CV-OPSD: Self-Evolving Long-Context Reasoners via Zero-Annotation Clean-View On-Policy Self-Distillation

Anonymous authors
Paper under double-blind review

Abstract

Models often answer correctly when the relevant clip or span is presented in isolation, yet fail when the same evidence is embedded in a long input. We study this gap as a problem of evidence localization under bounded computation. Our method, *clean-view on-policy self-distillation* (CV-OPSD), enables a long-context reasoner to self-evolve by pairing a full-input student with a live, stop-gradient teacher instantiated from the same model. The student generates the rollout; the teacher evaluates the same prefixes from a privileged view that reveals where the evidence lies without providing an external answer. As the shared model improves, its clean-view posterior supplies an evolving learning signal without requiring a stronger teacher. In the fully self-generated setting, the examples and their evidence views are produced without human annotation. The view uses a question-relevant temporal interval for video or marked supporting spans in the identical document. We formalize why such views can increase information accessible to a bounded predictor even when they do not add answer information. Across 9B and 4B backbones, CV-OPSD improves performance across video and long-text evaluations, with gains of up to 4.8 and 5.0 points, respectively. Controlled comparisons with hard SFT, same-view voting, and alternative teacher inputs show that the useful signal comes from transferring an evidence-conditioned posterior on states visited by the full-input policy.

1 Introduction

Foundation models now support strong reasoning, multimodal understanding, and increasingly long contexts (Kimi Team, 2026; Team et al., 2026b;a). These advances make larger inputs available to models, but context capacity alone does not ensure that the relevant evidence will be found and used. Long-context models frequently fail on examples they can solve once the relevant evidence is isolated. A short interval may answer a video question that the full clip does not, while a fact chain may be easy to follow until it is placed among thousands of distractor tokens. Related failures appear in long-context QA (Bai et al., 2024b; Zhang et al., 2024b) and controlled retrieval settings (Hsieh et al., 2024; Maharana et al., 2024).

Existing approaches to long-context modeling largely focus on improved architectures (Ding et al., 2023), positional encodings (Chen et al., 2023; Ding et al., 2024), or additional long-context training data (An et al., 2024; Bai et al., 2024a). However, these settings are often determined during pretraining and require costly large-scale training, while data-centric approaches rely on expensive human annotations or trajectories synthesized by stronger models, limiting their scalability. An ideal method should instead efficiently improve evidence localization within long contexts without external supervision or a stronger teacher.

We introduce *Clean-View On-Policy Self-Distillation* (CV-OPSD). For each training example, the student receives the original input X, while a same-weight teacher receives a privileged view $C_Z(X)$ that exposes the location of relevant evidence. As illustrated in the right panel of Figure 1, a forward KL transfers the teacher distribution on student-generated prefixes while the teacher's advantage comes entirely from the input view. The teacher is instantiated from the updated student at every

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.

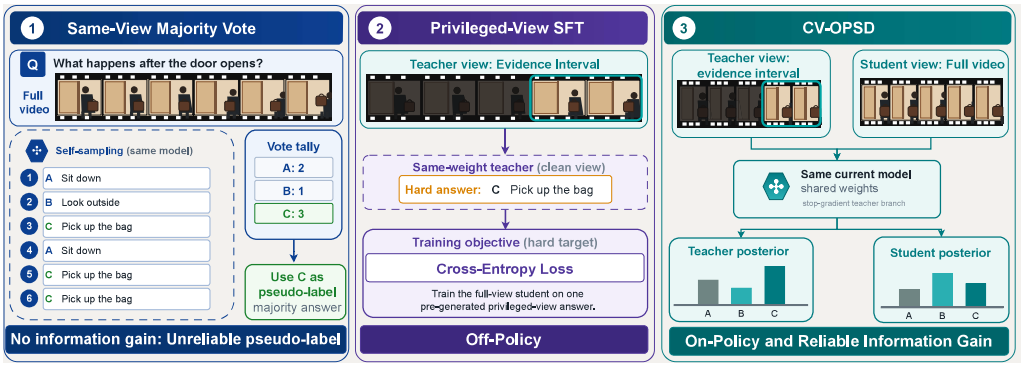

Figure 1: Three sources of self-training signal. Majority voting estimates a mode of the model's existing full-view distribution. SFT converts the privileged teacher into one off-policy hard answer. CV-OPSD evaluates the privileged posterior on student-generated prefixes.

step, allowing its privileged posterior to evolve with the full-context policy. We refer to this same-weight feedback process as *self-evolution*.

Empirically, CV-OPSD improves long-context reasoning across video and text and across both Qwen3.5-9B and Qwen3-VL-4B-Thinking (Yang et al., 2025). On video, it gains 4.8 points on LVBench (Wang et al., 2025) with human-annotated evidence intervals, while fully self-produced intervals retain a 2.8-point gain. Improvements on Video-MME (Fu et al., 2025) and the out-of-distribution MVBench (Li et al., 2024) show that the learned behavior transfers across video benchmarks. On text, pure privileged-view distillation improves BABILong (Kuratov et al., 2024), AA-LCR (Artificial Analysis, 2025), and LongBench v2 (Bai et al., 2024c), including gains of 8.9 points on BABILong and 5.0 points on AA-LCR for the 4B model.

Figure 1 contrasts CV-OPSD with two common forms of self-training supervision. Majority voting repeatedly samples from the same full input without introducing any additional source of information. As a result, it cannot reliably correct confident mistakes already present in the model, nor can it provide useful supervision on problems far beyond the model's current capability. Moreover, many questions do not yield a stable majority signal at all, which further reduces training efficiency. Privileged-view SFT addresses the information bottleneck by giving the teacher cleaner evidence, but remains strongly off-policy. The student is trained on pre-generated target trajectories whose prefixes it may not naturally produce when reasoning from the original, non-privileged input, limiting how effectively the supervision transfers to the actual inference process. CV-OPSD combines privileged information with on-policy distillation. The teacher receives a cleaner context and thus provides a more reliable learning signal, while supervision is applied to prefixes generated by the student's current full-view policy. This keeps training focused on the states the student actually encounters and makes the privileged supervision more effective for native reasoning.

Our theoretical analysis treats long-context prediction as evidence selection followed by conditional reasoning. It distinguishes raw information from information accessible to a bounded predictor, explaining how a marker or index can improve the teacher posterior without adding new facts. The teacher supplies computational supervision from evidence already present in the input, without requiring a stronger model or an external answer.

Our contributions are:

- We propose a zero-annotation, same-weight on-policy distillation method in which a live clean-view posterior evolves with the full-context policy, enabling self-evolution without a stronger teacher or privileged inputs at inference.
- We evaluate the method across video and long text, including a length-preserving textual view, cross-model replication, out-of-distribution tests, and controlled comparisons with majority voting, hard SFT, and answer-conditioned teaching.
- We formulate evidence localization as the bottleneck in long-context prediction and analyze privileged views through evidence sufficiency and information accessible to bounded predictors.

2 Privileged Evidence as Computational Supervision

The central premise of CV-OPSD is that supervision can change a model's computation without changing the answer it must produce. We develop this premise in three stages. First, long-context prediction is written as latent evidence selection followed by conditional reasoning. Second, privileged views are characterized by the information they make accessible to a bounded predictor rather than by raw context length. Third, this perspective separates evidence-conditioned teaching from same-view consensus.

2.1 Long-context prediction as latent evidence selection

Let Q be a question, $X = (x_1, \ldots, x_L)$ a long input, Y its answer, and $Z \in \mathcal{Z}$ a latent selector. The selected evidence is $E = X_Z$ and the remaining content is nuisance $N = X_{\bar{Z}}$. We assume the answer is approximately conditionally independent of nuisance once evidence is known.

Assumption 1 (Approximate evidence sufficiency)**.** *There exists Z such that*

$$I(Y; X \mid E, Q) \le \epsilon_{\text{suf}}, \qquad E = X_Z. \tag{1}$$

An ideal predictor marginalizes over possible evidence locations,

$$p(Y \mid X, Q) = \sum_{z \in \mathcal{Z}} p(Y \mid X_z, Q, z) p(z \mid X, Q). \tag{2}$$

The two factors encode different capabilities. The selector $p(z \mid X, Q)$ allocates computation to evidence, whereas $p(Y \mid X_z, Q, z)$ performs the task once that evidence is available. A privileged view $V = C_Z(X)$ intervenes on the former by making Z explicit through temporal selection or span marking. The teacher is consequently asked to solve the same prediction problem under a better-conditioned search process.

Suppose a bounded model selects the correct evidence with probability $1 - \delta(X)$ and has conditional answer loss at most ℓ_E after correct localization, while a miss incurs at most $\ell_{\max}$. Then

$$\mathbb{E}[\ell(Y, f(X, Q))] \le \ell_E + \mathbb{E}[\delta(X)]\ell_{\max} + O(\epsilon_{\text{suf}}). \tag{3}$$

The privileged teacher approximately suppresses the localization-dependent term. Its posterior therefore serves as a dense surrogate for how the full-view predictor should behave after successful evidence allocation. Distillation transfers this behavior without adding Z to the deployed input.

2.2 Accessible information under bounded computation

If V is a deterministic compression of X, ordinary Shannon information obeys $I(Y; V) \le I(Y; X)$. When the view includes an additional selector,

$$I(Y; X, Z \mid Q) = I(Y; X \mid Q) + I(Y; Z \mid X, Q), \tag{4}$$

and the final term measures genuine privileged information. This account is incomplete, however, for length-preserving views and deterministic self-produced selectors. In those cases the raw facts need not change at all; what changes is the ease with which a restricted predictor can exploit them. We capture this distinction through information accessible to an inference family $\mathcal{F}$:

$$I_{\mathcal{F}}(Y; U \mid Q) = H(Y \mid Q) - \inf_{f \in \mathcal{F}} \mathbb{E}[-\log f(Y \mid U, Q)]\,. \tag{5}$$

Accordingly, a bounded transformer may extract more accessible information from $C_Z(X)$ than from X, even when the view preserves every input element. The privileged view changes the computation induced by the input rather than the underlying facts.

Attention dilution provides one concrete mechanism. Let a_i denote normalized attention allocated to position i, and define evidence mass $A_E = \sum_{i \in Z} a_i$. Before localization, the expected mass on evidence can scale as $|Z|/L$. A marker introduces a salience bias $b > 0$:

$$A'_E = \frac{e^b \sum_{i \in Z} e^{s_i}}{e^b \sum_{i \in Z} e^{s_i} + \sum_{j \notin Z} e^{s_j}} \ge A_E. \tag{6}$$

Temporal selection additionally removes nuisance terms from the denominator. Both modality-specific operators increase effective evidence bandwidth despite making different changes to raw input length.

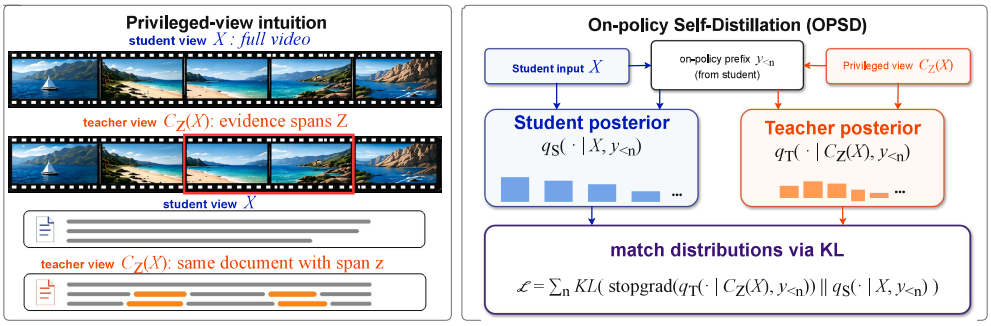

Figure 2: CV-OPSD for long video and long text. The video teacher receives the question-relevant temporal interval, while the text teacher receives the identical document with supporting spans marked. A current same-weight teacher under stop-gradient scores the student's rollout, and only the full-view student is retained for inference.

2.3 Why consensus is not evidence supervision

Let $\hat{Y}_{1:K} \sim q_\theta(\cdot \mid X, Q)$ be independent model samples. Conditioned on (X, Q, θ), their generation does not access the unknown ground-truth answer, so

$$I(Y; \hat{Y}_{1:K} \mid X, Q, \theta) = 0. \tag{7}$$

Voting can reduce sampling variance when the correct answer already has substantial probability, but it cannot reveal evidence that the underlying conditional computation repeatedly misses. It should therefore be most reliable on already-consistent examples and least informative precisely where localization supervision is needed. Our probe follows this pattern. Only 37% of items achieve a $\geq 50\%$ majority among eight samples, while the privileged teacher remains 42.9%-accurate on the no-majority subset.

This yields the conceptual distinction used in this work. A selector may contribute genuinely additional information, as in Eq. 4, or may reorganize existing information but becomes computationally accessible, as in Eq. 5. Both provide a training asymmetry that same-view resampling lacks.

3 Clean-View On-Policy Self-Distillation

3.1 Constructing paired full and privileged views

Each training example provides a full input X, a question Q, and a selector Z used only to construct the privileged view $C_Z(X)$. Student and teacher share the same model parameters and answer the same question. Their only difference is evidence access. The student receives the deployment input, whereas the teacher receives a view in which the relevant location is easier to identify. Figure 2 shows the video and text constructions used in our experiments.

Temporal evidence. For long video, the student processes the full clip, while the teacher receives only the evidence interval associated with the question. We study both human-annotated evidence intervals used to construct questions (Yu et al., 2025) and intervals produced during self-synthesis.

Textual evidence. For long documents, we generate multi-fact question–answer chains and insert their facts into PG19 passages (Rae et al., 2020). The student receives the resulting document without annotations. The teacher receives the same document content and all of its distractors, with each supporting span enclosed by evidence markers and a short instruction explaining the markers. No distractor is removed, so the two branches differ only in salience.

Across modalities, $C_Z(X)$ is required to expose a location rather than an answer. This constraint keeps the teacher on the original task and distinguishes privileged evidence from the answer-conditioned shortcut evaluated in Section 6.

3.2 On-policy privileged-posterior transfer

The student samples a trajectory $\hat{y} \sim q_\theta(\cdot \mid X, Q)$. At every visited prefix $\hat{y}_{<t}$, the full and privileged views score the same next-token vocabulary. We minimize the forward KL from the privileged distribution to the full-view distribution,

$$\mathcal{L}_{\mathrm{FKL}}(\theta) = \mathbb{E}_{\hat{y}\sim q_\theta(\cdot|X,Q)} \sum_t \mathrm{KL}[\mathrm{sg}\, q_\theta(\cdot \mid C_Z(X), Q, \hat{y}_{<t}) \,\|\, q_\theta(\cdot \mid X, Q, \hat{y}_{<t})], \tag{8}$$

where sg prevents gradients from entering the teacher branch. Both distributions are evaluated on prefixes generated from the full input. This removes the prefix mismatch of distilling a privileged-view sequence and retains the teacher's uncertainty instead of reducing it to one decoded answer.

The privileged posterior uses the student's current parameter values under no-gradient evaluation. There is no separate teacher optimization or moving-average update in the executed training path.

Equation 8 is the complete objective for both video and long-text experiments. No sequence-level reward, answer label, or reference-policy regularizer is used. The objective therefore isolates privileged-posterior transfer from reward optimization and hard-answer imitation.

3.3 Co-evolving posteriors and stability

Because the teacher uses current student parameters, its targets evolve during training. Updating the shared model allows the privileged posterior to improve alongside the full-context policy. We select checkpoints on held-out data and compare frozen and regularized alternatives in Section 6.1.

4 Experimental Setup

4.1 Training configurations

Models. Qwen3.5-9B is evaluated across video and long-text reasoning. Qwen3-VL-4B-Thinking provides a cross-architecture replication in both settings. Every run starts from the corresponding public base checkpoint. The vision tower is frozen, and the deployed architecture is unchanged. Training uses four student rollouts per prompt and a top-k approximation to the forward KL; optimizer and sequence settings are given in Appendix A.

Video data. The annotation-based construction uses questions grounded in human-annotated evidence intervals (Yu et al., 2025) to form paired views. The student receives the full video, while the teacher receives the union of question-specific evidence intervals. A fully self-generated variant instead constructs its own questions and evidence intervals. Evaluation videos are not included in either training pool.

Long-text data. We synthesize approximately 1.2K bAbI-style multi-fact chains (Weston et al., 2015), with 12–25 facts and short answers, and plant the facts at separated positions in PG19 passages. The student sees the unmarked passage; the teacher sees the same passage with all supporting facts marked. The carrier passages used for training are separated from those used for evaluation.

4.2 Evaluation protocol and comparisons

Benchmarks. Video evaluation uses LVBench (Wang et al., 2025) and Video-MME (Fu et al., 2025); MVBench (Li et al., 2024) provides the OOD video evaluation because it focuses on short videos. Text evaluation uses BABILong (Kuratov et al., 2024), AA-LCR (Artificial Analysis, 2025), and LongBench v2 (Bai et al., 2024c). AA-LCR contains 100 human-authored open-answer questions over document sets averaging roughly 100K tokens; each item requires multi-document retrieval and multi-step reasoning. LongBench v2 contains 503 multiple-choice questions with contexts ranging from 8K to 2M words and covers six realistic categories, including document QA, long in-context learning, dialogue history, code repositories, and structured data.

Scoring and controls. BABILong uses deterministic answer matching. Video-MME, MVBench, and LongBench v2 use their official multiple-choice answer extraction; LVBench uses a fixed LLM

judge; and AA-LCR LLM judging against its official open answers. Base and trained checkpoints use the same evaluation protocol. The main comparison is paired against the initialization. Privileged-view SFT ablates the value of on-policy distributional supervision by replacing it with hard off-policy targets. We also compare against same-view voting for self-consistency-based filtering and answer-conditioned teaching for solution-side privileged distillation.

5 Main Results

Evidence conditioning transfers across video model scales. On LVBench, the 9B model improves from 43.8 to 48.6 when the privileged view uses human-annotated evidence intervals associated with the questions (Table 1). Fully self-produced intervals reach 46.2, retaining a 2.4-point gain without human localization metadata. The self-produced construction also transfers to Qwen3-VL-4B-Thinking, raising LVBench accuracy from 44.7 to 46.2. These gains support the premise that a model can learn from its own evidence-conditioned predictions. Improving the teacher's access to relevant events provides useful supervision even when teacher and student share the same weights.

Table 1: Main results on long-video and long-text benchmarks (accuracy, %). **Bold** marks CV-OPSD; a dash denotes an unevaluated model–setting combination.

(a) Long video

		Qwen3.5-9B			Qwen3-VL-4B-Thinking		
Benchmark	Clean-view source	Base	**Ours**	Δ	Base	**Ours**	Δ
LVBench	Human-annotated intervals	43.8	**48.6**	+4.8	–	–	–
LVBench	Self-produced intervals	43.8	**46.2**	+2.4	44.7	**46.2**	+1.5
Video-MME	Self-produced intervals	71.6	**73.2**	+1.6	62.9	**64.8**	+1.9

(b) Long text

		Qwen3.5-9B			Qwen3-VL-4B-Thinking		
Benchmark	Decoding setting	Base	**Ours**	Δ	Base	**Ours**	Δ
BABILong	No thinking	61.9	**64.7**	+2.8	–	–	–
BABILong	Thinking	56.1	**64.9**	+8.8	47.8	**56.7**	+8.9
AA-LCR	Thinking	63.0	**66.0**	+3.0	14.0	**19.0**	+5.0
LongBench v2	Thinking	53.0	**54.9**	+1.9	33.0	**36.8**	+3.8

The self-produced construction uses only 579 examples, compared with 6,015 in the human-annotated construction. Its positive results on both backbones show that useful clean-view supervision can be obtained from a small, model-generated training pool. Questions and their evidence intervals are produced together, giving the teacher a localized view without requiring a human to identify the relevant event for each training example. This makes the zero-annotation setting practical while preserving the same full-video inference task.

On Video-MME, Qwen3.5-9B improves from 71.6 to 73.2, and Qwen3-VL-4B-Thinking improves from 62.9 to 64.8. For Qwen3-VL-4B-Thinking, accuracy improves by 1.9 points on short videos, 3.1 on medium videos, and 0.7 on long videos. Improvement across these groups shows that the learned behavior transfers to videos of different durations. Together with LVBench, these results support evidence-conditioned training as a way to improve the full-video policy beyond the questions used to construct its clean views.

Evidence supervision improves native long-text reasoning. On BABILong, pure privileged-view distillation raises Qwen3.5-9B accuracy from 61.9 to 64.7 in no-thinking mode and from 56.1 to 64.9 in thinking mode. Qwen3-VL-4B-Thinking improves from 47.8 to 56.7, a similar gain of 8.9 points. The text teacher receives the identical document with supporting spans marked, including all distractors. Improvement under this length-preserving construction supports our account of computational supervision. Identifying where to look makes the same model a useful teacher, even when the input facts and context length are preserved.

The 9B thinking gain of 8.8 points brings its final accuracy to 64.9, close to the 64.7 achieved without thinking. Both modes improve from the same initialization, showing that evidence-conditioned supervision benefits explicit reasoning as well as direct answer prediction. The similar thinking gain

on 4B supports this result across model scales. In each case, the student learns from the teacher's marked view and answers from the original, unmarked passage at inference.

Separated evidence remains useful across context lengths. The largest task gains in the 9B BABILong breakdown occur on qa3 (+5.0) and qa5 (+4.6), which require multi-fact state tracking, as detailed in Table 4. Four of the five tasks improve. This pattern supports the use of marked supporting facts as supervision for composing evidence distributed across a document. The student must recover those facts from the unmarked passage, so the teacher's location cues are transferred into a behavior that remains useful after the cues are removed.

Averaging over the five tasks, the gain is 0.6 points at 0K and between 3.6 and 5.4 points from 4K through 64K. The larger improvements in the presence of distractors are consistent with the evidence-selection bottleneck described in Section 2. Gains at intermediate lengths also show that the benefit persists across the tested context range. At 128K, which is beyond the training regime, accuracy still rises from 47.0 to 50.4. The learned use of supporting evidence continues to help as the surrounding context grows even under out-of-distribution setting.

Synthetic evidence supervision transfers to real documents. AA-LCR provides a substantially longer, real-document test. The 9B model improves from 63.0 to 66.0, while 4B improves from 14.0 to 19.0. Its questions require reasoning across document sets averaging roughly 100K tokens. LongBench v2 shows consistent transfer across scale, with 9B improving from 53.0 to 54.9 and 4B from 33.0 to 36.8. These results follow training on approximately 1.2K self-produced multi-fact chains. They show that supervision constructed from synthetic evidence dependencies can improve reasoning over real documents without using their benchmark questions or answers during training.

Transfer also extends across answer formats. AA-LCR requires open answers, whereas LongBench v2 uses multiple-choice questions and covers a broader collection of long-context tasks. Improvement on both benchmarks supports learning how to use relevant evidence beyond the short-answer format used during synthesis. The gains on the two backbones further show that this form of self-distillation remains useful at different levels of initial capability.

Overall. Video isolates temporal support, while text preserves distractors and adds a location cue. Gains under both constructions support evidence access as a source of same-weight supervision that improves the full-context policy at inference.

6 Analysis and Ablations

6.1 Does the teacher actually self-evolve?

CV-OPSD uses the updated student as its next teacher. To test whether this feedback improves learning, we compare it with a teacher frozen at initialization. As shown in Figure 3, the live-teacher run reaches 48.6 on LVBench, 5.4 points above the frozen-teacher run. Regularized variants also peak below the live teacher. Screening accuracy rises steadily from 58.33 to 61.11 over training. Together, the frozen-teacher comparison and the training trajectory support self-evolution through continued teacher updates as the full-context policy improves.

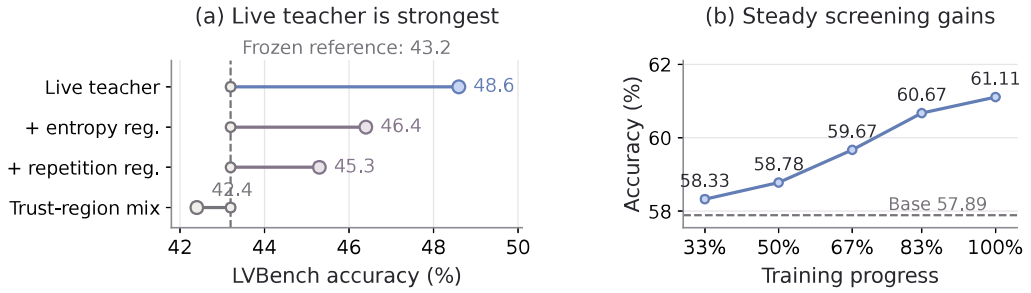

Figure 3: An evolving teacher improves self-distillation. (a) The live teacher outperforms the frozen teacher by 5.4 points and exceeds all regularized variants. (b) Screening accuracy rises throughout training, from 58.33 to 61.11 with the 57.89% base shown as a dashed line.

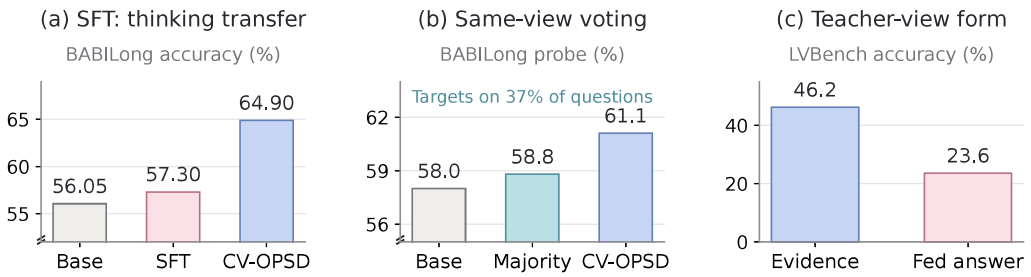

Figure 4: Baseline and teacher-view ablations. On-policy soft targets transfer to native thinking more effectively than hard SFT; majority voting provides sparse supervision; and answer-conditioned teaching is unstable on video.

6.2 What makes clean-view supervision effective?

Off-policy SFT. We greedily decode the same privileged teacher into hard-answer pairs and train for matched exposure. Results shows that SFT transfers weakly to native thinking (57.30 versus 64.90 for CV-OPSD). This 7.6-point huge gap supports transferring soft targets on student prefixes rather than imitating one privileged-view sequence.

Majority voting. Eight-sample same-view voting improves the matched BABILong probe only from 58.0 to 58.8, compared with 61.1 for CV-OPSD. A majority exists for only 37% of questions, leaving difficult, high-disagreement examples without targets even though the privileged-view predictor remains 42.9%-accurate. These results reveal that majority voting without information gain does not lead to a reliable model improvements, which align with our previous analysis.

Evidence versus fed answer. Holding the model, data, and optimization fixed, an evidence-conditioned teacher reaches 46.2 on LVBench, whereas feeding the model's own answer as privileged information reduces accuracy to 23.6 and induces repetitive hedging once the hint is removed. To understand this gap, we examine where the token-level KL is concentrated during training. Answer-conditioned OPSD produces its largest teacher–student discrepancies on answer-bearing tokens, teaching the policy to exploit a train-only hint. Because CV-OPSD never reveals the answer, its discrepancies instead arise while the rollout identifies and reasons from relevant events, directing updates toward evidence-grounded inference.

6.3 Does CV-OPSD improve evidence localization?

We test whether CV-OPSD improves long-video accuracy primarily through evidence localization using the two complementary diagnostics shown in Figure 5. On Charades-STA (Gao et al., 2017), the 4B temporal-grounding score rises from 51.93 to 54.48. We then remove evidence search by supplying both models with the same correct interval on LVBench. Their accuracies converge to 67.49 and 67.81, compared with 44.7 and 46.2 on full videos. The advantage contracts from 1.50 to a merely 0.32 points when localization is solved for both models.

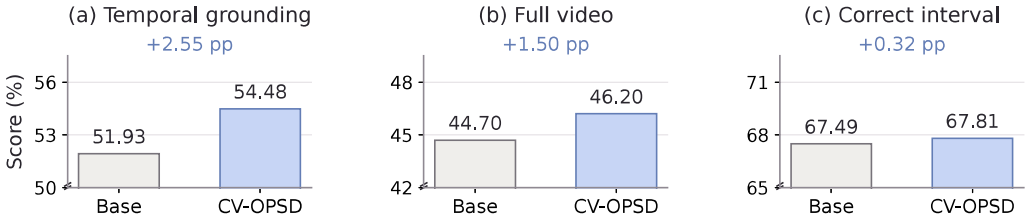

Figure 5: Localization accounts for most of the 4B gain. (a) Charades-STA improves by 2.55 points. (b,c) The LVBench advantage shrinks from 1.50 to 0.32 points with correct intervals. These panels use equal axis spans; slashes mark nonzero baselines.

CV-OPSD is better when asked to localize evidence directly, while its answer advantage nearly disappears when both models receive the correct interval. Together, these diagnostics support the conclusion that most of the full-video gain comes from improved evidence localization.

6.4 Does self-evolution generalize beyond the training regime?

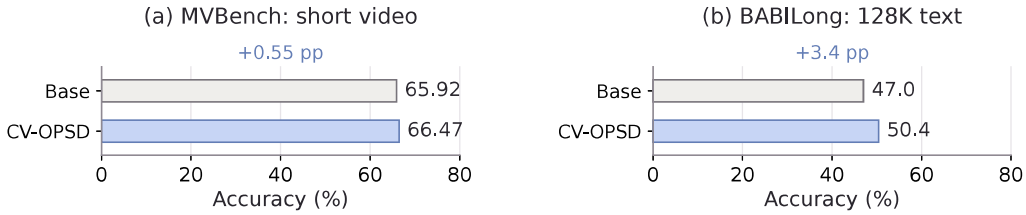

Figure 6: Gains persist across context-regime shifts, with +0.55 points on short-video MVBench after long-video training and +3.4 points on BABILong at 128K, beyond the training context length.

The video model is trained on long videos but improves from 65.92 to 66.47 on MVBench, a short-video benchmark. The text model improves from 47.0 to 50.4 at 128K, beyond its training context length. As shown in Figure 6, both gains persist outside the training regime, proving the strong generalization ability of our method. More details are given in Appendix A.

7 Related Work

On-policy distillation and self-distillation. Classical distillation supervises on teacher rollouts, whereas OPD supervises student rollouts (Hinton et al., 2015; Vapnik & Vashist, 2009; Agarwal et al., 2024). Variants study divergence design (Gu et al., 2024; Ko et al., 2024), black-box and multimodal transfer (Ye et al., 2025; Bousselham et al., 2025), or entropy, representations, and peer rollouts (Jin et al., 2026; Yang et al., 2026a; Yu et al., 2026). The limitation is unchanged: capability comes from a separate, stronger teacher with compatible reasoning patterns (Li et al., 2026).

OPSD reuses the same weights (Zhao et al., 2026), using translations or target regions (Liu et al., 2026; Zhang et al., 2026), uncertainty or filtering (Ke et al., 2026; Shen et al., 2026), contrastive hints in RLCSD (Pan et al., 2026), or agentic rewards (Lu et al., 2026; Yang et al., 2026b). It removes the stronger model but still typically requires privileged solution-side information.

Self-evolution. Voting and filtering turn model outputs into supervision. Self-consistency and STaR-style methods aggregate or retain sampled traces (Wang et al., 2023; Zelikman et al., 2022; Huang et al., 2023); ReST and Video-STaR filter trajectories (Gulcehre et al., 2023; Singh et al., 2023; Zohar et al., 2024). Quiet-STaR internalizes latent reasoning (Zelikman et al., 2024); ReST-MCTS, V-STaR, and rStar-Math add search or verification (Zhang et al., 2024a; Hosseini et al., 2024; Guan et al., 2025). Confidence and semantic voting provide cheaper filters (Jang et al., 2025; Jiang et al., 2025). Self-reward and model judges provide generated evaluations (Yuan et al., 2024; Wu et al., 2024a; Kumar et al., 2024); related self-training systems use the same principle (Wu et al., 2026; Wang et al., 2024). Their weakness is that selection cannot add evidence absent from the samples, and may collapse onto confident errors (Shafayat et al., 2025).

Adversarial methods use preference self-play (Chen et al., 2024; Wu et al., 2024b), competing critics (Cheng et al., 2024; Chen et al., 2025), or self-proposed tasks (Liu et al., 2025; Zhao et al., 2025). Without informative external consequences, the game may merely sharpen the output distribution(Zweiger et al., 2025; He et al., 2026).

8 Limitations and Conclusion

A remaining limitation is that the video setting still requires a source video corpus and a procedure for constructing candidate evidence intervals. It is therefore zero-annotation, but not entirely free of external data dependencies in the strictest sense.

Overall, CV-OPSD uses an evidence-conditioned posterior from the same evolving weights to improve full-context evidence localization and reasoning, without a stronger teacher or human answer labels. Experiments confirm its effectiveness and generalization.

AI USE STATEMENT

Generative AI tools were used to produce the self-synthesized training data described in the paper and to assist with language editing and formatting of the manuscript. The authors reviewed the AI-assisted material, verified the reported claims and results, and take responsibility for the final content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. Make your llm fully utilize the context, 2024. URL https://arxiv.org/abs/2404.16811.

Artificial Analysis. AA-LCR: Artificial analysis long context reasoning. https://artificialanalysis.ai/articles/announcing-aa-lcr, 2025.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models, 2024a. URL https://arxiv.org/abs/2401.18058.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. In *Annual Meeting of the Association for Computational Linguistics*, 2024b.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024c.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation, 2025. URL https://arxiv.org/abs/2510.23497.

Jiaqi Chen, Bang Zhang, Ruotian Ma, Peisong Wang, Xiaodan Liang, Zhaopeng Tu, Xiaolong Li, and Kwan-Yee K. Wong. Spc: Evolving self-play critic via adversarial games for llm reasoning, 2025. URL https://arxiv.org/abs/2504.19162.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models, 2024. URL https://arxiv.org/abs/2401.01335.

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Zheng Yuan, Yong Dai, Lei Han, Nan Du, and Xiaolong Li. Self-playing adversarial language game enhances llm reasoning, 2024. URL https://arxiv.org/abs/2404.10642.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens, 2023. URL https://arxiv.org/abs/2307.02486.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Chaoyou Fu et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Conference on Computer Vision and Pattern Recognition*, 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*, 2017.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models, 2024. URL https://arxiv.org/abs/2306.08543.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking, 2025. URL https://arxiv.org/abs/2501.04519.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling, 2023. URL https://arxiv.org/abs/2308.08998.

Bingxiang He, Yuxin Zuo, Zeyuan Liu, Shangziqi Zhao, Zixuan Fu, Junlin Yang, Cheng Qian, Kaiyan Zhang, Yuchen Fan, Ganqu Cui, Xiusi Chen, Youbang Sun, Xingtai Lv, Xuekai Zhu, Li Sheng, Ran Li, Huan ang Gao, Yuchen Zhang, Bowen Zhou, Zhiyuan Liu, and Ning Ding. How far can unsupervised rlvr scale llm training?, 2026. URL https://arxiv.org/abs/2603.08660.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners, 2024. URL https://arxiv.org/abs/2402.06457.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models?, 2024. URL https://arxiv.org/abs/2404.06654.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve, 2023. URL https://arxiv.org/abs/2210.11610.

Hyosoon Jang, Yunhui Jang, Sungjae Lee, Jungseul Ok, and Sungsoo Ahn. Self-training large language models with confident reasoning, 2025. URL https://arxiv.org/abs/2505.17454.

Chunyang Jiang, Yonggang Zhang, Yiyang Cai, Chi-Min Chan, Yulong Liu, Mingming Chen, Wei Xue, and Yike Guo. Semantic voting: A self-evaluation-free approach for efficient llm self-improvement on unverifiable open-ended tasks, 2025. URL https://arxiv.org/abs/2509.23067.

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, 2026. URL https://arxiv.org/abs/2603.07079.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning, 2026. URL https://arxiv.org/abs/2605.13255.

Kimi Team. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models, 2024. URL https://arxiv.org/abs/2402.03898.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning, 2024. URL https://arxiv.org/abs/2409.12917.

Yuri Kuratov et al. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, 2024.

Kunchang Li et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *Conference on Computer Vision and Pattern Recognition*, 2024.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026. URL https://arxiv.org/abs/2604.13016.

Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, Wee Sun Lee, and Natasha Jaques. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning, 2025. URL https://arxiv.org/abs/2506.24119.

Yihong Liu, Raoyuan Zhao, Michael A. Hedderich, and Hinrich Schütze. Crosslingual on-policy self-distillation for multilingual reasoning, 2026. URL https://arxiv.org/abs/2605.09548.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Self-distilled agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2605.15155.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents, 2024. URL https://arxiv.org/abs/2402.17753.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation, 2026. URL https://arxiv.org/abs/2606.11709.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. Can large reasoning models self-train?, 2025. URL https://arxiv.org/abs/2505.21444.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified opsd: On-policy self-distillation without losing how to think, 2026. URL https://arxiv.org/abs/2607.02234.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models, 2023. URL https://arxiv.org/abs/2312.06585.

V Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, Xijun Liu, Wenmeng Yu, Weihan Wang, Wei Li, Shuaiqi Duan, Sheng Yang, Ruiliang Lv, Mingdao Liu, Lihang Pan, Ke Ning, Junhui Ji, Jinjiang Wang, Jing Chen, Jiazheng Xu, Jiale Zhu, Jiale Cheng, Ji Qi, Guobing Gan, Guo Wang, Cong Yao, Zijun Dou, Zihao Zhou, Zihan Wang, Zhiqi Ge, Zhijie Li, Zhenyu Hou, Zhao Xue, Zehui Wang, Zehan Qi, Zehai He, Yutao Zhang, Yusen Liu, Yukuo Cen, Yuchen Li, Yuan Wang, Yu Yang, Yongbin Liu, Yijian Lu, Yifan Xu, Yanzi Wang, Yanxiao Zhao, Yanfeng Wang, Yadong Xue, Yabo Xu, Xinyu Zhang, Xinyu Liu, Xiao Liu, Wenyi Zhao, Wenkai Li, Tianyu Tong, Tianshu Zhang, Shudan

Zhang, Shengdong Yan, Qinkai Zheng, Mingde Xu, Licheng Bao, lat Long long, Jiaxing Xu, Jiaxin Fan, Jiawen Qian, Jiali Chen, Jiahui Lin, Jiadai Sun, Haozhi Zheng, Haoran Wang, Haochen Li, Hanyu Lai, Han Xu, Fan Yang, Dan Zhang, Da Yin, Chuangxin Zhao, Chengcheng Wu, Boyan Shi, Bowen Lv, Bowei Jia, Bo Li, Bin Chen, Baoxu Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-5v-turbo: Toward a native foundation model for multimodal agents, 2026a. URL https://arxiv.org/abs/2604.26752.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026b. URL https://arxiv.org/abs/2507.01006.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. In *Neural Networks*, 2009.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught evaluators, 2024. URL https://arxiv.org/abs/2408.02666.

Weihan Wang et al. LVBench: An extreme long video understanding benchmark. In *International Conference on Computer Vision*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merri"enboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks, 2015.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge, 2024a. URL https://arxiv.org/abs/2407.19594.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment, 2024b. URL https://arxiv.org/abs/2405.00675.

Zhengxian Wu, Kai Shi, Chuanrui Zhang, Zirui Liao, Jun Yang, Ni Yang, Qiuying Peng, Luyuan Zhang, Hangrui Xu, Tianhuang Su, Zhenyu Yang, Haonan Lu, and Haoqian Wang. When models judge themselves: Unsupervised self-evolution for multimodal reasoning, 2026. URL https://arxiv.org/abs/2603.21289.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, and Gang Chen. Oprd: On-policy representation distillation, 2026a. URL https://arxiv.org/abs/2606.06021.

Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, and Jianhua Tao. Opid: On-policy skill distillation for agentic reinforcement learning, 2026b. URL https://arxiv.org/abs/2606.26790.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models, 2025. URL https://arxiv.org/abs/2511.10643.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, Zhenxiang Li, Zhongying Tu, Conghui He, Yu Qiao, Yali Wang, Yi Wang, and Limin Wang. VRBench: A benchmark for multi-step reasoning in long narrative videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, and Yu Hu. Multi-rollout on-policy distillation via peer successes and failures, 2026. URL https://arxiv.org/abs/2605.12652.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2024. URL https://arxiv.org/abs/2401.10020.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. Quiet-star: Language models can teach themselves to think before speaking, 2024. URL https://arxiv.org/abs/2403.09629.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search, 2024a. URL https://arxiv.org/abs/2406.03816.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞bench: Extending long context evaluation beyond 100k tokens, 2024b. URL https://arxiv.org/abs/2402.13718.

Yan Zhang, Daiqing Wu, Huawen Shen, Can Ma, and Yu Zhou. Learn where to click from yourself: On-policy self-distillation for gui grounding, 2026. URL https://arxiv.org/abs/2605.00642.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data, 2025. URL https://arxiv.org/abs/2505.03335.

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, 2026. URL https://arxiv.org/abs/2601.18734.

Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-Levy. Video-star: Self-training enables video instruction tuning with any supervision, 2024. URL https://arxiv.org/abs/2407.06189.

Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyürek, Yoon Kim, and Pulkit Agrawal. Self-adapting language models, 2025. URL https://arxiv.org/abs/2506.10943.