From Gradient Surgery to Bottleneck Regularization: A Unified Framework for Combining Verifiable-Reward RL and On-Policy Self-Distillation

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning with verifiable rewards (RLVR) gives reasoning models a clean signal about whether a rollout is correct, but says little about which tokens made it correct. On-policy self-distillation (OPSD) offers the opposite trade-off. An answer-conditioned copy of the model can provide dense token-level supervision, but that privileged teacher can also exploit information the student will not have at test time. Recent work further shows that teacher-student likelihood shifts need not track token value.

We study how to combine these signals without letting privileged supervision override the verifier. CIBO-CRISP, our primary method, uses token-level REINFORCE with a bounded, polarity-aware, sign-preserving credit multiplier, a softly gated teacher-student divergence penalty, and an exponential-moving-average stability anchor. CRISP and VACS-CRISP are simpler intermediate constructions that expose the design choices leading to the final objective.

On Qwen2.5-Math-1.5B, CIBO-CRISP averages 62.7% Pass@1 on MATH-500 across two seeds, compared with 44.6% for GRPO, while remaining close on GSM8K (74.1% versus 72.4%). In matched seed-0 ablations, the full method reaches 63.6% on MATH-500 versus 60.8% with fixed beta and 60.2% with the privileged-divergence term removed. The theoretical claims are deliberately narrower than the practical algorithm. This is because we give a standard fixed-surrogate nonconvex-SGD result and an algebraic leakage-bookkeeping identity, but do not treat the implemented KL term as a mutual-information estimator or teacher likelihood shifts as causal token values.

1. Introduction

RLVR is attractive because its reward is externally checkable, that is, for a mathematical problem, the final answer either passes the verifier or it does not. Its weakness is temporal resolution. A single scalar reward is broadcast across a long reasoning trace, so decisive steps and incidental tokens receive the same sequence-level advantage. Recent work has attacked this credit-assignment problem with entropy-based decompositions, counterfactual span attribution, and token-specific baselines [2, 7, 8].

OPSD approaches the same problem from another direction. The model is conditioned on privileged information unavailable at inference time (here, the ground-truth answer) and the resulting distribution supervises the unconditioned policy [11]. This provides dense token-level guidance, but it also creates a mismatch. The teacher can rely on information the student will not have at test time. Several 2026 studies report degradation, bias, or weak correspondence between privileged likelihood shifts and outcome-relevant token value [1, 3, 4, 5, 6, 22, 23]. Heakl et al., for example, find full-KL distillation baselines below the untrained model on their evaluation suite [6], while recent diagnostics caution against reading likelihood shifts as token credit [22, 23].

Combining RL and self-distillation introduces a second practical issue: both losses update the

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

same parameters, and their gradients can disagree. PCGrad and CAGrad study symmetric reconciliation between peer objectives [9, 10]. Primary/auxiliary asymmetry is also established. PAGrad protects the RL/PPO gradient and modifies only a conflicting auxiliary gradient [20], while Bloop develops a broader primary-auxiliary formulation with projected auxiliary updates and an EMA estimate [21]. We therefore do not present asymmetric projection itself as new. Our question is narrower; how should verifier-grounded RL be combined with a privileged self-teacher?

We approach that question in three steps:

- CRISP(**C**ritic-free **R**eward-**I**ntegrated **S**elf-distillation **P**olicy **O**ptimisation) combines globally normalized policy-gradient RL with correctness-gated self-distillation, reconciling the two gradients with a symmetric PCGrad-like projection. The gate is simply a conservative way to limit privileged supervision, we do not claim it is the unique solution adopted by prior work.
- VACS-CRISP(**V**ariance-**A**ware **C**ausal **S**urgery **CRISP)** moves the modulation to individual tokens. It uses a detached, sign-aware teacher/student evidence weight, replaces the hard gate with a sigmoid gate, and treats RL as the primary objective. The RL gradient is left untouched, while a conflicting distillation gradient is projected. The projection pattern follows prior asymmetric gradient-surgery work [20, 21]. The contribution here is its use with privileged self-distillation, token-level evidence, and adaptive mixing.
- CIBO-CRISP(**C**ausal **I**nformation **B**ottleneck **O**ptimization **CRISP**) is the primary method. It removes explicit gradient surgery and instead optimizes one bottleneck-regularized objective, i.e, a token-level REINFORCE with a bounded, polarity-aware, sign-preserving multiplier, a softly gated teacher-student divergence term, and an EMA stability anchor. For a fixed surrogate objective, this admits the standard nonconvex-SGD guarantee stated in Section 4. The separate leakage identity is only bookkeeping for an explicitly defined abstract leakage functional. It does not turn the implemented KL term into a conditional-mutual-information estimator.

Section 2 places these choices in context, Section 3 introduces notation, and Section 4 develops the three constructions and the limited theoretical results. Sections 5 and 6 describe the final 1.5B experiments and results. Section 7 discusses what these results do and do not establish.

2. Related Work

RL with verifiable rewards and token-level credit. Most RLVR implementations still attach one rollout-level advantage to every generated token, whether the advantage is globally batch-normalized, REINFORCE++-style, or normalized within a GRPO rollout group. He et al. derive an entropy-based limit on token-level credit information [2]. Khandoga et al. estimate span importance by counterfactual perturbation [7]. Li et al. derive a gradient-energy baseline for variance reduction [8]. These methods recover structure from the RL signal itself. VACS-CRISP and CIBO-CRISP instead use disagreement between a privileged teacher and the unconditioned student, which is convenient but less directly tied to value. Recent diagnostics reinforce that distinction [22, 23].

Self-distillation with privileged teachers. Zhao et al. formalize answer-conditioned OPSD, where a privileged copy of the model supervises the unconditioned policy [11]. RLSD uses a stop-gradient evidence ratio to modulate a verifier-anchored advantage [1], making it the closest published precedent to the sampled-token modulation used here. CEPO replaces a single privileged reference with contrastive accepted/rejected evidence [6]. DemoPSD distills toward a disagreement-modulated reverse-KL barycentric target [4]. SDPO broadens privileged conditioning to textual feedback [3]. SDPG combines group-relative advantages, on-policy distillation, reference-policy KL, and an advantage-sign gate [5]. Two recent diagnostics are especially relevant. He et al. show that privileged shifts can concentrate on substitutable surface tokens [22], while Nguyen et al. distinguish likelihood change from action value and report near-chance outcome discrimination in one AIME setting [23]. We therefore use teacher/student disagreement as a bounded heuristic signal, not as an identified token value.

Gradient conflict. PCGrad [10] motivates projecting conflicting task gradients, while CAGrad develops a symmetric conflict-averse construction with convergence theory [9]. CRISP uses a simultaneous two-gradient PC-style rule rather than the canonical sequential/random-order PCGrad implementation. For VACS-CRISP, the more relevant precedents are asymmetric. PAGrad protects the main RL/PPO gradient and projects only a conflicting auxiliary gradient [20], and Bloop formalizes a primary/auxiliary hierarchy with EMA-smoothed projection [21]. VACS-CRISP borrows that hierarchy but combines it with privileged-teacher token modulation, soft gating, and an adaptive mixing rule. Bloop's theorem does not directly cover this update.

Bottleneck regularization and mutual information. Variational MI bounds can be useful when the estimator is part of the model, but they also make claim direction easy to get wrong [12]. CLUB, for example, constructs a contrastive upper bound under an explicit probabilistic setup [13]. CIBO-CRISP does not build such an estimator. Its gated teacher-student KL is a practical divergence regularizer intended to limit privileged-teacher influence, not a certified information bottleneck. The optimization discussion therefore stays at the level of the surrogate loss and the general biased-SGD literature [16].

Mean teachers and KL direction. CIBO-CRISP's EMA anchor follows the weight-averaged Mean Teacher idea [24], with related uses in vision self-distillation [14] and LLM continual learning [15]. The anchor has a different role from the privileged teacher. It stabilizes the student's own trajectory over time. Separately, Yao et al. argue for reverse rather than forward KL in a related weak-to-strong setting [17]. We retain forward KL here for continuity with OPSD-style distillation, but treat the direction as an empirical design choice rather than settled doctrine.

Benchmark caveat. Wu et al. document contamination concerns for the Qwen2.5 family on mathematical benchmarks, including MATH-500/AIME-style evaluation [18]. Because our experiments use the same model family, we emphasize matched comparisons between training methods rather than interpreting absolute scores as uncontaminated measures of reasoning.

3. Preliminaries

Let π_θ be the policy, $x \sim D$ a prompt, $y \sim \pi_\theta(\cdot|x)$ a rollout of length T, and $R\left(y, y^*\right) \in \{0, 1\}$ a verifier reward against gold answer y^* (the framework extends to $[0, 1]$-valued rewards). For a batch of N rollouts with rewards $r_1, \ldots, r_N$, the global advantage is

$$\hat{A}_i = \frac{r_i - \mu_r}{\sigma_r + \varepsilon}, \qquad \mu_r = \frac{1}{N}\sum_j r_j, \quad \sigma_r = \sqrt{\frac{1}{N}\sum_j \left(r_j - \mu_r\right)^2}.$$

The *teacher* is the same weights θ conditioned on privileged context $\tilde{x} = \left(x, y^*\right)$: $p_T = \pi_\theta\left(\cdot|\tilde{x}, y_{<t}\right)$; the *student* is $p_S = \pi_\theta\left(\cdot|x, y_{<t}\right)$. Given a student rollout y, the per-token forward KL, teacher supervising student, is

$$D_t = \sum_v p_T(v)\left[\log p_T(v) - \log p_S(v)\right], \qquad p_T \; stop - gradiented.$$

A *correctness gate,* omitting or down-weighting D_t for rollouts the student already gets right, is a mechanism CRISP shares in spirit with several methods in Section 2's leakage discussion, implemented as a hard indicator (CRISP), a smooth sigmoid (VACS-CRISP, CIBO-CRISP), or an RL-advantage-sign criterion (SDPG [5]).

4. Method

CRISP and VACS-CRISP are best read as stepping stones to CIBO-CRISP rather than as three co-equal proposals. They expose three choices separately, i.e., sequence- versus token-level credit, hard versus soft gating, and explicit gradient surgery versus a single regularized

objective. Table 1 summarizes the progression.

Table 1.

	Credit	**Gate**	**Combination**
CRISP	sequence-level $\hat{A}_i$	hard ($r_i < 1$)	symmetric PC-Grad
VACS-CRISP	token-level, sign-aware exp. weight	soft (sigmoid)	asymmetric surgery
CIBO-CRISP	token-level, polarity-aware tanh multiplier	soft (sigmoid)	None, single objective

4.1 CRISP

CRISP uses a globally normalized policy-gradient objective with the same sequence-level advantage on every token,

$$L_{RL} = -\frac{1}{N}\sum_i \hat{A}_i \frac{1}{T_i}\sum_t log\pi_\theta\left(y_i^t | x_i, y_i^{<t}\right), \qquad L_{SD} = \frac{\sum_i g_i \sum_t D_t(i)}{max\left(\sum_i g_i 1\right)}, \quad g_i = 1\left[r_i < 1\right].$$

The RL term is plain policy gradient. There is no separate π_{old} policy or PPO clipping ratio in the implementation. A cosine-annealed $\lambda(t) = \lambda_{max} cos\left(\frac{\pi t}{2T}\right)$ (clipped to $\left[0, \lambda_{max}\right]$) weights the two terms, combined via PC-Grad: with $g_{RL} = \nabla_\theta L_{RL}$, $g_{SD} = \nabla_\theta L_{SD}$, if $g_{RL} \cdot g_{SD} < 0$ both are projected simultaneously from the original pair,

$$g_{RL}' = g_{RL} - \frac{g_{RL} \cdot g_{SD}}{\|g_{SD}\|^2} g_{SD}, \qquad g_{SD}' = g_{SD} - \frac{g_{RL} \cdot g_{SD}}{\|g_{RL}\|^2} g_{RL},$$

(unchanged otherwise), and the update direction is $g_{RL}' + \lambda(t) g_{SD}'$.

4.2 VACS-CRISP

Per token, $\Delta_t = sg\left[logp_T\left(y_t\right) - logp_S\left(y_t\right)\right]$ (both sides detached, the student's only gradient path is the explicit log-prob factor below), and

$$w_t = exp\left(sign\left(\hat{A}_i\right)\Delta_t\right) = \left(\frac{p_T\left(y_t\right)}{p_S\left(y_t\right)}\right)^{sign\left(\hat{A}_i\right)}, \quad w_t \leftarrow clip\left(w_t, 1 - \epsilon_w, 1 + \epsilon_w\right), \quad \widetilde{A}_t = \hat{A}_i\left[(1 - \rho) + \rho w_t\right],$$

This weighting has the intended polarity. Teacher-endorsed tokens are amplified on positive-advantage rollouts, but are penalized less strongly rather than boosted on negative-advantage rollouts. The RL term is plain per-token REINFORCE. There is no PPO ratio.

$$L_{RL} = -\frac{1}{N}\sum_i \frac{1}{T_i}\sum_t \widetilde{A}_t log\pi_\theta\left(y_i^t | x_i, y_i^{<t}\right).$$

The gate is soft, $\gamma_i = \sigma\left(\beta_{gate}\left(r_i - r_{th}\right)\right)$, $L_{SD,i} = \left(1 - \gamma_i\right)\sum_t D_t(i)$. Surgery is primary/auxiliary asymmetric, i.e., g_{RL} is never modified, only g_{SD} is projected when conflicting: This pattern has prior art in PAGrad and related primary/auxiliary projection methods [20, 21]; the distinction here is the privileged-teacher signal and the adaptive weighting around it.

$$\widetilde{g}_{SD} = g_{SD} - 1\left[g_{RL} \cdot g_{SD} < 0\right] \frac{g_{RL} \cdot g_{SD}}{\|g_{RL}\|^2} g_{RL}, \qquad g_{combined} = g_{RL} + \lambda_{eff} \widetilde{g}_{SD},$$

with $\lambda_{eff} = \lambda_{base} \cdot clip\left(\sqrt{\|\underline{g_{RL}}\|^2 / \left(\|\underline{g_{SD}}\|^2 + \epsilon\right)}, m_{min}, m_{max}\right) \cdot \frac{1+clip(cos\phi,-1,1)}{2}$
(EMA-tracked, committed once per optimizer step), shrinking when the distillation gradient dominates in magnitude *or* conflicts more strongly.

4.3 CIBO-CRISP (PRIMARY METHOD)

$$L_{CIBO}(\theta) = L_{RL}(\theta) + \beta L_{priv}(\theta) + \alpha_{anchor} L_{anchor}(\theta),$$

a single scalar objective, one backward pass, no gradient surgery. With $\hat{c}_t = sg\left[logp_T\left(y_t|x, y^*, y_{<t}\right) - logp_S\left(y_t|x, y_{<t}\right)\right]$ (same quantity as Δ_t above),

$$\widetilde{A}_t = \hat{A}_i\left(1 + \lambda tanh\left(sign\left(A_i\right) c_t\right)\right), \quad 0 \le \lambda \le 1,$$

which is bounded and strictly sign-preserving, $\left(1 + \lambda tanh\left(sign\left(A_i\right) c_t\right)\right) \in (1 - \lambda, 1 + \lambda) \subset (0, 2)$, so the verifier-supplied advantage cannot change sign. The polarity factor makes the teacher-student evidence directional. Positive evidence strengthens reinforcement on positive-advantage rollouts but attenuates the penalty on negative-advantage rollouts. Negative evidence does the converse. L_{RL} takes the same per-token REINFORCE form as Section 4.2. The privileged-divergence regularizer uses the same soft gate,

$$L_{priv}(\theta) = \frac{1}{N} \sum_i \left(1 - \gamma_i\right) \sum_t D_t(i),$$

serving as a heuristic penalty on teacher–student divergence. It is not a mutual-information estimator and should not be interpreted as one. An EMA shadow $\theta_{EMA} \leftarrow \gamma_{EMA} \theta_{EMA} + \left(1 - \gamma_{EMA}\right)\theta$, updated once per optimizer step, supplies a stability anchor scored under the *student's own* context:

$$L_{anchor}(\theta) = - E\left[\sum_t \sum_v p_{EMA}\left(v|x, y_{<t}\right) logp_S\left(v|x, y_{<t}\right)\right], \qquad p_{EMA} \text{ computed with no gradient.}$$

β may optionally adapt once per optimizer step via a bounded rate-distortion-inspired sign rule (fast reward EMA R^-_t, fast/reference KL EMAs $\underline{KL}_t$, $\underline{KL}^{ref}_t$, baselined $\underline{KL}^{base}_t = \underline{KL}_t - \underline{KL}^{ref}_t$): Define the baselined KL change by $b_t = \underline{KL}_t - \underline{KL}^{ref}$ and the controller score by $s_t = \Delta b_t - \Delta R^-_t$. The adaptive update is

$$\beta_{t+1} = clip\left(\beta_t\left(1 + \alpha\, sign\left(s_t\right)\right),\ \beta_{min}, \beta_{max}\right),$$

a bounded sign controller inspired by a rate-distortion interpretation rather than a literal rate-distortion ratio.

Conditional optimization result and idealized leakage bookkeeping. The convergence statement is intentionally separated from the leakage-motivation discussion. For the theorem, freeze the gate/credit statistics and the EMA target so that one optimizer phase corresponds to a fixed differentiable surrogate objective F_β with fixed β. Assume (A1) F_β is lower-bounded by F_* and L-smooth, (A2) the stochastic gradient g_t is conditionally unbiased, $E\left[g_t|\theta_t\right] = \nabla F_\beta\left(\theta_t\right)$, with $E\left[\|g_t - \nabla F_\beta\left(\theta_t\right)\|^2|\theta_t\right] \le \sigma^2$, and (A3) the constant step size is $\eta \le 1/L$.

> **Theorem 1** (finite-horizon convergence for a fixed surrogate). *Under (A1)–(A3), SGD with* $\theta_{t+1} = \theta_t - \eta g_t$ *satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} E\|\nabla F_{\beta}(\theta_t)\|^2 \le \frac{2(F_{\beta}(\theta_0)-F_*)}{\eta T} + L\eta\sigma^2.$$

In particular, choosing the stated step size gives an inverse-square-root rate in T when the latter choice is feasible. Throughout this paper, an ε-stationary point means a point for which the expected squared gradient norm is at most ε. Under this squared-gradient convention, the bound corresponds to $O(1/\varepsilon^2)$ stochastic-gradient steps up to problem-dependent constants; under a convention based on $E\|\nabla F\| \le \varepsilon$, the corresponding rate statement would differ.

Proposition 2 (bookkeeping identity for an idealized leakage-regularized objective). Let $J(\theta)=E[R]$, with $0\le J(\theta)\le 1$, and let $\Lambda(\theta)\ge 0$ denote any explicitly defined leakage functional. Define the idealized objective

$$F_{\beta}^{exact}(\theta) = -J(\theta) + \beta\Lambda(\theta), \qquad \beta > 0.$$

For any θ_0, θ^*, *with* $\Delta F = F_{\beta}^{exact}(\theta_0) - F_{\beta}^{exact}(\theta^*)$, *one has the identity*

$$\Lambda(\theta^*) = \Lambda(\theta_0) + \frac{J(\theta^*)-J(\theta_0)-\Delta F}{\beta},$$

and hence the bound

$$\Lambda(\theta^*) \le \Lambda(\theta_0) + \frac{1-J(\theta_0)-\Delta F}{\beta}.$$

Thus an objective drop alone does not certify low leakage. The identity also depends on the initialization value $\Lambda(\theta_0)$ and the initial reward. An absolute leakage statement requires an independently justified definition and bound for Λ. Proposition 2 supplies no mutual-information interpretation by itself.

Theorem 1 is a standard nonconvex-SGD statement for a fixed surrogate objective. It does **not** claim that the deployed algorithm has an unbiased gradient. Stop-gradient credit weights, sampled gates, and the moving EMA target make the practical update a biased and time-varying surrogate-gradient procedure. Proposition 2 is only an algebraic bookkeeping identity for an idealized leakage-regularized objective, not a convergence or leakage-reduction theorem for the deployed algorithm. The implemented L_{priv} remains a heuristic teacher-student divergence penalty and is not assigned a mutual-information meaning.

Ablation surface. Each design choice above is independently togglable, i.e., CRISP's $\lambda(t)$ (including $\lambda_{max} = 0$) and PC-style projection versus direct sum, VACS-CRISP's token credit ($\rho = 0$ removes the teacher-dependent modulation and recovers the sequence-level advantage $\hat{A}_i$ at each token. It does not reproduce a PPO-clipped objective. CRISP itself uses plain policy gradient in the implementation), soft-vs-hard gate, and symmetric-vs-asymmetric surgery, CIBO-CRISP 's fixed-vs-adaptive β, $\alpha_{anchor} = 0$, $\lambda = 0$, and the L_{priv} divergence-term sum-vs-mean reduction.

5. Experiments

Setup. We post-train Qwen2.5-Math-1.5B with 4-bit loading and LoRA (rank 64, alpha 128, no dropout) on 2 x NVIDIA T4 GPUs. All reported training runs use NuminaMath-CoT for 300 optimizer steps, a learning rate of 3e-6 with 10 warmup steps, and gradient clipping at norm 1.0. Training rollouts use a 512-token prompt cap, a 1,280-token response cap, temperature 0.8, and top-p 0.95.

Methods. The main comparison is CIBO-CRISP against GRPO and standalone OPSD. CIBO-CRISP, GRPO and OPSD are run with seeds 0 and 1. CRISP and VACS-CRISP are retained as derivational intermediate constructions rather than additional full-training baselines.

Final CIBO configuration. All CIBO runs use the polarity-aware credit multiplier from Section 4.3 and token-mean reduction for the privileged-divergence term. We set the credit

coefficient to 0.25, anchor weight to 0.01, and EMA decay to 0.995. Adaptive beta starts at 0.10, uses a 0.01 sign-controller step, and begins adapting after a 20-step reference warmup. The fixed-beta ablation holds beta at 0.10. The no-IB ablation sets beta to zero. Both ablations are full 300-step seed-0 runs, so their comparison to the full seed-0 CIBO run is budget matched.

Evaluation. We report Pass@1 on GSM8K, MATH-500, AIME 2024, AIME 2025, and HMMT 2025. Every checkpoint is evaluated with the same deterministic decoding protocol (temperature 0, top-p 1, no sampling) and the same 1,280-token response cap. The evaluator uses 2,000 bootstrap resamples for item-level confidence intervals. For the 30-problem AIME/HMMT sets, we emphasize raw counts and percentages rather than treating a bootstrap interval from so few items as a precise uncertainty estimate.

6. Results

Main comparison. CIBO-CRISP shows its clearest advantage on MATH-500. Across two seeds it averages 62.7% Pass@1, compared with 44.6% for GRPO, an 18.1-point gap. Both CIBO seeds (63.6% and 61.8%) exceed both GRPO seeds (45.4% and 43.8%). On GSM8K the difference is much smaller(74.1% for CIBO versus 72.4% for GRPO).

The item-level MATH-500 bootstrap intervals tell the same story: CIBO seed 0 is 63.6% [59.2, 67.8] and seed 1 is 61.8% [57.6, 65.8], while GRPO seed 0 is 45.4% [41.0, 49.6] and seed 1 is 43.8% [39.4, 48.0]. These intervals quantify benchmark-item uncertainty within each checkpoint. With only two training seeds, they should not be read as a precise estimate of run-to-run variance.

Table 2. Final Pass@1 (%). Two-seed rows report the mean, with seed 0 and seed 1 in parentheses.

Method	Seeds	GSM8K	MATH-500	AIME 2024	AIME 2025
CIBO-CRISP	0,1	74.1 (75.4,72.8)	62.7 (63.6,61.8)	10.0 (10.0,10.0)	3.3 (6.7,0.0)
GRPO	0,1	72.4 (75.0,69.9)	44.6 (45.4,43.8)	0.0 (0.0,0.0)	0.0 (0.0,0.0)
OPSD	0, 1	11.8(11.9, 11.8)	14.3(18.4, 10.2)	0.0(0.0, 0.0)	0.0(0.0, 0.0)

On AIME 2024, CIBO solves 3/30 problems in each seed, while both GRPO seeds solve none. On AIME 2025, CIBO seed 0 solves 2/30 and seed 1 solves none; all GRPO runs remain at 0/30. Every reported method and ablation scores 0/30 on HMMT 2025 under this protocol, so we do not draw method-level conclusions from HMMT. The 30-item sets are useful as stress tests, but too small for strong statistical claims.

OPSD. The completed OPSD runs with seeds 0 and 1 performs much worse than either RLVR-based method (11.9% and 11.8% on GSM8K and 18.4% and 10.2% on MATH-500).

Table 3. Seed-0 CIBO ablations, all trained for 300 steps. Pass@1 (%).

Variant	GSM8K	MATH-500	AIME 2024	AIME 2025
Full CIBO (adaptive beta)	75.4	63.6	10.0	6.7
Fixed beta = 0.10	73.0	60.8	6.7	0.0
No IB (beta = 0)	75.0	60.2	6.7	0.0

Ablations. On MATH-500, full CIBO improves by 2.8 points over fixed beta and by 3.4 points over the no-IB variant. The picture is weaker on GSM8K, where no-IB is within half a point of full CIBO. We therefore interpret the ablations conservatively. The bottleneck term and adaptive controller help on the harder MATH-500 setting in this seed, but the evidence does not support a claim that every component improves every benchmark.

7. Discussion and Limitations

- VACS-CRISP remains an intermediate construction without its own convergence theorem. Asymmetric primary/auxiliary projection already has precedent in PAGrad and Bloop [20, 21], but their results do not directly cover VACS-CRISP's conflict-only projection, privileged-teacher auxiliary signal, and adaptive scaling.
- The theory for CIBO-CRISP is also narrower than the deployed algorithm. Theorem 1

assumes a fixed surrogate, fixed beta, frozen gate/credit statistics, and an unbiased stochastic gradient, whereas the practical update uses stop-gradient credit weights, sampled gates, a moving EMA target, and adaptive beta. Proposition 2 is only an algebraic identity for an abstract leakage functional; it does not certify the implemented KL penalty.

- **KL direction remains untested.** We use forward KL throughout, despite a relevant argument for reverse KL in weak-to-strong supervision [17].
- The privileged-divergence term is a heuristic regularizer, not a certified mutual-information quantity. A formal information-theoretic claim would require a separately specified probabilistic construction and a justified estimator or upper bound, such as a CLUB-style construction [13].
- Reduction convention. The final CIBO-CRISP runs use token-mean reduction for the privileged-divergence term, matching the response-token normalization of the RL loss and avoiding the sequence-length amplification observed with sequence-sum reduction. CRISP and VACS-CRISP remain derivational constructions with their original sequence-sum convention, so their loss scales should not be compared directly to the final CIBO implementation.
- **Benchmark contamination.** Published analyses raise contamination concerns for the Qwen2.5 family on mathematical benchmarks, with particularly direct evidence around MATH-500/AIME-style evaluation [18]. We therefore treat the matched relative gaps between methods as more informative than any one absolute score.
- **Scale.** All experiments use a 1.5B-parameter backbone. The results establish behavior at this scale only. They do not imply that the same ranking will hold for substantially larger models.

Appendix C expands on these limitations.

8. Conclusion

RLVR and privileged self-distillation solve different parts of the supervision problem. The verifier gives a sparse but trustworthy outcome signal, while the teacher gives dense but potentially unusable guidance. CIBO-CRISP combines them with a polarity-aware token multiplier, a gated teacher-student divergence term, and an EMA anchor, avoiding the explicit gradient surgery used in the intermediate CRISP and VACS-CRISP constructions. On Qwen2.5-Math-1.5B, CIBO-CRISP averages 62.7% on MATH-500 across two seeds versus 44.6% for GRPO, while the GSM8K gap is much smaller (74.1% versus 72.4%). The seed-0 ablations point in the same direction on MATH-500: 63.6% for full CIBO, 60.8% with fixed beta, and 60.2% without the privileged-divergence term. These results support the practical combination, but not a strong information-theoretic interpretation of it. The KL direction is untested, the teacher-disagreement signal remains heuristic, and the Qwen2.5 benchmark suite carries contamination risk. CRISP and VACS-CRISP should therefore be viewed as derivational tools, while the empirical claim rests on the final CIBO objective and its matched GRPO comparison.

9. AI Usage

We used generative AI tools as a writing assistant for grammatical checks, for consolidating mathematical results (the final proofs were done by hand), and as a coding assistant. The codebase was thoroughly reviewed by the authors before final experiment execution.

10. References

[1] Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, Nan Duan. “Self-Distilled RLVR.” arXiv:2604.03128, 2026.

[2] Yuhang He, Haodong Wu, Siyi Liu, Hongyu Ge, Hange Zhou, Keyi Wu, Zhuo Zheng, Qihong Lin, Zixin Zhong, Yongqi Zhang. “Rethinking Token-Level Credit Assignment in RLVR: A Polarity-Entropy Analysis.” arXiv:2604.11056, 2026.

[3] Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin,

Andreas Krause. “Reinforcement Learning via Self-Distillation.” ICML 2026. arXiv:2601.20802, 2026.

[4] Yunhe Li, Hao Shi, Wenhao Liu, Mengzhe Ruan, Hanxu Hou, Zhongxiang Dai, Shuang Qiu, Linqi Song. “DemoPSD: Disagreement-Modulated Policy Self-Distillation.” arXiv:2607.02502, 2026.

[5] Yifeng Liu, Shiyuan Zhang, Yifan Zhang, Quanquan Gu. “Self-Distilled Policy Gradient.” arXiv:2606.04036, 2026.

[6] Ahmed Heakl, Abdelrahman M. Shaker, Youssef Mohamed, Rania Elbadry, Omar Fetouh, Fahad Shahbaz Khan, Salman Khan. “CEPO: RLVR Self-Distillation using Contrastive Evidence Policy Optimization.” arXiv:2605.19436, 2026.

[7] Mykola Khandoga, Rui Yuan, Vinay Kumar Sankarapu. “Beyond Uniform Credit: Causal Credit Assignment for Policy Optimization.” arXiv:2602.09331, 2026.

[8] Yingru Li, Jiawei Xu, Ziniu Li, Jiacai Liu, Wei Liu, Yuxuan Tong, Longtao Zheng, Zhenghai Xue, Yaxiang Zhang, Tianle Cai, Ge Zhang, Qian Liu, Baoxiang Wang. “The Optimal Token Baseline: Variance Reduction for Long-Horizon LLM-RL.” arXiv:2602.07078, 2026.

[9] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, Qiang Liu. “Conflict-Averse Gradient Descent for Multi-Task Learning.” NeurIPS 34, 2024. arXiv:2110.14048v2.

[10] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, Chelsea Finn. “Gradient Surgery for Multi-Task Learning.” NeurIPS 33, 2020. arXiv:2001.06782.

[11] Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, Aditya Grover. “Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models.” arXiv:2601.18734, 2026.

[12] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, George Tucker. “On Variational Bounds of Mutual Information.” ICML, PMLR 97:5171–5180, 2019.

[13] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, Lawrence Carin. “CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information.” ICML, PMLR 119:1779–1788, 2020.

[14] Mattia Scardecchia. “Unsupervised Transformer Pre-Training for Images: Self-Distillation, Mean Teachers, and Random Crops.” arXiv:2510.03606, 2025.

[15] Idan Shenfeld, Mehul Damani, Jonas Hübotter, Pulkit Agrawal. “Self-Distillation Enables Continual Learning.” ICML 2026, PMLR 306. arXiv:2601.19897, 2026. Best Paper Award, ICLR 2026 Workshop on Lifelong Agents.

[16] Ahmad Ajalloeian, Sebastian U. Stich. “On the Convergence of SGD with Biased Gradients.” arXiv:2008.00051, 2021.

[17] Wei Yao, Wenkai Yang, Ziqiao Wang, Yankai Lin, Yong Liu. “Revisiting Weak-to-Strong Generalization in Theory and Practice: Reverse KL vs. Forward KL.” Findings of ACL, 2025. arXiv:2502.11107.

[18] Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, Qi Zhang. “Reasoning or Memorization? Unreliable Results of Reinforcement Learning Due to Data Contamination.” AAAI 2026. arXiv:2507.10532.

[19] Thomas M. Cover, Joy A. Thomas. *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[20] Byungchan Ko, Jungseul Ok. “Efficient Scheduling of Data Augmentation for Deep Reinforcement Learning.” NeurIPS 35, 2022, 33289–33301. arXiv:2206.00518.

[21] Yu-Guan Hsieh, James Thornton, Eugene Ndiaye, Michal Klein, Marco Cuturi, Pierre Ablin. “Careful with that Scalpel: Improving Gradient Surgery with an EMA.” ICML, PMLR 235, 2024. arXiv:2402.02998.

[22] Qiangqiang He, Zhongheng Wu, ZiJian Wang. “Not All Tokens Deserve Equal Credit: Counterfactual Sensitivity Credit Reallocation for Long-CoT Reasoning.” arXiv:2607.27888, 2026.

[23] Xuan-Phi Nguyen, Shrey Pandit, Yiran Zhao, Anurag Koul, Zeyu Liu, Shafiq Joty. “Privileged Likelihood Is Not Automatically Value: Three Checks for Token Credit in On-Policy Self-Distillation.” arXiv:2608.09263, 2026.

[24] Antti Tarvainen, Harri Valpola. “Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results.” NeurIPS 30, 2017.

APPENDIX

A. EXTENDED RELATED WORK

A.1 RL with Verifiable Rewards and Token-Level Credit Assignment. He et al. partition tokens using reward polarity and predictive entropy and derive an entropy ceiling on the conditional information available for token-level credit [2]. Khandoga et al. estimate reasoning-span importance through counterfactual perturbations [7], and Li et al. derive a token-specific baseline from cumulative gradient-energy structure [8]. These approaches infer token structure from the RL process itself. VACS-CRISP and CIBO-CRISP instead reuse disagreement between a privileged teacher and the unconditioned student. That signal is inexpensive once self-distillation is already present, but its semantics are weaker. Recent work questions whether either its magnitude or sign reliably tracks outcome value [22, 23].

A.2 On-Policy Self-Distillation with Privileged Teachers, and the Leakage Problem. Zhao et al. formalize OPSD: the model conditioned on privileged information y^* acts as teacher $\pi_\theta(\cdot|x, y^*)$, the same model without that context acts as student $\pi_\theta(\cdot|x)$, and a full per-token forward-KL loss pulls the student toward the teacher over an on-policy student rollout [11]. RLSD (Yang et al.) diagnoses leakage explicitly and proposes a stop-gradient evidence ratio $P_T^+(y_t)/P_S(y_t)$, evaluated only at the sampled token, modulating the *magnitude* of an otherwise-standard advantage while the advantage's sign/direction remains anchored to the verifier [1], which is a close published precedent to VACS-CRISP's sampled-token modulation mechanism. VACS-CRISP differs in explicitly folding the *sign of the sequence-level advantage* into the exponent used to build the ratio-derived weight (so a teacher-endorsed token is boosted when reinforcing a correct trajectory but penalized less strongly, rather than boosted, inside an incorrect one) and in retaining a separate, gated full-distribution KL term reconciled against the RL term via gradient surgery, rather than folding everything into the advantage alone. CEPO (Heakl et al.) argues RLSD's single-reference ratio still cannot distinguish a decisive reasoning step from filler, since both can look equally "surprising" relative to the model's own baseline, and instead constructs a *contrastive* score contrasting evidence from a correct-answer-conditioned teacher with an alternative teacher constructed from rejected rollouts in the same batch [6]. This avoids adding a separate rollout-generation stage while changing what the privileged signal compares. Their reported full-KL baseline degradation is evidence that uncritical distribution matching can be harmful, but it does not establish one universally necessary gating rule. The principle CRISP's correctness gate, VACS-CRISP's soft gate, and CIBO-CRISP's soft-gated privileged-divergence term encode different ways to limit the influence of privileged supervision. DemoPSD (Li et al.) independently names the same problem "privileged information leakage" (building explicitly on Hübotter et al.'s SDPO as their OPSD baseline) and proposes distilling toward a reverse-KL barycenter between student and teacher, modulated by teacher–student disagreement, rather than the teacher outright [4]. SDPO (Hübotter et al.) generalizes the privileged-conditioning idea by conditioning the teacher on richer textual feedback rather than the ground-truth answer, and distilling the feedback-informed teacher into the unconditioned policy [3], reframing answer-conditioning as one instance of a broader privileged-conditioning recipe. SDPG (Liu et al.) combines group-relative advantages, full-vocabulary on-policy distillation, and an explicit reference-policy KL regularizer, gating the distillation term by the sign of the RL advantage itself rather than by raw verifier correctness or a smoothed version of it [5], a third distinct gating criterion alongside CRISP's hard gate and VACS-CRISP/CIBO-CRISP 's soft one. Positioned against this family, one distinguishing design choice across the two intermediate constructions and CIBO-CRISP is to keep RL and distillation as two separate, explicitly-combined loss terms (or, for CIBO-CRISP , to unify them into one objective whose terms remain individually traceable to RL/distillation roles) rather than collapsing the fix into a single modified advantage, making the *combination mechanism itself* a first-class object of study. Concurrent diagnostics [22, 23] further show that privileged likelihood shifts may be dominated by sensitivity or context effects rather than action value, which is why this paper avoids a causal interpretation of the modulation signal. CIBO-CRISP retains this reward-polarity convention but replaces the exponential modulation with a bounded tanh multiplier, so privileged evidence cannot reverse the verifier-supplied advantage.

A.3 Gradient Conflict in Multi-Objective and Multi-Task Optimization. PCGrad [10] removes conflicting gradient components symmetrically, while CAGrad [9] optimizes a symmetric conflict-averse objective. CRISP uses a two-gradient PC-style construction computed from the original pair. Asymmetric primary/auxiliary projection is not new. PAGrad leaves the main PPO gradient untouched and projects only a conflicting auxiliary gradient [20], while Bloop casts the problem as a primary/auxiliary hierarchy, smooths the primary direction with an EMA, and proves convergence for its own stochastic method [21]. VACS-CRISP adopts the hierarchy but differs in the privileged-teacher signal, token-level modulation, soft gate, conflict-only projection, and adaptive mixing rule. Bloop's theorem therefore should not be read as a guarantee for VACS-CRISP.

A.4 Mutual Information Estimation and Bottleneck Regularization. Poole et al. survey variational MI bounds and their bias-variance trade-offs [12], while CLUB gives a contrastive upper-bound construction [13]. The distinction matters here because CIBO-CRISP does not implement either. Its soft-gated forward KL measures teacher-student divergence under the privileged/unprivileged contexts; it is neither an MI estimator nor an MI upper bound. Proposition 2 therefore keeps the leakage functional abstract. In particular, we do not identify it with conditional mutual information when the gold answer is deterministic given the full problem. A genuine information-theoretic guarantee would require an explicit random-variable construction and a justified estimator or bound.

A.5 Self-Distillation with a Mean Teacher. Mean Teacher averages model weights over time to create a slowly moving target [24]. Related EMA-teacher ideas appear in vision self-distillation [14] and, more recently, LLM continual learning [15]. CIBO-CRISP uses the same motif for a different role than the privileged teacher. The answer-conditioned teacher supplies training-only guidance, whereas the EMA copy regularizes drift of the student itself.

A.6 Weak-to-Strong Supervision and the Direction of KL. Yao et al. argue that reverse KL can be preferable in a related weak-to-strong setting because it lets the student avoid modes supported only by a weaker supervisor [17]. Our setting is not identical. The privileged teacher is more informed rather than simply weaker, but the argument is relevant enough that KL direction should be treated as an open design choice. We use forward KL throughout for continuity with the OPSD-style objective and do not claim it is universally preferable.

A.7 Benchmark Reliability for the Qwen2.5 Family. Wu et al. document contamination concerns for this model family on mathematical benchmarks in the GSM8K/MATH-500/AIME family [18]. Our final experiments therefore use identical backbones, prompts, decoding budgets, and evaluation code across methods, and we interpret relative gaps more strongly than absolute benchmark scores.

B. Proofs

B.1 Proof of Theorem 1

Let $F = F_{\beta}$. By L-smoothness and the update $\theta_{t+1} = \theta_t - \eta g_t$,

$$F(\theta_{t+1}) \leq F(\theta_t) - \eta \langle \nabla F(\theta_t), g_t \rangle + \frac{L\eta^2}{2} \| g_t \|^2.$$

Conditioning on θ_t and applying unbiasedness gives

$$E_t\left[F(\theta_{t+1})\right] \leq F(\theta_t) - \eta \|\nabla F(\theta_t)\|^2 + \frac{L\eta^2}{2}\left(\|\nabla F(\theta_t)\|^2 + \sigma^2\right).$$

Because $\eta \leq 1/L$,

$$E_t\left[F(\theta_{t+1})\right] \leq F(\theta_t) - \frac{\eta}{2}\|\nabla F(\theta_t)\|^2 + \frac{L\eta^2}{2}\sigma^2.$$

Taking total expectation, summing from $t = 0$ to $T - 1$, and using $F(\theta_T) \geq F_*$ yields

$$\frac{\eta}{2}\sum_{t=0}^{T-1} E\|\nabla F(\theta_t)\|^2 \leq F(\theta_0) - F_* + \frac{L\eta^2 T}{2}\sigma^2.$$

Dividing by ηT proves

$$\frac{1}{T}\sum_{t=0}^{T-1} E\|\nabla F(\theta_t)\|^2 \le \frac{2(F(\theta_0)-F_*)}{\eta T} + L\eta\sigma^2. \qquad \blacksquare$$

B.2 Proof of Proposition 2

By definition,

$$F_\beta^{exact}(\theta) = -J(\theta) + \beta\Lambda(\theta),$$

so

$$F_\beta^{exact}(\theta_0) - F_\beta^{exact}(\theta^*) = J(\theta^*) - J(\theta_0) + \beta\left(\Lambda(\theta_0) - \Lambda(\theta^*)\right).$$

Rearranging gives the stated identity,

$$\Lambda(\theta^*) = \Lambda(\theta_0) + \frac{J(\theta^*)-J(\theta_0)-\Delta F}{\beta}.$$

Since $0 \le J(\theta^*) \le 1$,

$$\Lambda(\theta^*) \le \Lambda(\theta_0) + \frac{1-J(\theta_0)-\Delta F}{\beta},$$

which proves the bound.

Interpretation. Proposition 2 is intentionally modest. It says that a drop in an idealized leakage-regularized objective is not enough, by itself, to establish small leakage. The initialization and reward terms also matter. The practical privileged-divergence penalty is a teacher-student KL heuristic and is not given an exact conditional-mutual-information interpretation.

B.3 Token-level credit: variance-reduction sketch

With $\tilde{A}_t = \hat{A}\left(1 + \lambda tanh\left(sign(\hat{A})\,\hat{c}_t\right)\right)$, the law of total variance gives $Var[\tilde{A}_t] = E[Var(\tilde{A}_t|\hat{A})] + Var(E[\tilde{A}_t|\hat{A}])$. Under the unverified assumption that $\hat{c}_t$ concentrates modulation on reward-relevant tokens, token-level credit may reduce the variance of the aggregate gradient estimator relative to assigning one scalar advantage to every token. This remains a hypothesis to be tested empirically, not a theorem; concurrent results [22, 23] provide concrete reasons the assumed alignment between privileged likelihood modulation and reward relevance may fail.

C. Extended Discussion and Limitations

No convergence guarantee for the specific VACS update. PAGrad and Bloop establish relevant primary/auxiliary precedents [20, 21], but VACS-CRISP differs in ways that matter for analysis: projection is conflict-only, the auxiliary signal comes from a privileged teacher, and the mixing weight adapts alongside token modulation. We do not show that Bloop's assumptions or fixed points transfer to this update.

CIBO-CRISP's guarantees are narrower than the configuration used in the experiments. Theorem 1 assumes a fixed surrogate, fixed beta, frozen gate/credit statistics, and unbiased stochastic gradients. The deployed algorithm instead uses stop-gradient credit weights, sampled gates, an EMA target, and adaptive beta. Proposition 2 is only an algebraic identity for an abstract leakage functional. This gap between the analyzed surrogate and the practical optimizer is a real limitation, not something hidden by stronger terminology.

KL direction is not ablated. Section 2 discusses a concrete argument for reverse KL in a related weak-to-strong setting [17]. We use forward KL throughout and are not aware of a controlled, otherwise-matched comparison for the exact privileged-teacher setting studied here.

The privileged-divergence term is a proxy rather than a certified information-theoretic bound. We choose the soft-gated forward KL because it is continuous with the self-distillation losses used in Sections 4.1-4.2, not because it is a mutual-information upper bound. CLUB-style estimators [13] illustrate the extra probabilistic machinery that a formal MI claim would require. We therefore use the phrase bottleneck regularization rather than certified information bottleneck.

Reduction convention. The final CIBO-CRISP experiments use token-mean reduction for the privileged-divergence term, while the RL objective is also normalized by response-token count. This removes the direct response-length amplification produced by the earlier sequence-sum CIBO configuration. CRISP and VACS-CRISP remain defined with sequence-sum distillation in the derivation and were not part of the final compute-budget comparison.

Scope of the privileged signal. All three constructions condition the teacher on the gold final answer. Hübotter et al. generalize privileged conditioning to richer textual feedback [3]. Whether the credit weighting and gates used here transfer to that setting, or need to be redesigned, remains untested.

Benchmark contamination. As noted in Section 2 [18], the Qwen2.5 family carries documented contamination risk on benchmarks including MATH-500 and AIME. Every method in our study uses the same backbone and evaluation pipeline, which makes matched relative comparisons more trustworthy than any one absolute score, but it does not eliminate contamination as a limitation.

Compute scale. The study is deliberately a 1.5B-parameter comparison. It provides evidence at that scale, not a claim about frontier-sized reasoning models.