VERIFIABLE ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has become a powerful approach for improving language-model reasoning, but its supervision is inherently sparse: a sequence-level verifier reward is propagated only through sampled trajectories and provides no direct signal for alternative actions. On-policy distillation (OPD) offers a complementary form of supervision by providing dense teacher feedback over the full next-token distribution, but this signal is not grounded in verified task outcomes. Simply co-optimizing the two objectives does not couple these signals: verifier feedback remains sparse and does not inform the dense teacher supervision. We introduce Verifiable On-Policy Distillation (VOPD), which transfers sparse verifier feedback into dense, verifier-grounded supervision. Our key observation is that the implicit-reward view of OPD places teacher supervision and verifier feedback in a common reward space while retaining a reward signal available for every next-token action. Under this view, VOPD projects verifier preference onto the teacher's implicit-reward direction and Rao–Blackwellizes the resulting signal over the full vocabulary, propagating verifier-grounded supervision to actions that were never sampled. The resulting objective can also be interpreted as adaptively extrapolating the distillation target according to teacher–verifier agreement. Across two teacher–student configurations and eight mathematical reasoning benchmarks, VOPD consistently improves both final performance and learning efficiency over vanilla OPD and other on-policy distillation baselines.

1 INTRODUCTION

Recent advances in reinforcement learning with verifiable rewards (RLVR) have substantially improved language-model reasoning (Shao et al., 2024; DeepSeek-AI, 2025), especially in domains such as mathematics and code where outcomes can be checked automatically. However, its supervision is inherently sparse: a sequence-level verifier reward is propagated only through sampled trajectories and provides no direct signal for alternative actions (Setlur et al., 2025; Yang et al., 2026a). As a result, verified feedback is well grounded in task outcomes but makes inefficient use of each model-generated trajectory (Setlur et al., 2025).

On-policy distillation (OPD) provides a complementary form of supervision (Agarwal et al., 2024; Lu & Lab, 2025). By querying a stronger teacher at student-visited prefixes, OPD provides dense distributional feedback over the full next-token vocabulary, including unsampled actions Li et al. (2026); Hübotter et al. (2026), enabling substantially more learning from each generated trajectory. Yet this signal reflects the teacher's preferences rather than verified task outcomes; when teacher preference and verifier feedback disagree, OPD can reinforce actions that do not improve task success (Lin et al., 2026).

These complementary strengths suggest a promising direction: combining RLVR and OPD to obtain supervision that is both verifier-grounded and dense Yang et al. (2026a); Lin et al. (2026). However, simply optimizing the two objectives together does not achieve this Lin et al. (2026); Ding et al. (2026). The teacher signal remains independent of verified outcomes and can therefore conflict with the verifier, while the verifier signal remains restricted to sampled trajectories and does not reshape the dense supervision over alternative actions. Consequently, directly adding verifier-based RL to OPD yields only marginal improvements (Lin et al., 2026; Ding et al., 2026). Recent methods have explored tighter combinations of teacher and verifier feedback (Yang et al., 2026a; Pan et al., 2026; Lin et al., 2026; Hou et al., 2026), but either restrict verifier-guided supervision to sampled actions or use predefined reweighting rules to modulate the dense teacher signal. What is missing

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.

is a principled representation through which sparse verifier feedback can be converted into dense action-level supervision.

To address this, we introduce VOPD, which transfers sparse verifier feedback into dense, verifier-grounded supervision through the teacher's implicit reward. Our key observation is that the implicit-reward view of OPD provides exactly the representation needed for this transfer. Reverse-KL distillation toward a teacher can equivalently be viewed as maximizing the teacher's implicit reward

$$d(s, a) = \log \frac{p_T(a \mid s)}{p_{\text{ref}}(a \mid s)}$$

under KL regularization to a reference policy (Rafailov et al., 2023; Yuan et al., 2024; Yang et al., 2026b). This view places teacher supervision and verifier feedback in a common reward space while retaining a crucial advantage of OPD: the teacher's implicit reward can be evaluated for every next-token action at each prefix. In VOPD, we exploit this implicit-reward view to transfer verifier preference into the dense teacher-reward signal. For each prompt, we accumulate the teacher's implicit rewards along student-generated responses and compare them with their verifier outcomes. We then project the verifier advantage onto the teacher-reward direction, extracting the component of verifier preference that can be represented through the teacher's implicit reward. Because the projected verifier signal remains a scalar reweighting of the token-wise implicit reward $d(s, a)$, it can be evaluated for every action in the vocabulary. We can therefore Rao–Blackwellize the corresponding sampled-action estimator over the full vocabulary, yielding dense, verifier-grounded supervision without additional verifier evaluations. The resulting objective also admits an equivalent adaptive-extrapolation interpretation. Rather than distilling toward a fixed teacher distribution, VOPD shifts the target along the teacher–reference log-probability direction by an amount determined by teacher–verifier agreement. Positive agreement strengthens movement along the teacher-reward direction, while weak or negative agreement suppresses or reverses it.

Empirically, we evaluate VOPD across two teacher–student settings and eight mathematical reasoning benchmarks, covering both same-size and strong-to-weak distillation. Compared with other on-policy distillation methods, VOPD consistently improves both final performance and learning efficiency. These results support our central premise: sparse verifier feedback can be substantially more effective when transferred through the teacher's implicit-reward representation than when optimized as a separate sparse objective.

2 Related Work

Reinforcement Learning with Verifiable Rewards (RLVR). RLVR has become a widely used paradigm for improving language-model reasoning by optimizing directly against automatically checkable outcomes (Shao et al., 2024; DeepSeek-AI, 2025). Early RL-based post-training commonly builds on Proximal Policy Optimization (PPO) (Schulman et al., 2017), while more recent reasoning systems often use variants such as GRPO that avoid training a separate value model. These methods use verifier-derived sequence-level advantages to reinforce successful trajectories. While effective, this supervision is inherently sparse: verifier feedback is available only for sampled responses and is typically shared across all tokens within a trajectory.

On-Policy Distillation (OPD). On-policy distillation transfers the behavior of a stronger teacher into a weaker, often smaller, student (Agarwal et al., 2024; Lu & Lab, 2025). OPD minimizes the reverse KL from the student distribution to the teacher distribution on trajectories generated by the student itself, making the optimization on-policy in a manner analogous to RLVR. When teacher logits are available, this objective provides dense supervision over the full next-token vocabulary at each visited prefix rather than only the sampled action. A growing line of work Yang et al. (2026b); Penaloza et al. (2026) studies how to improve or reinterpret this dense distillation signal. In particular, G-OPD (Yang et al., 2026b) views reverse-KL OPD through the teacher's implicit reward relative to a reference policy and extrapolates along this teacher–reference direction with a fixed strength. Our method builds on the same implicit-reward view, but uses verifier feedback to adapt how strongly this direction is followed for each prompt group rather than relying on a fixed extrapolation factor.

Combining Distillation with Verifier Feedback. Several recent methods combine teacher supervision with verifier-grounded learning. RLSD and RLCSD (Yang et al., 2026a; Pan et al., 2026) use verifier outcomes to determine the direction of the policy update, while teacher–student likelihood

differences provide finer-grained token-level credit magnitudes. Although this improves credit assignment within a sampled trajectory, the verifier-guided signal remains tied to the sampled actions rather than extending to the full vocabulary. A complementary line of work preserves the dense teacher direction while using verifier outcomes to modulate its strength. OPDVR (Lin et al., 2026) gates teacher-derived updates according to trajectory correctness, while Uni-OPD (Hou et al., 2026) adjusts teacher rewards based on the relative teacher scores of correct and incorrect rollouts. These constructions provide intuitive ways to incorporate outcome feedback into dense distillation, but rely on predefined rules for translating sequence-level verifier outcomes into modifications of the teacher signal. In contrast, VOPD derives this interaction directly by projecting verifier preference onto the teacher's implicit-reward direction and extends the resulting signal to full-vocabulary supervision without changing its expected local update.

Another popular way to exploit verifier feedback is through self-distillation (Zhao et al., 2026; Hübotter et al., 2026; Shenfeld et al., 2026), where the student bootstraps itself by using a hindsight-informed copy as the teacher. These methods incorporate verifier outcomes or feedback as additional context to the teacher-side prefix, then distill the resulting teacher logits back into the original student. Although these methods also leverage verifier feedback, they do so along a largely orthogonal axis to VOPD: rather than changing how verifier feedback reshapes a fixed teacher signal, they change the teacher distribution through hindsight context. In principle, VOPD can be applied in the same self-distillation setting to further improve how this hindsight-informed teacher signal is transferred.

3 PRELIMINARIES

Let p_θ denote the student policy, p_T the teacher, and p_{ref} a fixed reference policy, all defined over a shared vocabulary $\mathcal{V}$. For each prompt x, we sample a group of n responses $\mathcal{G}_x = \{Y_i\}_{i=1}^n$ from the current student, p_θ. We write $s_{it} = (x, y_{i,<t})$ for the prefix preceding token y_{it} and $V_i = V(x, Y_i)$ for the verifier reward assigned to response Y_i.

Verifier-Based Policy Optimization. A direct way to align a policy with verifiable task outcomes is policy-gradient reinforcement learning. A simple and widely used instance is Group Relative Policy Optimization (GRPO), which samples multiple responses to the same prompt and constructs an advantage by comparing their verifier rewards (Shao et al., 2024). Specifically, with group mean $\bar{V} = \frac{1}{n}\sum_{i=1}^n V_i$, and standard deviation $\sigma_V = \sqrt{\frac{1}{n}\sum_{i=1}^n (V_i - \bar{V})^2}$, for groups with $\sigma_V > 0$, GRPO defines the group-relative verifier advantage

$$A_i^V = \frac{V_i - \bar{V}}{\sigma_V}. \tag{1}$$

Suppressing clipping and importance-weighting details, the corresponding policy-gradient objective takes the form

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{x,i,t}\left[A_i^V \log p_\theta(y_{it} \mid s_{it})\right]. \tag{2}$$

The resulting supervision is grounded in verified task outcomes, but remains sparse: it is available only for sampled responses, and the same sequence-level advantage is applied uniformly to all sampled steps in each trajectory.

On-Policy Distillation and Implicit Reward. On-policy distillation provides a complementary way to transfer the behavior of a stronger teacher into a student. At each prefix s visited by the student, OPD minimizes the reverse-KL divergence

$$\ell_{\text{OPD}}(\theta; s) = D_{\text{KL}}(p_\theta(\cdot \mid s) \,\|\, p_T(\cdot \mid s)), \tag{3}$$

providing distributional supervision over the entire next-token vocabulary (Agarwal et al., 2024; Lu & Lab, 2025).

To place OPD in a reward-based form, define the teacher's implicit reward relative to the reference as $d(s,a) = \log \frac{p_T(a|s)}{p_{\text{ref}}(a|s)}$. The reverse-KL objective satisfies

$$\ell_{\text{OPD}}(\theta; s) = D_{\text{KL}}(p_\theta(\cdot \mid s) \,\|\, p_{\text{ref}}(\cdot \mid s)) - \mathbb{E}_{a\sim p_\theta(\cdot|s)}\left[d(s,a)\right]. \tag{4}$$

Thus, reverse-KL OPD can equivalently be viewed as maximizing the teacher's implicit reward $d(s,a)$ under KL regularization to the reference (Yuan et al., 2025; Yang et al., 2026b). Unlike verifier

feedback, which is available only on sampled trajectories, $d(s, a)$ can be evaluated for every $a \in \mathcal{V}$ at each visited prefix, providing dense supervision over alternative actions. This implicit-reward view therefore puts OPD and verifier-based RL in a common reward-based language.

4 VERIFIABLE ON-POLICY DISTILLATION

We introduce VOPD, which converts sparse verifier feedback into verifier-grounded dense supervision using the teacher's implicit reward. Starting from the additive combination of OPD and verifier-based policy optimization, VOPD projects the verifier advantage onto the teacher implicit-reward direction, maps the projected signal back to the scale of the implicit reward, and Rao–Blackwellizes the resulting sampled-action signal over the full vocabulary.

4.1 BEYOND CO-OPTIMIZATION

A direct combination of OPD and RLVR is to co-optimize both objectives

$$\mathcal{L}_{\text{add}} = \mathbb{E}_{x,s}\left[\ell_{\text{OPD}}(\theta; s)\right] + \lambda \mathcal{L}_{\text{GRPO}}, \tag{5}$$

where $\lambda \geq 0$ controls the strength of verifier-based supervision relative to ordinary OPD. However, simply adding the GRPO objective has limited effect because its supervision is substantially sparser and coarser than the dense distributional signal provided by OPD. This discrepancy is also reflected empirically: verifier-only RL requires far more optimization steps to reach comparable accuracy, and directly adding GRPO yields only marginal gains over vanilla OPD.

The implicit-reward formulation of OPD in Equation (4) suggests a natural way to integrate the two signals more tightly. Instead of keeping verifier feedback as a separate sparse objective, we express its teacher-representable component through the teacher's implicit reward. This allows verifier information to inherit OPD's dense action-level structure.

4.2 PROJECTING VERIFIER FEEDBACK ONTO THE TEACHER'S IMPLICIT REWARD

The verifier reward V_i is defined for an entire response, whereas the teacher's implicit reward $d(s, a)$ is defined for individual token actions. To place them at the same level, we aggregate the latter along each sampled response:

$$D_i = \sum_{t=1}^{|Y_i|} d(s_{it}, y_{it}) = \log \frac{p_T(Y_i \mid x)}{p_{\text{ref}}(Y_i \mid x)}. \tag{6}$$

We refer to D_i as the cumulative teacher's implicit reward. Crucially, this sum is itself an exact response-level implicit reward: if Y_i is treated as a single action, D_i measures the teacher's preference for the entire response relative to the reference.

Each sampled response is therefore associated with two response-level signals: the verifier reward V_i and the cumulative teacher's implicit reward D_i. Because the two rewards can differ in offset and scale, we compare them through standardized within-group advantages. GRPO already defines A_i^V in Equation (1); analogously, we standardize the cumulative teacher rewards within each prompt group:

$$A_i^D = \frac{D_i - \bar{D}}{\sigma_D}, \tag{7}$$

where $\bar{D}$ and σ_D denote the group mean and standard deviation of $\{D_i\}_{i=1}^n$.

Centering removes prompt-specific offsets, while standardization removes arbitrary differences in reward scale. Both transformations preserve the within-group ordering of responses, so $\mathbf{A}^V$ and $\mathbf{A}^D$ describe the relative preferences induced by the verifier and teacher over alternative attempts at the same prompt on a common scale.

We seek to transfer verifier preference through the teacher's implicit reward. To this end, we restrict the transferred signal to scalar multiples of the teacher-reward direction, $b\,\mathbf{A}^D$ for $b \in \mathbb{R}$. This restriction is what makes dense transfer possible: scaling the sequence-level teacher reward by b scales each additive token reward $d(s_{it}, y_{it})$ by the same factor. The resulting rescaled sampled-action signal can therefore easily be extended later to the full vocabulary.

Within this family, we choose the signal that best matches the observed verifier advantages:

$$c_x = \arg\min_b \frac{1}{n} \sum_{i=1}^{n} \left(A_i^V - bA_i^D\right)^2. \tag{8}$$

The solution is

$$c_x = \frac{1}{n} \sum_{i=1}^{n} A_i^V A_i^D = \frac{\frac{1}{n}\sum_i (V_i - \bar{V})(D_i - \bar{D})}{\sigma_V \sigma_D} \in [-1, 1], \tag{9}$$

and the projected verifier advantage is

$$\widehat{A}_i^V = c_x A_i^D. \tag{10}$$

$\widehat{A}_i^V$ is the component of verifier preference that can be represented through the teacher-reward direction. Since both A_i^V and A_i^D are standardized, the projection factor c_x is the empirical correlation between verifier reward and cumulative teacher's implicit reward within the prompt group. Positive c_x indicates that responses preferred more strongly by the teacher also tend to achieve better verified outcomes, so the teacher-reward direction should be reinforced; negative c_x indicates that verified outcomes favor movement against that direction, so it should instead be opposed.

4.3 From Projected Verifier Feedback to Full-Vocabulary Supervision

The projected verifier advantage $\widehat{A}_i^V = c_x A_i^D$ is expressed in standardized advantage coordinates. Substituting the definition of A_i^D gives

$$\widehat{A}_i^V = \frac{c_x}{\sigma_D}(D_i - \bar{D}). \tag{11}$$

Using this quantity directly would make the effective strength of the transferred signal scale inversely with σ_D. We therefore rescale it by σ_D, placing it on the same scale as the cumulative teacher reward:

$$\sigma_D \widehat{A}_i^V = c_x(D_i - \bar{D}). \tag{12}$$

During the policy update, the rollout statistics and c_x are held fixed. The term $-c_x\bar{D}$ is therefore a response-independent baseline and does not affect the policy-gradient expectation. Hence, the transferred sequence-level signal is equivalent to

$$c_x D_i = c_x \sum_t d(s_{it}, y_{it}). \tag{13}$$

Following the immediate-token approximation used in OPD (Fu et al., 2026), we treat each visited prefix s as fixed and replace the future implicit-reward return with the current-token reward $d(s, y)$. For a sampled next token $y \sim p_\theta(\cdot \mid s)$, this gives the sampled-action gradient estimator

$$\widehat{g}(s, y) = c_x d(s, y) \nabla_\theta \log p_\theta(y \mid s). \tag{14}$$

This is exactly the sampled-action estimator of the teacher-reward term underlying OPD, rescaled by the verifier-derived coefficient c_x.

Crucially, this signal need not remain restricted to the sampled action. At every visited prefix, the teacher and reference provide $d(s, a) = \log \frac{p_T(a|s)}{p_{\text{ref}}(a|s)}$ for every $a \in \mathcal{V}$. We can therefore integrate out the sampled next-token action. Specifically, we define the Rao–Blackwellized estimator as the conditional expectation of $\widehat{g}$ given the visited prefix:

$$\begin{aligned} \widetilde{g}(s) &:= \mathbb{E}_{a \sim p_\theta(\cdot|s)} \left[\widehat{g}(s, a) \mid s\right] \\ &= \mathbb{E}_{a \sim p_\theta(\cdot|s)} \left[c_x d(s, a) \nabla_\theta \log p_\theta(a \mid s)\right] \\ &= c_x \sum_{a \in \mathcal{V}} d(s, a) \nabla_\theta p_\theta(a \mid s) \\ &= \nabla_\theta \left[c_x \sum_{a \in \mathcal{V}} p_\theta(a \mid s) d(s, a)\right]. \end{aligned} \tag{15}$$

Algorithm 1 VOPD: Verifiable On-Policy Distillation

Require: Student p_θ, teacher p_T, reference p_{ref}, training question set $\mathcal{D} = \{x_j\}_{j=1}^N$, verifier V, group size n, verifier weight λ

for each training iteration **do**
  Sample a batch of prompts x from $\mathcal{D}$
  for each prompt x **do**
    // Step 1: On-policy rollout
    Sample n responses $\{Y_i\}_{i=1}^n \sim p_\theta(\cdot \mid x)$
    Obtain verifier rewards $V_i = V(x, Y_i)$
    // Step 2: Teacher–verifier projection
    for $i = 1, \ldots, n$ **do**
      Evaluate p_T and p_{ref} on prefixes of Y_i
      $D_i \leftarrow \sum_t \log \frac{p_T(y_{it}|s_{it})}{p_{\text{ref}}(y_{it}|s_{it})}$
    end for
    Compute $A_i^V = (V_i - \bar{V})/\sigma_V$ and $A_i^D = (D_i - \bar{D})/\sigma_D$
    $c_x \leftarrow \frac{1}{n}\sum_{i=1}^n A_i^V A_i^D$ ▷ Projection coefficient
  end for
  // Step 3: Verifiable distillation update
  Compute the gradient ascent direction

$$g_{\text{VOPD}} = \nabla_\theta \mathbb{E}_{x,s}\left[-D_{\text{KL}}(p_\theta(\cdot \mid s) \,\|\, p_T(\cdot \mid s)) + \lambda c_x \sum_{a\in\mathcal{V}} p_\theta(a \mid s) \log \frac{p_T(a \mid s)}{p_{\text{ref}}(a \mid s)}\right]$$

  Update $\theta \leftarrow \theta + \eta g_{\text{VOPD}}$
end for

By construction, $\widetilde{g}(s)$ is the conditional expectation of the sampled-action estimator. Hence, by the law of total expectation, $\mathbb{E}_s[\widetilde{g}(s)] = \mathbb{E}_{s,a}[\widehat{g}(s,a)]$. Thus, Rao–Blackwellization preserves the expected local gradient while integrating out next-token sampling variance, and in doing so turns the verifier-derived signal from a sampled-action update into dense, verifier-grounded supervision over the full vocabulary without additional verifier evaluations.

Moreover, Equation equation 15 shows that this Rao–Blackwellized estimator is exactly the gradient of the full-vocabulary objective

$$j_{\text{VOPD}}(\theta; s) = c_x \sum_{a\in\mathcal{V}} p_\theta(a \mid s) d(s,a). \tag{16}$$

4.4 The VOPD Objective

We can now return to the co-optimization objective in Equation equation 5. Rather than retaining the sparse GRPO term, VOPD replaces it with the projected verifier signal derived above and Rao–Blackwellized it over the full vocabulary. Combining Equation equation 16 with the ordinary OPD loss gives

$$\begin{aligned}\ell_{\text{VOPD}}(\theta; s) = {} & D_{\text{KL}}(p_\theta(\cdot \mid s) \,\|\, p_T(\cdot \mid s)) \\ & - \lambda c_x \sum_{a\in\mathcal{V}} p_\theta(a \mid s) \log \frac{p_T(a \mid s)}{p_{\text{ref}}(a \mid s)}.\end{aligned} \tag{17}$$

The full training objective is therefore $\mathcal{L}_{\text{VOPD}}(\theta) = \mathbb{E}_{x,s}\left[\ell_{\text{VOPD}}(\theta; s)\right]$.

Adaptive Extrapolation View. The objective also admits an equivalent target-distribution interpretation. Up to a normalization constant independent of θ, Equation (17) is also equivalent to the reverse-KL distillation toward the teacher-extrapolation target

$$q_x(a \mid s) \propto p_T(a \mid s)^{1+\lambda c_x} p_{\text{ref}}(a \mid s)^{-\lambda c_x}. \tag{18}$$

Under this view, λc_x thus controls how far the distillation target moves along the teacher–reference log-probability direction: a positive c_x extrapolates beyond the teacher, while a negative c_x correspondingly moves the target back toward the reference. This connects VOPD to ExOPD (Yang

Table 1: **Performance under 4B same-size distillation.** The teacher is the RL-trained Qwen3-4B and the student is Qwen3-4B. Best performance is **bolded**.

Method	AIME24	AIME25	AMC23	HMMT25-Feb	HMMT25-Nov	MATH-500	Minerva	Olympiad-Bench	Average
Teacher	72.50	62.29	95.00	37.71	43.75	94.25	41.64	66.17	64.16
Student	20.83	20.21	6.41	11.46	7.50	82.25	30.51	48.63	35.97
OPD	69.58	65.83	**96.09**	39.58	46.04	94.05	40.44	66.06	64.71
OPD + GRPO	71.04	63.12	95.47	40.00	47.50	93.90	40.62	66.32	64.75
ExOPD	70.62	66.04	95.62	39.79	48.54	94.05	40.44	**66.88**	65.25
VOPD	**73.33**	**68.12**	95.94	**41.04**	**50.42**	**94.55**	**40.99**	66.65	**66.38**

et al., 2026b): ExOPD uses a fixed positive extrapolation strength regardless of verifier feedback, whereas VOPD adapts the extrapolation strength for each prompt group according to the observed teacher–verifier agreement.

Computational Overhead. Algorithm 1 summarizes the full training procedure. Compared with ordinary OPD, VOPD additionally requires a frozen reference model and evaluates its logits on the sampled prefixes, introducing extra model memory and one additional forward evaluation over each rollout batch. Beyond this reference-model evaluation, the remaining computation is lightweight: forming the cumulative teacher rewards and group-level coefficients c_x requires only simple reductions over already-computed quantities, with no additional rollouts or verifier evaluations. Since on-policy training is typically dominated by rollout generation, the extra reference evaluation contributes only a small amount of computational overhead.

5 Experiments

5.1 Experiment Setup

Distillation Settings. We focus our experiments on mathematical reasoning and evaluate two teacher–student configurations spanning both same-size and strong-to-weak distillation. For the same-size setting, we use the RL-trained Qwen3-4B (Yang et al., 2025) from Yang et al. (2026b) as the teacher and distill it into Qwen3-4B non-thinking mode. For the strong-to-weak setting, we distill the same teacher into the smaller Qwen3-1.7B model. We set the reference policy p_{ref} to the student's initialization and keep it frozen throughout training.

Baselines. We compare against three on-policy distillation baselines. **OPD** performs standard reverse-KL on-policy distillation on student-generated prefixes. **OPD+GRPO** co-optimizes the group-relative verifier objective with the OPD objective, corresponding to the naive combination in Equation (5). **Ex-OPD** (Yang et al., 2026b) extrapolates along the teacher–reference direction with a fixed strength, and can be viewed as a verifier-agnostic special case of our method with $c_x = 1$.

Training Setting. All methods are trained on DAPO-Math-17k (Yu et al., 2025) using student-generated on-policy rollouts. We a constant learning rate and one optimization update per rollout batch. All distillation methods employ the top-k distillation approximation: optimizing over the student's top-16 tokens at each prefix. For ExOPD and our VOPD, we set $\lambda = 0.25$ unless otherwise specified. We train each run until validation accuracy converges. We implement all methods in VERL (Sheng et al., 2024) with FSDP (Zhao et al., 2023) and use vLLM (Kwon et al., 2023) for rollout generation and offline evaluation. Each run uses 8 NVIDIA A100 GPUs.

Evaluation. We evaluate on eight mathematical reasoning benchmarks: AIME 2024 (AI-MO, 2024a), AIME 2025 (OpenCompass, 2025), AMC 2023 (AI-MO, 2024b), HMMT February 2025 (Balunović et al., 2025), HMMT November 2025 (Balunović et al., 2025), MATH-500 (Lightman et al., 2024), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). We report *avg@*n, the average correctness of independent generations for each problem. All methods are evaluated from their final training checkpoint.

For additional implementation details, including hyperparameter choices, please refer to Section A.1.

5.2 Main Results

VOPD Consistently Improves Distillation Performance. We first compare the converged performance of different on-policy distillation objectives. As shown in Table 1, under 4B same-size

Table 2: **Performance under strong-to-weak distillation.** The teacher is the RL-trained Qwen3-4B and the student is Qwen3-1.7B. Best performance is **bolded**.

Method	AIME24	AIME25	AMC23	HMMT25-Feb	HMMT25-Nov	MATH-500	Minerva	Olympiad-Bench	Average
Teacher	72.50	62.29	95.00	37.71	43.75	94.25	41.64	66.17	64.16
Student	12.92	13.33	44.84	6.46	4.58	72.25	20.31	37.61	31.01
OPD	50.62	40.00	86.88	21.25	26.25	88.60	34.83	58.64	50.88
OPD + GRPO	49.58	42.50	85.31	21.88	26.67	89.65	33.18	59.01	50.97
ExOPD	52.71	41.04	**89.22**	**22.50**	28.54	88.95	35.57	60.20	52.34
VOPD	**57.29**	**42.71**	88.28	22.29	**29.38**	**90.20**	**36.03**	**60.39**	**53.32**

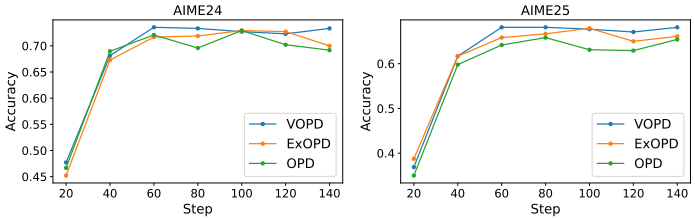


Figure 1: Validation accuracy over training in the 4B same-size distillation setting. Compared with vanilla OPD and fixed-extrapolation ExOPD, VOPD improves more rapidly during training and reaches higher converged accuracy.

λ	AIME24	AIME25
0.00	69.58	65.83
0.25	73.33	68.12
0.50	72.91	67.91
0.75	71.88	66.67

Table 3: Effect of the verifier-alignment strength λ in the 4B same-size setting.

distillation, VOPD achieves the strongest average performance across the eight mathematical reasoning benchmarks, improving the average accuracy from 64.71 for vanilla OPD to 66.38 and obtaining the best result on six of the eight benchmarks. Notably, the distilled student also exceeds the teacher's average performance (64.16) and the fixed-extrapolation ExOPD (65.25), indicating that verifier-grounded dense supervision can improve beyond direct imitation of the teacher. In contrast, co-optimizing OPD with GRPO yields only marginal improvement over vanilla OPD, suggesting that simply adding sparse verifier supervision does not provide the same benefit.

The gains persist under the more challenging strong-to-weak setting. As shown in Table 2, VOPD achieves the highest average performance of 53.32 and the highest accuracy on six of eight benchmarks. Although the trained 1.7B student still remains below the 4B teacher, VOPD consistently improves the effectiveness of transferring the teacher's reasoning capability to the smaller model. Together, these results show that the benefit of VOPD extends beyond same-size distillation and remains effective under substantial student–teacher capacity mismatch.

VOPD Improves Learning Efficiency. Beyond final accuracy, VOPD also improves more rapidly during training. Figure 1 shows validation accuracy on AIME 2024 and AIME 2025 throughout 4B same-size distillation. Compared with vanilla OPD and fixed-extrapolation ExOPD, VOPD separates from the baselines early in training and maintains stronger validation performance through convergence. This indicates that verifier-grounded dense supervision not only improves the final distilled model, but also makes more effective use of each on-policy rollout during training.

5.3 Ablations and Analyses

Effect of Verifier-Grounded Extrapolation Strength λ. We ablate the extrapolation scale λ in the 4B same-size setting on the AIME benchmarks. As shown in Table 3, $\lambda = 0.25$ performs best on both AIME 2024 and AIME 2025, substantially improving over ordinary OPD ($\lambda = 0$). Increasing λ further gradually degrades performance. This phenomenon can be explained by the adaptive-extrapolation view from which λ controls how far verifier agreement can move the target beyond the teacher distribution: a moderate value effectively exploits the verifier-grounded teacher signal, whereas an overly large value extrapolates too aggressively and over-amplifies teacher-preferred modes.

Analysis of Teacher–Verifier Agreement. A central premise of VOPD is that the teacher's implicit reward captures part, but not all, of the verifier's preference over candidate responses. If the two signals were unrelated, the projected coefficient c_x would concentrate near zero, and little verifier

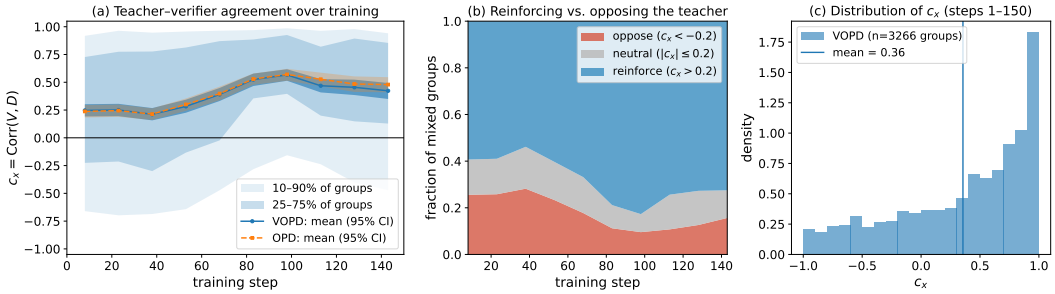


Figure 2: **Analysis of teacher–verifier agreement in practice.** (a) Mean of agreement coefficient c_x at each training step for VOPD and vanilla OPD. (b) Fraction of prompt groups that reinforce, remain neutral to, or oppose the teacher-reward direction over training. (c) Distribution of c_x across the entire training.

information could be transferred through the teacher-reward direction. Conversely, if they were perfectly agreed across prompt groups, a fixed extrapolation strength would largely suffice. We therefore examine the empirical agreement coefficient $c_x = \mathrm{Corr}(V, D)$ throughout training.

We study the strong-to-weak distillation setting and find evidence for both parts of the premise. We first investigate whether the teacher reward contains meaningful information about the verifier. The average c_x remains consistently positive throughout training Figure 2, indicating that the teacher's relative preferences, on average, agree with verifier outcomes. We observe the same pattern under vanilla OPD, showing that this general agreement is already present in ordinary distillation trajectories rather than being created by VOPD.

At the same time, this agreement is far from uniform across prompt groups. As shown in panels (b)–(c) of Figure 2, the estimated c_x values are broadly distributed: while most groups reinforce the teacher direction, a nontrivial fraction are approximately neutral or opposing. This means that, in practice, VOPD does not simply extrapolate positively everywhere: it identifies groups where teacher and verifier preferences disagree and correspondingly weakens or reverses the extrapolation. This adaptive behavior may help explain its performance advantage over fixed-extrapolation methods such as ExOPD.

6 Limitations

Like other forms of on-policy distillation, VOPD assumes access to a teacher distribution that is meaningfully stronger than the student, which may be unavailable in frontier settings. Recent self-distillation approaches offer a possible alternative through hindsight-informed teacher distributions, and extending VOPD to such settings is a promising direction. Our method also requires teacher logits to represent verifier feedback through the teacher's dense implicit reward, limiting direct application to closed-source teachers that expose only sampled outputs. Finally, VOPD requires an additional frozen reference model, increasing memory usage and adding one reference forward pass per rollout batch. The memory cost is inherent to the current formulation, while the additional computation is small relative to rollout generation.

7 Conclusion

In this work, we study how to combine the verifier-grounded supervision of RLVR with the dense teacher supervision of on-policy distillation. We introduce VOPD, which transfers verifier feedback through the dense implicit-reward signal underlying OPD. By placing teacher and verifier feedback in a common reward space, VOPD projects verifier preference onto the teacher's implicit-reward direction and Rao–Blackwellizes the resulting signal over the full vocabulary, converting sparse verifier feedback into dense, verifier-grounded supervision. Experiments show that VOPD consistently improves both final performance and learning efficiency over standard OPD and other on-policy distillation baselines, demonstrating its effectiveness in improving language-model reasoning.

AI Assistance Disclosure

We used generative AI tools to assist with research execution, including method implementation, experimental logging design, and the compilation and visualization of results. We also used generative AI to revise and polish author-written text and mathematical derivations for clarity, grammar, and presentation. Generative AI was not used to retrieve or discover related works, generate synthetic datasets, draft sections of the paper, or prove mathematical claims. All AI-assisted work was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

Reproducibility Statement

We are committed to ensuring the reproducibility of our work. In this paper, we provide detailed descriptions of VOPD, including its derivation, objective, training procedure, and evaluation protocol in the main text and appendix. We also report the teacher–student configurations, training data, hyperparameter settings, rollout and decoding configurations, and benchmark-specific evaluation details in Sections A.1 and 5.1. Additionally, we will release our code, together with instructions for training, evaluation, and reproducing the reported results, upon completion of the review process.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

AI-MO. Aime 2024. https://huggingface.co/datasets/AI-MO/aimo-validation-aime, 2024a.

AI-MO. Amc 2023. https://huggingface.co/datasets/AI-MO/aimo-validation-amc, 2024b.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, February 2025. URL https://matharena.ai/.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Yifan Ding, Xincheng Wei, Yoshua Y. Li, Ziheng Li, Yuquan Lu, Siyu Zhang, Dongsheng Ma, Rongxiang Weng, Xunliang Cai, and Yun Chen. Saf-opd: Stable advantage fusion for on-policy distillation, 2026. URL https://arxiv.org/abs/2607.29209.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=uZbZb058bA.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, Chengquan Zhang, Zhuotao Tian, Han Hu, Yi Yang, Fei Wu, and Hehe Fan. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe, 2026. URL https://arxiv.org/abs/2605.03677.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation, 2026. URL https://arxiv.org/abs/2601.20802.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=IFXTZERXdM7.

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

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Wenze Lin, Jiale Zhao, Xitai Jiang, Songde Rao, Yining Li, Shenzhi Wang, Bingxiang He, and Gao Huang. On-policy distillation with verifiable reward, 2026. URL https://arxiv.org/abs/2608.24696.

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

OpenCompass. Aime 2025. https://huggingface.co/datasets/opencompass/AIME2025, 2025.

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.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models, 2026. URL https://arxiv.org/abs/2602.04942.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=A6Y7AqlzLW.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning, 2026. URL https://arxiv.org/abs/2601.19897.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026a. URL https://arxiv.org/abs/2604.03128.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation, 2026b. URL https://arxiv.org/abs/2602.12125.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*, 2024.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=8ThnPFhGm8.

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.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel, 2023. URL https://arxiv.org/abs/2304.11277.

A Appendix

A.1 Implementation Details

Training Data. All methods are trained on DAPO-Math-17k. Following the prompt alignment used in prior OPD work (Li et al., 2026), we modify the answer-format instruction to require the final answer in `\boxed{}` form, matching the format used during post-training of the teachers, and adapt the verifier accordingly. The verifier extracts the content of the last `\boxed{}` in the response, normalizes both prediction and ground truth, and checks equivalence by normalized string matching, symbolic equivalence with SymPy, and a `math_verify`-based fallback, producing a binary reward in $\{0, 1\}$.

Additional Training Details. We cap prompts at 1,024 tokens and use the response-length limits reported in Table 5; more than 70% of student rollouts terminate within these limits. Allowing responses to finish is particularly important for VOPD, since responses truncated before producing a final answer receive a zero verifier signal. After each rollout, the frozen teacher and reference models score the student-generated prefixes by teacher forcing; no teacher-generated trajectories are used during distillation. Following prior OPD implementations, all distillation methods approximate the vocabulary expectation using the student's top-k tokens at each prefix, with $k = 16$.

Training uses bfloat16 mixed precision under FSDP. DeepSeek-1.5B weights are loaded in float32, while Qwen3 weights are loaded in bfloat16. Rollout generation and evaluation use vLLM in bfloat16 for all settings.

We train each run until validation accuracy converges. This takes 140 training steps in the 4B same-size setting, and 160 steps in the strong-to-weak setting. These runs all cover less than one pass over DAPO-Math-17k, so no training prompt is encountered twice.

For OPD+GRPO, the sequence-level GRPO advantage is broadcast unchanged to all tokens in the sampled response, with no length normalization or discounting.

Hyperparameters. Table 5 summarizes the training and method-specific hyperparameters for both distillation settings. Unless otherwise specified, we use the same hyperparameters across methods within each setting.

Evaluation. The complete decoding configuration, benchmark sizes, and number of samples are reported in Table 4. We scale the number of samples per problem n across benchmarks so that each benchmark uses approximately the same amount of evaluation compute.

Table 4: Evaluation protocol.

Benchmark	#Problems	Samples per problem n
AIME 2024	30	16
AIME 2025	30	16
AMC 2023	40	16
HMMT Feb. 2025	30	16
HMMT Nov. 2025	30	16
MATH-500	500	4
Minerva Math	272	4
OlympiadBench	674	4
Decoding: temperature 0.7, top-p 0.95, max 31,744 new tokens, vLLM		
Metric: avg@n (mean accuracy over samples), rule-based `\boxed{}` grading		
Checkpoint: final training step		

Table 5: Training hyperparameters for the three distillation settings. Entries spanning all columns are shared.

Distillation Setting	4B Same-size	Strong-to-weak
Student	Qwen3-4B	Qwen3-1.7B
Teacher	Qwen3-4B-RL-Math	Qwen3-4B-RL-Math
Reference π_{ref}	Qwen3-4B	Qwen3-1.7B
Training steps	140	160
Thinking mode	non-thinking	
Group size G	4	
Max response length	12,288	
Weight precision	bf16	
Training data	DAPO-Math-17k (`\boxed{}` prompt), 17,917 prompts	
Batch Size (Prompts per step)	64	
Minibatch Size	64	
Max prompt length	1,024	
Rollout temperature / top-p	1.0 / 1.0	
Updates per step / PPO epochs	1 / 1 (fully on-policy)	
Optimizer	AdamW, $\beta = (0.9, 0.999)$, wd $= 0.01$	
Learning rate / schedule	2×10^{-6}, constant, no warm-up	
Gradient clipping	1.0	
KL penalty / entropy bonus	none / none	
Loss aggregation	token-mean	
Top-k OPD	$k = 16$, student top-k	
Teacher temperature	1.0	
ExOPD λ	0.25	
VOPD λ	0.25	