OWLET	**73.3**	**71.9**	**66.2**	**36.3**	**29.8**	**32.1**	**58.6**	**60.4**

whether evidence-guided post-training remains effective across a second backbone family, different model sizes, and generation with or without an explicit reasoning trace. The baselines and evaluation

Table 10: Accuracy (%) for Qwen3.5 in non-thinking mode at 64 input frames, with the two Video-MME 2 question groups reported separately. *Logic* and *Rel.* are the group types and *All* combines them by question count. We **bold** the best result for each backbone and benchmark.

Backbone Model	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2			MLVU	Avg.
					Logic	Rel.	All		
Qwen3.5-0.8B	Vanilla	49.4	51.8	48.9	17.4	14.9	15.8	41.9	41.6
	SFT	49.5	52.2	49.5	17.2	14.9	15.7	42.1	41.8
	GRPO	50.8	53.9	50.4	19.2	18.1	18.5	43.3	43.4
	OPSD	50.9	53.9	51.4	18.7	16.4	17.2	44.1	43.5
	OWLET	**54.9**	**58.6**	**55.9**	**22.6**	**18.4**	**19.9**	**45.2**	**46.9**
Qwen3.5-4B	Vanilla	64.7	67.4	58.0	27.1	23.4	24.7	50.3	53.0
	SFT	65.1	68.0	58.4	27.5	24.6	25.6	50.7	53.6
	GRPO	65.2	68.3	59.8	28.2	25.1	26.2	50.8	54.1
	OPSD	65.5	68.9	58.3	28.3	24.2	25.6	51.1	53.9
	OWLET	**68.3**	**72.2**	**63.4**	**31.8**	**26.8**	**28.5**	**58.8**	**58.2**
Qwen3.5-9B	Vanilla	65.0	67.1	61.9	31.2	22.5	25.6	52.7	54.4
	SFT	65.9	67.2	62.5	32.4	23.3	26.5	53.7	55.2
	GRPO	66.3	67.7	63.4	31.9	24.1	26.8	53.5	55.5
	OPSD	66.1	68.7	62.9	32.7	23.5	26.7	53.3	55.5
	OWLET	**70.7**	**72.8**	**65.2**	**33.2**	**26.6**	**28.9**	**61.6**	**59.8**
Qwen3.5-27B	Vanilla	71.4	70.2	64.2	33.6	28.8	30.4	61.5	59.5
	SFT	71.8	70.5	64.2	33.4	29.1	30.6	61.8	59.8
	GRPO	72.0	71.1	64.5	35.2	30.4	32.1	61.1	60.2
	OPSD	71.4	70.9	65.1	34.1	28.9	30.7	62.9	60.2
	OWLET	**76.1**	**76.2**	**68.8**	**37.6**	**32.9**	**34.6**	**66.7**	**64.5**

Table 11: Video-MME accuracy (%) by video duration (Qwen3.5-0.8B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Vanilla	57.1	41.5	36.7	45.1
	SFT	57.4	41.3	37.9	45.5
	GRPO	59.5	42.7	38.2	46.8
	OPSD	57.9	42.5	38.3	46.2
	OWLET	**61.8**	**47.9**	**41.3**	**50.3**
32	Vanilla	58.7	46.7	39.2	48.2
	SFT	59.6	46.7	40.2	48.8
	GRPO	61.0	49.0	40.2	50.1
	OPSD	59.2	48.1	41.4	49.6
	OWLET	**67.0**	**51.5**	**43.2**	**53.9**
64	Vanilla	62.4	47.3	38.5	49.4
	SFT	62.1	47.1	39.4	49.5
	GRPO	63.9	48.8	39.6	50.8
	OPSD	64.4	47.9	40.5	50.9
	OWLET	**69.1**	**52.7**	**43.0**	**54.9**

protocol follow Appendix E.2. At inference, all methods use uniformly sampled student frames; the evidence-privileged teacher view is used only during training.

Table configuration. Tables 15–17 report the non-thinking results, and Tables 21–23 report the thinking results. Each table fixes the mode and frame budget and compares the five methods within each backbone. The Video-MME duration breakdowns appear in Tables 18–20 and Tables 24–26, respectively. The metric definitions and aggregation conventions are the same as in Appendix E.4.

Table 12: Video-MME accuracy (%) by video duration (Qwen3.5-4B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Vanilla	69.1	54.6	48.1	57.2
	SFT	70.3	55.9	48.6	58.3
	GRPO	69.8	56.2	50.5	58.8
	OPSD	70.3	56.1	50.6	59.0
	OWLET	**73.7**	**59.9**	**55.0**	**62.8**
32	Vanilla	73.4	58.8	52.1	61.4
	SFT	74.2	60.4	53.2	62.6
	GRPO	74.5	61.6	53.5	63.2
	OPSD	74.7	60.6	53.6	62.9
	OWLET	**79.0**	**64.7**	**57.4**	**67.0**
64	Vanilla	76.7	62.9	54.5	64.7
	SFT	77.3	63.2	54.7	65.1
	GRPO	77.1	64.1	54.3	65.2
	OPSD	78.1	63.2	55.3	65.5
	OWLET	**80.1**	**67.6**	**57.3**	**68.3**

Table 13: Video-MME accuracy (%) by video duration (Qwen3.5-9B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Vanilla	69.1	54.8	50.7	58.2
	SFT	70.2	54.0	52.4	58.9
	GRPO	70.4	58.0	52.1	60.2
	OPSD	70.7	56.7	52.6	60.0
	OWLET	**71.2**	**60.8**	**54.6**	**62.2**
32	Vanilla	75.1	58.8	52.1	62.0
	SFT	75.3	59.5	52.9	62.6
	GRPO	77.1	59.1	53.4	63.2
	OPSD	75.7	60.3	53.7	63.2
	OWLET	**79.6**	**67.3**	**56.1**	**67.7**
64	Vanilla	76.6	63.8	54.5	65.0
	SFT	77.6	64.7	55.3	65.9
	GRPO	78.2	65.4	55.3	66.3
	OPSD	77.4	65.0	55.9	66.1
	OWLET	**82.0**	**70.7**	**59.4**	**70.7**

All improvements below are percentage-point differences from the strongest of SFT, GRPO, and OPSD for the specified metric, model size, mode, and frame budget.

Results in non-thinking mode. OWLET achieves the highest accuracy on all five benchmarks and the highest macro-average in each of the nine non-thinking configurations. Its macro-average advantage ranges from 3.1 to 6.5 points. For the 2B model, the macro-average rises from 44.7 at 16 frames to 47.4 at 32 frames and 49.2 at 64 frames, exceeding the corresponding strongest baselines by 4.8, 5.5, and 6.3 points. The 4B model also shows large gains: at 16 frames, it reaches 50.3 compared with 43.8 for GRPO, giving the largest macro-average margin of 6.5 points. The advantage remains positive at 8B, where the margins are 3.1, 3.6, and 3.5 points across the three budgets. These results show that the improvement extends across all evaluated Qwen3-VL sizes, with larger margins for the 2B and 4B models in this mode.

Table 14: Video-MME accuracy (%) by video duration (Qwen3.5-27B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Vanilla	74.5	63.5	56.5	64.8
	SFT	74.3	62.8	57.1	64.7
	GRPO	75.6	64.5	57.9	66.0
	OPSD	75.4	64.1	58.0	65.8
	OWLET	**77.9**	**67.7**	**59.0**	**68.2**
32	Vanilla	79.7	67.5	59.6	68.9
	SFT	80.1	66.5	60.3	69.0
	GRPO	81.3	68.2	61.7	70.4
	OPSD	82.0	67.9	62.3	70.7
	OWLET	**84.0**	**71.1**	**64.8**	**73.3**
64	Vanilla	84.1	70.2	59.8	71.4
	SFT	84.0	70.8	60.6	71.8
	GRPO	85.1	70.4	60.5	72.0
	OPSD	84.2	70.0	60.0	71.4
	OWLET	**87.8**	**74.8**	**65.8**	**76.1**

Table 15: Accuracy (%) for Qwen3-VL in non-thinking mode at 16 input frames, with the two Video-MME 2 question groups reported separately. *Logic* and *Rel.* are the group types and *All* combines them by question count. We **bold** the best result for each backbone and benchmark.

Backbone Model	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2			MLVU	Avg.
					Logic	Rel.	All		
Qwen3-VL-2B	Vanilla	45.2	53.1	42.9	14.4	14.3	14.3	34.8	38.1
	SFT	45.8	53.1	43.4	15.4	14.9	15.1	35.8	38.6
	GRPO	47.0	55.4	44.2	15.8	15.9	15.9	37.2	39.9
	OPSD	46.5	53.2	45.3	15.9	15.5	15.6	36.7	39.5
	OWLET	**52.4**	**61.2**	**50.7**	**19.9**	**18.1**	**18.7**	**40.6**	**44.7**
Qwen3-VL-4B	Vanilla	48.7	60.1	48.5	14.6	15.3	15.0	37.3	41.9
	SFT	50.1	60.9	49.9	14.2	16.2	15.5	37.1	42.7
	GRPO	50.4	62.3	49.5	16.7	18.2	17.7	39.1	43.8
	OPSD	50.2	60.8	49.9	15.3	16.7	16.2	38.9	43.2
	OWLET	**58.9**	**69.8**	**55.3**	**23.8**	**19.0**	**20.7**	**47.0**	**50.3**
Qwen3-VL-8B	Vanilla	57.9	63.0	53.2	26.2	21.5	23.1	40.3	47.5
	SFT	59.2	63.5	54.6	27.3	22.1	23.9	40.0	48.3
	GRPO	59.2	64.6	53.8	27.3	21.6	23.6	42.1	48.7
	OPSD	59.1	64.2	54.2	27.1	22.6	24.2	41.6	48.7
	OWLET	**60.7**	**68.2**	**58.5**	**29.8**	**23.6**	**25.8**	**45.6**	**51.8**

Results in thinking mode. The same ordering holds in all nine thinking configurations: OWLET leads the baselines on every benchmark and in macro-average accuracy. Using the reported macro-averages, its advantage ranges from 2.6 to 4.3 points. At 64 frames, the 2B, 4B, and 8B models reach 51.0, 56.6, and 56.7, improving on the strongest baselines by 2.9, 2.6, and 2.7 points, respectively. The strongest baseline varies across configurations: SFT leads at 2B with 16 frames, OPSD at 2B with 32 frames, and GRPO at 2B with 64 frames. The advantage of OWLET therefore persists across different baseline orderings and when the student produces an explicit reasoning trace. This supports the usefulness of visual-evidence supervision for thinking models as well as for models evaluated in non-thinking mode.

Temporal understanding and long-video reasoning. The per-benchmark results show that the macro-average improvement is supported by complementary video-understanding tasks. In non-thinking mode at 2B with 64 frames, OWLET scores 62.2 on MVBench, 56.0 on LongVideoBench,

Table 16: Accuracy (%) for Qwen3-VL in non-thinking mode at 32 input frames, with the two Video-MME 2 question groups reported separately. *Logic* and *Rel.* are the group types and *All* combines them by question count. We **bold** the best result for each backbone and benchmark.

Backbone Model	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2			MLVU	Avg.
					Logic	Rel.	All		
Qwen3-VL-2B	Vanilla	49.4	53.8	44.3	16.1	15.3	15.6	37.1	40.0
	SFT	50.3	54.8	46.9	17.1	16.4	16.6	38.8	41.5
	GRPO	51.8	55.0	45.4	18.2	16.6	17.2	40.2	41.9
	OPSD	51.0	55.3	46.4	18.1	16.4	17.0	38.1	41.6
	OWLET	**56.2**	**60.9**	**54.9**	**22.7**	**18.1**	**19.7**	**45.3**	**47.4**
Qwen3-VL-4B	Vanilla	54.3	61.4	51.1	16.4	15.1	15.6	45.1	45.5
	SFT	55.7	62.7	54.0	17.9	17.9	17.9	45.9	47.2
	GRPO	55.0	64.2	51.8	18.8	15.1	16.4	47.3	46.9
	OPSD	55.8	62.8	52.5	18.5	17.0	17.5	46.2	47.0
	OWLET	**62.4**	**71.5**	**58.3**	**25.9**	**21.5**	**23.1**	**50.4**	**53.1**
Qwen3-VL-8B	Vanilla	62.1	64.2	55.1	28.6	22.0	24.3	45.6	50.3
	SFT	63.2	64.4	57.2	28.6	23.1	25.0	46.6	51.3
	GRPO	62.7	65.9	54.4	29.2	23.4	25.4	46.5	51.0
	OPSD	63.1	64.8	55.4	29.4	23.2	25.4	47.1	51.1
	OWLET	**65.9**	**70.1**	**60.5**	**31.9**	**23.9**	**26.7**	**51.2**	**54.9**

Table 17: Accuracy (%) for Qwen3-VL in non-thinking mode at 64 input frames, with the two Video-MME 2 question groups reported separately. *Logic* and *Rel.* are the group types and *All* combines them by question count. We **bold** the best result for each backbone and benchmark.

Backbone Model	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2			MLVU	Avg.
					Logic	Rel.	All		
Qwen3-VL-2B	Vanilla	49.2	53.9	45.1	16.8	17.3	17.1	40.3	41.1
	SFT	50.1	54.6	46.8	18.1	18.0	18.0	41.8	42.3
	GRPO	51.4	55.0	47.6	17.9	19.7	19.1	41.7	42.9
	OPSD	50.7	54.0	47.1	18.1	18.2	18.2	42.1	42.4
	OWLET	**58.3**	**62.2**	**56.0**	**22.8**	**20.2**	**21.1**	**48.2**	**49.2**
Qwen3-VL-4B	Vanilla	56.4	63.0	51.3	17.6	16.4	16.8	50.1	47.5
	SFT	57.7	64.4	52.6	17.9	16.5	17.0	51.2	48.6
	GRPO	58.5	64.4	53.2	18.3	17.5	17.8	52.6	49.3
	OPSD	57.4	63.6	53.3	18.3	16.9	17.4	51.5	48.6
	OWLET	**64.7**	**71.0**	**60.8**	**28.1**	**22.8**	**24.7**	**56.9**	**55.6**
Qwen3-VL-8B	Vanilla	66.0	64.2	56.5	30.2	24.1	26.3	51.8	53.0
	SFT	66.5	65.3	58.0	30.6	25.1	27.0	52.9	53.9
	GRPO	67.5	65.0	57.1	31.2	25.5	27.5	52.2	53.9
	OPSD	67.0	65.1	57.2	31.5	24.3	26.8	53.2	53.9
	OWLET	**69.4**	**70.6**	**62.8**	**33.5**	**26.8**	**29.2**	**55.2**	**57.4**

and 48.2 on MLVU, exceeding the strongest baseline on each benchmark by 7.2, 8.4, and 6.1 points. In thinking mode at 8B with 32 frames, it reaches 71.8 on MVBench and 53.0 on MLVU, compared with strongest-baseline scores of 66.9 and 45.5. These gains span temporal perception, reasoning about referred content within a long context, and tasks that combine broad video understanding with access to specific details. Their occurrence in both modes is consistent with the teacher view's use of global context together with question-conditioned local evidence.

Logic and relevance in Video-MME 2. Both question subsets improve over the strongest baseline in every model–mode–budget configuration. In non-thinking mode at 4B with 64 frames, OWLET reaches 28.1 on logic and 22.8 on relevance, compared with strongest-baseline scores of 18.3 and 17.5. The respective gains of 9.8 and 5.3 points show that the combined improvement is supported by both subsets, although their margins differ. In thinking mode at 8B with 64 frames, logic accuracy rises to 29.5 from 28.3 and relevance accuracy to 26.6 from 24.0, giving gains of 1.2 and 2.6 points.

Table 18: Video-MME accuracy (%) by video duration for Qwen3-VL in non-thinking mode (Qwen3-VL-2B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Vanilla	58.4	43.1	34.2	45.2
	SFT	59.0	43.4	34.9	45.8
	GRPO	59.8	44.9	36.3	47.0
	OPSD	59.9	44.6	35.1	46.5
	OWLET	**64.4**	**51.4**	**41.4**	**52.4**
32	Vanilla	65.1	48.0	35.0	49.4
	SFT	66.0	49.2	35.7	50.3
	GRPO	67.4	50.2	37.7	51.8
	OPSD	66.4	50.0	36.7	51.0
	OWLET	**69.0**	**56.0**	**43.7**	**56.2**
64	Vanilla	66.7	47.6	33.4	49.2
	SFT	67.3	48.5	34.5	50.1
	GRPO	69.2	50.5	34.4	51.4
	OPSD	67.9	49.1	35.1	50.7
	OWLET	**72.7**	**57.2**	**45.0**	**58.3**

Table 19: Video-MME accuracy (%) by video duration for Qwen3-VL in non-thinking mode (Qwen3-VL-4B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Vanilla	62.2	45.9	37.9	48.7
	SFT	64.2	47.7	38.4	50.1
	GRPO	64.7	46.8	39.8	50.4
	OPSD	64.4	47.0	39.2	50.2
	OWLET	**69.1**	**58.4**	**49.2**	**58.9**
32	Vanilla	68.9	50.4	43.8	54.3
	SFT	71.1	50.9	45.1	55.7
	GRPO	69.3	51.5	44.3	55.0
	OPSD	70.5	52.2	44.8	55.8
	OWLET	**74.7**	**59.9**	**52.8**	**62.4**
64	Vanilla	70.1	53.4	45.9	56.4
	SFT	71.2	54.9	46.9	57.7
	GRPO	72.0	55.7	47.8	58.5
	OPSD	71.6	54.3	46.3	57.4
	OWLET	**77.1**	**63.6**	**53.4**	**64.7**

The subset breakdown thus reveals variation in which questions benefit most while retaining a positive improvement on both types under the common per-question accuracy protocol.

Performance across video durations. The Video-MME duration tables show that OWLET outperforms the baselines on short, medium, and long videos in all eighteen configurations. In non-thinking mode at 2B with 64 frames, its short, medium, and long accuracies are 72.7, 57.2, and 45.0, exceeding the strongest baselines in the corresponding splits by 3.5, 6.7, and 9.9 points. In thinking mode at 4B with 32 frames, the three scores are 75.1, 60.6, and 53.3, with margins of 1.8, 3.9, and 4.8 points. The larger gains on long videos in these configurations are consistent with retaining global temporal context around the selected evidence. The improvements on shorter videos additionally show that evidence-guided training remains useful when the relevant information lies within a smaller temporal span.

Table 20: Video-MME accuracy (%) by video duration for Qwen3-VL in non-thinking mode (Qwen3-VL-8B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Vanilla	68.4	55.1	50.0	57.9
	SFT	69.6	57.2	50.9	59.2
	GRPO	70.9	54.9	51.9	59.2
	OPSD	69.5	56.4	51.4	59.1
	OWLET	**72.2**	**57.8**	**52.2**	**60.7**
32	Vanilla	74.0	58.0	54.1	62.1
	SFT	74.6	60.0	55.0	63.2
	GRPO	75.4	58.2	54.5	62.7
	OPSD	75.3	58.4	55.5	63.1
	OWLET	**77.3**	**63.2**	**57.1**	**65.9**
64	Vanilla	78.7	64.0	55.3	66.0
	SFT	79.2	64.5	55.8	66.5
	GRPO	80.3	65.5	56.6	67.5
	OPSD	80.0	65.0	55.9	67.0
	OWLET	**82.3**	**68.0**	**58.0**	**69.4**

Table 21: Accuracy (%) for Qwen3-VL in thinking mode at 16 input frames, with the two Video-MME 2 question groups reported separately. *Logic* and *Rel.* are the group types and *All* combines them by question count. We **bold** the best result for each backbone and benchmark.

Backbone Model	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2			MLVU	Avg.
					Logic	Rel.	All		
Qwen3-VL-2B	Base model	49.3	60.5	45.1	17.8	15.3	16.2	35.1	41.3
	SFT	51.0	61.5	46.0	19.4	16.8	17.7	37.1	42.7
	GRPO	50.9	61.8	46.0	18.3	15.2	16.3	36.3	42.3
	OPSD	50.0	61.3	46.6	19.9	16.6	17.8	35.2	42.2
	OWLET	**54.2**	**65.9**	**52.8**	**22.1**	**17.4**	**19.1**	**43.0**	**47.0**
Qwen3-VL-4B	Base model	54.5	64.2	55.4	23.1	18.6	20.2	41.3	47.1
	SFT	56.0	64.8	55.5	23.8	20.4	21.6	42.8	48.1
	GRPO	55.5	65.5	56.3	22.5	19.5	20.6	42.0	48.0
	OPSD	55.5	65.9	56.4	24.5	19.1	21.0	42.6	48.3
	OWLET	**59.1**	**69.7**	**58.2**	**25.1**	**21.0**	**22.4**	**47.4**	**51.4**
Qwen3-VL-8B	Base model	55.3	65.3	54.0	23.1	19.1	20.5	40.0	47.0
	SFT	55.8	66.1	55.1	23.5	19.5	20.9	41.7	47.9
	GRPO	57.0	66.6	54.4	24.0	20.4	21.7	41.8	48.3
	OPSD	56.4	66.6	55.5	24.0	19.5	21.1	41.3	48.2
	OWLET	**60.5**	**70.5**	**57.0**	**27.2**	**22.1**	**23.8**	**47.0**	**51.8**

Robustness across frame budgets. For each of the six model–mode combinations, the macro-average of OWLET increases as the student budget grows from 16 to 32 and then to 64 frames. For example, the 4B thinking model reaches 51.4, 54.5, and 56.6, and retains an advantage over the strongest baseline at every budget. Individual benchmark scores can respond differently to additional frames: the 2B thinking model's MVBench accuracy changes from 67.2 at 32 frames to 66.8 at 64 frames, while its macro-average increases from 49.4 to 51.0. The consistent result across budgets is therefore the improvement over baselines evaluated at the same frame budget. This pattern indicates that evidence-guided post-training remains useful both when the student view is sparse and when more frames are available.

Comparison between reasoning modes. The comparison between thinking and non-thinking modes varies with the backbone and frame budget. At 2B with 64 frames, OWLET reaches a macro-average of 51.0 in thinking mode and 49.2 in non-thinking mode. At 8B with the same frame budget, the ordering reverses: the non-thinking model scores 57.4 and the thinking model 56.7.

Table 22: Accuracy (%) for Qwen3-VL in thinking mode at 32 input frames, with the two Video-MME 2 question groups reported separately. *Logic* and *Rel.* are the group types and *All* combines them by question count. We **bold** the best result for each backbone and benchmark.

Backbone Model	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2			MLVU	Avg.
					Logic	Rel.	All		
Qwen3-VL-2B	Base model	52.2	61.0	51.0	20.1	16.1	17.5	38.0	44.0
	SFT	53.3	62.1	52.3	20.3	16.7	18.0	38.1	44.8
	GRPO	53.5	61.2	51.9	21.9	17.4	19.0	39.1	44.9
	OPSD	53.1	62.1	52.9	21.3	17.4	18.8	39.1	45.2
	OWLET	**57.4**	**67.2**	**55.0**	**25.5**	**19.4**	**21.5**	**46.0**	**49.4**
Qwen3-VL-4B	Base model	58.9	67.0	57.3	25.2	20.3	22.0	49.4	50.9
	SFT	59.4	68.0	57.9	25.8	20.4	22.3	49.4	51.4
	GRPO	58.5	66.7	57.2	25.5	20.2	22.1	49.1	50.7
	OPSD	59.5	67.3	57.5	26.2	20.8	22.7	49.0	51.2
	OWLET	**63.0**	**71.5**	**60.2**	**28.9**	**23.4**	**25.3**	**52.4**	**54.5**
Qwen3-VL-8B	Base model	61.2	66.1	58.2	26.1	20.1	22.2	45.0	50.5
	SFT	62.1	66.9	59.1	26.2	20.7	22.6	45.5	51.3
	GRPO	61.8	66.0	58.8	27.3	21.3	23.4	44.4	50.9
	OPSD	61.6	66.9	58.5	26.4	20.9	22.8	45.3	51.0
	OWLET	**64.9**	**71.8**	**60.6**	**29.1**	**24.1**	**25.8**	**53.0**	**55.2**

Table 23: Accuracy (%) for Qwen3-VL in thinking mode at 64 input frames, with the two Video-MME 2 question groups reported separately. *Logic* and *Rel.* are the group types and *All* combines them by question count. We **bold** the best result for each backbone and benchmark.

Backbone Model	Method	Video-MME	MVBench	LongVideoBench	Video-MME 2			MLVU	Avg.
					Logic	Rel.	All		
Qwen3-VL-2B	Base model	55.5	63.1	51.4	20.1	18.1	18.8	46.0	47.0
	SFT	56.4	65.2	52.1	20.6	19.5	19.9	46.2	48.0
	GRPO	56.7	64.0	52.4	21.2	19.2	19.9	47.3	48.1
	OPSD	56.5	64.7	52.0	20.9	18.8	19.5	47.1	48.0
	OWLET	**60.3**	**66.8**	**55.6**	**25.6**	**21.3**	**22.8**	**49.6**	**51.0**
Qwen3-VL-4B	Base model	61.8	69.1	58.1	26.1	21.1	22.8	52.1	52.8
	SFT	63.2	70.3	58.8	27.5	21.3	23.5	52.7	53.7
	GRPO	62.5	70.7	59.7	26.8	21.2	23.2	53.8	54.0
	OPSD	63.0	69.6	58.9	26.6	21.6	23.4	52.1	53.4
	OWLET	**66.2**	**73.8**	**61.1**	**30.0**	**23.7**	**25.9**	**56.0**	**56.6**
Qwen3-VL-8B	Base model	63.0	68.7	59.9	26.9	22.3	23.9	48.8	52.9
	SFT	64.3	70.0	60.1	28.0	24.0	25.4	49.9	53.9
	GRPO	64.2	70.2	60.5	28.3	23.8	25.4	49.6	54.0
	OPSD	64.2	69.4	61.4	27.5	22.3	24.1	49.5	53.7
	OWLET	**66.4**	**73.5**	**63.0**	**29.5**	**26.6**	**27.6**	**53.0**	**56.7**

In both cases, OWLET improves on the strongest baseline within each mode. Together with the Qwen3.5 results, these comparisons cover both evaluated backbone families and reasoning modes, with different margins across configurations.

E.6 Additional Motivation Evidence Across Models and Frame Budgets

The main text presents the motivation diagnostics for Qwen3.5-9B with a 64-frame student view. This subsection reports the same two supervision gaps on four additional configurations: Qwen3.5-0.8B at 16 and 64 frames, Qwen3.5-27B at 64 frames, and Qwen3-VL-8B in thinking mode at 64 frames. The left panel measures concentration of the teacher–student absolute log-probability gap across response tokens, and the right panel follows output KL and hidden divergence during a 300-step output-only distillation diagnostic.

Table 24: Video-MME accuracy (%) by video duration for Qwen3-VL in thinking mode (Qwen3-VL-2B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Base model	61.5	46.3	40.1	49.3
	SFT	63.5	48.5	41.1	51.0
	GRPO	63.0	47.5	42.2	50.9
	OPSD	62.6	46.9	40.5	50.0
	OWLET	**66.0**	**52.8**	**43.8**	**54.2**
32	Base model	65.2	50.4	41.0	52.2
	SFT	66.2	52.0	41.8	53.3
	GRPO	66.1	51.6	42.7	53.5
	OPSD	66.2	51.9	41.2	53.1
	OWLET	**70.3**	**54.4**	**47.3**	**57.4**
64	Base model	71.5	52.3	42.5	55.5
	SFT	72.6	52.5	44.2	56.4
	GRPO	73.4	53.8	42.9	56.7
	OPSD	72.0	52.6	44.8	56.5
	OWLET	**74.6**	**59.3**	**47.0**	**60.3**

Table 25: Video-MME accuracy (%) by video duration for Qwen3-VL in thinking mode (Qwen3-VL-4B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Base model	65.4	52.1	46.0	54.5
	SFT	67.1	53.4	47.5	56.0
	GRPO	66.9	52.9	46.7	55.5
	OPSD	66.2	53.0	47.4	55.5
	OWLET	**70.3**	**56.5**	**50.6**	**59.1**
32	Base model	72.4	56.1	48.2	58.9
	SFT	73.2	56.6	48.3	59.4
	GRPO	72.1	54.9	48.4	58.5
	OPSD	73.3	56.7	48.5	59.5
	OWLET	**75.1**	**60.6**	**53.3**	**63.0**
64	Base model	74.2	61.1	50.3	61.8
	SFT	75.8	62.2	51.6	63.2
	GRPO	74.5	61.4	51.5	62.5
	OPSD	75.4	62.3	51.4	63.0
	OWLET	**78.3**	**65.8**	**54.6**	**66.2**

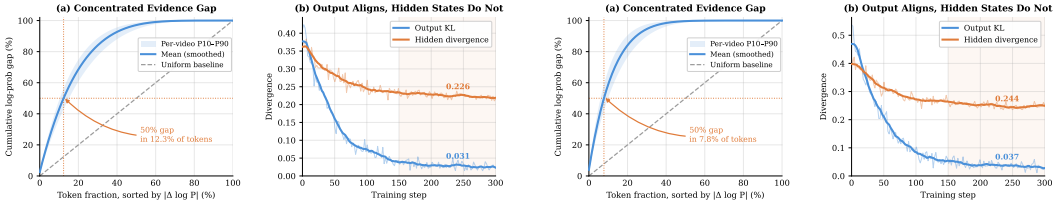
Concentrated evidence sensitivity across budgets and scales. For Qwen3.5-0.8B, increasing the student budget from 16 to 64 frames moves the 50%-gap point from 12.3% to 7.8% of response tokens. At 64 frames, Qwen3.5-27B reaches the same point at 5.2% of tokens. In each setting, a small fraction of response tokens carries half of the gap.

The pattern transfers to another backbone and reasoning mode. The Qwen3-VL-8B thinking configuration places 50% of its gap on 5.7% of response tokens at 64 frames. Its curve has the same qualitative shape as the Qwen3.5 curves, with a small fraction of tokens accounting for a large fraction of the view-dependent likelihood shift.

Output alignment does not remove the hidden-state gap. The right panels show the complementary supervision gap. At the final plotted step, output KL lies between 0.031 and 0.068 across the four configurations, whereas hidden divergence remains between 0.226 and 0.489. For example, Qwen3.5-0.8B at 16 frames reaches 0.031 output KL but retains 0.226 hidden divergence; Qwen3.5-

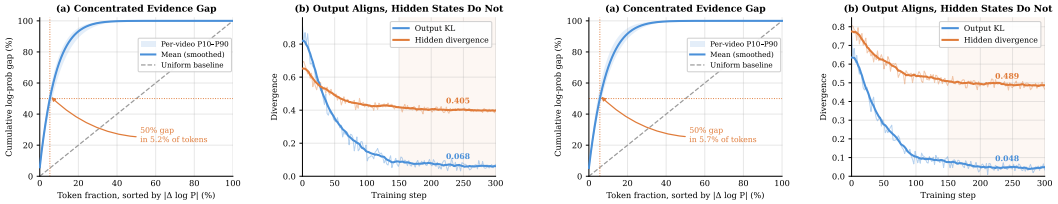
Table 26: Video-MME accuracy (%) by video duration for Qwen3-VL in thinking mode (Qwen3-VL-8B). **All** averages the three splits. We **bold** the best result for each frame budget and duration.

Frames	Method	Short	Medium	Long	All
16	Base model	66.1	52.6	47.2	55.3
	SFT	67.0	52.8	47.7	55.8
	GRPO	68.4	54.4	48.2	57.0
	OPSD	67.8	53.5	47.8	56.4
	OWLET	**71.4**	**58.8**	**51.2**	**60.5**
32	Base model	73.2	57.2	53.2	61.2
	SFT	73.8	58.4	54.2	62.1
	GRPO	73.7	57.9	53.7	61.8
	OPSD	73.7	57.7	53.5	61.6
	OWLET	**76.6**	**62.5**	**55.7**	**64.9**
64	Base model	75.5	62.5	51.1	63.0
	SFT	76.9	63.2	52.7	64.3
	GRPO	76.8	63.5	52.2	64.2
	OPSD	76.9	63.6	52.2	64.2
	OWLET	**78.1**	**66.2**	**54.8**	**66.4**



(a) Qwen3.5-0.8B, non-thinking, 16 frames.

(b) Qwen3.5-0.8B, non-thinking, 64 frames.



(c) Qwen3.5-27B, non-thinking, 64 frames.

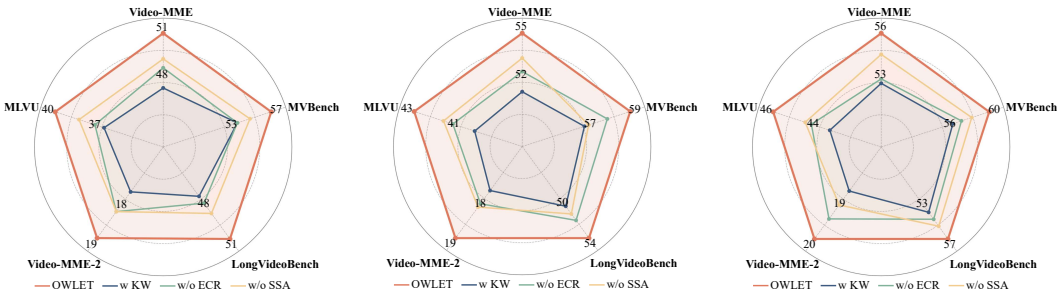
(d) Qwen3-VL-8B, thinking, 64 frames.

Figure 7: **Additional motivation diagnostics across four configurations.** Each subfigure contains the same two diagnostics: the left panel reports the cumulative absolute teacher–student log-probability gap, with the marked point giving the response-token fraction that carries 50% of the gap; the right panel reports output KL and hidden divergence during 300 steps of output-only distillation.

27B at 64 frames reaches 0.068 and 0.405, respectively; and Qwen3-VL-8B thinking reaches 0.048 and 0.489. The exact scales vary with the model and budget, but output-only matching consistently leaves a non-zero representation discrepancy in these diagnostics. This is the empirical gap that motivates adding SSA to the token-level evidence signal used by ECR.

E.7 Ablation Across Models and Frame Budgets

The ablations in this subsection extend the Qwen3.5-9B comparison in Section 5.3 to Qwen3.5-0.8B, 4B, and 27B in non-thinking mode, and to Qwen3-VL-2B, 4B, and 8B in both non-thinking and thinking modes. Each model–mode combination is considered at student frame budgets of $N \in \{16, 32, 64\}$, giving 27 configurations in the appendix. The comparisons address three parts of



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 8: **Ablation results for Qwen3.5-0.8B.** Panels (a)–(c) use 16, 32, and 64 student frames, respectively. The five axes report accuracy (%) on Video-MME, MVBench, LongVideoBench, Video-MME 2, and MLVU. The legend defines the key-window, w/o ECR, and w/o SSA variants described in the ablation text; evaluation uses uniformly sampled student frames.

the design: the temporal context supplied by the teacher view, the allocation of positive credit across response tokens, and the alignment of the representations formed under the two views.

Comparison setup. Within each model–mode–budget configuration, the full method and its ablations share the training data, optimizer settings, student rollout budget, and evaluation protocol. Each variant changes one component while retaining the others. The student generates responses from the uniformly sampled view, and every evaluation uses N uniformly sampled frames.

Ablation variants. The four traces in each panel correspond to the following configurations:

- **OWLET** uses the global-plus-local teacher view together with ECR and SSA. ECR redistributes positive response-level advantages according to the evidence contrast of Equation 4, and SSA aligns response-end hidden states on verifier-accepted responses through Equation 6.
- **w/ Key Windows** replaces the global-plus-local teacher view with question-conditioned local windows, while retaining ECR and SSA.
- **w/o ECR** removes the evidence-dependent token weights while retaining the global-plus-local teacher view and SSA. In Equation 4, the token advantages reduce to $\widetilde{A}_t^i = A^i$.
- **w/o SSA** sets $\lambda_{\mathrm{SSA}} = 0$ in Equation 7, while preserving the global-plus-local teacher view and ECR.

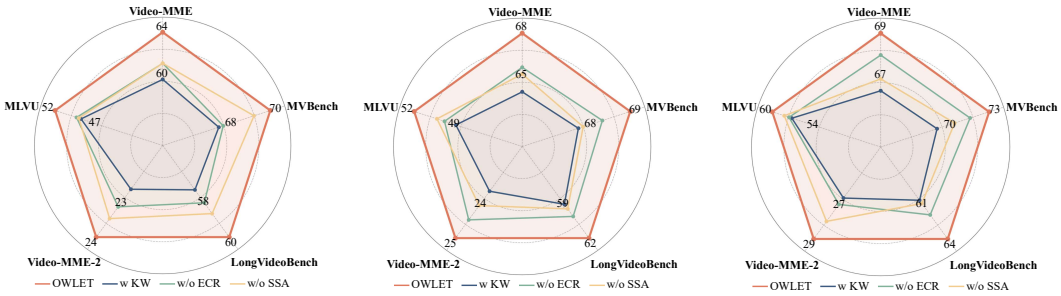
Removing both ECR and SSA would recover the GRPO objective of Equation 1; the reported ablations remove one component at a time.

Comparisons across configurations. Within each figure, the three panels hold the model and reasoning mode fixed while varying the student frame budget; the remaining figures vary model size and reasoning mode.

Across the grouped plots, removing ECR or SSA lowers accuracy by different amounts across tasks, and replacing the global-plus-local teacher view with key windows is the strongest reduction in many panels. Increasing the student frame budget changes absolute accuracy and individual margins, while the full method remains the leading configuration across the evaluated Qwen3.5 and Qwen3-VL panels.

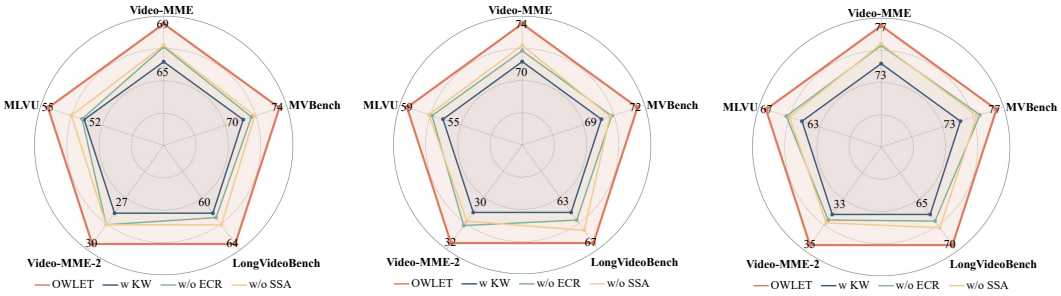
F Additional Analyses

This section reports nine analyses covering alternative distillation objectives, teacher and SSA settings, token-level credit, qualitative responses, training cost and response length, and post-training behavior beyond the standard recipe. Unless stated otherwise, every analysis uses Qwen3.5-9B with a 64-frame student view and the training configuration of Appendix E.3.



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 9: **Ablation results for Qwen3.5-4B.** Panels (a)–(c) correspond to 16, 32, and 64 uniformly sampled student frames.



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 10: **Ablation results for Qwen3.5-27B.** Panels (a)–(c) correspond to 16, 32, and 64 uniformly sampled student frames.

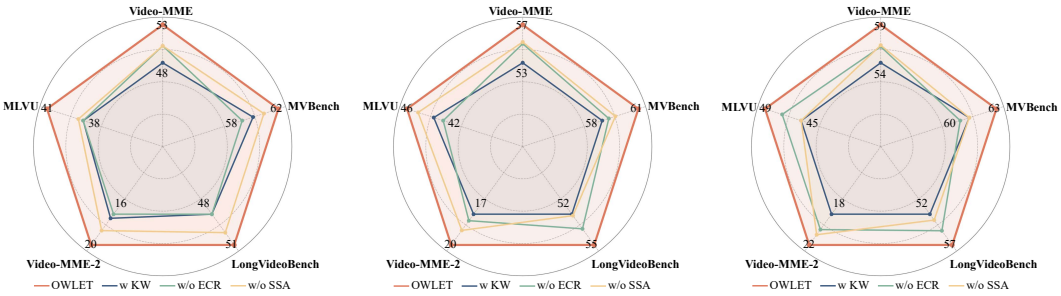
F.1 Comparison with On-Policy Distillation Variants

Section 5.2 compares OWLET against OPSD with the token-level KL objective of Equation 2. Here we compare alternative output divergences and distillation–GRPO schedules on Qwen3.5-9B with 64 uniformly sampled student frames. All methods receive the same training data and student rollout budget, and all teacher-assisted objectives use the teacher view combining global context with local windows. Evaluation uses the ordinary student view without evidence annotations or teacher computation.

Compared objectives. We retain the standard OPSD objective as the reference and consider three variants:

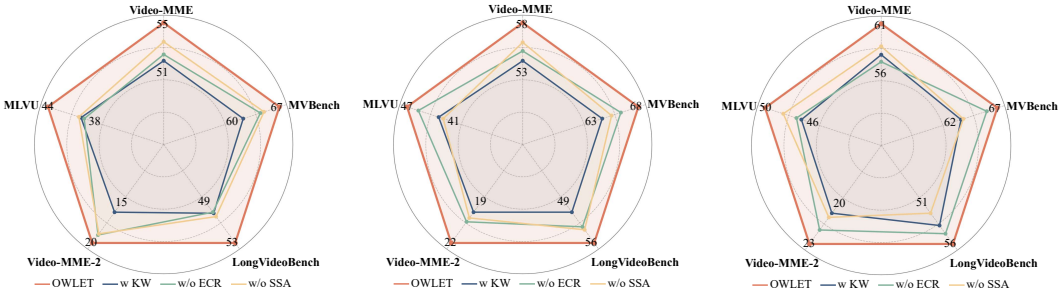
- **OPSD-JSD** replaces the student-to-teacher KL in Equation 2 with the generalized Jensen–Shannon divergence, using an interpolation coefficient of $\alpha = 0.5$. This comparison changes the output divergence while preserving on-policy sampling from the student.
- **OPSD+GRPO** combines the distillation loss with the GRPO objective of Equation 1, using an equal weight. The GRPO term applies a response-level advantage uniformly across tokens; it does not use evidence-aware credit redistribution or hidden-state alignment.
- **OPSD-Anneal** combines distillation with GRPO using equal initial weights for the two objectives, and linearly decays the distillation weight to zero, so that training ends in pure GRPO. This variant tests a schedule that uses teacher supervision early and leaves GRPO as the sole objective at the end.

These variants vary how output-distribution supervision is defined or combined with the response-level reward.



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 11: **Ablation results for Qwen3-VL-2B in non-thinking mode.** Panels (a)–(c) correspond to 16, 32, and 64 uniformly sampled student frames.



(a) 16 frames (b) 32 frames (c) 64 frames

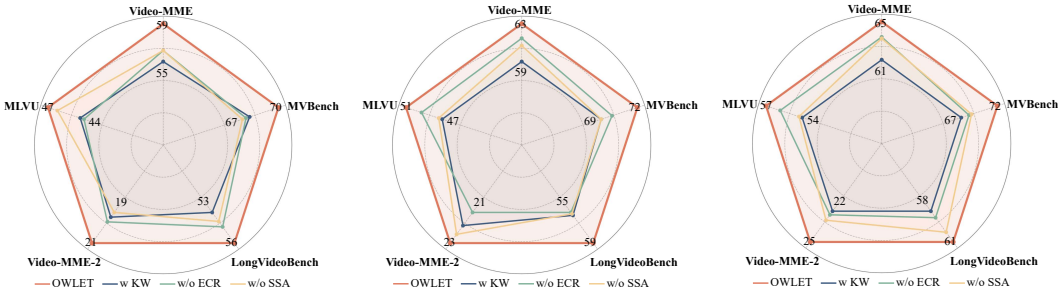
Figure 12: **Ablation results for Qwen3-VL-2B in thinking mode.** Panels (a)–(c) correspond to 16, 32, and 64 uniformly sampled student frames.

Overall comparison. Table 27 shows that the three modifications have different effects. OPSD-Anneal achieves the highest macro-average among the distillation configurations, at 56.3, followed by standard OPSD at 55.5, OPSD+GRPO at 55.4, and OPSD-JSD at 54.9. Thus, neither using the Jensen–Shannon divergence nor adding GRPO yields an aggregate improvement over standard OPSD in this setting. Annealing improves the macro-average, but OWLET remains ahead at 59.8 and obtains the highest accuracy on every benchmark. All benchmark differences discussed below are absolute percentage-point differences.

Effect of the divergence choice. OPSD-JSD scores lower than standard reverse-KL OPSD on four of the five benchmarks: Video-MME decreases from 66.1 to 65.3, MVBench from 68.7 to 68.2, LongVideoBench from 62.9 to 61.5, and Video-MME 2 from 26.7 to 25.9. MLVU is the exception, increasing from 53.3 to 53.7. The resulting macro-average is 54.9, compared with 55.5 for standard OPSD.

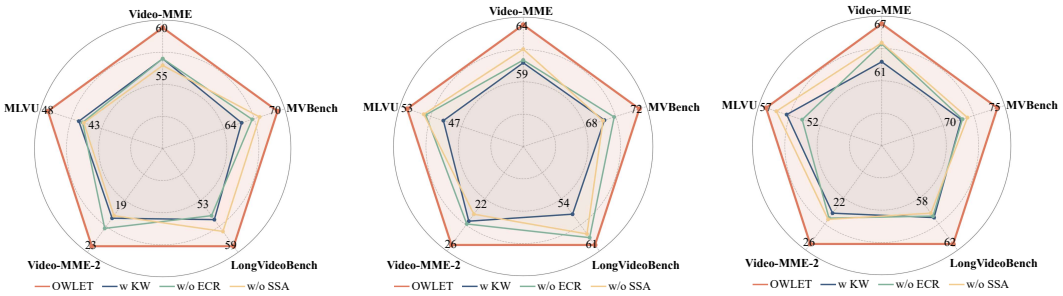
Combining distillation with GRPO. OPSD+GRPO improves over standard OPSD on Video-MME, LongVideoBench, and Video-MME 2 by 0.2, 0.4, and 0.3 points, respectively. These gains are offset by decreases of 0.9 points on MVBench and 0.6 points on MLVU, leaving the macro-average at 55.4, close to the 55.5 of standard OPSD. Adding the GRPO term changes which benchmarks benefit without improving the overall average in this run.

Annealing the distillation objective. OPSD-Anneal outperforms both OPSD-JSD and OPSD+GRPO on all five benchmarks and is the strongest distillation variant by macro-average. Relative to standard OPSD, it improves Video-MME by 1.0 point, LongVideoBench by 0.5, Video-MME 2 by 0.4, and MLVU by 2.0, while MVBench decreases slightly from 68.7 to 68.5. The macro-average rises from 55.5 to 56.3.



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 13: **Ablation results for Qwen3-VL-4B in non-thinking mode.** Panels (a)–(c) correspond to 16, 32, and 64 uniformly sampled student frames.



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 14: **Ablation results for Qwen3-VL-4B in thinking mode.** Panels (a)–(c) correspond to 16, 32, and 64 uniformly sampled student frames.

Comparison with OWLET. Against OPSD-Anneal, OWLET improves Video-MME, MVBench, LongVideoBench, Video-MME 2, and MLVU by 3.6, 4.3, 1.8, 1.8, and 6.3 points, respectively. Its macro-average is 59.8, compared with 56.3 for OPSD-Anneal, for a 3.5-point advantage.

F.2 Teacher Parameterization

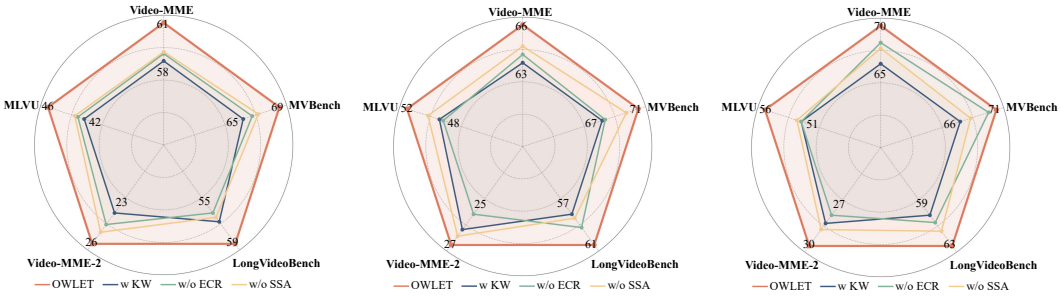
The teacher of OWLET shares the student's architecture and supplies token scores and hidden-state targets from the teacher view combining global context with local windows. We compare the default fixed teacher with an exponential moving average (EMA) teacher on Qwen3.5-9B with 64 student frames. The two settings use the same teacher-view construction, training data, student rollout budget, and ECR and SSA objectives; only the teacher parameters differ.

Fixed and EMA teachers. The fixed teacher retains $\bar{\theta} = \theta_0$ throughout the entire training run, where θ_0 is the common checkpoint used to initialize the teacher and student. The EMA variant instead maintains an exponential moving average of the student parameters with decay 0.99:

$$\bar{\theta} \leftarrow 0.99\,\bar{\theta} + 0.01\,\theta.$$

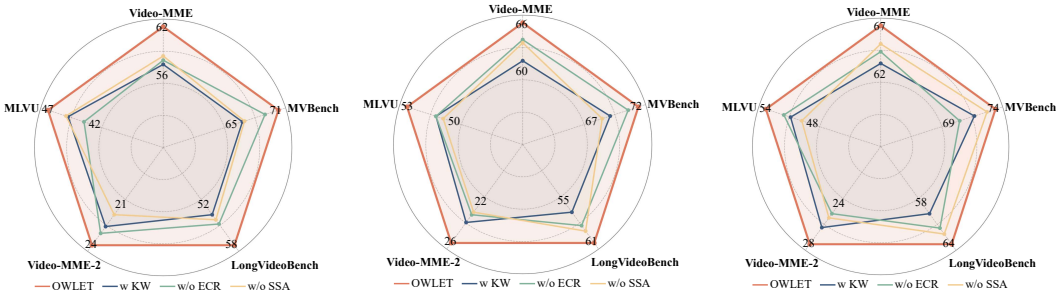
In both settings, the teacher scores student-generated responses under the privileged view and supplies the teacher term in the evidence contrast of Equation 3 and the representation targets for SSA in Equation 6. The fixed teacher's scores and hidden states are recomputed for each student response, while its weights remain unchanged. Evaluation uses only the student with uniformly sampled frames.

Results across benchmarks. Table 28 shows that the fixed teacher outperforms the EMA teacher on all five benchmarks. Its accuracies are 70.7, 72.8, 65.2, 28.9, and 61.6 on Video-MME, MVBench, LongVideoBench, Video-MME 2, and MLVU, respectively, compared with 69.3, 70.5, 64.7, 27.9, and 60.8 for EMA. The corresponding advantages are 1.4, 2.3, 0.5, 1.0, and 0.8 percentage points. The largest difference occurs on MVBench, while LongVideoBench shows the smallest difference.



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 15: **Ablation results for Qwen3-VL-8B in non-thinking mode.** Panels (a)–(c) correspond to 16, 32, and 64 uniformly sampled student frames.



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 16: **Ablation results for Qwen3-VL-8B in thinking mode.** Panels (a)–(c) correspond to 16, 32, and 64 uniformly sampled student frames.

The macro-average is 59.8 for the fixed teacher and 58.6 for EMA, a difference of 1.2 points. Thus, retaining the initial teacher weights is favored across the evaluated tasks, with the magnitude of the advantage varying by benchmark.

Effect on the evidence contrast. With the fixed teacher, the first term in Equation 3 is evaluated under θ_0 and the privileged view, while the second uses the evolving rollout-policy parameters θ_{old} and the uniform view on the same question, response token, and prefix. At initialization the weights coincide; after student updates, the contrast includes both visual and parameter differences. With EMA, the teacher side also changes as student parameters are incorporated into the average, changing the token scores used by ECR.

Effect on representation targets. Teacher parameterization also affects SSA because the response-end target states are computed by the teacher. The fixed setting uses θ_0 for these targets, whereas the EMA setting uses the evolving averaged parameters; both apply SSA only to verifier-accepted responses and backpropagate through the student.

Choice of the default teacher. The fixed teacher improves all five benchmarks over EMA with decay 0.99, increasing the macro-average from 58.6 to 59.8. We use it as the default throughout the paper.

F.3 Statistics of Advantage Signs and Highly Weighted Tokens

ECR redistributes only positive trajectory credit, so its reach depends on how often responses actually receive positive advantages, and its usefulness depends on which tokens end up carrying large weights. We measure both during the training run of Section 5.2.

Table 27: Accuracy (%) of on-policy distillation variants and OWLET on Qwen3.5-9B at 64 student frames. All methods use the same teacher view, training data, and student rollout budget. Avg. is the unweighted mean over the five benchmarks. The best result in each column is shown in bold.

Method	Video-MME	MVBench	LongVideoBench	Video-MME 2	MLVU	Avg.
OPSD	66.1	68.7	62.9	26.7	53.3	55.5
OPSD-JSD	65.3	68.2	61.5	25.9	53.7	54.9
OPSD+GRPO	66.3	67.8	63.3	27.0	52.7	55.4
OPSD-Anneal	67.1	68.5	63.4	27.1	55.3	56.3
OWLET	**70.7**	**72.8**	**65.2**	**28.9**	**61.6**	**59.8**

Table 28: Effect of the teacher parameterization on Qwen3.5-9B at 64 input frames. The fixed teacher retains the initial checkpoint throughout training, while the EMA teacher averages student parameters with decay 0.99 under the same teacher view and training configuration. Scores are accuracies (%), and Avg. is the unweighted mean over the five benchmarks.

Teacher $\bar{\theta}$	Video-MME	MVBench	LongVideoBench	Video-MME 2	MLVU	Avg.
EMA (decay 0.99)	69.3	70.5	64.7	27.9	60.8	58.6
Fixed (default)	**70.7**	**72.8**	**65.2**	**28.9**	**61.6**	**59.8**

Advantage signs. Table 29 reports, over four phases of training, the fraction of rollouts with positive, negative, and zero group-normalized advantages. With binary rewards and $G = 8$ rollouts per prompt, zero advantages occur when all responses in a group receive the same reward. Verifier-accepted responses from these groups can still contribute to SSA.

The first phase contains 28.5% positive-advantage rollouts, 7.2% negative-advantage rollouts, and 64.3% zero-advantage rollouts. By the final phase, these fractions become 4.7%, 2.2%, and 93.1%, respectively. Thus, the fraction carrying a nonzero group-normalized signal falls from 35.7% to 6.9%, while the zero-advantage fraction increases by 28.8 percentage points. The positive fraction, which is the part of the policy signal that ECR reweights through Equation 4, decreases by 23.8 percentage points across training; the negative fraction decreases in parallel by 5.0 points. Averaging the four equally sized phases gives approximately 16.4% positive, 4.8% negative, and 78.9% zero advantages over the analyzed run.

Table 29: Fractions (%) of rollouts with positive, negative, and zero advantages across training phases on Qwen3.5-9B at 64 input frames.

Advantage	Steps 0–250	250–500	500–750	750–1,000
$A^i > 0$	28.5	21.8	10.6	4.7
$A^i < 0$	7.2	5.8	3.8	2.2
$A^i = 0$	64.3	72.4	85.6	93.1

Which tokens receive large weights. To characterize where the redistributed credit goes, we use a sample of 100 question–video examples from the training split and evaluate the same diagnostic at three checkpoints. From each example, we retain the positively advantaged student responses and rank their valid response tokens by the redistributed weights $\widetilde{A}_t^i$. We compare the highest-weight and lowest-weight deciles and pool the selected token instances across the three checkpoints. Prompt tokens, padding, and special control tokens are excluded. The table therefore reports token-level composition, rather than the fraction of responses assigned to each category; longer responses contribute more selected token instances.

The category definitions are operational and mutually exclusive. After separating punctuation and formatting from lexical spans, we apply the following rules:

- **Answer-committing tokens** are option letters, answer numbers, or equivalent answer values that occur in an explicit answer context such as "Answer: B" or "the correct option is B". The same

letter or number elsewhere in an explanation is not assigned to this category solely because of its surface form.

- **Temporal markers** are words or multiword expressions that indicate order, duration, repetition, or change over time, including "before", "after", "then", "first", "later", "again", and "at the end".
- **Event verbs and nouns** describe visually observable actions, interactions, changes, or processes, such as "pours", "opens", "jumps", "collision", and "movement". Auxiliary and modal verbs are excluded from this category.
- **Objects, places, attributes** identify visible entities or their properties, including people, objects, locations, colors, shapes, counts, and visible states, such as "kitchen", "container", "red", and "three".
- **Function words, connectives, and auxiliaries** provide grammatical or discourse structure rather than visual content, including articles, determiners, conjunctions, prepositions, pronouns, auxiliary and modal verbs, negation, and discourse connectives such as "the", "is", "and", "because", and "therefore".
- **Punctuation and formatting** consist of punctuation marks, line breaks, whitespace-only pieces, and other formatting symbols without lexical content.
- **Others** is the fallback category for proper names, generic adverbs, unmatched subwords, or tokens whose category cannot be determined reliably by the predefined rules.

When a lexical span matches more than one rule, the tagger uses the following priority: explicit answer context, temporal expression, function word or connective, event expression, visual entity or attribute, and finally *Others*. BPE pieces belonging to the same decoded lexical span inherit that span's category. This priority makes the assignment deterministic and ensures that every selected token receives exactly one label.

We compare the highest- and lowest-weight token deciles from 100 training examples at three checkpoints. Table 30 reports the category proportions after pooling the selected token instances across checkpoints.

The top decile contains 74.3% tokens from the four evidence- or answer-oriented categories: temporal markers (17.3%), event verbs and nouns (20.7%), objects, places and attributes (21.0%), and answer-committing tokens (15.3%). The corresponding share in the bottom decile is 24.8%. Event verbs and nouns increase from 5.2% in the bottom decile to 20.7% in the top decile, while objects, places and attributes increase from 7.5% to 21.0%.

The complementary pattern appears for linguistic scaffolding. Function words, connectives and auxiliaries account for 4.8% of the top decile and 23.1% of the bottom decile, while punctuation and formatting account for 3.1% and 25.4%, respectively. The *Others* category accounts for 17.8% of the top decile and 26.7% of the bottom decile.

The highest-weight decile contains more temporal, event, visual-entity, and answer tokens than the lowest-weight decile. The redistributed weight is computed from the centered evidence contrast in Equation 4, so these proportions identify where ECR concentrates training weight.

F.4 Sensitivity to the SSA Weight

We vary the SSA weight λ_{SSA} in Equation 7 on Qwen3.5-9B in non-thinking mode with 64 uniformly sampled student frames. The fixed teacher, privileged teacher view, ECR objective, and remaining training settings are held constant. Table 31 evaluates $\lambda_{\text{SSA}} \in \{0, 0.01, 0.02, 0.05, 0.1\}$. A weight of 0 removes the representation-alignment term while retaining ECR, and 0.05 is the default setting. We report accuracy on each benchmark and the unweighted macro-average to examine both the overall sensitivity and differences across tasks.

The macro-average accuracies are 58.7%, 58.8%, 59.2%, 59.8%, and 59.4% as λ_{SSA} increases from 0 to 0.1. The default 0.05 gives the highest average and improves all five benchmarks over disabling SSA, with gains of 1.1, 1.7, 0.6, 0.5, and 1.7 points on Video-MME, MVBench, LongVideoBench, Video-MME 2, and MLVU.

The best individual scores occur at different weights: 0.02 for LongVideoBench, 0.05 for MVBench, 0.1 for Video-MME and MLVU, and a tie between 0.02 and 0.05 for Video-MME 2. At 0.1, the

Table 30: Composition (%) of tokens in the top and bottom weight deciles among positively advantaged responses. Each response token is assigned to one of seven mutually exclusive categories; tokens that do not match the predefined rules are assigned to *Others*.

Category	Top decile	Bottom decile	Examples
Temporal markers	17.3	6.2	after, then, at the end
Event verbs and nouns	20.7	5.2	pours, jumps, collision
Objects, places, attributes	21.0	7.5	kitchen, red container, three
Answer-committing tokens	15.3	5.9	option letters, final numbers
Function words, connectives, and auxiliaries	4.8	23.1	the, is, and, therefore
Punctuation and formatting	3.1	25.4	commas, line breaks
Others	17.8	26.7	proper names, generic adverbs, unmatched subwords

Table 31: Sensitivity to the SSA weight on Qwen3.5-9B in non-thinking mode with 64 uniformly sampled student frames. Scores are accuracies (%), and Avg. is the unweighted mean over the five benchmarks. All other training settings are held constant; $\lambda_{\mathrm{SSA}} = 0$ disables SSA while retaining ECR, and 0.05 is the default weight.

λ_{SSA}	**Video-MME**	**MVBench**	**LongVideoBench**	**Video-MME 2**	**MLVU**	**Avg.**
0 (w/o SSA)	69.6	71.1	64.6	28.4	59.9	58.7
0.01	70.5	71.1	64.8	27.9	59.5	58.8
0.02	70.3	71.5	**65.3**	**28.9**	59.9	59.2
0.05 (default)	70.7	**72.8**	65.2	**28.9**	61.6	**59.8**
0.1	**70.9**	71.3	64.8	28.3	**61.9**	59.4

macro-average falls to 59.4% as MVBench, LongVideoBench, and Video-MME 2 decline relative to the default. The default remains within 0.3 points of the best score on every benchmark.

F.5 Qualitative Comparison Across Post-Training Methods

We compare the responses of the Vanilla checkpoint and the checkpoints obtained with SFT, GRPO, OPSD, and OWLET on two video questions. For each example, all five models receive the same video frames, question, and answer options under the ordinary student inference protocol. The comparison therefore examines how different post-training methods affect the answers produced from a common input. The privileged teacher view is used only during training; the OWLET checkpoint receives no additional evidence at inference.

The two examples cover the global-temporal and local-evidence question types discussed in Section 3.1 and defined in Appendix D.3. We select one example of each type on which OWLET answers correctly while Vanilla, SFT, GRPO, and OPSD answer incorrectly. Figures 17 and 18 show the questions, video overviews, and model responses. These selected cases show differences in the use of temporal and visual evidence; aggregate performance on the two question types is reported separately in Appendix F.8.

Case 1: Global temporal ordering. Figure 17 asks which sequence of three scenes matches the video. The correct answer is C: people camping in the evening, followed by people queuing in sunny weather, and finally a crowd raising their hands and cheering near dusk. All four options contain the same three scenes and differ in their order. Answering therefore requires relating events across separated parts of the video; identifying the activities individually is insufficient to distinguish the alternatives. The lighting descriptions help identify the scenes, but the required order is their occurrence in the video.

The baseline responses exhibit two different ordering errors. Vanilla and GRPO choose D, placing the sunny queuing scene before the evening camping scene. SFT and OPSD choose B, placing the

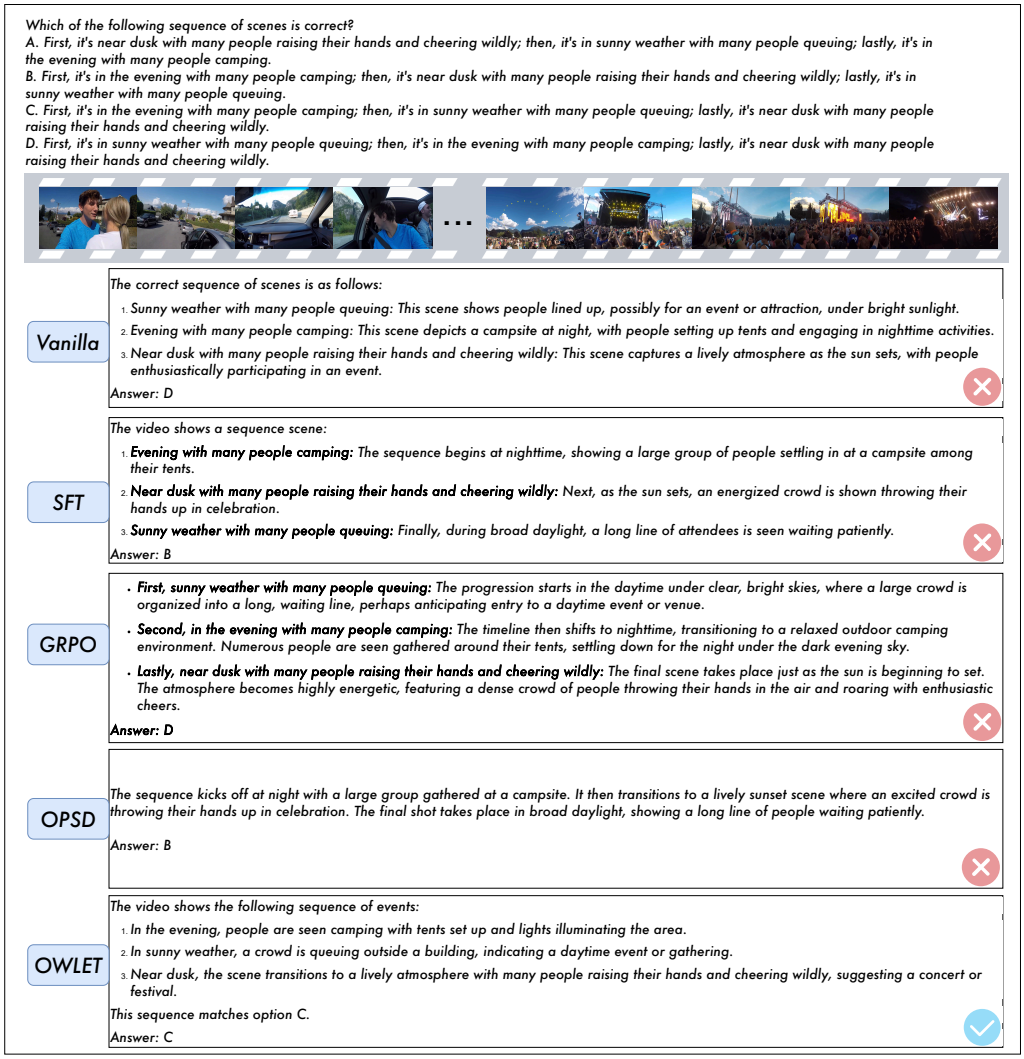


Figure 17: **Global-temporal case.** Responses from Vanilla and checkpoints trained with SFT, GRPO, OPSD, and OWLET to the same video input and question. Vanilla and GRPO select D, and SFT and OPSD select B; OWLET recovers the correct camping–queuing–cheering order and selects C.

cheering scene before the queuing scene. All four responses mention the three target activities, yet arrange them into an incorrect sequence. Their explanations also elaborate on the individual scenes, describing tents, waiting attendees, or an excited crowd, without resolving the temporal relation that determines the answer. In comparison, OWLET gives the correct camping–queuing–cheering sequence and selects C. The example shows a global-temporal distinction: a response can describe the relevant events fluently while failing to preserve their order. The difference among the displayed answers lies in how these events are connected across time.

Case 2: Local evidence grounding. Figure 18 asks which additional item is visible when prepared meat is spooned onto mashed potatoes in a glass dish, at the moment associated with the subtitle "Just pour the potatoes over minced meat!!". The options are a knife, a pot, green onion, and cheese, with B (pot) as the correct answer. The action and subtitle specify the local moment to inspect, while the requested object lies in the surrounding scene. Resolving this question requires associating an object with that particular moment. General knowledge of the recipe, or the presence of an ingredient elsewhere in the video, does not determine the answer.



Figure 18: **Local-evidence case.** Responses to the same video input and question about the object visible at a specified cooking step. Vanilla, SFT, and GRPO select C (green onion), and OPSD selects A (knife); OWLET identifies the pot and selects the correct answer B.

Vanilla, SFT, and GRPO all select C (green onion), but their explanations differ. Vanilla claims that all four candidate items are visible and then chooses green onion, without explaining why it is the appropriate choice. SFT describes green onions on the counter, and GRPO makes the same identification while explicitly excluding the pot. OPSD instead chooses A (knife), claiming that the knife is visible nearby and that the remaining items are outside the frame. Thus, several responses refer directly to the requested cooking step and provide specific spatial descriptions, but still identify the wrong object. OWLET identifies a cooking pot in the background and selects B. The case shows the need to ground an object description in the question-specified temporal context.

Relation to evidence-guided post-training. The two cases show complementary evidence requirements: ordering separated events and identifying a detail at a specific moment. OWLET answers both correctly under the same student inputs used by the baselines.

F.6 Training-Time Comparison

Table 32 compares the training cost of the three on-policy methods, GRPO, OPSD, and OWLET, on Qwen3.5-9B in non-thinking mode with 64 uniformly sampled student frames. All three methods use eight H200 GPUs and a budget of 1,000 updates. We report the average time per step for Rollout and Update, their sum, and the accumulated step time over this training budget.

Overall training cost. OPSD has the lowest measured cost at 848.00 seconds per step, followed by GRPO at 914.29 seconds and OWLET at 948.55 seconds. Over 1,000 updates, these correspond to 235.6, 254.0, and 263.5 hours of accumulated step time, respectively. Relative to GRPO, OWLET adds 3.7% (approximately 9.5 hours); relative to OPSD, it adds 11.9% (approximately 27.9 hours).

Table 32: On-policy training cost on eight H200 GPUs for Qwen3.5-9B at 64 student frames, averaged over the first 1,000 updates. Update covers the non-rollout portion of each timed step. Total (s/step) is the sum of Rollout and Update; Total (h) gives the accumulated step time.

Method	Rollout (s/step)	Update (s/step)	Total (s/step)	Total (h)
GRPO	406.99	507.30	914.29	254.0
OPSD	410.89	437.11	848.00	235.6
OWLET	434.31	514.24	948.55	263.5

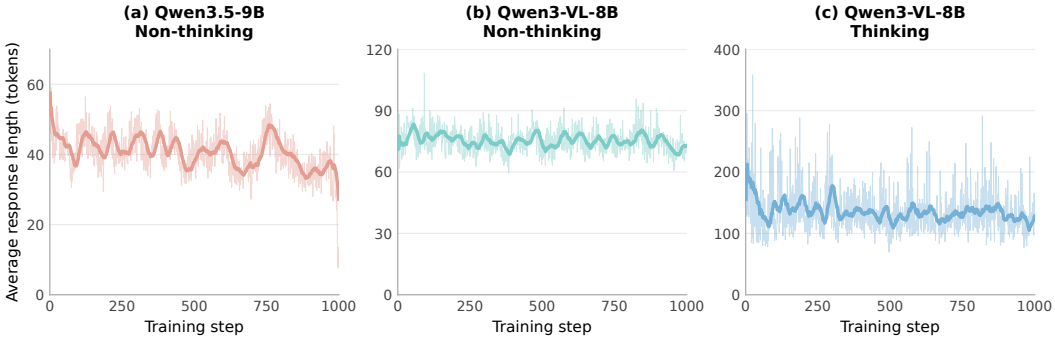


Figure 19: Response length during the first 1,000 training steps of OWLET with 64 student frames: (a) Qwen3.5-9B in non-thinking mode, (b) Qwen3-VL-8B-Instruct in non-thinking mode, and (c) Qwen3-VL-8B-Thinking. Thin, light curves show the mean number of non-padding generated tokens across all 128 responses at each step; thick curves show 25-step trailing moving averages, using the available history for the first 24 steps. Thinking lengths include the generated reasoning trace and final answer. The panels use separate vertical scales, and all per-step means are shown.

Rollout and update costs. Relative to GRPO, OWLET spends 27.32 more seconds per step on rollout and 6.94 more seconds on update. Relative to OPSD, the rollout difference is 23.42 seconds and the update difference is 77.13 seconds.

Training and inference implications. The comparison quantifies the cost of running each method for the same number of updates. For OWLET, the fixed teacher supplies token scores and hidden-state targets only during training. At inference, the student uses the standard uniformly sampled view without invoking the teacher or constructing the privileged teacher view.

F.7 Response Length during On-Policy Training

We compare response-length dynamics across three OWLET runs: Qwen3.5-9B in non-thinking mode, Qwen3-VL-8B-Instruct in non-thinking mode, and Qwen3-VL-8B-Thinking. Each run uses 64 uniformly sampled student frames, 16 prompts per step, and $G = 8$ responses per prompt. We analyze the 1,000 training steps of each run, covering 128,000 responses per run and 384,000 responses in total. All generated responses are included, irrespective of verifier correctness. Every analyzed step is matched to a rollout file containing all 128 decoded responses. For the thinking model, length includes the generated reasoning trace as well as the final answer. Figure 19 shows the trajectories, and Table 33 summarizes their scales and changes.

Qwen3.5-9B: late shortening and direct answers. The Qwen3.5-9B run shows an early reduction followed by intermediate fluctuations and a sharper decrease near the end. Steps 1–50 and 51–100 average 44.7 and 40.0 tokens, respectively. The final ten steps average 17.2 tokens, with step 1,000 averaging 11.0. The decoded rollouts connect this late shortening to a change in response format. We count a response as answer-only when its entire decoded text consists of `Answer: X`, where `X` is a single uppercase option letter, allowing whitespace and an optional terminal full stop. Such responses account for 22 of 6,400 rollouts (0.34%) over steps 901–950, but 821 of 1,280 rollouts (64.1%) over

Table 33: Mean response length in generated tokens for the three runs in Figure 19. Overall averages all 1,000 steps. Change is the percentage difference between the final and first 100-step means, computed before rounding.

Model	Mode	Steps 1–100	Steps 901–1,000	Overall	Change (%)
Qwen3.5-9B	Non-thinking	42.4	33.9	40.4	−20.0
Qwen3-VL-8B	Non-thinking	77.5	72.0	75.2	−7.1
Qwen3-VL-8B	Thinking	145.6	123.7	135.2	−15.0

the final ten steps. Thus, the late decrease coincides with a marked shift toward direct answer-only outputs.

Qwen3-VL-8B-Instruct: relatively stable response length. The non-thinking Qwen3-VL run maintains a more consistent response length, averaging 75.2 tokens overall. Its consecutive 100-step means range from 72.0 to 77.5 tokens, with the final block 7.1% shorter than the first. The final ten steps average 70.9 tokens, and no decoded response matches the answer-only pattern defined above.

Qwen3-VL-8B-Thinking: early compression with continued reasoning output. The thinking run produces longer responses, averaging 135.2 generated tokens including the reasoning trace and final answer. Its mean falls from 157.0 tokens over steps 1–50 to 134.1 over steps 51–100, followed by a rebound to 146.5 over steps 101–200. Thereafter, the trajectory continues to fluctuate: the 100-step averages over steps 401–1,000 lie between 123.7 and 136.5 tokens. The final 100 steps are 15.0% shorter than the first 100, but the final ten steps still average 133.8 tokens. This run shows early shortening followed by fluctuations at a lower typical length, and differs from the sharp late decrease observed for Qwen3.5-9B.

Joint interpretation. Across the three runs, mean response length decreases between the first and final 100 steps. The trajectories differ: Qwen3-VL-Instruct remains relatively stable, Qwen3-VL-Thinking shortens early, and Qwen3.5-9B shifts toward answer-only responses near the end of training.

F.8 Global-Temporal Accuracy after Post-Training

Section 3.1 shows that restricting the input to annotated evidence favors questions that can be answered from local moments, while broader temporal questions require additional context. We use the same diagnostic split to examine this trade-off after post-training. The Vanilla and Key Windows rows are input-view controls from the common initial Qwen3.5-9B checkpoint. GRPO, OPSD, and OWLET are evaluated at their final checkpoints with uniformly sampled student frames. We report overall accuracy together with the held-out global-temporal and local-evidence subsets defined in Appendix D.3.

The two input-view controls reproduce the expected specialization trade-off. Vanilla obtains 50.5% overall, with 51.8% on global-temporal questions and 49.2% on local-evidence questions. Key Windows increases the overall score to 57.8% and the local-evidence score to 65.4%, but its global-temporal score is only 50.2%. Concentrating all frames around question-relevant moments is therefore effective for the local subset, yet it does not preserve the broader temporal information needed by the global subset.

As shown in Table 34, all trained baselines improve over the Vanilla model on both subsets. GRPO reaches 55.8% overall, 53.4% on global-temporal questions, and 58.2% on local-evidence questions. OPSD reaches 54.5%, 52.4%, and 56.6%, respectively. OWLET reaches 59.2% overall, with 58.6% global-temporal and 59.8% local-evidence accuracy. Relative to GRPO, OWLET gains 3.4 points overall, 5.2 points on global-temporal questions, and 1.6 points on local-evidence questions. Relative to OPSD, the corresponding gains are 4.7, 6.2, and 3.2 points. The larger margins on the global-temporal subset indicate that the improvement is not limited to resolving the localized event; OWLET also retains the broader temporal information that the student needs under uniform sampling.

OWLET also provides a more balanced result than the Key Windows control. Its overall accuracy is 1.4 points higher, and its global-temporal accuracy is 8.4 points higher, while Key Windows remains 5.6 points higher on local-evidence questions. Thus, the strongest local-evidence score in this table does not translate into the strongest overall score: the two subsets are equally represented in the diagnostic split, so the global-temporal improvement substantially affects the final average.

Table 34: Accuracy (%) on the held-out diagnostic split before and after post-training, overall and by question type, for Qwen3.5-9B at 64 input frames. The Vanilla and Key Windows rows use the same initial checkpoint with uniformly sampled and question-conditioned Key Windows views, respectively; GRPO, OPSD, and OWLET are evaluated at their final checkpoints under the uniformly sampled student view.

Method	Overall	Global-temporal	Local-evidence
Vanilla	50.5	51.8	49.2
Key Windows	57.8	50.2	**65.4**
GRPO	55.8	53.4	58.2
OPSD	54.5	52.4	56.6
OWLET	**59.2**	**58.6**	59.8

F.9 Extended Training Schedules

The main experiments train for 1,000 steps. We continue the Qwen3.5-9B run at 64 input frames with the same training data, optimizer and learning rate, rollout budget of 16 prompts with $G = 8$ responses per step, clipping range, rule-based verifier, global-plus-local teacher view, and ECR and SSA settings. The teacher stays fixed at the initial checkpoint. We save checkpoints at 1,000, 1,200, 1,500, 1,800, 2,000, 2,200, 2,500, 2,800, and 3,000 steps, and evaluate each with the main-text protocol on Video-MME, MVBench, LongVideoBench, Video-MME 2, and MLVU. The schedule ends at 3,000 steps because of the available compute budget.

Table 35 reports the results. Macro-average accuracy increases from 59.8 at 1,000 steps to 61.3 at 3,000 steps, with the highest observed score of 61.5 at 2,800 steps. Between the initial and final checkpoints, LongVideoBench gains 2.3 points, Video-MME 2.2, MLVU 1.7, Video-MME 2 0.7, and MVBench 0.6; the five benchmarks reach their highest observed scores at different checkpoints. The mean absolute change between adjacent checkpoints is 0.30 points for the macro-average and ranges from 0.41 to 0.67 points across benchmark columns.

The additional 2,000 updates yield a 1.5-point macro-average gain. Using the per-step cost measured in Appendix F.6, they correspond to approximately 527 hours on eight H200 GPUs. The main comparison uses a common 1,000-update budget for GRPO, OPSD, and OWLET, and the main-text results continue to use the fixed 1,000-step checkpoint for every backbone and frame budget.

Table 35: Per-benchmark accuracy (%) of OWLET checkpoints under an extended training schedule, for Qwen3.5-9B at 64 input frames. Checkpoints are saved at nine steps between the main-text budget of 1,000 and 3,000, spaced 200–300 steps apart, and each is evaluated with the protocol of Table 1 on the same five benchmarks; the 1,000-step row reproduces the corresponding OWLET entry there. **Avg.** is the macro-average over the five benchmarks, computed before rounding. The last row gives the change from 1,000 to 3,000 steps.

Training steps	**Video-MME**	**MVBench**	**LongVideoBench**	**Video-MME 2**	**MLVU**	**Avg.**
1,000 (main text)	70.7	72.8	65.2	28.9	61.6	59.8
1,200	70.5	73.1	65.5	28.4	60.8	59.7
1,500	71.0	73.6	65.5	28.6	60.5	59.8
1,800	72.3	73.0	64.8	29.5	61.0	60.1
2,000	72.0	73.6	66.1	29.2	62.7	60.7
2,200	70.9	73.3	66.9	29.1	62.9	60.6
2,500	71.5	73.0	67.2	29.5	62.5	60.7
2,800	72.8	73.9	67.7	30.0	63.1	61.5
3,000	72.9	73.4	67.5	29.6	63.3	61.3
Δ (3,000 vs. 1,000)	+2.2	+0.6	+2.3	+0.7	+1.7	+1.5