LESS DATA, BETTER TIMING: STUDENT–CURRICULUM COUPLING FOR VLM ON-POLICY DISTILLATION IN TEMPORAL VIDEO GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides dense supervision directly on student-generated trajectories, making it an effective post-training strategy for vision-language models in temporal video grounding (TVG). However, existing pipelines typically construct the training curriculum from a fixed teacher and the initial student state, implicitly assuming that selected examples retain positive supervision value throughout optimization. We show that supervision trustworthiness and supervision necessity are distinct yet coupled: the former concerns target credibility, while the latter varies with the student's current task competence; together, they shape supervision value. Building on this coupled view, we introduce **Student–Curriculum Coupling (SCC)**, a closed-loop framework in which a compact Anchor–Frontier curriculum defines the candidate supervision space and the evolving student dynamically determines its active subset. Supervision can therefore be activated, suspended, or reactivated as competence changes, concentrating teacher computation and optimization on current task-level deficits. Across three TVG benchmarks, SCC achieves a 5.1% relative improvement in mean recall over Video-OPD on its original curriculum, while using 60.0% fewer training examples and reducing training time by 50.4%. Ablations support the complementary roles of capability-structured curriculum design and student-dependent supervision in achieving these gains. Together, these results establish SCC as a data- and compute-efficient framework for TVG post-training, delivering stronger temporal grounding by aligning trustworthy supervision with the student's evolving learning needs.

1 INTRODUCTION

Temporal video grounding (TVG) localizes the segment in an untrimmed video that corresponds to a natural-language query (Gao et al., 2017). As vision-language models (VLMs) have become increasingly capable of jointly reasoning over video and language, recent work has cast TVG as generative temporal prediction, with target intervals produced directly as timestamps (Huang et al., 2024a; Wang et al., 2025). However, precise localization, particularly at event boundaries, remains challenging, motivating task-specific post-training (Wang et al., 2026). Among these approaches, on-policy distillation (OPD) trains the student on its own generated trajectories using dense token-level teacher feedback (Lu & Thinking Machines Lab, 2025; Agarwal et al., 2024; Li et al., 2026c). Video-OPD applies this paradigm to TVG (Li et al., 2026a). Beyond the distillation objective itself, the effectiveness of OPD for TVG also depends on how teacher supervision is allocated across training examples, highlighting the importance of curriculum design and adaptive data selection (Bengio et al., 2009; Chen et al., 2023).

In OPD for TVG, curriculum construction can use teacher reliability and the performance gap between a fixed teacher and the initial student to select informative examples (Li et al., 2026a). Selected examples typically receive supervision whenever presented, carrying an implicit *persistent-value assumption* that their supervision remains useful throughout optimization. As student competence evolves, however, initial suitability alone cannot determine whether supervision remains necessary. We therefore distinguish *supervision trustworthiness*, which concerns whether the teacher provides a credible target for an example, from *supervision necessity*, which concerns whether the current student still exhibits a task-level deficit on that example. Together, these factors motivate a

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.

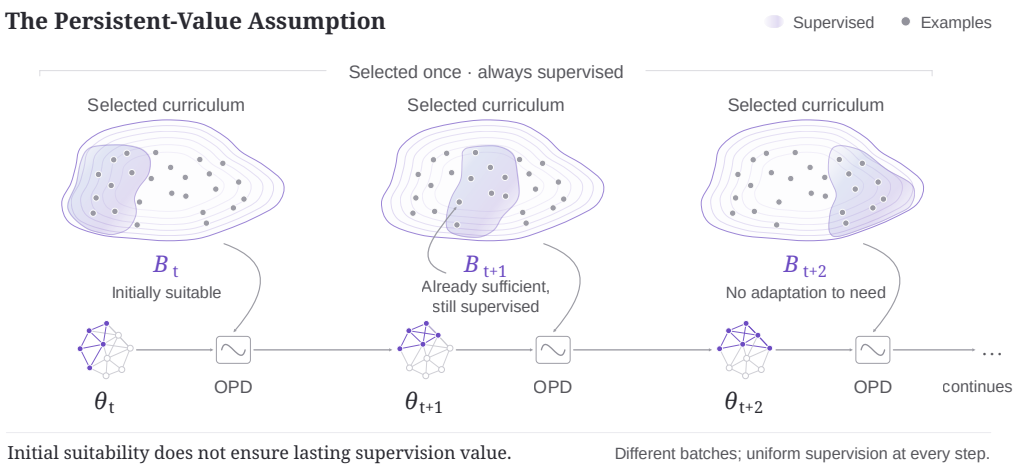

Figure 1: **The persistent-value assumption.** All examples in each scheduled batch (shaded) receive OPD supervision as the student evolves, including those that already satisfy the task criterion.

coupled view in which curriculum composition and supervision realization are jointly designed to align trustworthy supervision with the student's changing learning needs.

Building on this coupled view, we introduce **Student–Curriculum Coupling (SCC)**, a closed-loop OPD framework that jointly designs curriculum composition and supervision realization. SCC constructs a compact Anchor–Frontier (AF) curriculum around the initial student's capability profile, preserving trustworthy supervision opportunities for capability acquisition and stabilization. These opportunities contribute to optimization only when the current student exhibits a task-level deficit, so that candidate-space composition and current student competence jointly determine active supervision. OPD updates in turn reshape the student's supervision needs, which guide subsequent curriculum realization and close the feedback loop.

We evaluate SCC on Charades-TimeLens (Zhang et al., 2025), ActivityNet-TimeLens (Zhang et al., 2025), and QVHighlights-TimeLens (Zhang et al., 2025). Compared with Video-OPD on its Teacher-Validated Disagreement Focusing (TVDF) curriculum (Li et al., 2026a), SCC achieves a 5.1% relative improvement in mean recall across the three benchmarks and intersection-over-union (IoU) thresholds of 0.3, 0.5, and 0.7. These gains are obtained with 60.0% fewer training examples and 50.4% shorter training time.

Our contributions are threefold:

- We identify the implicit persistent-value assumption in TVG OPD and develop a coupled view that links supervision trustworthiness with the student's evolving supervision necessity.
- We introduce SCC, a closed-loop OPD framework that couples a compact AF candidate space with student-dependent supervision realization.
- We demonstrate consistent gains in TVG accuracy and training efficiency over Video-OPD on TVDF. Ablations isolate the roles of candidate-space composition and student-dependent realization, while training-dynamics analysis tracks supervision as student competence evolves.

2 Problem Formulation and Motivation

Our analysis focuses on OPD in VLM post-training for TVG. Section 2.1 identifies the persistent-value assumption illustrated in Figure 1; Section 2.2 formalizes the non-stationary marginal value of supervision; and Section 2.3 distinguishes supervision trustworthiness from necessity.

2.1 The Assumption of Persistent Supervision Value

For a TVG example $x = (v, q)$, where v is an untrimmed video and q is a natural-language query, let $a^\star(x)$ denote the ground-truth temporal interval and $a_M(x)$ the interval predicted by model M under a fixed decoding protocol. A common curriculum-construction strategy evaluates each candidate using a fixed teacher and the initial student, and selects examples according to their TVG

performance:

$$q_M(x) = \mathrm{tIoU}(a_M(x), a^\star(x)), \qquad \mathcal{D}_{\mathrm{cand}} = \{x \in \mathcal{P} : c(q_T(x), q_{S,0}(x)) = 1\}. \tag{1}$$

Here, $\mathrm{tIoU}(\cdot,\cdot)$ denotes temporal intersection over union, which measures the overlap between two temporal intervals. Accordingly, $q_M(x) \in [0,1]$ is the TVG score of model M on example x. $q_T(x)$ and $q_{S,0}(x)$ denote the TVG scores of the fixed teacher and the student before optimization, respectively. $\mathcal{P}$ is the candidate data pool, $c : [0,1]^2 \to \{0,1\}$ is the selection rule, and $\mathcal{D}_{\mathrm{cand}} \subseteq \mathcal{P}$ is the resulting candidate curriculum.

This construction selects examples with a credible teacher signal and a targeted level of difficulty for the initial student. However, $q_{S,0}(x)$ reflects the initial student's performance on x, not a fixed property of the example. Once x is included in $\mathcal{D}_{\mathrm{cand}}$, its curriculum membership remains unchanged even as optimization alters the student's competence and the potential learning value of x. Continuing to treat x as supervision-bearing whenever it is scheduled therefore introduces an implicit assumption of persistent supervision value.

2.2 The Non-Stationary Marginal Value of Supervision

To formalize how supervision value changes with the student, let θ_t denote the student parameters immediately before training step t, and let $J_{\mathrm{TVG}}(\theta)$ denote the expected TVG generalization performance of the student parameterized by θ. For an example x encountered at step t, let θ_{t+1}^{+x} denote the student parameters after an update that includes the OPD loss associated with x. Similarly, let θ_{t+1}^{-x} denote the parameters after the corresponding update without the OPD loss associated with x. All other examples and optimization conditions are held fixed. We define the marginal supervision value of x as:

$$\Delta_t(x) = \mathbb{E}\left[J_{\mathrm{TVG}}\left(\theta_{t+1}^{+x}\right) - J_{\mathrm{TVG}}\left(\theta_{t+1}^{-x}\right) \mid \theta_t, x\right]. \tag{2}$$

The expectation is taken over training stochasticity, with the same realization used for the updates with and without OPD supervision on x. This quantity characterizes the effect of supervising x and is not assumed to be directly observable during training.

Even when the video, query, annotation, and teacher remain fixed, $\Delta_t(x)$ can vary because the student-generated trajectory, the resulting OPD signal, and the model's response to that signal all depend on θ_t. Supervision may support capability acquisition before the student can localize the queried event, lose value once the student produces a satisfactory interval, and become useful again if later updates degrade that capability. The marginal supervision value is therefore non-stationary and reflects both the student's training history and its current interaction with the example.

2.3 Supervision Trustworthiness and Necessity

We characterize the state dependence of supervision value through *supervision trustworthiness* $\mathcal{T}(x)$ and *supervision necessity* $\mathcal{N}_t(x)$. The former indicates whether the teacher provides a credible target for example x, while the latter indicates whether the current student still exhibits a task-level deficit:

$$\mathcal{T}(x) = \mathbf{1}\{q_T(x) \geq \tau_T\}, \qquad \mathcal{N}_t(x) = \mathbf{1}\{q_{S,t}(x) < \tau_S\}. \tag{3}$$

$\mathbf{1}\{\cdot\}$ denotes the indicator function. $q_{S,t}(x)$ is the student's TVG score on x at step t, and τ_T and τ_S are thresholds for supervision trustworthiness and student competence, respectively. $\mathcal{T}(x)$ captures the fixed teacher–example relationship, whereas $\mathcal{N}_t(x)$ captures the evolving student–example relationship. The necessity indicator is a task-level proxy and does not directly observe $\Delta_t(x)$.

The two factors are distinct but coupled in supervision allocation. When both hold, a trustworthy target is available for an unresolved student outcome. Trustworthiness without necessity may render further supervision redundant (Figure 1), while necessity without trustworthiness provides no credible distillation target. Curriculum composition therefore determines the availability of trustworthy supervision, while current student competence determines its necessity during optimization.

3 Student–Curriculum Coupling

Building on the coupled view developed in Section 2, we introduce Student–Curriculum Coupling, a closed-loop OPD framework for TVG. Section 3.1 reviews the OPD objective, and Section 3.2 introduces the Anchor–Frontier candidate supervision space. Section 3.3 defines student-dependent realization, while Section 3.4 presents the coupled objective and efficient execution. Figure 2 illustrates the framework. Additional theoretical details are provided in Appendix A.

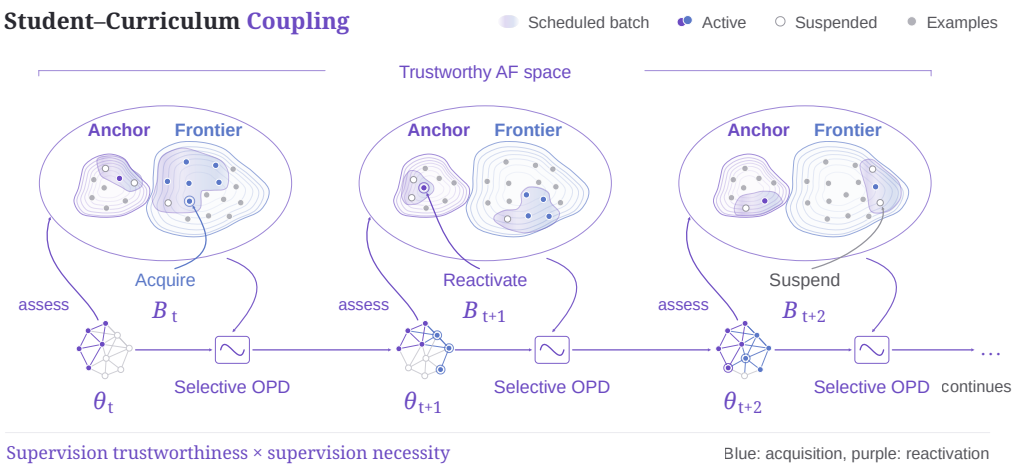

Figure 2: **Student–Curriculum Coupling.** Shaded regions denote successive scheduled batches within the AF space. Current-student assessment determines active and suspended supervision, supporting capability acquisition and Anchor reactivation. Selective OPD updates the student, whose new state guides supervision in subsequent batches, closing the loop.

3.1 On-Policy Distillation

On-policy distillation optimizes a student VLM π_θ using token-level supervision from a teacher VLM π_T whose parameters remain fixed. The student generates trajectories from its current policy, and the teacher supplies target next-token distributions at the corresponding prefixes (Agarwal et al., 2024; Li et al., 2026a). Given a TVG input $x = (v, q)$, the rollout policy π_{θ_t} at training step t generates an autoregressive trajectory of length L_t:

$$y_t = (y_{t,1}, \ldots, y_{t,L_t}) \sim \pi_{\theta_t}(\cdot \mid x).$$

At position k, let $h_{t,k} = (x, y_{t,<k})$ denote the current prefix, and define the student and teacher next-token distributions as $p_{\theta,k}(\cdot) = \pi_\theta(\cdot \mid h_{t,k})$ and $p_{T,k}(\cdot) = \pi_T(\cdot \mid h_{t,k})$, respectively. The frozen teacher scores the student-generated trajectory without a separate teacher rollout.

Using reverse Kullback–Leibler (KL) divergence for token-level alignment (Gu et al., 2024; Li et al., 2026c), we define the per-example objective under the fixed rollout distribution as:

$$\mathcal{L}_{\mathrm{OPD},t}(\theta; x) = \mathbb{E}_{y_t \sim \pi_{\theta_t}(\cdot|x)} \left[\sum_{k=1}^{L_t} D_{\mathrm{KL}}(p_{\theta,k} \,\|\, p_{T,k}) \right]. \tag{4}$$

Here, θ_t denotes the fixed rollout-policy parameters, while θ is the optimization variable initialized from θ_t. The rollout distribution remains fixed during differentiation, and gradients propagate through the student next-token probabilities.

Following Video-OPD (Li et al., 2026a), we optimize an importance-weighted sampled-token surrogate $\ell_{\mathrm{OPD}}(x, y_t; \theta, \theta_t, \pi_T)$, with teacher-derived signals held fixed during the update. At $\theta = \theta_t$, its gradient matches that of Equation 4 in expectation over student rollouts. Appendix A.2 provides the surrogate and derives this local gradient identity.

3.2 Anchor–Frontier Candidate Space

To support capability acquisition and stabilization within a compact supervision space, we introduce the Anchor–Frontier (AF) design. It combines examples already supported by the initial student with examples that offer substantial learning headroom. Both groups contain only examples with trustworthy supervision, satisfying $\mathcal{T}(x) = 1$ as defined in Equation 3.

Anchors retain examples on which the initial student already performs well. They preserve access to corrective supervision if the student's performance on these examples later declines. From the candidate pool $\mathcal{P}$, we select N_A such examples:

$$\mathcal{A} = \mathrm{Select}_{N_A}\left(\{x \in \mathcal{P} : \mathcal{T}(x) = 1,\ q_{S,0}(x) \geq \tau_A\}\right). \tag{5a}$$

Frontiers complement this coverage with examples on which the initial student has low competence, providing opportunities for capability acquisition. We select N_F such examples:

$$\mathcal{F} = \mathrm{Select}_{N_F}\left(\{x \in \mathcal{P} : \mathcal{T}(x) = 1,\ q_{S,0}(x) < \tau_F\}\right). \tag{5b}$$

Here, N_A and N_F are the selection budgets, while τ_A and τ_F define the high- and low-competence regions. The operator $\mathrm{Select}_N(\cdot)$ returns a subset of N examples. The thresholds satisfy $\tau_F < \tau_A$, leaving the intermediate competence range outside the candidate space. The candidate space, $\mathcal{D}_{\mathrm{AF}} = \mathcal{A} \cup \mathcal{F}$, remains fixed during training. Section 3.3 describes how its supervision is realized as the student evolves.

3.3 Student-Dependent Realization

We organize the AF candidate space into a fixed training schedule $\mathcal{S} = (\mathcal{B}_0, \ldots, \mathcal{B}_{K-1})$, where $\mathcal{B}_t \subseteq \mathcal{D}_{\mathrm{AF}}$ contains the examples assigned to step t. Here, K counts scheduled training steps, including those that produce no optimizer update. Scheduled batch sizes may vary across steps. The schedule determines when each example is assessed, while the current student determines whether it receives OPD supervision.

At step t, the student generates one trajectory $y_{t,i} \sim \pi_{\theta_t}(\cdot \mid x_i)$ for each $x_i \in \mathcal{B}_t$. We assess the current prediction using:

$$q_{S,t}(x_i; y_{t,i}) = \mathrm{tIoU}(a(y_{t,i}), a^\star(x_i)), \tag{6}$$

where $a(y_{t,i})$ is the temporal interval decoded from the trajectory. Applying the supervision-necessity criterion from Section 2.3 yields the effective batch:

$$\mathcal{B}_t^{\mathrm{eff}} = \{(x_i, y_{t,i}) : x_i \in \mathcal{B}_t,\ q_{S,t}(x_i; y_{t,i}) < \tau_S\}. \tag{7}$$

Each retained trajectory is reused for teacher scoring and OPD optimization.

This rule supports capability acquisition, supervision suspension, and Anchor reactivation. A Frontier whose current score is below the criterion receives supervision for capability acquisition. For any candidate that meets the criterion, OPD supervision is suspended. For an Anchor with $q_{S,0}(x_i) \geq \tau_S$ at curriculum construction, a current score $q_{S,t}(x_i; y_{t,i}) < \tau_S$ reactivates its supervision to support capability stabilization. Reactivation is thus defined relative to the initial competence assessment.

3.4 Closed-Loop OPD Optimization

For each nonempty scheduled batch $\mathcal{B}_t$, we aggregate the per-example OPD surrogate from Section 3.1 over the effective batch:

$$\mathcal{L}_t^{\mathrm{SCC}}(\theta) = \frac{1}{n_t} \sum_{(x_i, y_{t,i}) \in \mathcal{B}_t^{\mathrm{eff}}} \ell_{\mathrm{OPD}}(x_i, y_{t,i}; \theta, \theta_t, \pi_T). \tag{8}$$

Here, θ is initialized from θ_t for the current update. The normalization denominator n_t is preset for each step, independently of the number of assigned or retained examples. Appendix A.3 derives the equivalent masked formulation and explains the effect of this normalization. The rollout policy and effective-batch membership remain fixed during optimization, with no gradients through sampling, temporal decoding, or competence assessment.

Under this objective, every scheduled example requires a student rollout for assessment, but teacher scoring, OPD loss evaluation, and backpropagation are restricted to $\mathcal{B}_t^{\mathrm{eff}}$. The compact AF space limits its rollout workload, while student-dependent realization reduces the subsequent supervision cost. If $\mathcal{B}_t^{\mathrm{eff}} = \varnothing$, the optimizer and weight-decay updates are skipped, leaving the student parameters and optimizer state unchanged as the schedule advances.

Let $\mathcal{G}$ denote the sampling, assessment, and selection procedure in Section 3.3, so that $\mathcal{B}_t^{\mathrm{eff}} = \mathcal{G}(\mathcal{B}_t; \theta_t)$. Let $\mathcal{U}_{\mathrm{SCC}}$ denote the update induced by Equation 8, with optimizer state and learning-rate settings left implicit. The update and subsequent realization satisfy:

$$\begin{aligned} \theta_{t+1} &= \mathcal{U}_{\mathrm{SCC}}\left(\theta_t, \mathcal{B}_t, \mathcal{B}_t^{\mathrm{eff}}; \pi_T\right), \\ \mathcal{B}_{t+1}^{\mathrm{eff}} &= \mathcal{G}(\mathcal{B}_{t+1}; \theta_{t+1}). \end{aligned} \tag{9}$$

The second line applies whenever a subsequent scheduled batch exists.

Equation 9 closes the loop between student optimization and supervision realization, while the candidate space and schedule remain fixed.

Table 1: **Evaluation on three TVG benchmarks.** Published benchmark results are taken from Video-OPD (Li et al., 2026a). Bold values indicate the best performance among the non-proprietary methods shown in the table.

	Charades-TimeLens			ActivityNet-TimeLens			QVHighlights-TimeLens		
Models for Evaluation	**R@0.3**	**R@0.5**	**R@0.7**	**R@0.3**	**R@0.5**	**R@0.7**	**R@0.3**	**R@0.5**	**R@0.7**
Proprietary Models									
GPT-4o (Hurst et al., 2024)	60.6	44.5	23.5	55.2	41.4	25.8	69.0	54.8	38.5
GPT-5 (Singh et al., 2025)	59.3	42.0	22.0	57.4	44.9	30.4	72.4	60.4	46.4
Gemini-2.0-Flash (Comanici et al., 2025)	66.4	53.5	27.1	62.9	54.0	37.7	76.2	66.4	48.3
Gemini-2.5-Flash (Comanici et al., 2025)	68.7	56.1	30.6	66.8	57.5	41.3	78.2	69.4	55.0
Gemini-2.5-Pro (Comanici et al., 2025)	74.1	61.1	34.0	72.3	64.2	47.1	84.1	75.9	61.1
Open-Source Models									
VideoChat-Flash-7B (Li et al., 2026b)	60.2	37.9	17.8	35.5	21.8	10.5	45.2	30.6	16.7
Qwen2.5-VL-7B (Bai et al., 2025b)	58.1	35.1	18.2	47.2	32.5	20.2	55.0	41.7	29.3
VideoChat-R1-7B (Li et al., 2025b)	51.9	30.8	11.7	35.0	23.9	11.3	29.3	19.1	9.4
Time-R1-7B (Wang et al., 2026)	57.9	32.0	16.9	44.8	31.0	19.0	65.8	51.5	36.1
TVG-R1-7B (Chen et al., 2025)	44.5	23.6	12.4	46.7	31.0	18.6	55.8	41.2	28.0
VideoChat-R1.5-7B (Yan et al., 2026)	46.4	24.0	10.4	40.6	25.3	16.4	62.2	44.5	28.3
MiMo-VL-7B (Xiaomi LLM-Core Team, 2025)	57.9	42.6	20.5	49.3	38.7	22.4	57.1	42.6	28.4
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	61.7	41.5	23.1	41.2	30.7	20.0	46.6	38.2	29.5
Post-Training Frameworks									
OP-RKD [Qwen3-VL-8B] (Li et al., 2026a)	67.3	47.0	27.8	59.1	44.9	30.6	70.7	57.4	44.5
OP-FKD [Qwen3-VL-8B] (Li et al., 2026a)	66.9	47.3	27.6	58.6	44.2	30.5	71.2	58.9	46.5
GRPO [Qwen3-VL-8B] (Li et al., 2026a)	72.7	44.4	27.6	58.6	42.7	32.1	69.8	53.0	41.5
Video-OPD [Qwen3-VL-8B] (Li et al., 2026a)	73.1	45.8	32.4	60.5	45.6	35.8	73.8	60.3	50.4
SCC [Qwen3-VL-8B] (ours)	**74.0**	**48.1**	**33.0**	**64.8**	**49.2**	**38.2**	**76.8**	**63.9**	**54.3**

4 EXPERIMENTS

Section 4.1 describes the training setup, AF curriculum construction, and evaluation protocol. Section 4.2 reports the main accuracy and efficiency results, while Section 4.3 decomposes SCC by separately evaluating candidate-space composition and student-dependent realization. Additional results and implementation details are provided in Appendix B.

4.1 EXPERIMENTAL SETUP

Training Details. We use Qwen3-VL-8B-Instruct (Bai et al., 2025a) as the student. For direct comparison with Video-OPD (Li et al., 2026a), we use its released Qwen3-VL-32B teacher, trained with group relative policy optimization (GRPO). Following the same configuration, we set the maximum input video token length to 8,192, sample videos at 2 FPS, cap the number of frames at 768, and set the maximum video-frame token length to 768. We use a learning rate of 1×10^{-6} and a 79-step training schedule with variable scheduled global batch sizes capped at 32 (Appendix B.2). Each scheduled example uses a single on-policy rollout, and we follow Video-OPD for rollout sampling, video preprocessing, and the remaining optimization settings. We set $\tau_S = 0.7$ for student-dependent realization throughout the main experiments; sensitivity to this criterion is examined in Appendix B.7.

AF curriculum construction. Following the AF design in Section 3.2, we construct the candidate supervision space from TimeLens-100K (Zhang et al., 2025), using public data from HiREST (Zala et al., 2023), QuerYD (Oncescu et al., 2021), CosMo-Cap (Wang et al., 2024a), InternVid-VTime (Wang et al., 2024b; Huang et al., 2024a), and DiDeMo (Hendricks et al., 2017). We first retain teacher-qualified examples with $q_T(x) \geq 0.7$, and then select 100 Anchor examples satisfying $q_{S,0}(x) \geq 0.7$ and 900 Frontier examples satisfying $q_{S,0}(x) < 0.3$. The resulting 1,000-example curriculum is assigned to a 79-step training schedule. We exclude duplicate annotations and videos overlapping with the TVG evaluation sets. Source composition and training-schedule details are provided in Appendix B.2, with sensitivity to the Anchor–Frontier ratio examined in Appendix B.10.

Evaluation Benchmarks and Metrics. Our evaluation focuses on TVG performance and training efficiency. We evaluate on Charades-STA (Gao et al., 2017), ActivityNet (Caba Heilbron et al., 2015), and QVHighlights (Lei et al., 2021), using the corrected temporal annotations from TimeLens (Zhang et al., 2025). Following prior work (Li et al., 2025a; Wang et al., 2025; 2024c), we report mean IoU (mIoU) and Recall at IoU thresholds of 0.3, 0.5, and 0.7. We assess training efficiency by curriculum size, the number of teacher-scored OPD routes, and training time.

4.2 Main Results

Temporal Video Grounding. Table 1 compares SCC with representative proprietary models, open-source VLMs, and post-training frameworks across three TVG benchmarks. Our method achieves the best performance among the non-proprietary methods in Table 1 across all reported recall metrics and consistently outperforms Video-OPD. Comparison with TimeLens-8B reveals benchmark-dependent trade-offs under different training regimes (Appendix B.4).

Accuracy and Training Efficiency. Table 2 reports curriculum scale and supervision routing, while Figure 3 summarizes accuracy and training time. For comparison, we additionally apply Video-OPD to the same AF curriculum used by SCC. With the same 1,000 examples and 1,004 scheduled rollouts, SCC requires only 492 teacher-scored OPD routes, a 51.0% reduction from Video-OPD. With AF held fixed, replacing uniform OPD with student-dependent routing yields an approximately 2.1% relative improvement in mean recall and reduces training time by 15.3%. Compared with Video-OPD on TVDF, SCC uses 60.0% fewer curriculum examples and achieves higher endpoint mIoU in approximately half the training time.

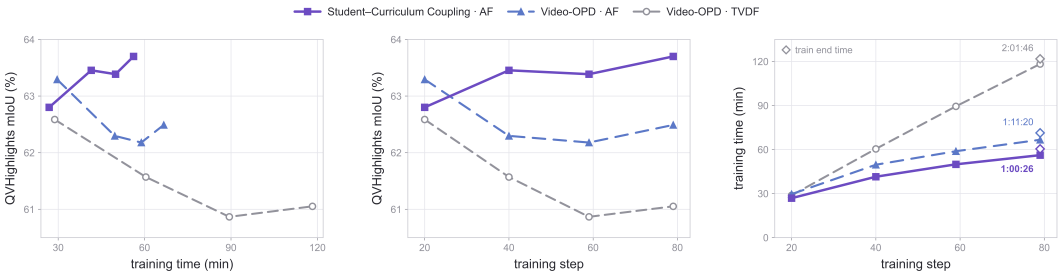

Figure 3: **Accuracy–efficiency comparison.** QVHighlights mIoU versus training time (left), QVHighlights mIoU across training steps (middle), and cumulative training time (right). Diamonds mark training completion. All times are measured on eight NVIDIA A100 GPUs.

Table 2: **Training data and supervision routing.** All runs use 79 scheduled training steps. Presentations / Rollouts counts student rollouts; Teacher / OPD Routes counts rollouts receiving teacher scoring and OPD optimization; Suspended Routes counts rollouts not routed to teacher scoring or OPD after current-student assessment.

Method	Curriculum	Data and Schedule		Supervision Routing		
		# Examples	Presentations / Rollouts	Teacher / OPD Routes	Suspended Routes	Route Rate
Video-OPD	TVDF	2,500	2,504	2,504	0	100.0%
Video-OPD	AF	1,000	1,004	1,004	0	100.0%
SCC	**AF**	**1,000**	**1,004**	**492**	**512**	**49.0%**

4.3 Decomposing Student–Curriculum Coupling

SCC combines a capability-structured candidate space with student-dependent realization of its supervision. Table 3 decomposes their contributions by varying one component while holding the other fixed. All trained variants share the same student initialization and frozen teacher and are trained for 79 scheduled steps.

Candidate-Space Composition. Table 3 (a) compares candidate curricula under the same student-dependent realization. AF performs best across all metrics and benchmarks. Its advantage over the matched-size Random curriculum supports capability-structured selection, while the comparison with Frontier-only highlights the complementary role of Anchors. AF also outperforms the larger TVDF curriculum, indicating that student-dependent realization benefits from a well-structured candidate space. Sensitivity to the Anchor–Frontier ratio and candidate-space scale is examined in Appendices B.10 and B.11, respectively.

Student-Dependent Realization. Table 3 (b) fixes the AF candidate space and removes student dependence from curriculum realization, applying OPD supervision uniformly across scheduled examples. This change reduces every reported metric across all three benchmarks. The consistent

Table 3: **Decomposition of Student–Curriculum Coupling.** Base Model denotes Qwen3-VL-8B-Instruct before post-training, with scores taken from Video-OPD (Li et al., 2026a). All other results are obtained using the same local evaluation pipeline. The upper block varies the candidate space with student-dependent realization fixed; the lower block fixes AF and compares student-dependent realization with uniform OPD supervision. TVDF uses the original 2,500-example curriculum from Video-OPD (Li et al., 2026a), while Random, Frontier-only, and AF each use 1,000 examples. ✓ denotes student-dependent realization, and × denotes uniform OPD supervision over all scheduled examples. ⋆ marks benchmarks re-annotated with TimeLens. Values are percentages rounded to one decimal place; bold marks the best result for each metric within each block.

Candidate Space	Student-Dep. Realization	Charades⋆ mIoU	Charades⋆ R@0.3	Charades⋆ R@0.5	Charades⋆ R@0.7	ActivityNet⋆ mIoU	ActivityNet⋆ R@0.3	ActivityNet⋆ R@0.5	ActivityNet⋆ R@0.7	QVHighlights⋆ mIoU	QVHighlights⋆ R@0.3	QVHighlights⋆ R@0.5	QVHighlights⋆ R@0.7
Base Model	–	42.9	61.7	41.5	23.1	30.4	41.2	30.7	20.0	36.9	46.6	38.2	29.5
(a) Candidate-Space Composition													
TVDF	✓	52.1	72.7	46.2	32.9	49.3	62.1	47.4	37.4	61.2	73.6	60.2	50.9
Random	✓	51.6	73.0	45.9	31.7	48.8	62.3	46.4	36.4	60.9	73.8	59.8	50.0
Frontier-only	✓	52.0	73.1	46.3	32.5	49.5	63.1	47.4	37.3	62.3	75.3	62.0	51.9
AF	✓	**52.8**	**74.0**	**48.1**	**33.0**	**50.6**	**64.8**	**49.2**	**38.2**	**63.7**	**76.8**	**63.9**	**54.3**
(b) Student-Dependent Realization													
AF	×	52.3	73.3	47.1	32.8	49.6	63.4	47.8	37.2	62.5	75.3	62.6	52.1
AF	✓	**52.8**	**74.0**	**48.1**	**33.0**	**50.6**	**64.8**	**49.2**	**38.2**	**63.7**	**76.8**	**63.9**	**54.3**

decline shows that candidate-space composition alone is insufficient and that effective coupling requires supervision to respond to the student's current competence. Further analyses on TVDF and static routing are provided in Appendices B.6 and B.12, respectively.

5 Training Dynamics of Student–Curriculum Coupling

Figure 4 traces the coupled evolution of student competence and the effective curriculum, highlighting the complementary roles of Anchors and Frontiers during training. Anchor competence remains high in the early stage, and most Anchor presentations are suspended, limiting additional optimization in already-supported capability regions. When an Anchor falls below the task criterion, its supervision is reactivated, allowing Anchors to support capability stabilization. Frontiers, meanwhile, account for most active OPD routes, and their competence rises with overall student progress. The AF curriculum therefore combines stabilization of existing capabilities with acquisition along the learning frontier.

As more Frontier examples reach the criterion, cumulative acquisition rises to 46.9% of the 900 unique Frontier examples. Over the same period, the update rate declines, suspension becomes more prevalent, and effective-batch OPD loss decreases through the main acquisition phase. Later in training, OPD supervision becomes increasingly sparse, with occasional empty effective batches and updates concentrated on the remaining deficits. This progression reflects the closed loop underlying SCC, in which the current student shapes the effective curriculum, the resulting OPD updates alter the student, and the updated state governs supervision in subsequent batches.



Figure 4: **Training dynamics of Student–Curriculum Coupling.** Current student competence for Anchors, Frontiers, and the full AF curriculum (left); realized OPD updates, suspensions, and Anchor reactivations (middle); and effective-batch OPD loss with cumulative Frontier acquisition (right). The shaded interval marks a representative drop in Anchor competence and the resulting reactivation of supervision.

6 Related Work

VLMs for Temporal Video Grounding. VLM-based TVG research has focused on both training strategies and temporal modeling. Training approaches range from boundary-aware instruction tuning and time-aware objectives (Huang et al., 2024a; Wang et al., 2024c) to reinforcement-based post-training for temporal localization and spatiotemporal perception (Wang et al., 2026; Li et al., 2025b; Chen et al., 2025). Complementary work makes temporal structure explicit through timestamp-aware visual encoding (Ren et al., 2024) and dedicated temporal representations (Huang et al., 2024b; Wang et al., 2025), with TRACE modeling events as sequences of timestamps, saliency scores, and captions (Guo et al., 2025). To handle longer videos, TimeSuite combines efficient token processing with grounded tuning (Zeng et al., 2025), while ReVisionLLM adopts recursive coarse-to-fine localization (Hannan et al., 2025). Alongside these modeling advances, TimeLens examines annotation quality and training strategies, providing curated training data and re-annotated benchmarks for more reliable evaluation (Zhang et al., 2025).

On-Policy Distillation. OPD trains on student-generated sequences with teacher feedback at the same visited prefixes, reducing distribution mismatch between training and inference (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025). MiniLLM adopts reverse KL to avoid overestimating low-probability teacher outputs (Gu et al., 2024), while entropy-aware OPD supplements this objective with forward KL when teacher entropy is high to preserve generation diversity (Jin et al., 2026). Alongside objective design, OPD analyses examine how compatible reasoning patterns and teacher capabilities novel to the student affect knowledge transfer, linking successful distillation to progressive alignment on high-probability tokens (Li et al., 2026c). Video-OPD extends this paradigm to TVG, providing dense token-level supervision by scoring student-generated trajectories without separate teacher rollouts (Li et al., 2026a). Building on this formulation, our work identifies the implicit persistent-value assumption and develops a coupled view of OPD.

Curriculum Learning and Data Selection. Curriculum learning organizes training examples by difficulty or learning structure (Bengio et al., 2009), while classical self-paced learning adapts an easy-to-hard curriculum alongside model updates (Kumar et al., 2010). Online hard example mining and Selective-Backprop instead prioritize high-loss examples (Shrivastava et al., 2016; Jiang et al., 2019). For language models, DoReMi learns domain-mixture weights through proxy-model optimization (Xie et al., 2023), while RegMix predicts effective mixtures from smaller-scale experiments (Liu et al., 2025). Skill-It adapts sampling to skill dependencies and current learning progress (Chen et al., 2023). In TVG, curriculum design encompasses temporal-task complexity and difficulty-aware data construction for reinforcement post-training (Wang et al., 2025; Chen et al., 2025; Wang et al., 2026). Video-OPD combines teacher validation with teacher–student disagreement for candidate selection (Li et al., 2026a). Compared with prior work on sample prioritization, SCC couples supervision trustworthiness with evolving student competence, jointly designing curriculum composition and supervision realization in OPD for VLM-based TVG.

7 Conclusion

This work shows how supervision aligned with evolving task competence can improve both accuracy and efficiency in VLM post-training. Distinguishing supervision trustworthiness from necessity leads to a coupled view of OPD, linking curriculum composition to the student's changing learning needs. SCC realizes this view through a compact AF-defined space of trustworthy supervision opportunities, with current student competence guiding their selective activation. Across three TVG benchmarks, it achieves higher grounding accuracy than Video-OPD on TVDF with fewer training examples, fewer teacher-scored trajectories, and shorter training time. Together with ablations and training-dynamics analyses, these results highlight the value of jointly designing curriculum composition and supervision timing.

Despite these gains, SCC assesses supervision necessity using tIoU between a single sampled prediction and the ground-truth interval. This task-level proxy does not directly estimate the marginal effect of supervising an example on generalization, and routing decisions may be sensitive to sampling variability. Extending this annotation-dependent assessment to tasks without comparable ground truth also requires alternative signals. Beyond these assessment limitations, our evaluation focuses primarily on TVG post-training within one VLM family. Future work should develop more direct, less annotation-dependent estimates of supervision value and evaluate this coupled design across additional model families and tasks.

AI USE STATEMENT

Generative AI tools assisted with manuscript refinement, discussions of derivations and proof arguments, feedback on methodology and experimental design, and interpretation of results. The authors remain responsible for all scientific claims, references, experimental results, and other content of the submission.

ETHICS STATEMENT

This work studies temporal video grounding using existing video datasets and pretrained models described in Section 4.1. The method may inherit biases and limitations from these resources. Improved video localization can support retrieval and accessibility but may also facilitate privacy-invasive video analysis. Deployment should account for consent, privacy, dataset bias, and localization errors, particularly in applications involving people. Use or redistribution of the data and models remains subject to their original licenses and access conditions. The reported benchmark results do not establish suitability for safety-critical deployment.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 describe the method, model configurations, video preprocessing, training settings, and evaluation protocol. Appendix A provides supporting derivations, and Appendix B details the algorithm, curriculum construction, training schedule, and supplementary experiments.

ACKNOWLEDGMENTS

We acknowledge the creators of TimeLens-100K and TimeLens-Bench (Zhang et al., 2025), which are released under the License Term of TimeLens. These resources are used solely for academic research, with no commercial or production use. We also thank the contributors to the underlying video datasets for making these resources available.

REFERENCES

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

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

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48, 2009. doi: 10.1145/1553374.1553380.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-It! a data-driven skills framework for understanding and training language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ruizhe Chen, Tianze Luo, Zhiting Fan, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. Datasets and recipes for video temporal grounding via reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 983–992. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-industry.66.

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

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

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5267–5275, 2017.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. TRACE: Temporal grounding video LLM via causal event modeling. In *International Conference on Learning Representations*, 2025.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. ReVisionLLM: Recursive vision-language model for temporal grounding in hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19012–19022, 2025. doi: 10.1109/CVPR52734.2025.01771.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5803–5812, 2017.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024a.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. LITA: Language instructed temporal-localization assistant. In *Computer Vision – ECCV 2024*, pp. 202–218. Springer, 2024b. doi: 10.1007/978-3-031-73039-9_12.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Angela H Jiang, Daniel L-K Wong, Giulio Zhou, David G Andersen, Jeffrey Dean, Gregory R Ganger, Gauri Joshi, Michael Kaminksy, Michael Kozuch, Zachary C Lipton, et al. Accelerating deep learning by focusing on the biggest losers. *arXiv preprint arXiv:1910.00762*, 2019.

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

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, volume 23, pp. 1189–1197, 2010.

Jie Lei, Tamara L. Berg, and Mohit Bansal. QVHighlights: Detecting moments and highlights in videos via natural language queries. *arXiv preprint arXiv:2107.09609*, 2021.

Jiaze Li, Yaya Shi, Zongyang Ma, Haoran Xu, Feng cheng, Huihui Xiao, Ruiwen Kang, Fan Yang, Tingting Gao, and Di Zhang. iMOVE: Instance-motion-aware video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23959–23975, 2025a.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-OPD: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. *arXiv preprint arXiv:2602.02994*, 2026a.

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

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*, 2025b.

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 *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026b.

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*, 2026c.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, 2025.

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

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

Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. QuerYD: A video dataset with high-quality text and audio narrations. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2265–2269, 2021.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024. doi: 10.1109/CVPR52733.2024.01357.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

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

Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. COSMO: Contrastive streamlined multimodal model with interleaved pre-training. *arXiv preprint arXiv:2401.00849*, 2024a.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 959–975. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.50.

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. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024b.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawk-Eye: Training video-text LLMs for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024c.

Xiaomi LLM-Core Team. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69798–69818, 2023.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. VideoChat-R1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oğuz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23056–23065, 2023.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *International Conference on Learning Representations*, 2025.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs. *arXiv preprint arXiv:2512.14698*, 2025.

APPENDIX

A THEORETICAL ANALYSIS

This appendix develops the theoretical and optimization analysis underlying SCC. Section A.1 analyzes the state dependence of marginal supervision value and distinguishes task-criterion satisfaction from distributional alignment. Section A.2 derives the sampled-token OPD estimator, and Section A.3 formalizes student-dependent realization, objective normalization, and selective execution.

A.1 SUPERVISION NECESSITY AND MARGINAL VALUE

Local marginal effects. Using the paired updates defining $\Delta_t(x)$, let:

$$d_t(x) = \theta_{t+1}^{+x} - \theta_{t+1}^{-x}. \tag{10}$$

The two updates share all other batch contributions, normalization, initial optimizer state, and training randomness. Thus, $d_t(x)$ captures the optimizer's response to including the on-policy distillation (OPD) contribution of x. To expose the state dependence of supervision value, we relate this displacement to the local change in population temporal video grounding (TVG) performance.

Suppose J_{TVG} has an L_J-Lipschitz gradient on a common convex neighborhood containing all paired updates, and the expectations below are finite. The following expansion is conditional on this regularity assumption, which need not hold for metrics based on discrete decoding. A first-order expansion gives:

$$\begin{aligned} \Delta_t(x) &= \mathbb{E}\big[\big\langle \nabla J_{\mathrm{TVG}}(\theta_{t+1}^{-x}), d_t(x)\big\rangle \,\big|\, \theta_t, x\big] + R_t(x), \\ |R_t(x)| &\le \frac{L_J}{2}\mathbb{E}\big[\|d_t(x)\|_2^2 \,\big|\, \theta_t, x\big]\,. \end{aligned} \tag{11}$$

Here, $R_t(x)$ is the conditional expected first-order remainder. For each paired update, integrating the gradient along $\theta_{t+1}^{-x} + s\, d_t(x)$ for $s \in [0,1]$ and applying Lipschitz continuity bounds the remainder by $L_J\|d_t(x)\|_2^2/2$ in absolute value. Taking conditional expectations yields the stated bound.

The leading term measures the alignment of the supervision-induced displacement with the local direction of increasing population performance. In OPD, this displacement reflects the student's sampled prefixes, the corresponding distillation signals, and the optimizer state. Both the displacement and the performance gradient can vary as the student evolves, allowing supervision from a fixed teacher on a fixed example to have changing marginal effects.

Task-criterion satisfaction and distributional alignment. A realized prediction can satisfy the task-level criterion while distributional alignment remains incomplete. Consider a one-step model with two output symbols encoding a correct interval y_c and an incorrect interval y_w. Let the student and teacher distributions be $p_S = (\alpha, 1-\alpha)$ and $p_T = (\beta, 1-\beta)$, respectively, with $\alpha, \beta \in (1/2, 1)$ and $\alpha \neq \beta$. For $\tau_S \le 1$, a student rollout producing y_c has temporal intersection over union (tIoU) equal to one and satisfies the task-level competence criterion used for routing. Nevertheless,

$$\begin{aligned} D_{\mathrm{KL}}(p_S\|p_T) &= \alpha \log\frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta} > 0, \\ \frac{\partial}{\partial\alpha} D_{\mathrm{KL}}(p_S\|p_T) &= \log\frac{\alpha(1-\beta)}{\beta(1-\alpha)} \neq 0. \end{aligned} \tag{12}$$

Thus, satisfying the task criterion need not eliminate the distributional alignment signal. The criterion evaluates the realized task outcome, while the Kullback–Leibler (KL) divergence compares predictive distributions.

Implications for supervision allocation. These observations motivate separating the availability of trustworthy supervision from its realization during training. Initial suitability alone does not guarantee persistent marginal value, and a residual distillation discrepancy does not establish a deficit in the realized task outcome. In Student–Curriculum Coupling (SCC), the necessity indicator uses current task performance to guide supervision activation without directly estimating $\Delta_t(x)$. Within this allocation principle, the Anchor–Frontier (AF) space retains trustworthy opportunities for capability acquisition and stabilization.

A.2 SAMPLED-TOKEN OPD ESTIMATOR

Section 3.1 defines token-level reverse-KL alignment under a fixed rollout distribution. Here, we present the sampled-token surrogate used for optimization (Li et al., 2026a) and establish its local gradient identity with Equation 4. The derivation assumes bounded rollout lengths and smooth, positive token probabilities over a finite vocabulary.

Sampled-token supervision. At training step t, the rollout policy π_{θ_t} generates:

$$y_t = (y_{t,1}, \ldots, y_{t,L_t}) \sim \pi_{\theta_t}(\cdot \mid x),$$

with prefix $h_{t,k} = (x, y_{t,<k})$ at position k. For a vocabulary token z, define:

$$p_{t,k}(z) = \pi_{\theta_t}(z \mid h_{t,k}), \qquad q_{t,k}(z) = p_{T,k}(z) = \pi_T(z \mid h_{t,k}),$$

the rollout-policy and frozen-teacher probabilities at the same prefix. The sampled-token supervision signal is:

$$r_{t,k} = \log q_{t,k}(y_{t,k}) - \log p_{t,k}(y_{t,k}). \tag{13}$$

Since $y_{t,k}$ is sampled from $p_{t,k}$ conditional on $h_{t,k}$,

$$\mathbb{E}_{y_{t,k} \sim p_{t,k}} \left[-r_{t,k} \mid h_{t,k} \right] = D_{\mathrm{KL}}(p_{t,k} \,\|\, q_{t,k}) \,. \tag{14}$$

Thus, $-r_{t,k}$ is an unbiased single-token estimate of the reverse-KL value at that prefix (Li et al., 2026c).

Importance-weighted surrogate. During the update, θ_t remains fixed, while θ denotes the optimization variable initialized from θ_t. The token-level likelihood ratio is:

$$\rho_{t,k}(\theta) = \frac{\pi_\theta(y_{t,k} \mid h_{t,k})}{\pi_{\theta_t}(y_{t,k} \mid h_{t,k})}. \tag{15}$$

Following Video-OPD (Li et al., 2026a), the per-example surrogate is:

$$\ell_{\mathrm{OPD}}(x, y_t; \theta, \theta_t, \pi_T) = -\sum_{k=1}^{L_t} \mathrm{sg}[r_{t,k}] \, \rho_{t,k}(\theta), \tag{16}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. Holding the sampled trajectory, rollout probabilities, and supervision signals fixed gives:

$$\nabla_\theta \ell_{\mathrm{OPD}} = -\sum_{k=1}^{L_t} \mathrm{sg}[r_{t,k}] \, \rho_{t,k}(\theta) \, \nabla_\theta \log \pi_\theta(y_{t,k} \mid h_{t,k}). \tag{17}$$

At the rollout point $\theta = \theta_t$, the likelihood ratio equals one, and its gradient satisfies:

$$\nabla_\theta \rho_{t,k}(\theta)|_{\theta=\theta_t} = \nabla_\theta \log \pi_\theta(y_{t,k} \mid h_{t,k})|_{\theta=\theta_t} \,.$$

Relation to the fixed-rollout objective. For a fixed prefix $h_{t,k}$, write $p_{\theta,k}(z) = \pi_\theta(z \mid h_{t,k})$, so that $p_{\theta_t,k} = p_{t,k}$. Differentiating the token-level reverse KL yields:

$$\begin{aligned} \nabla_\theta D_{\mathrm{KL}}(p_{\theta,k} \,\|\, q_{t,k}) &= \mathbb{E}_{z \sim p_{\theta,k}} \left[\left(\log \frac{p_{\theta,k}(z)}{q_{t,k}(z)} + 1 \right) \nabla_\theta \log p_{\theta,k}(z) \right] \\ &= \mathbb{E}_{z \sim p_{\theta,k}} \left[\log \frac{p_{\theta,k}(z)}{q_{t,k}(z)} \nabla_\theta \log p_{\theta,k}(z) \right] . \end{aligned} \tag{18}$$

The second equality follows from the score-function identity $\mathbb{E}_{z \sim p_{\theta,k}}[\nabla_\theta \log p_{\theta,k}(z)] = 0$. Combining Equations 13, 17, and 18 at $\theta = \theta_t$ gives:

$$\begin{aligned} &\mathbb{E}_{y_{t,k} \sim p_{t,k}} \left[\nabla_\theta \left(-\mathrm{sg}[r_{t,k}] \rho_{t,k}(\theta) \right) \mid h_{t,k} \right] \big|_{\theta=\theta_t} \\ &\quad = \nabla_\theta D_{\mathrm{KL}}(p_{\theta,k} \,\|\, q_{t,k})|_{\theta=\theta_t} \,. \end{aligned} \tag{19}$$

Taking expectation over the fixed rollout policy and summing over generated positions establishes the connection to the objective in Section 3.1:

$$\begin{aligned} &\mathbb{E}_{y_t \sim \pi_{\theta_t}(\cdot|x)} \left[\nabla_\theta \ell_{\mathrm{OPD}}(x, y_t; \theta, \theta_t, \pi_T)|_{\theta=\theta_t} \right] \\ &\quad = \nabla_\theta \mathcal{L}_{\mathrm{OPD},t}(\theta; x)|_{\theta=\theta_t} \,. \end{aligned} \tag{20}$$

This identity establishes gradient matching at $\theta = \theta_t$ with the rollout distribution fixed during differentiation. The full sequence-level reverse KL additionally differentiates prefix visitation probabilities, yielding long-horizon score-function terms (Gu et al., 2024). Section 3.4 aggregates the per-example surrogate above over the effective batch with fixed routing decisions.

A.3 Objective Normalization and Efficient Execution

The gradient identity in Appendix A.2 applies to the base OPD surrogate before task-dependent routing. SCC optimizes the masked surrogate in Equation 8, with routing decisions held fixed during each update. Because retention depends on the sampled trajectory, it changes both the frequency of supervision and the trajectory gradients contributing to the expected update.

Anchor and Frontier membership is fixed by the initial student assessment. For presentation i at step t, the current trajectory $y_{t,i}$ determines the supervision mask:

$$m_{t,i} = \mathbf{1}\{q_{S,t}(x_i; y_{t,i}) < \tau_S\}. \tag{21}$$

The task score uses the ground-truth interval, and token-level signals are teacher-derived. Scores at or above τ_S suspend supervision. Retained trajectories are reused for teacher scoring and OPD without additional sampling.

The same rule applies to both curriculum regions. Retained Frontiers receive acquisition-oriented supervision. An Anchor $x_i \in \mathcal{A}$ is reactivated when $q_{S,0}(x_i) \geq \tau_S$ and $m_{t,i} = 1$. Anchor selection uses τ_A; reactivation compares initial and current performance against τ_S and can occur on the first scheduled presentation. Acquisition and reactivation both enter the effective batch and the supervised count. Routing statistics count supervised and suspended presentations, including repeated occurrences; curriculum size counts unique examples.

For a nonempty scheduled batch, let n_t denote the preset normalization denominator and $M_t = \sum_{i=1}^{|\mathcal{B}_t|} m_{t,i}$ the number of supervised presentations. Let $\ell_{t,i}(\theta)$ abbreviate the per-example surrogate $\ell_{\mathrm{OPD}}(x_i, y_{t,i}; \theta, \theta_t, \pi_T)$ defined in Equation 16. The coupled objective in Equation 8 has the equivalent masked form:

$$\mathcal{L}_t^{\mathrm{SCC}}(\theta) = \frac{1}{n_t} \sum_{i=1}^{|\mathcal{B}_t|} m_{t,i} \ell_{t,i}(\theta) = \frac{1}{n_t} \sum_{i:m_{t,i}=1} \ell_{t,i}(\theta). \tag{22}$$

With the sampled trajectories, teacher signals, and masks held fixed during the update, its gradient is:

$$\nabla_\theta \mathcal{L}_t^{\mathrm{SCC}}(\theta) = \frac{1}{n_t} \sum_{i:m_{t,i}=1} \nabla_\theta \ell_{t,i}(\theta). \tag{23}$$

Thus, evaluating only the retained terms is algebraically equivalent to masking the full scheduled batch, provided that the denominator remains n_t. Teacher scores and OPD losses for suspended trajectories need not be computed.

For $M_t > 0$, define the active-only mean as $\mathcal{L}_t^{\mathrm{active}}(\theta) = M_t^{-1} \sum_{i:m_{t,i}=1} \ell_{t,i}(\theta)$. Then:

$$\mathcal{L}_t^{\mathrm{SCC}}(\theta) = \frac{M_t}{n_t} \mathcal{L}_t^{\mathrm{active}}(\theta). \tag{24}$$

Active-only normalization would increase each retained term's weight by n_t/M_t relative to the coupled objective. Normalization by n_t preserves the weight $1/n_t$, so suspension removes supervision contributions without amplifying the remaining terms. The total retained weight is M_t/n_t. This scaling applies to the objective and its gradient, but does not imply proportional scaling of parameter updates under adaptive optimization.

When $M_t = 0$, the masked objective and its gradient are zero. The optimizer and weight-decay updates are explicitly skipped, preserving the student parameters and optimizer state while the training schedule advances. All scheduled presentations still require student generation and assessment; selective execution removes only the subsequent teacher scoring and OPD computation for suspended trajectories.

B Implementation Details and Additional Experiments

Sections B.1 and B.2 describe the SCC algorithm, AF construction, and training schedule. The remaining subsections report supplementary accuracy–efficiency results and analyses of alternative objectives, student-dependent realization beyond AF, competence-criterion sensitivity, teacher choice, checkpoint-wise performance, the Anchor–Frontier ratio, and broader video understanding.

B.1 ALGORITHM

Algorithm 1 summarizes SCC, from AF construction to student-dependent routing and selective OPD updates.

Algorithm 1 Student–Curriculum Coupling (SCC)

Require: Candidate pool $\mathcal{P}$, initial student π_{θ_0}, frozen teacher π_T, curriculum parameters $\tau_T, \tau_A, \tau_F, N_A, N_F$, competence criterion τ_S, and scheduled-step budget K
Ensure: Trained student π_{θ_K}

Evaluate $q_T(x)$ and $q_{S,0}(x)$ for each $x \in \mathcal{P}$
Construct $\mathcal{A}$ and $\mathcal{F}$ using Equation 5; set $\mathcal{D}_{\mathrm{AF}} \leftarrow \mathcal{A} \cup \mathcal{F}$
Construct and fix the schedule $\mathcal{S} = (\mathcal{B}_0, \ldots, \mathcal{B}_{K-1})$ over $\mathcal{D}_{\mathrm{AF}}$ and the normalization denominators $\{n_t\}_{t=0}^{K-1}$
for $t = 0, \ldots, K-1$ **do**
    Obtain the scheduled batch $\mathcal{B}_t$ from $\mathcal{S}$
    Sample $y_{t,i} \sim \pi_{\theta_t}(\cdot \mid x_i)$ for each $x_i \in \mathcal{B}_t$
    Decode each trajectory and compute $q_{S,t}(x_i; y_{t,i})$ using Equation 6
    Form the effective batch

$$\mathcal{B}_t^{\mathrm{eff}} \leftarrow \{(x_i, y_{t,i}) : x_i \in \mathcal{B}_t,\ q_{S,t}(x_i; y_{t,i}) < \tau_S\}$$

    if $\mathcal{B}_t^{\mathrm{eff}} \neq \varnothing$ **then**
        Score only the retained student trajectories with π_T
        Form $\mathcal{L}_t^{\mathrm{SCC}}(\theta)$ using Equation 8, with denominator n_t and fixed trajectories and routing decisions
        $\theta_{t+1} \leftarrow \mathcal{U}_{\mathrm{SCC}}\big(\theta_t, \mathcal{B}_t, \mathcal{B}_t^{\mathrm{eff}}; \pi_T\big)$
    else
        Skip optimizer and weight-decay updates; retain the optimizer state
        $\theta_{t+1} \leftarrow \theta_t$
    end if
end for
return π_{θ_K}

B.2 CURRICULUM CONSTRUCTION AND SCHEDULE DETAILS

Curriculum construction. We construct AF offline from TimeLens-100K (Zhang et al., 2025), using examples from HiREST (Zala et al., 2023), QuerYD (Oncescu et al., 2021), CosMo-Cap (Wang et al., 2024a), InternVid-VTime (Wang et al., 2024b; Huang et al., 2024a), and DiDeMo (Hendricks et al., 2017). Following the construction in Equation 5, we retain candidates with $q_T(x) \geq 0.7$ and select 100 Anchors with $q_{S,0}(x) \geq 0.7$ and 900 Frontiers with $q_{S,0}(x) < 0.3$. Candidates in the intermediate competence range $0.3 \leq q_{S,0}(x) < 0.7$ are excluded. We use 0.7 as a stringent task-level threshold for both teacher qualification and Anchor membership, so that qualified candidates have reliable teacher predictions and Anchors represent clearly supported initial capabilities. The Frontier threshold of 0.3 isolates examples with substantial learning headroom, while excluding the intermediate range creates a clear separation between stabilization and acquisition. Table 4 reports the resulting source composition.

Table 4: **Source composition of AF.** Unique counts refer to selected examples; presentations additionally include four Frontier repeats in the final scheduled step.

Source	Anchor	Frontier	Unique	Presentations
CosMo-Cap	49	437	486	489
InternVid-VTime	26	235	261	262
QuerYD	11	106	117	117
DiDeMo	10	89	99	99
HiREST	4	33	37	37
Total	**100**	**900**	**1,000**	**1,004**

Training schedule. Per-step AF allocations follow a piecewise-linear profile with target shares of 100%, 70%, 30%, 10%, and 5% at steps 1, 20, 40, 60, and 79, respectively. The interpolated allocations are normalized to 1,000 unique examples and deterministically rounded to integer counts, with a target Anchor–Frontier ratio of 1:9. Example ordering and assignment within these allocations are randomized before training and fixed throughout optimization. The final step includes four repeated

Frontier presentations, yielding 1,004 presentations. For loss normalization, SCC uses $n_t = 32$ for steps 1–78 and $n_t = 8$ for step 79, independently of the number of assigned or retained AF presentations. Student-dependent realization selects trajectories for OPD without changing their assigned steps; steps with an empty effective batch advance the schedule without an optimizer update. Table 5 summarizes presentation counts within each evaluation-checkpoint interval.

Table 5: **Distribution of AF across the training schedule.** Counts denote presentations grouped by evaluation-checkpoint intervals. † includes four Frontier repeats in the final step.

Training Steps	Anchor	Frontier	Total
1–20	53	481	534
21–40	31	276	307
41–59	12	104	116
60–79	4	43†	47
Total	**100**	**904**†	**1,004**

B.3 Detailed Accuracy–Efficiency Analysis

Table 6 complements Figure 3 with training time and endpoint recall. We compare SCC with Video-OPD on its Teacher-Validated Disagreement Focusing (TVDF) curriculum (Li et al., 2026a) and on the same AF curriculum. Training time denotes wall-clock time on eight NVIDIA A100 GPUs, including initialization, optimization, and checkpoint saving. Mean recall at each threshold averages the corresponding metric across Charades-TimeLens (Zhang et al., 2025), ActivityNet-TimeLens (Zhang et al., 2025), and QVHighlights-TimeLens (Zhang et al., 2025).

SCC reduces training time by 50.4% relative to Video-OPD on TVDF and by 15.3% on the same AF curriculum, while improving mean recall at all three thresholds. The AF comparison shows that the efficiency gain persists when the candidate curriculum is held fixed.

AF construction uses teacher and initial-student predictions scored against ground-truth intervals. The resulting tIoU scores determine teacher qualification and the Anchor and Frontier candidate regions, from which examples are selected under the prescribed budgets. Once these scores are available, selection requires no further model inference, teacher scoring of student-generated prefixes, or token-level divergence estimation.

Table 6: **Training time and endpoint recall.** All three methods are trained and evaluated locally. Training time on eight NVIDIA A100 GPUs is reported as hours:minutes:seconds. Recall is averaged across the three TimeLens benchmarks and reported as percentages to two decimal places. Bold marks the shortest time and highest recall at each threshold. Means are computed before rounding.

Method	Training Time	Mean R@0.3	Mean R@0.5	Mean R@0.7
Video-OPD (Li et al., 2026a) (TVDF)	2:01:46	69.13	50.57	39.53
Video-OPD (Li et al., 2026a) (AF)	1:11:20	70.69	52.53	40.70
SCC (AF)	**1:00:26**	**71.83**	**53.72**	**41.84**

B.4 Comparison with TimeLens

Table 7 compares SCC with TimeLens-8B (Zhang et al., 2025) and single-round Video-OPD (Li et al., 2026a). TimeLens-8B undergoes TVG supervised fine-tuning followed by GRPO on 12,000 examples, while Video-OPD and SCC use OPD curricula of 2,500 and 1,000 examples, respectively. SCC outperforms single-round Video-OPD across all reported metrics. Compared with TimeLens-8B, SCC achieves higher scores on all ActivityNet-TimeLens metrics and higher mIoU, R@0.5, and R@0.7 on QVHighlights-TimeLens, including an R@0.7 of 54.3 versus 51.8. TimeLens-8B retains an advantage on Charades-TimeLens and QVHighlights-TimeLens R@0.3, indicating benchmark-dependent trade-offs across these training regimes.

Table 7: **Comparison with TimeLens-8B.** Baseline results are taken from Video-OPD (Li et al., 2026a); values are percentages; bold indicates the best result in each column.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
Method	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
TimeLens-8B	**53.3**	**74.6**	**49.5**	**33.4**	49.3	63.8	48.0	36.4	63.0	**77.8**	63.4	51.8
Video-OPD (Round 1)	52.0	73.1	45.8	32.4	47.3	60.5	45.6	35.8	61.0	73.8	60.3	50.4
SCC	52.8	74.0	48.1	33.0	**50.6**	**64.8**	**49.2**	**38.2**	**63.7**	76.8	**63.9**	**54.3**

B.5 Alternative Post-Training Objectives on AF

Table 8 compares SCC with alternative post-training objectives (Li et al., 2026a) on the same AF curriculum. SCC achieves the strongest R@0.7 on all three benchmarks and the best result across all Charades-TimeLens metrics. GRPO performs best on R@0.3 and R@0.5 for QVHighlights-TimeLens and on R@0.5 for ActivityNet-TimeLens, indicating different trade-offs between coarse retrieval and precise temporal localization. This comparison focuses on endpoint quality under method-specific training schedules. OP-FKD, OP-RKD, and GRPO are trained for 32 steps. GRPO runs for one epoch with a global prompt batch size of 32, sampling eight rollouts per example at temperature 1.0.

Table 8: **Alternative post-training objectives on AF.** All methods use the AF curriculum and are trained and evaluated using our local pipeline. Off-policy forward- and reverse-KL distillation (OP-FKD and OP-RKD), together with GRPO, follow the objective definitions in Video-OPD (Li et al., 2026a). All values are percentages, and bold indicates the best result in each column.

	Charades-TimeLens			ActivityNet-TimeLens			QVHighlights-TimeLens		
Method	**R@0.3**	**R@0.5**	**R@0.7**	**R@0.3**	**R@0.5**	**R@0.7**	**R@0.3**	**R@0.5**	**R@0.7**
OP-FKD (Li et al., 2026a)	72.20	47.43	31.28	62.22	46.71	35.76	75.47	62.36	50.49
OP-RKD (Li et al., 2026a)	69.61	47.01	29.91	61.47	46.67	35.73	74.95	62.04	51.14
GRPO (Li et al., 2026a)	67.98	47.81	27.27	63.76	**49.78**	36.31	**77.16**	**65.61**	51.53
SCC	**73.95**	**48.14**	**33.01**	**64.78**	49.18	**38.24**	76.77	63.85	**54.25**

B.6 Student-Dependent Realization on TVDF

To assess student-dependent realization beyond the AF candidate space, we retain the original TVDF curriculum and enable only the current-student competence gate with $\tau_S = 0.7$. Of 2,504 scheduled presentations, 955 receive teacher scoring and OPD updates, and 1,549 are suspended. Table 9 compares the variant at checkpoint 79 with the published Video-OPD results (Li et al., 2026a).

The variant achieves higher R@0.7 on all three benchmarks and improves all three recall metrics on ActivityNet-TimeLens. Small decreases occur at R@0.3 on Charades-TimeLens and at R@0.3 and R@0.5 on QVHighlights-TimeLens. These results support the applicability of student-dependent realization beyond AF, with the most consistent gains under the stricter localization criterion.

Table 9: **Student-dependent realization on TVDF.** Both methods use the TVDF curriculum. Video-OPD results are taken from Li et al. (2026a); the student-dependent variant is trained and evaluated locally at checkpoint 79. Values are recall percentages rounded to one decimal place; bold indicates the higher value in each column.

	Charades-TimeLens			ActivityNet-TimeLens			QVHighlights-TimeLens		
Method	**R@0.3**	**R@0.5**	**R@0.7**	**R@0.3**	**R@0.5**	**R@0.7**	**R@0.3**	**R@0.5**	**R@0.7**
Video-OPD (Li et al., 2026a)	**73.1**	45.8	32.4	60.5	45.6	35.8	**73.8**	**60.3**	50.4
Video-OPD + Student-Dep. Realization	72.7	**46.2**	**32.9**	**62.1**	**47.4**	**37.4**	73.6	60.2	**50.9**

B.7 Sensitivity to the Student-Competence Criterion

We vary the student-competence criterion over $\tau_S \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ while keeping the AF curriculum, student initialization, teacher, data order, and optimization protocol fixed. The main setting, $\tau_S = 0.7$, was fixed before benchmark evaluation and was not selected from this sweep. We set the main criterion to $\tau_S = 0.7$ to apply the same stringent task-level standard when determining whether supervision remains necessary during training. All runs are evaluated at checkpoint 79, and a scheduled occurrence receives OPD supervision when $q_{S,t} < \tau_S$. Table 10 reports the resulting supervision routing, while Table 11 reports complete TVG performance.

Table 10: **Realized supervision routing under different student-competence criteria.** All settings use the same 1,004 scheduled presentations. The shaded row denotes the main setting.

τ_S	**OPD Routes**	**Suspended Routes**	**OPD Route Rate**
0.5	397	607	39.54%
0.6	438	566	43.63%
0.7	**492**	**512**	**49.00%**
0.8	563	441	56.08%
0.9	646	358	64.34%

Table 11: **TVG performance under different student-competence criteria.** All benchmarks use the TimeLens annotations. All settings use AF; "–" denotes uniform OPD supervision, while the remaining rows vary τ_S. Values are percentages. The shaded row denotes the setting used in the main experiments, and bold indicates the best result in each column.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
τ_S	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
-	52.26	73.33	47.13	32.80	49.61	63.40	47.84	37.20	62.49	75.34	62.62	52.11
0.5	52.51	73.71	47.81	32.68	50.52	64.20	48.82	38.22	63.17	**76.83**	62.23	52.56
0.6	52.41	73.68	47.07	32.47	50.30	63.84	48.36	38.18	63.30	76.18	62.69	53.41
0.7	52.79	73.95	**48.14**	33.01	**50.63**	**64.78**	**49.18**	**38.24**	**63.70**	76.77	**63.85**	**54.25**
0.8	**52.85**	**74.13**	47.79	**33.51**	50.11	63.44	48.20	38.07	62.85	75.86	62.88	52.56
0.9	52.40	73.39	47.81	32.35	49.67	63.20	47.62	37.38	61.93	74.56	62.10	51.66

Increasing τ_S monotonically increases the fraction of teacher-scored OPD routes, from 39.54% at $\tau_S = 0.5$ to 64.34% at $\tau_S = 0.9$, whereas endpoint performance remains non-monotonic. The main setting, $\tau_S = 0.7$, achieves the best result on eight of the twelve reported metrics, including all four ActivityNet-TimeLens metrics and three of the four QVHighlights-TimeLens metrics. Charades-TimeLens slightly favors $\tau_S = 0.8$ on mean intersection over union (mIoU), R@0.3, and R@0.7. Performance declines at $\tau_S = 0.9$ despite its higher supervision coverage, indicating that routing more examples through OPD does not necessarily improve the final student.

B.8 Robustness to Teacher Choice

We assess robustness to teacher choice using Qwen3-VL-4B and Qwen3-VL-8B (Bai et al., 2025a) teachers trained with the same GRPO recipe on a separately constructed 2,500-example difficulty-weighted curriculum. Construction follows the Time-R1 and TimeLens procedures (Wang et al., 2026; Zhang et al., 2025) adopted in Video-OPD (Li et al., 2026a), but the selected example set differs from that used to train its released 32B teacher. Each teacher supervises a Qwen3-VL-8B student on the original AF candidate set without teacher-specific re-filtering, with all other student-training settings held constant. We reuse the 32B-qualified AF set without assuming that every example remains qualified under the 4B/8B teachers. Table 12 reports teacher and student performance across the three TVG benchmarks.

At the final scheduled step, students outperform their respective teachers in mIoU across all six teacher–benchmark pairings, supporting robustness to teacher choice. The 4B-supervised student leads on QVHighlights and ActivityNet, and the 8B-supervised student on Charades-STA, with no uniform benefit from greater teacher size. Differences in training procedures and, for the 4B teacher, model size preclude strictly matched before–after comparisons.

Table 12: **Robustness to teacher choice.** GRPO teachers and their Qwen3-VL-8B students are evaluated at final checkpoints using TimeLens annotations. Student runs share the same AF configuration and differ only in teacher choice. mIoU and recall are percentages; Δ mIoU denotes the student's gain over its teacher in percentage points. Bold marks the higher value for each metric within each teacher–student pair.

Teacher	Model	Benchmark	mIoU	R@0.3	R@0.5	R@0.7	Δ mIoU
4B GRPO	Direct teacher	QVHighlights	61.61	76.51	61.65	50.49	–
4B GRPO	**SCC student**	QVHighlights	**64.92**	**77.87**	**65.41**	**55.09**	+3.31
4B GRPO	Direct teacher	Charades-STA	51.35	**73.15**	47.40	30.90	–
4B GRPO	**SCC student**	Charades-STA	**52.39**	72.58	**47.90**	**33.66**	+1.04
4B GRPO	Direct teacher	ActivityNet	49.83	64.82	48.49	37.02	–
4B GRPO	**SCC student**	ActivityNet	**51.45**	**65.29**	**49.91**	**39.20**	+1.62
8B GRPO	Direct teacher	QVHighlights	60.55	74.24	60.68	49.84	–
8B GRPO	**SCC student**	QVHighlights	**62.98**	**75.86**	**63.01**	**53.54**	+2.43
8B GRPO	Direct teacher	Charades-STA	52.20	73.65	48.32	31.64	–
8B GRPO	**SCC student**	Charades-STA	**53.05**	**73.86**	**48.68**	**33.81**	+0.85
8B GRPO	Direct teacher	ActivityNet	50.32	64.73	49.71	37.44	–
8B GRPO	**SCC student**	ActivityNet	**51.11**	**65.11**	**49.73**	**38.84**	+0.79

B.9 Checkpoint-Wise TVG Performance

Table 13 summarizes TVG performance across training checkpoints. The final checkpoint achieves the highest mIoU on Charades-TimeLens and QVHighlights-TimeLens, along with the highest

R@0.7 on QVHighlights-TimeLens. ActivityNet-TimeLens reaches its best scores at checkpoints 40 and 59, with modest declines at the final checkpoint.

Table 13: **Checkpoint-wise TVG performance of SCC.** All checkpoints are taken from the main training run. All values are percentages, and bold indicates the best checkpoint for each metric within each benchmark.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
Checkpoint	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**
20	48.24	68.66	45.97	28.84	49.39	63.93	49.07	37.31	62.80	**77.35**	**64.50**	51.59
40	52.75	**74.04**	47.78	32.92	**50.93**	64.67	**49.47**	**38.64**	63.46	76.77	63.27	53.28
59	52.73	73.60	47.81	**33.24**	50.86	**65.16**	49.31	**38.64**	63.39	76.31	63.40	53.28
79	**52.79**	73.95	**48.14**	33.01	50.63	64.78	49.18	38.24	**63.70**	76.77	63.85	**54.25**

B.10 Sensitivity to the Anchor–Frontier Ratio

We examine how the Anchor/Frontier ratio affects realized supervision and endpoint performance within a fixed 1,000-example candidate space. The main 1:9 configuration is compared with 3:7 and 1:1 alternatives. The main 1:9 ratio was fixed before benchmark evaluation and was not selected from this comparison. Each configuration follows the same student-dependent realization rule and is evaluated at checkpoint 79. Table 14 reports the supervision routing and TVG performance.

The region-specific routing pattern remains stable across the tested ratios. Only 11.0–13.6% of Anchor presentations receive OPD supervision, compared with 53.2–53.9% of Frontier presentations. As the Anchor share increases, the overall OPD route rate decreases from 49.0% to 41.8% and 33.8%. The candidate-space ratio therefore directly shapes the realized supervision workload under the same routing rule.

The main 1:9 ratio achieves the best result on 11 of the 12 TVG metrics. Both alternative ratios use fewer teacher-scored OPD routes than the main configuration and yield lower endpoint performance overall. Together, these results show that the Anchor–Frontier ratio controls an accuracy–workload trade-off, with the Frontier-heavy main configuration providing the strongest endpoint performance among the tested ratios.

Table 14: **Sensitivity to the Anchor–Frontier ratio.** Each candidate space contains 1,000 unique examples. Ratios denote the numbers of unique Anchor and Frontier examples. Panel (a) reports routing over 1,004 scheduled presentations, including repeated occurrences in the training schedule. Panel (b) reports TVG performance at checkpoint 79. Values are percentages. The shaded rows denote the main 1:9 configuration, and bold values in panel (b) indicate the best result in each column.

(a) Realized supervision routing.

	Anchor			Frontier			Overall		
A/F ratio	**OPD Routes**	**Suspended**	**Rate**	**OPD Routes**	**Suspended**	**Rate**	**OPD Routes**	**Suspended**	**Rate**
1:9	11	89	11.00%	481	423	53.21%	492	512	49.00%
3:7	41	260	13.62%	379	324	53.91%	420	584	41.83%
1:1	68	433	13.57%	271	232	53.88%	339	665	33.76%

(b) Endpoint TVG performance.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
A/F ratio	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**
1:9	**52.79**	**73.95**	**48.14**	**33.01**	**50.63**	**64.78**	**49.18**	38.24	**63.70**	**76.77**	**63.85**	**54.25**
3:7	52.35	73.63	47.04	32.77	49.95	63.49	47.98	38.04	62.32	74.69	62.04	52.50
1:1	52.43	73.36	47.10	32.80	50.39	64.11	48.29	**38.36**	62.86	75.86	62.43	52.95

B.11 Sensitivity to Candidate-Space Scale

We examine sensitivity to candidate-space scale by comparing the main 1,000-example AF curriculum with 1,500-, 2,000-, and 2,500-example variants, all maintaining an Anchor–Frontier ratio of 1:9. All runs use 79 scheduled steps with global batch sizes capped at 32. Larger curricula preserve the original 1,000 examples at their scheduled positions and add teacher-qualified examples through a deterministic, source-matched allocation balanced across vacant positions. All unique examples are presented, with four final-step repeats yielding 1,004, 1,504, 2,004, and 2,504 presentations, respectively. Table 15 reports TVG performance at checkpoint 79.

Expanding the candidate space does not consistently improve endpoint performance. The main 1,000-example configuration achieves the best result on nine of the twelve metrics, including all four on QVHighlights-TimeLens. The 1,500-example variant yields small gains in Charades-TimeLens R@0.3 and ActivityNet-TimeLens mIoU and R@0.7. These results show that a compact AF curriculum can define an effective supervision space for OPD.

Table 15: **Sensitivity to candidate-space scale.** Candidate-space size counts unique examples. All configurations use an Anchor–Frontier ratio of 1:9 and are evaluated at checkpoint 79. Values are percentages rounded to two decimal places. The shaded row denotes the main configuration, and bold indicates the best result in each column.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
Candidate-Space	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**
1,000	**52.79**	73.95	**48.14**	**33.01**	50.63	**64.78**	**49.18**	38.24	**63.70**	**76.77**	**63.85**	**54.25**
1,500	52.69	**74.19**	47.58	32.65	**50.81**	64.40	49.07	**38.56**	62.40	75.41	61.71	51.79
2,000	52.63	74.07	47.61	32.83	50.07	63.82	48.13	37.91	62.70	75.67	62.43	53.08
2,500	52.00	73.51	47.04	31.91	50.38	64.11	48.82	38.31	62.76	76.38	62.82	52.43

B.12 Additional Controls on Student-Dependent Realization

To better understand SCC's gains, we evaluate a budget-matched static routing baseline and a variant using effective-batch normalization. Table 16 reports results at checkpoint 79.

Budget-matched static routing. We fix a routing mask before training to assign teacher scoring and OPD to 492 of the 1,004 scheduled presentations, matching SCC's realized supervision budget. Constrained random sampling selects 49 distinct Anchors and 443 distinct Frontiers, preserving the 1:9 Anchor–Frontier ratio and within-region source proportions up to integer rounding. Selected examples retain their original positions, with per-step supervision counts matched using only aggregate counts from the reference SCC run. The mask remains fixed throughout the control run, and loss normalization uses the same per-step denominators as SCC. SCC achieves higher mIoU on all three benchmarks and outperforms this control on 11 of the 12 metrics.

Effective-batch normalization. This experiment retains student-dependent routing and averages the OPD loss over retained presentations using the active-only normalization defined in Appendix A.3. The main SCC configuration performs better on most metrics.

Table 16: **Controls on student-dependent realization.** All runs use AF and are evaluated at checkpoint 79. Static routing fixes 492 supervised presentations before training. SCC (effective-batch) averages loss over supervised presentations per batch. Values are percentages; bold marks column maxima and shading denotes SCC.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
Configuration	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**	**mIoU**	**R@0.3**	**R@0.5**	**R@0.7**
Video-OPD on AF	52.26	73.33	47.13	32.80	49.61	63.40	47.84	37.20	62.49	75.34	62.62	52.11
Static routing	52.21	72.79	47.19	**33.04**	49.65	63.36	47.29	37.44	62.12	74.95	61.39	51.53
SCC (effective-batch)	**52.82**	**74.31**	48.02	32.74	50.55	64.53	48.78	38.13	62.88	75.54	62.75	53.54
SCC	52.79	73.95	**48.14**	33.01	**50.63**	**64.78**	**49.18**	**38.24**	**63.70**	**76.77**	**63.85**	**54.25**

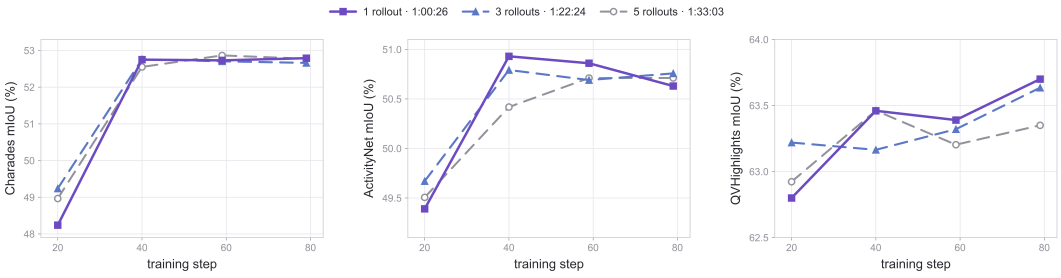

Figure 5: **Sensitivity to rollout count.** Benchmark mIoU during training with one, three, or five student rollouts per scheduled example. Total training times are 1:00:26, 1:22:24, and 1:33:03, respectively (hours:minutes:seconds).

B.13 Sensitivity to Rollout Count

We compare the default single-rollout assessment with three- and five-rollout variants using mean tIoU against the ground-truth interval. When this score falls below $\tau_S = 0.7$, only the first sampled trajectory receives teacher scoring and contributes to the OPD loss. Figure 5 reports benchmark mIoU across training checkpoints.

Additional rollouts provide modest gains at some checkpoints, including slightly higher endpoint mIoU on ActivityNet-TimeLens, but yield no consistent advantage across benchmarks. Training time increases with the rollout count. These results support single-rollout assessment as a practical choice, maintaining comparable overall TVG performance with shorter training time under the evaluated settings.

B.14 General Video Understanding

We further evaluate broader video understanding on TempCompass (Liu et al., 2024), MVBench (Li et al., 2024), and Video-MME (Fu et al., 2025), using accuracy as the evaluation metric. All post-training runs and evaluations use our local pipeline. The baselines follow the objective definitions in Video-OPD (Li et al., 2026a) and are trained on the same AF curriculum.

As shown in Table 17, SCC achieves the highest accuracy on all three benchmarks. Although the gains are modest, their consistency indicates that the improvements in temporal grounding do not compromise broader video-understanding capabilities.

Table 17: **Evaluation on broader video-understanding benchmarks.** Values are accuracy percentages, and bold indicates the best result in each column.

Method	TempCompass	MVBench	Video-MME
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	73.16	68.17	67.96
OP-RKD (Li et al., 2026a)	73.16	68.73	67.52
OP-FKD (Li et al., 2026a)	73.29	68.15	67.33
GRPO (Li et al., 2026a)	73.04	68.95	67.63
SCC	**73.35**	**69.08**	**68.56**