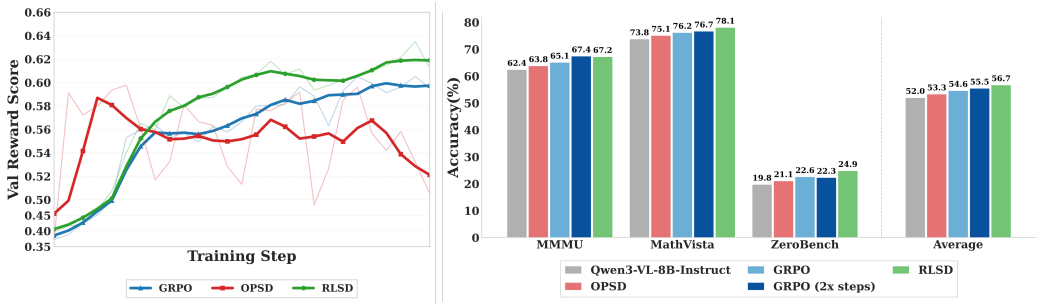
Self-Distilled RLVR

Chenxu Yang[♠♡]*, **Chuanyu Qin**[♠♡]*, **Qingyi Si**[♢]*, **Minghui Chen**[♠♡], **Naibin Gu**[♠♡], **Dingyu Yao**[♠♡], **Zheng Lin**[♠♡]†, **Weiping Wang**[♠], **Nan Duan**[♢], **Jiaqi Wang**[♢]
[♠]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[♡]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[♢]JD.COM
{yangchenxu,linzheng}@iie.ac.cn, siqingyi.phoebus@jd.com

Abstract

On-policy distillation (OPD) has become a popular training paradigm in the LLM community. This paradigm selects a larger model as the teacher to provide dense, fine-grained signals for each sampled trajectory, in contrast to reinforcement learning with verifiable rewards (RLVR), which only obtains sparse signals from verifiable outcomes in the environment. Recently, the community has explored on-policy self-distillation (OPSD), where the same model serves as both teacher and student, with the teacher receiving additional privileged information such as reference answers to enable self-evolution. This paper demonstrates that learning signals solely derived from the privileged teacher result in severe information leakage and unstable long-term training. Accordingly, we identify the optimal niche for self-distillation and propose **RLSD** (**RL**VR with **S**elf-**D**istillation). Specifically, we leverage self-distillation to obtain token-level policy differences for determining fine-grained update magnitudes, while continuing to use RLVR to derive reliable update directions from environmental feedback (e.g., response correctness). This enables RLSD to simultaneously harness the strengths of both RLVR and OPSD, achieving a higher convergence ceiling and superior training stability.



(a) Average performance dynamics on the validation set.

(b) Performance comparison on reasoning tasks.

Figure 1: Convergence dynamics of RLSD vs. baselines on Qwen3-VL-8B-Instruct. **(a)** OPSD peaks early and degrades, whereas RLSD inherits GRPO's stability while reaching a higher ceiling. **(b)** GRPO and RLSD report results at 200 training steps, while GRPO (2×steps) reports results at 400 steps; RLSD at 200 steps already surpasses GRPO trained for twice as many steps, demonstrating faster convergence. (Note: This figure focuses on training efficiency at the early stage; final converged performance is reported in Table 2, where both RLSD and GRPO achieve their best performance.)

* Equal Contribution.
† Zheng Lin is the corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of training paradigms for post-training LLMs.

Method	Trajectory	Efficiency	Leakage Risk	Signal	Direction Anchoring
SFT	✗Off-policy	✓High	✓N/A	✓ Rich	✗Teacher
RLVR (GRPO)	✓On-policy	✓High	✓N/A	✗ Weak	✓Environment
OPD	✓On-policy	✗Low	✓N/A	✓ Rich	✗Teacher
OPSD	✓On-policy	✓High	✗Severe	✓ Rich	✗Teacher
RLSD (Ours)	✓On-policy	✓High	✓N/A	✓ Rich	✓Environment

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) methods such as GRPO [1] have become a central paradigm for training large reasoning models [2, 3], where each trajectory receives only a single scalar signal determined by the outcome from the environment. On-policy distillation (OPD) [4, 5] complements this by leveraging a stronger teacher model to provide dense, token-level logits as learning signals along the student's own sampled trajectories, enriching the trajectory-level supervision to the token level and thereby achieving faster convergence. Recent work has shown that OPD from advanced teachers can match or even outperform RLVR [6], establishing it as an equally compelling paradigm (see Table 1 for a systematic comparison).

Despite its effectiveness, OPD relies on a separate, typically much larger teacher model, incurring substantial computational overhead. Moreover, since OPD requires computing token-level distributions over a shared vocabulary, the teacher and student models must share the same vocabulary, significantly reducing the practical scalability of this paradigm. On-policy self-distillation (OPSD) [7, 8] offers an appealing alternative: a single model serves as both teacher and student, where the teacher is the same model conditioned on *privileged information* r (such as a verified reasoning trace or environmental feedback) and the student operates solely on the input query. OPSD achieves several-fold improvements in token efficiency over GRPO without requiring any external model. However, we prove this efficiency gain is fragile: as illustrated by the red curve in Figure 1(a), performance peaks early and subsequently deteriorates, accompanied by systematic privileged information leakage, where the model explicitly appeals to an invisible reference solution during inference despite never having access to it (see Figure 2 for a representative example). These phenomena raise a natural question: *why does OPD work while OPSD fails?*

We trace the answer to a structural distinction between the two settings. In OPD, teacher and student observe the same input (information-symmetric), so the teacher's dense signal reflects superior reasoning under shared information access. In OPSD, the teacher conditions on privileged information that the student cannot observe (information-asymmetric), creating a fundamental mismatch. We prove that this asymmetry renders the OPSD objective *ill-posed*: it contains an irreducible mutual information gap $I(Y_t; R \mid X, Y_{<t}) > 0$ that the student can never eliminate, regardless of its capacity (Theorem 1). At the gradient level, we show that while the expected OPSD gradient is benign, the per-sample gradients carry an r-specific deviation whose variance is proportional to this mutual information. These theoretical findings are directly corroborated by our diagnostic experiments: the on-policy KL divergence between teacher and student stagnates under OPSD with no sustained reduction, whereas the same metric decreases steadily under OPD, as shown in Figure 3. Early in training, the beneficial gradient component dominates, producing rapid improvement; as the student approaches the teacher's marginal distribution, the deviation takes over, and its path-dependent accumulation drives the model toward encoding $x \to r$ correlations. This two-phase dynamic precisely accounts for the observed pattern of early gains followed by progressive degradation.

Our analysis pinpoints the root cause: in all distribution-matching formulations, the teacher's privileged evaluation $P_T(y_t \mid r)$ enters the gradient *direction*, making leakage structurally unavoidable regardless of how the distillation target is compressed. Yet the evidence ratio $P_T(y_t)/P_S(y_t)$ also carries a useful signal: it measures how much the privileged information revises the model's belief about each token. The challenge is therefore not to discard this signal, but to change how it is used. A key insight underlying our design is that the signals governing update *direction* and update *magnitude* have asymmetric requirements: **the directional signal can be sparse but must be reliable**, as an erroneous direction harms the policy; **the magnitude signal, by contrast, benefits from being as dense as possible to enable fine-grained discrimination among tokens**.

We propose **RL**VR with **S**elf-**D**istillation (**RLSD**), which instantiates this principle by repurposing the teacher from a generative target to a *magnitude evaluator*. Specifically, the environment reward determines the *direction* of each token's update (reinforcement or penalization), while the teacher's evidence ratio modulates only the *magnitude*. This decoupling anchors gradient directions exclusively to the reliable environment reward, while retaining the teacher's dense, token-level assessment for fine-grained credit discrimination across token positions. Notably, prior approaches to token-level credit assignment, such as value-function estimation in PPO [9] and various credit assignment methods [10, 11], typically require training auxiliary networks or incur substantial additional overhead, yet still produce noisy estimates. In contrast, self-distillation provides a natural and virtually cost-free source of per-token credit information, requiring only a single additional forward pass. As summarized in Table 1, RLSD is the only paradigm that simultaneously achieves on-policy training, high token efficiency, rich update signals, and environment-anchored optimization, while serving as a drop-in replacement for the uniform advantage in standard GRPO without any auxiliary loss or model. Our contributions are as follows:

- We identify the root cause of OPSD's failure through controlled experiments and formal analysis, proving that distribution matching under information asymmetry induces an irreducible gap that drives privileged information leakage through the gradient structure.
- We propose RLSD, a new training paradigm that unifies the strengths of RLVR and OPSD: the reliable environment reward governs update directions, while the privileged teacher provides rich, token-level modulation of update magnitudes.
- Extensive experiments across 24 benchmarks spanning multimodal, video, and text-only reasoning demonstrate that RLSD consistently outperforms both RLVR and self-distillation baselines, sustaining improvement beyond the point where self-distillation degrades.

2 Preliminaries

GRPO. Given a question x, a language model π_θ generates a response $y = (y_1, \ldots, y_T)$, and a verifier provides a binary reward $R(x, y) \in \{0, 1\}$. GRPO [1] samples a group of G responses $\{y^{(i)}\}_{i=1}^{G}$ per question and computes a group-relative sequence-level advantage $A^{(i)} = (R(x, y^{(i)}) - \mu_G)/\sigma_G$, where μ_G, σ_G are the group mean and standard deviation. The policy is updated via the clipped surrogate objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\min\Big(\rho_t^{(i)}A^{(i)},\ \text{clip}(\rho_t^{(i)}, 1{-}\epsilon, 1{+}\epsilon)A^{(i)}\Big)\right], \tag{1}$$

with $\rho_t^{(i)} = \pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)})/\pi_{\theta_{\text{old}}}(y_t^{(i)} \mid x, y_{<t}^{(i)})$. A key limitation is that all tokens within a response share the same advantage $A^{(i)}$, as the reward is only sequence-level.

OPD and OPSD. On-Policy Distillation (OPD) [4] addresses sparse rewards by having the student π_θ sample on-policy trajectories while a larger teacher $\pi_{\hat{\theta}}$ provides dense token-level supervision, at the cost of maintaining a separate model. On-Policy Self-Distillation (OPSD) [7, 8] eliminates this overhead by instantiating both roles from π_θ: the teacher's advantage comes not from greater capacity, but from conditioning on privileged information r (e.g., a verified reasoning trace). Both share the same objective, minimizing a per-token divergence D between teacher and student along the student's trajectory $\hat{y} \sim P_S(\cdot \mid x)$:

$$\mathcal{L}_{\text{OP(S)D}}(\theta) = \mathbb{E}_{(x,r),\,\hat{y}\sim P_S}\left[\frac{1}{|\hat{y}|}\sum_{t=1}^{|\hat{y}|}D(P_T \,\|\, P_S)\right], \tag{2}$$

where $P_S(\cdot \mid y_{<t}) \triangleq \pi_\theta(\cdot \mid x, y_{<t})$, and the two frameworks differ only in the teacher: $P_T^{\text{OPD}} \triangleq \pi_{\hat{\theta}}(\cdot \mid x, y_{<t})$ versus $P_T^{\text{OPSD}} \triangleq \pi_\theta(\cdot \mid x, r, y_{<t})$. Gradients flow only through P_S.

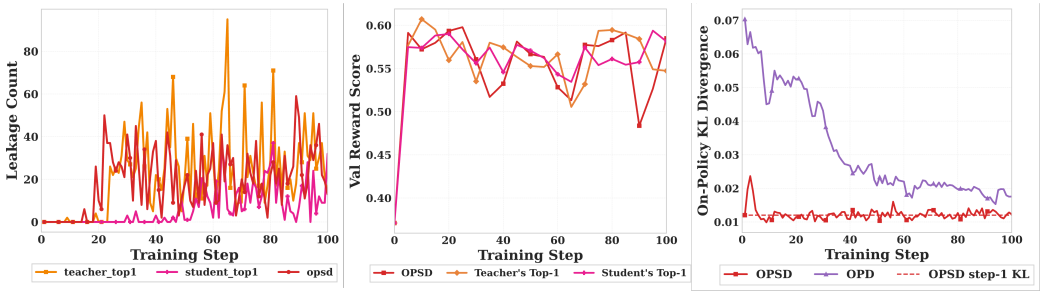
3 Why Does OPD Work While OPSD Fails?

3.1 Empirical Observations: Leakage and Performance Degradation

We document three empirical phenomena that motivate our formal analysis.

OPSD-tuned Model Output (✗):
I need to determine if the sample mean is within \$1 of the population mean. To do this, I must first calculate both means. . . . **<Omitted>** . . . Given that the **reference solution** uses 9 values, I will assume the sample has 9 prices . . . **<Omitted>**. . . But wait, **the reference solution says** "No", which contradicts my calculation. Let me re-examine. . . .

Figure 2: A representative example illustrating the privileged information leakage exhibited by the OPSD-trained model, where the model appeals to an invisible reference solution during inference.



(a) Leakage occurrence dynamics. **(b)** Performance on the validation set **(c)** KL Divergence comparison.

Figure 3: Leakage, KL divergence, and validation performance of OPSD and its ablated variants.

Privileged information leakage. OPSD-trained models systematically reference privileged information unavailable at inference time. Figure 2 presents a representative example in which the model explicitly appeals to an invisible "reference solution" during generation, and Figure 3(a) shows that the frequency of such leakage increases monotonically over training.

Performance degradation. Figure 3(b) reports validation accuracy peaking within the first 10–20 steps and subsequently declining, consistent with the escalating leakage.

KL divergence stagnation. Figure 3(c) contrasts the on-policy teacher–student KL divergence under OPD and OPSD. OPD's KL decreases steadily, reflecting genuine convergence; OPSD's drops briefly then plateaus near its initial value, suggesting an irreducible gap in the OPSD objective.

These observations raise two questions: *why* does OPSD induce leakage, and *why* does it fail to reduce the KL divergence? We answer both below.

3.2 OPSD's Failure: An Ill-Posed Objective

The irreducible mutual information gap. Let r denote the privileged information drawn from $P(r \mid x)$. Since any question x admits multiple semantically valid reasoning paths, $P(r \mid x)$ is non-degenerate. An optimal student that cannot condition on r should recover the *marginal* teacher distribution $\bar{P}_T(y_t) \triangleq \mathbb{E}_r[P_T(y_t \mid x, r, y_{<t})]$ via the law of total probability, yielding the *ideal* objective $\mathcal{L}^*(\theta) = \mathbb{E}_x[D_{\mathrm{KL}}(\bar{P}_T \,\|\, P_S)]$. However, OPSD enforces a *per-sample* match $P_S(\cdot \mid x) \to P_T(\cdot \mid x, r)$ for each concrete (x, r), forcing a conditionally independent parameterization to match a conditionally dependent target.

Theorem 1 (KL Decomposition)**.** *The OPSD objective and the ideal objective satisfy:*

$$\mathcal{L}_{\mathrm{OPSD}} \;=\; \mathcal{L}^* \;+\; I(Y_t;\, R \mid X,\, Y_{<t}). \tag{3}$$

The mutual information term is *independent of* θ: it is determined entirely by the teacher's conditional distributions and $P(r \mid x)$, and cannot be eliminated by any student within the feasible set. The global optimum is $P_S^* = \bar{P}_T$, at which the residual loss equals $I(Y_t; R \mid X, Y_{<t}) > 0$—a strictly positive, irreducible lower bound. This explains the KL stagnation in Figure 3(c) (the dashed line marks OPSD's KL at step 1, already near this floor) and, as we show next, this residual actively contaminates the gradient direction. Proof and extended discussion are in Appendix B.1.

Gradient structure: the mechanism of leakage. While Theorem 1 establishes that $I(Y_t; R \mid X)$ is θ-independent, which may suggest the *expected* OPSD gradient identical to the ideal marginal-matching gradient, *per-sample* gradients carry an r-specific deviation:

Proposition 1 (Per-Sample Gradient Decomposition)**.** *The per-sample gradient decomposes as*

$$g(\theta; r) = \underbrace{-\sum_v \bar{P}_T(v)\, \nabla_\theta \log P_S(v)}_{g^*(\theta)\text{: marginal matching}} + \underbrace{-\sum_v \big[P_T(v \mid r) - \bar{P}_T(v)\big]\, \nabla_\theta \log P_S(v)}_{\delta(\theta; r)\text{: } r\text{-specific deviation}}, \tag{4}$$

with $\mathbb{E}_r[\delta] = 0$ *and* $\mathbb{E}_r[\|\delta\|^2] = \sum_v \mathrm{Var}_r[P_T(v \mid r)] \cdot \|\nabla_\theta \log P_S(v)\|^2$*. The deviation vanishes when* $I(Y_t; R \mid X) = 0$ *and its variance increases monotonically with the mutual information.*

Although δ has zero mean, path-dependent optimizers (SGD, Adam) do not cancel these perturbations across iterations. Early in training, $\|g^*\| \gg \|\delta\|$ and the gradient drives genuine marginal matching, producing the steep rise in Figure 3(b). As $P_S \to \bar{P}_T$, $\|g^*\| \to 0$ while $\|\delta\|$ remains bounded away from zero (its variance is governed by the θ-independent mutual information). Parameter updates become dominated by δ, accumulating $x \to r$ correlations in θ and triggering self-reinforcing degradation. This two-phase dynamic precisely matches the late-stage decline in Figure 3(b) and the monotonically increasing leakage in Figure 3(a). Full derivation is provided in Appendix A.2.

Leakage bandwidth: a universal prediction. The gradient decomposition predicts that *any* variant in which $P_T(\cdot \mid r)$ enters the gradient direction must suffer from leakage, regardless of how the distillation target is compressed. To verify this, we ablate two variants that compress the target to a single token: *Teacher's Top-1* ($\arg\max_v P_T(v \mid r)$) and *Student's Top-1* ($\arg\max_v P_S(v)$). All three variants exhibit leakage (Figure 3(a,b)), with severity ordered by *leakage bandwidth*—the effective number of token positions at which r-specific information enters the gradient. We provide the full bandwidth analysis, irreducibility proof, and shared-parameter trilemma (objective stability, sustained improvement, and leakage-free training cannot hold simultaneously) in Appendices B.3 and B.4.

4 RLSD: Self-Distillation as RLVR's Wingman

The preceding analysis identifies a precise root cause: distribution matching fails because privileged information enters the *gradient direction*, contaminating the optimization trajectory. Yet the same quantity at the heart of this failure, the evidence ratio $P_T(y_t)/P_S(y_t)$, also carries a useful signal: it measures how much the privileged information revises the model's belief about each token. The challenge, then, is not to discard this signal, but to change *how* it is used. We propose RLVR with Self-Distillation (RLSD), which repurposes the teacher's role entirely: instead of serving as a generative target for distribution matching, the discrepancy between P_T and P_S is used as a token-level credit assignment signal within the policy gradient framework. **The privileged information influences only how much credit each token receives, not which tokens are reinforced or penalized, nor the direction of parameter updates**.

Step 1: Privileged information gain. Given a student-sampled trajectory $y = (y_1, \ldots, y_T)$, we compute the log-probability of each token under both the student context (x only) and the teacher context (x and r), and define the *privileged information gain* at each position:

$$\Delta_t = \mathrm{sg}(\log P_T(y_t) - \log P_S(y_t))\,, \tag{5}$$

where sg denotes the stop-gradient operator. Since teacher and student share the same model, Δ_t isolates the *marginal contribution of the privileged information* r to the prediction of y_t. A large positive Δ_t indicates that r strongly supports this token; a negative value indicates that r disfavors it. Crucially, Δ_t provides a dense, token-level signal that naturally reflects how much each token is informed by the privileged information, making it a principled and lightweight basis for fine-grained credit assignment within a trajectory. The stop-gradient ensures that Δ_t serves purely as a weighting signal and does not introduce auxiliary gradient pathways.

Step 2: Direction-aware evidence reweighting. We construct a per-token weight from the privileged information gain, modulated by the sign of the sequence-level advantage:

$$w_t = \exp(\mathrm{sign}(A) \cdot \Delta_t) = \left(\frac{P_T(y_t)}{P_S(y_t)}\right)^{\mathrm{sign}(A)}. \tag{6}$$

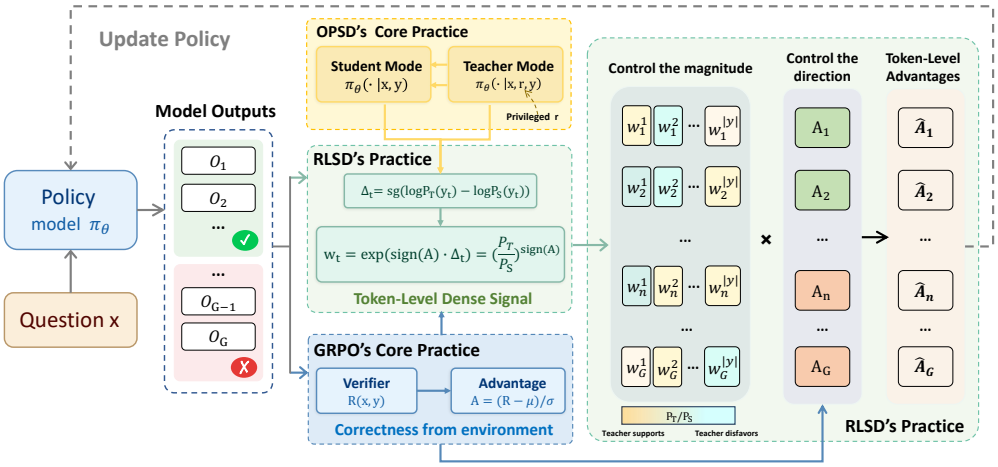


Figure 4: An overview of our RLSD method.

This formulation admits a natural Bayesian interpretation. $P_S(y_t)$ represents the model's *prior* assessment of token y_t, based solely on the question x. $P_T(y_t)$ represents the *posterior* assessment after observing the privileged information r. The ratio $P_T(y_t)/P_S(y_t)$ is therefore an *evidence ratio*: the factor by which the privileged information revises the model's belief about each token. Under mild modeling assumptions, this ratio can be shown to equal the Bayesian belief update $P(r \mid x, y_{\leq t})/P(r \mid x, y_{<t})$, i.e., the degree to which generating y_t increases the posterior probability that the privileged information r is consistent with the trajectory (Theorem 4 in Appendix B.5).

The $\mathrm{sign}(A)$ exponent implements direction-aware credit assignment. When $A > 0$, $w_t = P_T/P_S$: tokens that the privileged information supports receive larger weights, concentrating positive credit on the tokens most aligned with the correct reasoning trace. When $A < 0$, $w_t = P_S/P_T$: the ratio is inverted, so tokens that the privileged information *disfavors* bear greater blame, while tokens it supports receive attenuated punishment. Since $\exp(\cdot) > 0$ for all inputs, the weight is strictly positive, guaranteeing that the sign of the token-level advantage is never flipped by the reweighting. The environment reward retains exclusive authority over whether a trajectory is reinforced or penalized; the teacher only modulates the relative magnitude across tokens within a trajectory.

This design parallels the importance sampling ratio $\pi_\theta/\pi_{\text{old}}$ used in GRPO for policy updates. GRPO uses the ratio between current and old policies to control the *step size* of updates; RLSD uses the ratio between posterior and prior to control the *credit distribution* across tokens. Both are importance ratios operating within the same policy gradient framework, yielding a structurally unified formulation.

Step 3: Clipped credit assignment. Following the design philosophy of the clipped surrogate objective in GRPO, we clip the evidence weights to bound the maximum influence of single token:

$$\hat{A}_t = A \cdot \mathrm{clip}(w_t,\ 1-\epsilon_w,\ 1+\epsilon_w)\,, \tag{7}$$

where ϵ_w bounds the per-token credit deviation. The clipping in Eq. (7) serves an analogous role to the importance ratio clipping in GRPO, acting as trust-region constraints that stabilize training. In practice, to avoid an abrupt transition at the start of training, we linearly interpolate between the uniform advantage and the reweighted advantage over the training steps using $\lambda \in [0, 1]$, gradually shifting to the uniform advantage. The final objective of RLSD is as follows:

$$\mathcal{L}_{\text{RLSD}}(\theta) = \mathbb{E}\left\{\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\min\left[\hat{w}_t A^{(i)},\ \mathrm{clip}(\hat{w}_t, 1-\epsilon_w, 1+\epsilon_w)\,A^{(i)}\right]\right\}, \tag{8}$$

where $\hat{w}_t = (1-\lambda) + \lambda\, w_t$. RLSD serves as a drop-in replacement for the uniform advantage in standard GRPO, introducing no auxiliary loss and adding only one teacher forward pass per response, which is negligible relative to rollout generation. The full procedure is detailed in Algorithm 1 (Appendix D). From a broader perspective, RLSD, GRPO, and on-policy self-distillation can all be expressed as instances of a single token-level policy gradient template differing only in how the token-level advantage $\hat{A}_t$ is defined; we provide this unified view in Appendix C.

5 Experiment

5.1 Experimental Setup

Training data and benchmarks. Our primary evaluation focuses on **multimodal reasoning**, where we train on MMFineReason-123K [12], a challenging subset of MMFineReason-1.8M obtained via difficulty-based filtering, and evaluate on eight benchmarks spanning diverse mathematical and general reasoning capabilities. MathVista [13], MathVision [14], WeMath [15], and MathVerse [16] target mathematical reasoning in visual contexts; MMMU-Pro [17] and MMBench [18] probe broad multi-discipline knowledge and general multimodal competence; MM-Star [19] stresses fine-grained visual discrimination; and ZeroBench [20] provides a stress test for reasoning robustness against frontier-model failures. To assess generalization, we further evaluate RLSD on **video understanding** and **text-only reasoning**, each spanning eight additional benchmarks. We report accuracy (%) and the unweighted mean *Avg.*. Detailed dataset and benchmark descriptions are provided in Appendix E.

Models and Baselines. We conduct experiments on Qwen3-VL-8B-Instruct and Qwen3-VL-4B-Instruct [21] for multimodal reasoning, Qwen3-VL-8B-Instruct for video understanding, and Qwen3-8B (Non-thinking) for text-only reasoning. We compare RLSD against the following baselines. **GRPO** [1] is a standard RLVR method that estimates token-level advantages from sparse scalar outcome rewards at the sequence level, serving as our primary RL baseline. **OPSD** [7] is an on-policy self-distillation method where a single model acts as both teacher and student, with the teacher conditioned on privileged information (e.g., verified reasoning traces) to provide dense per-token supervision. **SDPO** [8] extends self-distillation to the reinforcement learning with rich feedback setting, treating the current model conditioned on environment feedback as a self-teacher to derive dense credit assignment signals. **GRPO+OPSD** is a straightforward combination baseline inspired by the MOPD approach proposed in MIMO-v2-Flash [6], which jointly optimizes reinforcement learning and distillation objectives. Specifically, we combine the GRPO loss with a distillation KL divergence loss via linear interpolation with a tuned weighting coefficient. Unlike MOPD, which relies on a separately trained expert model for distillation, we adopt the OPSD formulation by using the model's own rollout outputs as the distillation target, thereby eliminating the need for an external expert while still leveraging dense self-distillation signals alongside sparse environment rewards.

Implementation Details. Due to space constraints, we defer the full implementation details to Appendix F, which covers training configuration, learning rates, GRPO-based hyperparameters, RLSD-specific hyperparameters, privileged information construction, and hardware setup.

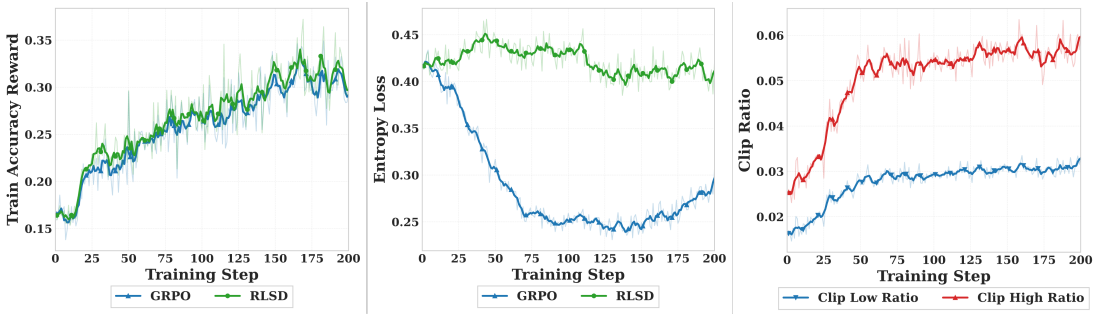
5.2 Experimental Results

Multimodal reasoning across model scales. Table 2 reports results across eight multimodal reasoning benchmarks on Qwen3-VL-Instruct at both 8B and 4B scales. RLSD achieves the highest average accuracy on *both* model sizes, outperforming the Base LLM by 4.78% / 2.34% and GRPO by 2.41% / 0.84% on 8B / 4B respectively, demonstrating that the benefits of decoupled credit assignment hold consistently across model capacities. Gains are most pronounced on challenging mathematical benchmarks such as MathVision (+7.30% over GRPO on 8B) and WeMath (+5.72%), where fine-grained discrimination of reasoning steps matters most. RLSD also consistently surpasses the self-distillation baselines OPSD and SDPO at both scales, validating our theoretical diagnosis (§3.2): rather than suffering from distribution-matching leakage, RLSD anchors the update direction to the verifier reward while extracting only the teacher's dense magnitude signal. Finally, RLSD outperforms the additive fusion baseline GRPO+OPSD by 2.56 points on 8B. Linearly combining bounded rewards with unbounded, high-variance KL losses incurs severe scale mismatch and destabilizes training; RLSD circumvents this through *multiplicative* modulation, exponentiating the teacher's signal into a dynamically bounded multiplier that mathematically guarantees sign preservation.

Generalization to other domains. To assess whether RLSD's gains transfer beyond static-image reasoning, we further evaluate on two additional domains. On **video understanding** (Table 3, eight benchmarks on Qwen3-VL-8B), RLSD outperforms GRPO by 1.25 points on average, with consistent gains across long-video and reasoning-heavy tasks where pivotal tokens are sparsely distributed and dense credit assignment is especially valuable. On **text-only reasoning** (Table 3, eight mathematical and scientific benchmarks on Qwen3-8B (Non-thinking)), RLSD further improves over GRPO by 1.98 points on average. Together, these results confirm that the underlying mechanism of RLSD applies broadly across modalities and model families.

Table 2: Multimodal reasoning results on the Qwen3-VL-Instruct models. **Bold** indicates the best result among all models. Avg. is computed over the eight selected benchmarks.

Method	MMMU-Pro	MathVista	MathVision	ZeroBench$_{sub}$	WeMath	MMBench	MM-Star	MathVerse	Avg.
Qwen3-VL-8B-Instruct									
Base LLM	51.75	73.80	47.37	19.76	54.10	89.79	71.83	54.61	57.88
GRPO	55.78	76.20	48.82	22.60	56.57	90.29	72.20	59.52	60.25
OPSD	53.58	75.10	47.53	21.06	54.95	89.70	71.57	58.03	58.94
SDPO	52.64	74.00	47.27	**25.15**	52.19	90.03	72.21	58.44	58.99
GRPO+OPSD	55.78	75.90	48.52	22.16	54.76	**90.53**	**73.53**	59.59	60.10
RLSD *(Ours)*	**57.73**	**78.50**	**56.12**	22.90	**62.29**	90.41	73.47	**59.82**	**62.66**
Qwen3-VL-4B-Instruct									
Base LLM	47.42	67.95	45.39	18.91	48.86	88.17	66.43	53.48	54.58
GRPO	48.70	68.75	49.10	19.21	50.57	**89.10**	67.57	55.62	56.08
OPSD	**49.23**	**70.60**	47.40	19.16	47.90	89.02	66.67	54.31	55.54
SDPO	48.48	69.30	47.93	19.61	46.00	88.99	67.43	53.23	55.12
GRPO+OPSD	49.04	69.95	48.29	19.76	47.71	88.72	67.83	55.93	55.90
RLSD *(Ours)*	48.69	69.70	**49.97**	**20.36**	**52.48**	89.06	**68.27**	**56.85**	**56.92**



(a) Training reward dynamics comparison. **(b)** Training entropy dynamics comparison. **(c)** Clip ratios on credit assignment

Figure 5: Training dynamics on the multimodel reasoning tasks.

5.3 Analysis

Training Dynamics. Figure 5 illustrates the training dynamics over 200 optimization steps. As shown in Figure 5(a), RLSD demonstrates a steeper initial ascent and converges to a higher accuracy reward ceiling compared to standard GRPO. Figure 5(b) reveals that GRPO suffers from rapid entropy collapse, whereas RLSD maintains a consistently higher entropy level by selectively strengthening critical reasoning tokens without uniformly suppressing alternatives at every position. Finally, Figure 5(c) confirms that the clipped credit assignment mechanism is actively engaged, with clip ratios stabilizing around 3%–6%, successfully bounding the teacher's per-token influence and acting as a trust-region constraint analogous to the importance ratio clipping in GRPO.

Ablation studies. We conduct comprehensive ablations on RLSD's design choices, with full results in Appendix H. We summarize the key findings here. **(i) Robustness to clipping threshold ϵ_w:** performance is stable across $\epsilon_w \in \{0.2, 0.3, 0.4\}$ (62.50–62.66 average), indicating that RLSD does not require careful tuning of the trust-region bound. **(ii) Robustness to mixing coefficient λ:** initial values $\lambda \in \{0.5, 0.75\}$ both yield ≈ 62.7 average, while $\lambda = 1.0$ slightly degrades to 62.19, consistent with our design intuition that overly aggressive teacher modulation is suboptimal. **(iii) Sensitivity to decay schedule:** the average accuracy degrades monotonically from 62.66 to 60.41 as the λ-decay step extends from 60 to 200. We attribute this to a dynamically narrowing teacher–student capability gap: as training progresses, the student's reasoning ability improves continuously through RLVR optimization, whereas the teacher does not directly benefit from this optimization. The effective informational advantage of the teacher therefore diminishes over time, and prolonging its guidance beyond the early phase contaminates the policy with increasingly stale credit signals. **(iv) Privileged information granularity:** replacing the verified chain-of-thought with only the ground-truth answer yields nearly identical performance (62.77 vs. 62.66), demonstrating

Table 3: Generalization of RLSD beyond multimodal reasoning. **(a)** Video understanding on Qwen3-VL-8B-Instruct across eight benchmarks. **(b)** Text-only reasoning on Qwen3-8B (Non-thinking) across eight mathematical and scientific benchmarks.

Method	VideoMME	Tempcompass	MotionBench	LongVB	VCRBench	LongVR	MMVU	VideoMathQA	Avg.
GRPO	**70.11**	75.98	57.01	64.85	44.00	78.61	60.40	50.33	62.66
RLSD *(Ours)*	69.44	**76.53**	**57.16**	**65.97**	**45.74**	**79.32**	**65.20**	**51.95**	**63.91**
Method	**GSM8K**	**MATH-500**	**AMC**	**AIME24**	**AIME25**	**Minerva**	**OlympiadBench**	**GPQA**	**Avg.**
GRPO	**96.00**	92.00	89.06	58.13	46.67	51.29	60.74	58.46	69.04
RLSD *(Ours)*	95.60	**93.20**	**92.66**	**61.46**	**47.50**	**53.13**	**65.93**	**58.71**	**71.02**

that RLSD's credit signal does not depend on expensive verified reasoning traces—a single answer suffices. This makes RLSD substantially more practical than OPSD and SDPO, which require verified CoT or successful prior rollouts.

Case Study. Figure 6 (Appendix I) provides a qualitative view of how RLSD redistributes sequence-level reward to the token level. In correct trajectories, larger credit concentrates on tokens that actually determine correctness, while generic narration is down-weighted; in incorrect trajectories, the strongest blame falls on the erroneous reasoning steps rather than neutral setup tokens. This pattern is consistent with the design goal of RLSD: the environment reward determines whether a trajectory is reinforced or penalized, while the privileged teacher modulates the relative magnitude of token-level credit according to each token's contribution to the final outcome.

6 Related Work

Credit assignment in RLVR. RLVR methods such as GRPO [1] provide only sequence-level signals, which is a fundamental limitation for long-horizon reasoning where fine-grained credit assignment is essential [2]. Process reward models and step-level value estimators address this with auxiliary supervision [22–27], but require additional reward modeling. A recent line of work instead reshapes token-level updates using model-internal proxies such as entropy, uncertainty, or attention dynamics [28–31, 10, 32, 33, 11]. Our method is complementary in motivation but differs in mechanism: rather than heuristic proxies, we extract token-level credit from self-distillation under privileged context, while anchoring update directions to the verifier reward.

On-policy distillation. On-policy distillation (OPD) provides dense token-level supervision via a stronger teacher [4, 5, 34, 6], but requires maintaining a separate, typically larger model. Self-distillation variants eliminate this overhead by letting the same model serve as both teacher and student, conditioned on privileged information such as reference solutions, verifier feedback, demonstrations, or conciseness instructions [7, 8, 35–38]. A concurrent work, TRRD [39], also identifies the conflict between distillation and reward maximization, but operates on the trust region anchor rather than the advantage magnitude. All these methods retain a distribution-matching objective. We depart from this paradigm: when the teacher conditions on information unavailable to the student, distribution matching becomes structurally ill-posed. RLSD therefore repurposes the privileged discrepancy as a scalar credit multiplier rather than a generative target.

7 Conclusion

In this work, we identified the fundamental limitation of on-policy self-distillation (OPSD): the information asymmetry between teacher and student renders the distribution-matching objective structurally ill-posed, inducing an irreducible mutual information gap that drives privileged information leakage and progressive performance degradation. We formalized this deficiency at both the objective level and the gradient level, explaining the empirically observed pattern of early gains followed by systematic collapse. Based on this analysis, we proposed RLSD, a new training paradigm that moves beyond distribution matching between the prior and posterior: instead of forcing the student to imitate the teacher's conditional distribution, RLSD repurposes the discrepancy between them to provide fine-grained control over per-token update magnitudes, while anchoring the update directions to the environment reward. This design simultaneously inherits the rich, token-level supervision of OPSD and the reliable environmental grounding of RLVR, unifying the strengths of both paradigms within a single framework. We discuss limitations and future directions in Appendix J.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62472419, 62472420)

References

[1] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.03300

[2] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[3] K. Team, T. Bai, Y. Bai, Y. Bao, S. H. Cai, Y. Cao, Y. Charles, H. S. Che, C. Chen, G. Chen, H. Chen, J. Chen, J. Chen, J. Chen, J. Chen, K. Chen, L. Chen, R. Chen, X. Chen, Y. Chen, Y. Chen, Y. Chen, Y. Chen, Y. Chen, Y. Chen, Y. Chen, Z. Chen, Z. Chen, D. Cheng, M. Chu, J. Cui, J. Deng, M. Diao, H. Ding, M. Dong, M. Dong, Y. Dong, Y. Dong, A. Du, C. Du, D. Du, L. Du, Y. Du, Y. Fan, S. Fang, Q. Feng, Y. Feng, G. Fu, K. Fu, H. Gao, T. Gao, Y. Ge, S. Geng, C. Gong, X. Gong, Z. Gongque, Q. Gu, X. Gu, Y. Gu, L. Guan, Y. Guo, X. Hao, W. He, W. He, Y. He, C. Hong, H. Hu, J. Hu, Y. Hu, Z. Hu, K. Huang, R. Huang, W. Huang, Z. Huang, T. Jiang, Z. Jiang, X. Jin, Y. Jing, G. Lai, A. Li, C. Li, C. Li, F. Li, G. Li, G. Li, H. Li, H. Li, J. Li, J. Li, J. Li, L. Li, M. Li, W. Li, W. Li, X. Li, X. Li, Y. Li, Y. Li, Y. Li, Y. Li, Z. Li, Z. Li, W. Liao, J. Lin, X. Lin, Z. Lin, Z. Lin, C. Liu, C. Liu, H. Liu, L. Liu, S. Liu, S. Liu, S. Liu, T. Liu, T. Liu, W. Liu, X. Liu, Y. Liu, Y. Liu, Y. Liu, Y. Liu, Y. Liu, Z. Liu, Z. Liu, E. Lu, H. Lu, Z. Lu, J. Luo, T. Luo, Y. Luo, L. Ma, Y. Ma, S. Mao, Y. Mei, X. Men, F. Meng, Z. Meng, Y. Miao, M. Ni, K. Ouyang, S. Pan, B. Pang, Y. Qian, R. Qin, Z. Qin, J. Qiu, B. Qu, Z. Shang, Y. Shao, T. Shen, Z. Shen, J. Shi, L. Shi, S. Shi, F. Song, P. Song, T. Song, X. Song, H. Su, J. Su, Z. Su, L. Sui, J. Sun, J. Sun, T. Sun, F. Sung, Y. Tai, C. Tang, H. Tang, X. Tang, Z. Tang, J. Tao, S. Teng, C. Tian, P. Tian, A. Wang, B. Wang, C. Wang, C. Wang, C. Wang, D. Wang, D. Wang, D. Wang, F. Wang, H. Wang, H. Wang, H. Wang, H. Wang, H. Wang, J. Wang, J. Wang, J. Wang, K. Wang, L. Wang, Q. Wang, S. Wang, S. Wang, S. Wang, W. Wang, X. Wang, X. Wang, Y. Wang, Y. Wang, Y. Wang, Y. Wang, Y. Wang, Y. Wang, Z. Wang, Z. Wang, Z. Wang, Z. Wang, Z. Wang, Z. Wang, C. Wei, M. Wei, C. Wen, Z. Wen, C. Wu, H. Wu, J. Wu, R. Wu, W. Wu, Y. Wu, Y. Wu, Y. Wu, Z. Wu, C. Xiao, J. Xie, X. Xie, Y. Xie, Y. Xin, B. Xing, B. Xu, J. Xu, J. Xu, J. Xu, L. H. Xu, L. Xu, S. Xu, W. Xu, X. Xu, X. Xu, Y. Xu, Y. Xu, Y. Xu, Z. Xu, Z. Xu, J. Yan, Y. Yan, G. Yang, H. Yang, J. Yang, K. Yang, N. Yang, R. Yang, X. Yang, X. Yang, Y. Yang, Y. Yang, Y. Yang, Z. Yang, Z. Yang, Z. Yang, H. Yao, D. Ye, W. Ye, Z. Ye, B. Yin, C. Yu, L. Yu, T. Yu, T. Yu, E. Yuan, M. Yuan, X. Yuan, Y. Yue, W. Zeng, D. Zha, H. Zhan, D. Zhang, H. Zhang, J. Zhang, P. Zhang, Q. Zhang, R. Zhang, X. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Z. Zhang, C. Zhao, F. Zhao, J. Zhao, S. Zhao, X. Zhao, Y. Zhao, Z. Zhao, H. Zheng, R. Zheng, S. Zheng, T. Zheng, J. Zhong, L. Zhong, W. Zhong, M. Zhou, R. Zhou, X. Zhou, Z. Zhou, J. Zhu, L. Zhu, X. Zhu, Y. Zhu, Z. Zhu, J. Zhuang, W. Zhuang, Y. Zou, and X. Zu, "Kimi k2.5: Visual agentic intelligence," 2026. [Online]. Available: https://arxiv.org/abs/2602.02276

[4] R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. R. Garea, M. Geist, and O. Bachem, "On-policy distillation of language models: Learning from self-generated mistakes," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=3zKtaqxLhW

[5] K. Lu and T. M. Lab, "On-policy distillation," *Thinking Machines Lab: Connectionism*, 2025, https://thinkingmachines.ai/blog/on-policy-distillation.

[6] C. Team, B. Xiao, B. Xia, B. Yang, B. Gao, B. Shen, C. Zhang, C. He, C. Lou, F. Luo, G. Wang, G. Xie, H. Zhang, H. Lv, H. Li, H. Chen, H. Xu, H. Zhang, H. Liu, J. Duo, J. Wei, J. Xiao, J. Dong, J. Shi, J. Hu, K. Bao, K. Zhou, L. Li, L. Zhao, L. Zhang, P. Li, Q. Chen, S. Liu, S. Yu, S. Cao, S. Chen, S. Yu, S. Liu, T. Zhou, W. Su, W. Wang, W. Ma, X. Deng, B. Mao, B. Ye, C. Cai, C. Wang, C. Zhu, C. Ma, C. Chen, C. Li, D. Zhu, D. Xiao, D. Zhang, D. Zhang, F. Liu, F. Yang, F. Shi, G. Wang, H. Tian, H. Wu, H. Qu, H. Yi, H. An, H. Guan, X. Zhang, Y. Song, Y. Yan, Y. Zhao, Y. Lai, Y. Gao, Y. Cheng, Y. Tian, Y. Wang, Z. Tang, Z. Tang, Z. Wen, Z. Song, Z. Zheng, Z. Jiang, J. Wen, J. Sun, J. Li, J. Xue, J. Xia, K. Fang, M. Zhu, N. Chen, Q. Tu, Q. Zhang, Q. Wang, R. Li, R. Ma, S. Zhang, S. Wang, S. Li, S. Gu, S. Ren, S. Deng, T. Guo, T. Lu, W. Zhuang, W. Zhang, W. Xiong, W. Huang, W. Yang, X. Zhang, X. Yong, X. Wang, X. Xie, Y. Jiang, Y. Yang, Y. He, Y. Tu, Y. Dong, Y. Liu, Y. Ma, Y. Yu, Y. Xiang, Z. Huang, Z. Lin, Z. Xu, Z. Chen, Z. Deng, Z. Zhang, and Z. Yue, "Mimo-v2-flash technical report," 2026. [Online]. Available: https://arxiv.org/abs/2601.02780

[7] S. Zhao, Z. Xie, M. Liu, J. Huang, G. Pang, F. Chen, and A. Grover, "Self-distilled reasoner: On-policy self-distillation for large language models," 2026. [Online]. Available: https://arxiv.org/abs/2601.18734

[8] J. Hübotter, F. Lübeck, L. Behric, A. Baumann, M. Bagatella, D. Marta, I. Hakimi, I. Shenfeld, T. K. Buening, C. Guestrin, and A. Krause, "Reinforcement learning via self-distillation," 2026. [Online]. Available: https://arxiv.org/abs/2601.20802

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[10] C. Xie, R. Pan, X. Wu, Y. Zhang, J. Fu, T. Gao, and G. Zhou, "Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning," *arXiv preprint arXiv:2510.10649*, 2025.

[11] Z. Li, L. Kang, F. Xiao, L. Xing, Q. Si, Z. Li, W. Gong, D. Yang, Y. Xiao, and H. Guo, "Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning," *arXiv preprint arXiv:2601.07408*, 2026.

[12] H. Lin, Z. Liu, Y. Zhu, C. Qin, J. Lin, X. Shang, C. He, W. Zhang, and L. Wu, "Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods," 2026. [Online]. Available: https://arxiv.org/abs/2601.21821

[13] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts," 2024. [Online]. Available: https://arxiv.org/abs/2310.02255

[14] K. Wang, J. Pan, W. Shi, Z. Lu, M. Zhan, and H. Li, "Measuring multimodal mathematical reasoning with math-vision dataset," 2024. [Online]. Available: https://arxiv.org/abs/2402.14804

[15] R. Qiao, Q. Tan, G. Dong, M. Wu, C. Sun, X. Song, Z. GongQue, S. Lei, Z. Wei, M. Zhang, R. Qiao, Y. Zhang, X. Zong, Y. Xu, M. Diao, Z. Bao, C. Li, and H. Zhang, "We-math: Does your large multimodal model achieve human-like mathematical reasoning?" 2024. [Online]. Available: https://arxiv.org/abs/2407.01284

[16] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, Y. Qiao *et al.*, "Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?" in *European Conference on Computer Vision*. Springer, 2024, pp. 169–186.

[17] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," 2024. [Online]. Available: https://arxiv.org/abs/2311.16502

[18] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *European conference on computer vision*. Springer, 2024, pp. 216–233.

[19] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin *et al.*, "Are we on the right way for evaluating large vision-language models?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 27 056–27 087, 2024.

[20] J. Roberts, M. R. Taesiri, A. Sharma, A. Gupta, S. Roberts, I. Croitoru, S.-V. Bogolin, J. Tang, F. Langer, V. Raina, V. Raina, H. Xiong, V. Udandarao, J. Lu, S. Chen, S. Purkis, T. Yan, W. Lin, G. Shin, Q. Yang, A. T. Nguyen, D. I. Atkinson, A. Baranwal, A. Coca, M. Dang, S. Dziadzio, J. D. Kunz, K. Liang, A. Lo, B. Pulfer, S. Walton, C. Yang, K. Han, and S. Albanie, "Zerobench: An impossible visual benchmark for contemporary large multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2502.09696

[21] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2511.21631

[22] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=v8L0pN6EOi

[23] P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui, "Math-shepherd: Verify and reinforce llms step-by-step without human annotations," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9426–9439.

[24] L. Luo, Y. Liu, R. Liu, S. Phatale, M. Guo, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng *et al.*, "Improve mathematical reasoning in language models by automated process supervision," *arXiv preprint arXiv:2406.06592*, 2024.

[25] G. Chen, M. Liao, C. Li, and K. Fan, "Step-level value preference optimization for mathematical reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 7889–7903.

[26] L. Zhang, A. Hosseini, H. Bansal, M. Kazemi, A. Kumar, and R. Agarwal, "Generative verifiers: Reward modeling as next-token prediction," *arXiv preprint arXiv:2408.15240*, 2024.

[27] G. Cui, L. Yuan, Z. Wang, H. Wang, Y. Zhang, J. Chen, W. Li, B. He, Y. Fan, T. Yu *et al.*, "Process reinforcement through implicit rewards," *arXiv preprint arXiv:2502.01456*, 2025.

[28] D. Cheng, S. Huang, X. Zhu, B. Dai, X. Zhao, Z. Zhang, and F. Wei, "Reasoning with exploration: An entropy perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 36, 2026, pp. 30 377–30 385.

[29] S. Wang, L. Yu, C. Gao, C. Zheng, S. Liu, R. Lu, K. Dang, X.-H. Chen, J. Yang, Z. Zhang *et al.*, "Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[30] M. Chen, G. Chen, W. Wang, and Y. Yang, "Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization," *arXiv preprint arXiv:2505.12346*, 2025.

[31] W. Sun, W. Yang, P. Jian, Q. Du, F. Cui, S. Ren, and J. Zhang, "Ktae: A model-free algorithm to key-tokens advantage estimation in mathematical reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2505.16826

[32] Y. Li, Z. Dong, Y. Sun, W. Wang, S. Xiong, Y. Luo, J. Liu, H. Lu, J. Wang, W. Su *et al.*, "Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization," *arXiv preprint arXiv:2510.13554*, 2025.

[33] X. Chen, X. Li, Z. Sun, and W. Yu, "Beyond high-entropy exploration: Correctness-aware low-entropy segment-based advantage shaping for reasoning llms," *arXiv preprint arXiv:2512.00908*, 2025.

[34] Y. Fu, H. Huang, K. Jiang, Y. Zhu, and D. Zhao, "Revisiting on-policy distillation: Empirical failure modes and simple fixes," 2026. [Online]. Available: https://arxiv.org/abs/2603.25562

[35] I. Shenfeld, M. Damani, J. Hübotter, and P. Agrawal, "Self-distillation enables continual learning," in *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026. [Online]. Available: https://openreview.net/forum?id=HlWA3V6iKF

[36] T. Ye, L. Dong, X. Wu, S. Huang, and F. Wei, "On-policy context distillation for language models," *arXiv preprint arXiv:2602.12275*, 2026.

[37] H. Sang, Y. Xu, Z. Zhou, R. He, Z. Wang, and J. Sun, "On-policy self-distillation for reasoning compression," *arXiv preprint arXiv:2603.05433*, 2026.

[38] E. Penaloza, D. Vattikonda, N. Gontier, A. Lacoste, L. Charlin, and M. Caccia, "Privileged information distillation for language models," in *The 1st Workshop on Scaling Post-training for LLMs*, 2026. [Online]. Available: https://openreview.net/forum?id=FbJu6NEBQR

[39] Z. Zhang, S. Jiang, Y. Shen, Y. Zhang, D. Ram, S. Yang, Z. Tu, W. Xia, and S. Soatto, "Reinforcement-aware knowledge distillation for llm reasoning," 2026. [Online]. Available: https://arxiv.org/abs/2602.22495

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[41] C. An, Z. Xie, X. Li, L. Li, J. Zhang, S. Gong, M. Zhong, J. Xu, X. Qiu, M. Wang, and L. Kong, "Polaris: A post-training recipe for scaling reinforcement learning on advanced reasoning models," 2025. [Online]. Available: https://hkunlp.github.io/blog/2025/Polaris

[42] M. Luo, S. Tan, J. Wong, X. Shi, W. Y. Tang, M. Roongta, C. Cai, J. Luo, L. E. Li, R. A. Popa, and I. Stoica, "Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl," https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2, 2025, notion Blog.

[43] W. Fu, J. Gao, X. Shen, C. Zhu, Z. Mei, C. He, S. Xu, G. Wei, J. Mei, J. Wang, T. Yang, B. Yuan, and Y. Wu, "Areal: A large-scale asynchronous reinforcement learning system for language reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2505.24298

[44] K. Feng, M. Zhang, H. Li, K. Fan, S. Chen, Y. Jiang, D. Zheng, P. Sun, Y. Zhang, H. Sun *et al.*, "Onethinker: All-in-one reasoning model for image and video," *arXiv preprint arXiv:2512.03043*, 2025.

[45] X. Li, Z. Yan, D. Meng, L. Dong, X. Zeng, Y. He, Y. Wang, Y. Qiao, Y. Wang, and L. Wang, "Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning," 2025. [Online]. Available: https://arxiv.org/abs/2504.06958

[46] K. Feng, K. Gong, B. Li, Z. Guo, Y. Wang, T. Peng, B. Wang, and X. Yue, "Video-r1: Reinforcing video reasoning in mllms," *arXiv preprint arXiv:2503.21776*, 2025.

[47] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, P. Chen, Y. Li, S. Lin, S. Zhao, K. Li, T. Xu, X. Zheng, E. Chen, C. Shan, R. He, and X. Sun, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," 2025. [Online]. Available: https://arxiv.org/abs/2405.21075

[48] Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, and L. Hou, "Tempcompass: Do video llms really understand videos?" 2024. [Online]. Available: https://arxiv.org/abs/2403.00476

[49] W. Hong, Y. Cheng, Z. Yang, W. Wang, L. Wang, X. Gu, S. Huang, Y. Dong, and J. Tang, "Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models," 2025. [Online]. Available: https://arxiv.org/abs/2501.02955

[50] H. Wu, D. Li, B. Chen, and J. Li, "Longvideobench: A benchmark for long-context interleaved video-language understanding," 2024. [Online]. Available: https://arxiv.org/abs/2407.15754

[51] Y. Qi, Y. Zhao, Y. Zeng, X. Bao, W. Huang, L. Chen, Z. Chen, J. Zhao, Z. Qi, and F. Zhao, "Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2504.07956

[52] Y. Chen, W. Huang, B. Shi, Q. Hu, H. Ye, L. Zhu, Z. Liu, P. Molchanov, J. Kautz, X. Qi, S. Liu, H. Yin, Y. Lu, and S. Han, "Scaling rl to long videos," 2025. [Online]. Available: https://arxiv.org/abs/2507.07966

[53] Y. Zhao, L. Xie, H. Zhang, G. Gan, Y. Long, Z. Hu, T. Hu, W. Chen, C. Li, J. Song, Z. Xu, C. Wang, W. Pan, Z. Shangguan, X. Tang, Z. Liang, Y. Liu, C. Zhao, and A. Cohan, "Mmvu: Measuring expert-level multi-discipline video understanding," 2025. [Online]. Available: https://arxiv.org/abs/2501.12380

[54] H. Rasheed, A. Shaker, A. Tang, M. Maaz, M.-H. Yang, S. Khan, and F. S. Khan, "Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos," 2025. [Online]. Available: https://arxiv.org/abs/2506.05349

[55] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems," 2021. [Online]. Available: https://arxiv.org/abs/2110.14168

[56] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," 2021. [Online]. Available: https://arxiv.org/abs/2103.03874

[57] AI-MO, "Amc 2023," 2024. [Online]. Available: https://huggingface.co/datasets/AI-MO/aimo-validation-amc

[58] MAA Committees, "Aime problems and solutions," https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

[59] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, and V. Misra, "Solving quantitative reasoning problems with language models," 2022. [Online]. Available: https://arxiv.org/abs/2206.14858

[60] C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun, "Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems," 2024. [Online]. Available: https://arxiv.org/abs/2402.14008

[61] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "Gpqa: A graduate-level google-proof q&a benchmark," 2023. [Online]. Available: https://arxiv.org/abs/2311.12022

[62] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu, "Hybridflow: A flexible and efficient rlhf framework," in *Proceedings of the Twentieth European Conference on Computer Systems*. ACM, Mar. 2025, p. 1279–1297. [Online]. Available: http://dx.doi.org/10.1145/3689031.3696075

[63] Y. Zheng, J. Lu, S. Wang, Z. Feng, D. Kuang, Y. Xiong, and R. Zhang, "Easyr1: An efficient, scalable, multi-modality rl training framework," https://github.com/hiyouga/EasyR1, 2025.

A Extended Analysis of OPSD's Failure

This appendix expands the formal analysis of §3.2. Appendix A.1 elaborates on the objective-level analysis, Appendix A.2 formalizes the two-phase training dynamics, and Appendix A.3 provides the full leakage bandwidth comparison.

A.1 Elaboration on the Irreducible Mutual Information Gap

We expand the discussion in §3.2 on why the mutual information gap in Theorem 1 is irreducible by the student's optimization.

On the non-degeneracy of $P(r \mid x)$. Even when each training instance x_i is paired with a single reference trace r_i, from the epistemic perspective of the student—which can neither observe r nor deterministically derive it from x—the privileged information remains an uncertain latent variable. $P(r \mid x)$ should therefore be treated as non-degenerate within the probabilistic modeling framework.

Formal statement of optimality. The optimal student policy that cannot condition on r recovers the marginal distribution

$$P_S^*(y_t \mid x, y_{<t}) = \mathbb{E}_{r \sim P(r|x,y_{<t})}[P_T(y_t \mid x, r, y_{<t})]\,, \tag{9}$$

and the ideal distillation objective is $\mathcal{L}^*(\theta) = \mathbb{E}_x[D_{\mathrm{KL}}(\bar{P}_T(\cdot) \,\|\, P_S(\cdot \mid x))]$. The OPSD objective enforces instead the per-sample match

$$\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_x\, \mathbb{E}_{r \sim P(r|x)}[D_{\mathrm{KL}}(P_T(\cdot \mid x, r) \,\|\, P_S(\cdot \mid x))]\,. \tag{10}$$

Why the gap is irreducible. Within the feasible set $\mathcal{F} = \{Q : Q(\cdot \mid x, y_{<t})$ does not condition on $r\}$, the global optimum is $P_S^* = \bar{P}_T$, at which the residual loss equals $I(Y_t; R \mid X, Y_{<t}) > 0$. This residual grows with the informativeness of the privileged signal and cannot be reduced through legitimate optimization.

How the residual drives leakage. Under a biased objective, the optimizer continues to receive a non-vanishing loss signal and is driven to absorb harmful noise into the model parameters. This residual term directly contaminates the gradient direction (formalized in Appendix A.2), steering parameter updates away from genuine reasoning improvement and toward encoding spurious $x \to r$ correlations. Since the student's architecture precludes direct conditioning on r, the only available pathway is to encode statistical correlations between x and r in θ, effectively learning a mapping $x \to r$. This is the mathematical origin of privileged information leakage. In contrast, OPD employs an external teacher whose predictions do not condition on privileged information inaccessible to the student, so the mutual information gap does not arise, and the KL divergence decreases steadily.

A.2 Two-Phase Training Dynamics

We provide the detailed analysis of the gradient dynamics summarized in §3.2.

Benign expected gradient. Since $I(Y_t; R \mid X)$ does not depend on θ, we have $\nabla_\theta \mathcal{L}_{\text{OPSD}} = \nabla_\theta \mathcal{L}^* = -\sum_v \bar{P}_T(v) \nabla_\theta \log P_S(v)$. At the population level, the OPSD gradient is identical to that of the ideal marginal matching objective.

Pathological per-sample gradients. In practice, optimization operates on concrete samples (x, r), with $g(\theta; r) = -\sum_{v \in \mathcal{V}} P_T(v \mid r) \cdot \nabla_\theta \log P_S(v)$. Property (i) of Proposition 1 ($\mathbb{E}_r[\delta] = 0$) may suggest that the deviation is innocuous on average; however, any optimizer that computes gradients on individual samples or mini-batches, such as SGD and Adam [40], is inherently path-dependent. Zero-mean perturbations in nonlinear optimization do not necessarily cancel over the course of training.

Phase 1 (early training). The student P_S is far from the teacher's marginal $\bar{P}_T$, so the beneficial component dominates: $\|g^*(\theta)\| \gg \|\delta(\theta; r)\|$. In this regime, the gradient predominantly drives marginal matching, and the student rapidly acquires general reasoning capabilities. This corresponds to the steep rise in validation accuracy observed during the first 10–20 steps in Figure 3(b).

Phase 2 (late training). As P_S approaches $\bar{P}_T$, the beneficial component $\|g^*(\theta)\|$ diminishes toward zero. The deviation component $\|\delta(\theta; r)\|$, however, remains bounded away from zero: its variance is governed by $I(Y_t; R \mid X)$, which is independent of θ and therefore does not decay with optimization progress. Parameter updates thus become increasingly dominated by δ, and the path-dependent accumulation of these perturbations drives the model toward regions of parameter space that encode $x \to r$ correlations, triggering a self-reinforcing degradation. This transition marks the onset of the performance decline in Figure 3(b) and the monotonically increasing leakage counts in Figure 3(a).

A.3 Leakage Bandwidth: Detailed Comparison Across Variants

This subsection elaborates the leakage bandwidth analysis introduced in §3.2.

Definition. We define *leakage bandwidth* as the effective number of token positions at which r-specific information enters the gradient direction.

Full OPSD. Operates over the entire vocabulary $\mathcal{V}$: the teacher's privileged preference $P_T(v \mid r)$ weights the gradient contribution of every token, yielding the widest bandwidth.

Teacher's Top-1. Collapses the target to a single token $\arg\max_v P_T(v \mid r)$ that is entirely determined by r, narrowing the bandwidth but producing the most concentrated injection of privileged information, which explains why it exhibits the *most severe* leakage.

Student's Top-1. Restricts the target support to $\arg\max_v P_S(v)$, yielding the narrowest bandwidth; yet the gradient weight $P_T(v_S^* \mid r)/P_S(v_S^*)$ at the selected token remains a function of r, so leakage persists, albeit at the lowest rate.

In all three cases, the directional dependence of the gradient on r is irreducible (proof in Appendix B.3), explaining the universal occurrence of leakage observed in Figure 3(a). The performance degradation in Figure 3(b) follows directly: stronger leakage accelerates the transition from Phase 1 to Phase 2 and produces more rapid decline.

B Deferred Proofs and Extended Analysis

B.1 Proof of Theorem 1 (KL Decomposition)

We suppress conditioning on $(x, y_{<t})$ throughout for notational clarity. Expanding $\mathcal{L}_{\text{OPSD}}$:

$$\mathcal{L}_{\text{OPSD}} = \mathbb{E}_r \left[\sum_v P_T(v \mid r) \log \frac{P_T(v \mid r)}{P_S(v)} \right]. \tag{11}$$

Inserting $\bar{P}_T(v)/\bar{P}_T(v)$ into the logarithm and separating terms:

$$= \mathbb{E}_r \left[\sum_v P_T(v \mid r) \log \frac{P_T(v \mid r)}{\bar{P}_T(v)} \right] + \mathbb{E}_r \left[\sum_v P_T(v \mid r) \log \frac{\bar{P}_T(v)}{P_S(v)} \right]. \tag{12}$$

The first term equals $\mathbb{E}_r[D_{\mathrm{KL}}(P_T(\cdot \mid r)\|\bar{P}_T)] = I(Y_t; R \mid X, Y_{<t})$ by the definition of conditional mutual information. For the second term, observe that $\log \frac{\bar{P}_T(v)}{P_S(v)}$ does not depend on r, so the expectation over r acts only on $P_T(v \mid r)$:

$$\mathbb{E}_r\left[\sum_v P_T(v \mid r) \log \frac{\bar{P}_T(v)}{P_S(v)}\right] = \sum_v \bar{P}_T(v) \log \frac{\bar{P}_T(v)}{P_S(v)} = D_{\mathrm{KL}}(\bar{P}_T \| P_S) = \mathcal{L}^*. \tag{13}$$

Combining both terms yields $\mathcal{L}_{\mathrm{OPSD}} = \mathcal{L}^* + I(Y_t; R \mid X, Y_{<t})$. □

B.2 Proof of Proposition 1 (Per-Sample Gradient Decomposition)

The decomposition $g(\theta; r) = g^*(\theta) + \delta(\theta; r)$ follows by adding and subtracting $\bar{P}_T(v)$ inside the sum:

$$\begin{aligned} g(\theta; r) &= -\sum_v P_T(v \mid r)\, \nabla_\theta \log P_S(v) \\ &= -\sum_v \bar{P}_T(v)\, \nabla_\theta \log P_S(v) - \sum_v [P_T(v \mid r) - \bar{P}_T(v)]\, \nabla_\theta \log P_S(v). \end{aligned} \tag{14}$$

Property (i). Since $\mathbb{E}_r[P_T(v \mid r)] = \bar{P}_T(v)$ by definition:

$$\mathbb{E}_r[\delta(\theta; r)] = -\sum_v \mathbb{E}_r[P_T(v \mid r) - \bar{P}_T(v)]\, \nabla_\theta \log P_S(v) = 0. \tag{15}$$

Property (ii). Since $\nabla_\theta \log P_S(v)$ is independent of r:

$$\begin{aligned} \mathbb{E}_r[\|\delta(\theta; r)\|^2] &= \mathbb{E}_r\left[\left\|\sum_v [P_T(v \mid r) - \bar{P}_T(v)]\, \nabla_\theta \log P_S(v)\right\|^2\right] \\ &= \sum_v \mathbb{E}_r\left[(P_T(v \mid r) - \bar{P}_T(v))^2\right] \|\nabla_\theta \log P_S(v)\|^2 \\ &\quad + \sum_{v \neq v'} \mathbb{E}_r[(P_T(v \mid r) - \bar{P}_T(v))(P_T(v' \mid r) - \bar{P}_T(v'))]\, \langle \nabla_\theta \log P_S(v), \nabla_\theta \log P_S(v') \rangle. \end{aligned} \tag{16}$$

The diagonal terms yield $\sum_v \mathrm{Var}_r[P_T(v \mid r)] \cdot \|\nabla_\theta \log P_S(v)\|^2$. The cross terms are generally non-zero but are bounded by the same variances via the Cauchy–Schwarz inequality. The stated result is exact for the diagonal contribution and serves as a lower bound for the full expression. When $I(Y_t; R \mid X) = 0$, $P_T(v \mid r) = \bar{P}_T(v)$ for all r, so $\delta \equiv 0$. □

B.3 Proof of Irreducibility of Leakage Across Pilot Variants

We derive the per-token gradient form for each variant and verify that the directional dependence on r is irreducible in all three cases. In each variant, the student's policy gradient at token position t along trajectory $\hat{y}$ can be expressed as $\hat{A}_t \cdot \nabla_\theta \log P_S(y_t \mid x, y_{<t})$, where $\hat{A}_t$ involves the teacher's privileged evaluation.

(i) **Full OPSD.** The gradient sums over the entire vocabulary:

$$g_t(\theta; r) = -\sum_{v \in \mathcal{V}} P_T(v \mid r)\, \nabla_\theta \log P_S(v). \tag{17}$$

Every token in $\mathcal{V}$ receives a gradient contribution weighted by $P_T(v \mid r)$, which encodes the teacher's privileged preference. The deviation $\delta_t = -\sum_v [P_T(v \mid r) - \bar{P}_T(v)]\, \nabla_\theta \log P_S(v)$ operates over the full vocabulary, yielding the widest leakage bandwidth.

(ii) **Teacher's Top-1.** The teacher distribution is collapsed to a point mass at $v_T^* = \arg\max_v P_T(v \mid r)$. The effective gradient becomes:

$$g_t(\theta; r) = -\nabla_\theta \log P_S(v_T^*), \tag{18}$$

where v_T^* is entirely determined by r. The deviation component $\delta_t = -[\nabla_\theta \log P_S(v_T^*) - \sum_v \bar{P}_T(v)\nabla_\theta \log P_S(v)]$ is concentrated on a single r-dependent token, making the gradient direction maximally sensitive to the specific realization of r.

(iii) **Student's Top-1.** The target is restricted to the student's most probable token $v_S^* = \arg\max_v P_S(v)$. The gradient weight at this token is proportional to $P_T(v_S^* \mid r)/P_S(v_S^*)$, which still depends on the teacher's privileged evaluation. Although the support is anchored to the student's distribution (narrowest bandwidth), the r-dependence of the gradient weight persists.

In all three variants, $\hat{A}_t$ involves $P_T(\cdot \mid r)$, ensuring that the deviation component δ carries r-specific information into the parameter update direction. By Proposition 1, the deviation has zero mean but strictly positive variance whenever $I(Y_t; R \mid X) > 0$, regardless of how the target distribution is compressed. The three variants differ in how many tokens receive non-zero gradient weight (bandwidth), but the directional dependence on r is common to all, establishing the irreducibility of leakage. □

B.4 The Impossibility Trilemma Under Shared Parameters

The analysis in §3.2 characterizes the per-step gradient pathology of distribution matching. In OPSD, the coupling runs deeper: teacher and student share a single set of parameters θ:

$$P_S^\theta(\cdot \mid x, y_{<t}) = \pi_\theta(\cdot \mid x, y_{<t}), \qquad P_T^\theta(\cdot \mid x, r, y_{<t}) = \pi_\theta(\cdot \mid x, r, y_{<t}). \tag{19}$$

Although P_T is treated with stop-gradient within each optimization step, every update to θ simultaneously alters both distributions. This entanglement couples the gradient-level leakage pathology with a macro-level training stability problem. We show that their interaction gives rise to an impossibility result.

B.4.1 Strategy A: Frozen Teacher

One natural remedy is to fix the teacher at initialization: $P_T^{\theta_0}$. The optimization objective at step k becomes:

$$\mathcal{L}_A(\theta_k) = \mathbb{E}_r\Big[D_{\mathrm{KL}}\Big(P_T^{\theta_0}(\cdot \mid r) \parallel P_S^{\theta_k}(\cdot \mid x)\Big)\Big]. \tag{20}$$

By Theorem 1, the global optimum is $P_S^* = \bar{P}_T^{\theta_0}$, the marginal distribution of the frozen teacher.

Proposition 2 (Capacity Ceiling)**.** *As $P_S^{\theta_k} \to \bar{P}_T^{\theta_0}$, the distillation signal vanishes: $\|\nabla_\theta \mathcal{L}_A(\theta_k)\| \to 0$. The student's representational capacity is strictly upper-bounded by the quality of the initial checkpoint θ_0, precluding further improvement even when the model possesses sufficient capacity to learn a superior policy.*

B.4.2 Strategy B: Online Teacher

Alternatively, the teacher may evolve in tandem with the student, using $P_T^{\theta_k}$ at each step k:

$$\theta_{k+1} = \theta_k - \eta\, \nabla_\theta \mathcal{L}_k(\theta_k), \quad \text{where} \quad \mathcal{L}_k(\theta) \triangleq \mathbb{E}_r\Big[D_{\mathrm{KL}}\Big(P_T^{\theta_k}(\cdot \mid r) \parallel P_S^{\theta}(\cdot \mid x)\Big)\Big]. \tag{21}$$

Although $P_T^{\theta_k}$ is treated as a fixed target within each step via stop-gradient, the update $\theta_k \to \theta_{k+1}$ shifts the teacher distribution for the subsequent iteration.

Theorem 2 (Training Instability Under Online Teacher)**.** *Define the joint objective $\mathcal{L}(\theta) \triangleq \mathbb{E}_r[D_{\mathrm{KL}}(P_T^\theta(\cdot \mid r) \| P_S^\theta(\cdot \mid x))]$. The change from step k to $k{+}1$ decomposes as:*

$$\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k) = \underbrace{\Delta_S}_{\leq 0 \text{ (student improvement)}} + \underbrace{\Delta_T}_{\text{sign uncontrolled (teacher drift)}}, \tag{22}$$

where Δ_S captures the student's progress under the current teacher and Δ_T captures the induced teacher drift. When $|\Delta_T| > |\Delta_S|$ with $\Delta_T > 0$, the true objective worsens despite each step performing valid gradient descent on its local surrogate.

Insert and subtract $\mathbb{E}_r[D_{\mathrm{KL}}(P_T^{\theta_k}(\cdot \mid r) \| P_S^{\theta_{k+1}})]$ in $\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k)$. The exact expressions are:

$$\Delta_S = \mathbb{E}_r\left[D_{\mathrm{KL}}\left(P_T^{\theta_k}(\cdot \mid r) \parallel P_S^{\theta_{k+1}}\right)\right] - \mathbb{E}_r\left[D_{\mathrm{KL}}\left(P_T^{\theta_k}(\cdot \mid r) \parallel P_S^{\theta_k}\right)\right], \tag{23}$$

$$\Delta_T = \mathbb{E}_r\left[D_{\mathrm{KL}}\left(P_T^{\theta_{k+1}}(\cdot \mid r) \parallel P_S^{\theta_{k+1}}\right)\right] - \mathbb{E}_r\left[D_{\mathrm{KL}}\left(P_T^{\theta_k}(\cdot \mid r) \parallel P_S^{\theta_{k+1}}\right)\right]. \tag{24}$$

$\Delta_S \leq 0$ follows from gradient descent on $\mathcal{L}_k$ with $P_T^{\theta_k}$ held fixed. The sign of Δ_T is uncontrolled: the same update that improves P_S simultaneously perturbs P_T, and the asymmetry of KL divergence provides no guarantee on the direction of this induced shift.

B.4.3 Self-Reinforcing Feedback

The gradient-level leakage of §3.2 and the teacher drift interact through a positive feedback loop.

Proposition 3 (Self-Reinforcing Feedback Loop)**.** *Define the model's sensitivity to privileged information as $S(\theta) \triangleq \mathbb{E}_r[D_{\mathrm{KL}}(P_T^{\theta}(\cdot \mid r) \| \bar{P}_T^{\theta})] = I(Y_t; R \mid X, Y_{<t})$. Under Strategy B, the following cycle operates:*

- *(i) The per-sample deviation $\delta(\theta; r)$ drives parameters toward r-predictive features.*
- *(ii) These features are encoded in the shared θ, enhancing the model's capacity to exploit r when operating as teacher.*
- *(iii) $S(\theta_{k+1}) \geq S(\theta_k)$, amplifying $\mathrm{Var}_r[P_T(v \mid r)]$.*
- *(iv) By Proposition 1(ii), the deviation variance grows, reinforcing step (i).*

By Proposition 1, $g(\theta; r) = g^*(\theta) + \delta(\theta; r)$ with $\mathbb{E}_r[\delta] = 0$. Each stochastic update $\theta_{k+1} = \theta_k - \eta\, g(\theta_k; r_k)$ introduces a perturbation $-\eta\, \delta(\theta_k; r_k)$ whose direction correlates with r_k. Over multiple steps with varying r_k, the path-dependent accumulation of these perturbations biases parameters toward regions that encode $x \to r$ correlations: parameter configurations for which δ aligns with such correlations yield lower subsequent loss, causing the optimizer to preferentially retain these drifts. The accumulated drift enriches the shared representation θ with features predictive of r given x. Since teacher and student share θ, these features become available to both roles simultaneously. At the subsequent step, the teacher $P_T^{\theta_{k+1}}(\cdot \mid x, r, y_{<t})$ more effectively leverages r through the newly encoded features, increasing the dispersion of $P_T(\cdot \mid r)$ across different realizations of r: $S(\theta_{k+1}) \geq S(\theta_k)$. By Proposition 1(ii), the deviation variance $\mathbb{E}_r[\|\delta\|^2]$ grows with $S(\theta)$, and the cycle restarts with a strictly larger deviation magnitude.

Let $\rho_k \triangleq \mathbb{E}_r[\|\delta(\theta_k; r)\|^2] / \mathbb{E}_r[\|g(\theta_k; r)\|^2]$ denote the fraction of gradient variance attributable to the deviation component. The feedback loop drives ρ_k monotonically upward: early in training, ρ_k remains small since $\|g^*\|$ dominates; as $\|g^*\| \to 0$ and $\|\delta\|$ grows via the loop, $\rho_k \to 1$. At this point, parameter updates are overwhelmingly determined by the leakage signal, and training collapses.

B.4.4 The Trilemma

The analyses of Strategies A and B, together with the gradient decomposition, yield the central impossibility result of on-policy self-distillation.

Theorem 3 (Impossibility Trilemma)**.** *In any distribution-matching framework where teacher and student share parameters (i.e., P_T and P_S are parameterized by the same θ), the following three properties cannot hold simultaneously:*

- *(a) **Objective stability**: the optimization target does not drift between successive steps ($\Delta_T = 0$).*
- *(b) **Sustained improvement**: the distillation signal does not vanish ($\|\nabla_\theta \mathcal{L}\| \not\to 0$).*
- *(c) **Leakage-free training**: the deviation component does not drive parameter drift.*

Strategy A (frozen teacher) satisfies (a) by construction ($P_T^{\theta_0}$ is fixed, hence $\Delta_T = 0$), but violates (b) by Proposition 2. Strategy B (online teacher) satisfies (b) since the teacher continually evolves to provide non-vanishing signal, but violates (a) by Theorem 2. Property (c) is violated under *either* strategy whenever $I(Y_t; R \mid X) > 0$: by Proposition 1(ii), the per-sample deviation variance $\mathbb{E}_r[\|\delta\|^2]$ is strictly positive, independently of the teacher management scheme. Under Strategy B, Proposition 3 further ensures that ρ_k is monotonically non-decreasing, so the leakage intensifies over the course of training. Hybrid strategies (e.g., periodic teacher snapshots) merely interpolate between (a) and (b): within each update interval, (a) holds but (b) progressively degrades as the

teacher becomes stale; refreshing the teacher restores (b) while disrupting (a). Property (c) remains unsatisfied throughout, as the underlying mutual information gap persists regardless of the snapshot schedule.

B.5 Bayesian Interpretation of RLSD Evidence Weights

This section provides a formal probabilistic derivation of the RLSD evidence ratio $w_t = P_T(y_t)/P_S(y_t)$ introduced in §4, establishing that it corresponds to a sequential Bayesian belief update.

B.5.1 Modeling Assumption

Assumption 1 (Consistent Conditional Approximation)**.** *The shared model π_θ provides a consistent approximation to the true conditional distributions in the following sense:*

$$P_S(y_t \mid x, y_{<t}) \triangleq \pi_\theta(y_t \mid x, y_{<t}) \approx P(y_t \mid x, y_{<t}), \tag{25}$$

$$P_T(y_t \mid x, r, y_{<t}) \triangleq \pi_\theta(y_t \mid x, r, y_{<t}) \approx P(y_t \mid x, r, y_{<t}). \tag{26}$$

When the model has sufficient capacity and the privileged information r is provided via in-context conditioning, Assumption 1 is reasonable: a single architecture can faithfully represent both the prior predictive distribution (without r) and the posterior predictive distribution (with r). This parallels the standard assumption in Bayesian deep learning that a sufficiently expressive model family can approximate the true posterior predictive distribution.

B.5.2 The Evidence Ratio as Belief Update

Theorem 4 (RLSD Weights as Belief Update Ratios)**.** *Under Assumption 1, the token-level evidence ratio satisfies the identity:*

$$w_t = \frac{P_T(y_t \mid x, r, y_{<t})}{P_S(y_t \mid x, y_{<t})} = \frac{P(r \mid x, y_{\le t})}{P(r \mid x, y_{<t})}, \tag{27}$$

i.e., w_t equals the ratio of the posterior belief in the privileged information r after observing token y_t to the belief before observing y_t.

Proof. Apply Bayes' theorem to the joint distribution $P(r, y_t \mid x, y_{<t})$:

$$P(y_t \mid x, r, y_{<t}) = \frac{P(r, y_t \mid x, y_{<t})}{P(r \mid x, y_{<t})} = \frac{P(r \mid x, y_{\le t}) \cdot P(y_t \mid x, y_{<t})}{P(r \mid x, y_{<t})}. \tag{28}$$

Dividing both sides by $P(y_t \mid x, y_{<t})$ and applying Assumption 1:

$$\frac{P_T(y_t \mid x, r, y_{<t})}{P_S(y_t \mid x, y_{<t})} \approx \frac{P(y_t \mid x, r, y_{<t})}{P(y_t \mid x, y_{<t})} = \frac{P(r \mid x, y_{\le t})}{P(r \mid x, y_{<t})}. \tag{29}$$

□

B.5.3 Interpretation

Theorem 4 reveals that the RLSD weight w_t is a *sequential Bayesian evidence ratio*: it quantifies the degree to which generating token y_t updates the belief that the privileged information r is consistent with the current trajectory.

Concretely, $w_t > 1$ indicates that generating y_t *strengthens* the posterior belief in r, constituting positive evidence for the correct reasoning path. For instance, when solving $2x + 3 = 7$, writing "$2x = 4$" substantially increases the belief that the trajectory leads to the correct answer $x = 2$; RLSD assigns such tokens elevated positive credit. $w_t < 1$ indicates that y_t *weakens* the belief in r, constituting negative evidence. An erroneous step such as "$5x = 7$" would sharply reduce the belief; in an incorrect trajectory, such tokens bear greater blame. $w_t = 1$ indicates that y_t is informationally neutral with respect to r, as is typical of formatting connectives (e.g., "therefore," "we have") that carry no reasoning content; RLSD assigns these tokens the baseline advantage.

B.5.4 Telescoping to Sequence-Level Evidence

Taking the product over all token positions yields a telescoping identity:

$$\prod_{t=1}^{T} w_t = \prod_{t=1}^{T} \frac{P(r \mid x, y_{\leq t})}{P(r \mid x, y_{<t})} = \frac{P(r \mid x, y)}{P(r \mid x)}. \tag{30}$$

The right-hand side is the sequence-level Bayesian evidence ratio: the total belief update after observing the complete trajectory y. The per-token weights $\{w_t\}_{t=1}^{T}$ thus provide a fine-grained *decomposition* of this sequence-level evidence into individual token contributions, elevating the granularity of credit assignment from the sequence level (as in GRPO) to the token level.

B.5.5 Contrast with OPSD: Logical Attribution vs. Behavioral Cloning

Theorem 4 clarifies a fundamental distinction between RLSD and distribution matching methods.

OPSD performs *behavioral cloning*: it requires the student to replicate the teacher's output distribution $P_T(\cdot \mid x, r)$ at every token position. If the teacher, having observed the reference solution, favors a particular phrasing (e.g., "according to the hint"), OPSD drives the student to adopt the same phrasing, even when it carries no reasoning value. The objective is to minimize the *distributional distance* between P_S and P_T.

RLSD performs *logical credit attribution*: it does not require the student to imitate any specific token choice of the teacher. The weight w_t measures a single property: whether the token generated by the student constitutes positive Bayesian evidence for the correct answer r. As long as a student-generated token increases the posterior belief in r ($w_t > 1$), it receives elevated credit regardless of whether it matches the teacher's preferred wording. RLSD is concerned not with *what* the student says, but with the *informational contribution* of what it says toward the correct derivation.

In information-theoretic terms, w_t quantifies the *per-token information gain* toward the correct answer. RLSD allocates credit accordingly: tokens that advance the reasoning are rewarded, tokens that derail it are penalized, and tokens that are informationally inert receive the baseline advantage. No imitation of the teacher's distributional shape is involved at any point.

B.5.6 Connection to the OPSD Gradient

It is instructive to note that the importance weight $P_T(v \mid r)/P_S(v)$ appearing in the OPSD per-sample gradient (Proposition 1) is mathematically identical to the evidence ratio w_t used in RLSD. This is not coincidental: both methods operate on the same underlying quantity, but employ it in fundamentally different ways.

In OPSD, P_T/P_S serves as a *gradient weight* that spans the entire vocabulary $\mathcal{V}$, driving P_S to match the shape of P_T. The per-sample deviation $[P_T(v \mid r) - \bar{P}_T(v)]$ is inseparable from the expected signal (Proposition 1), causing r-specific patterns to infiltrate parameter updates.

In RLSD, P_T/P_S serves as a *stop-gradient scalar credit multiplier* applied only to the on-policy sampled token y_t. It modulates the *magnitude* of the existing advantage but does not enter the gradient direction. The privileged information influences only how much credit each token receives, not which tokens are reinforced or penalized, nor the direction of parameter updates.

It is precisely this change in usage, from distribution matching to credit assignment, from gradient weight to scalar multiplier, that transforms the same mathematical quantity from the source of leakage into a tool for precise credit attribution.

B.5.7 Limitations of the Bayesian Approximation

The Bayesian interpretation (Theorem 4) relies on Assumption 1, which may not hold exactly in practice. When the model's in-context conditioning is imperfect, e.g., when r is excessively long or the model has limited capacity, the evidence ratio P_T/P_S becomes a noisy approximation of the true belief update ratio.

However, the structural guarantees of RLSD *do not depend on Assumption 1*. The direction anchoring property ($\text{sign}(\hat{A}_t) \equiv \text{sign}(A)$, §4) and the support set isolation (gradients act only on student-sampled tokens) are consequences of the algorithm design (stop-gradient, on-policy sampling, scalar

modulation, clipping) and hold regardless of the quality of the Bayesian approximation. Similarly, the leakage-free guarantee established in Appendix B.6 and the resolution of the impossibility trilemma (Appendix B.4) are purely structural results. This separation ensures that all theoretical guarantees remain valid even when the model is an imperfect Bayesian agent.

B.6 Formal Leakage-Free Guarantee for RLSD

Building on the Bayesian interpretation in Appendix B.5 and the diagnosis of distribution matching in §3.2, this section establishes that RLSD is structurally immune to privileged information leakage and verifies that it circumvents all three deficiencies identified in §3.2 and Appendix B.4.

B.6.1 Gradient Structure of RLSD

The RLSD policy gradient takes the form:

$$\nabla_\theta \mathcal{J}_{\text{RLSD}}(\theta) = \mathbb{E}_x \, \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y^{(i)}|} \sum_{t=1}^{|y^{(i)}|} \hat{A}_t^{(i)} \cdot \nabla_\theta \log \pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)}) \right], \tag{31}$$

where the token-level advantage is constructed as specified in Eqs. (5)–(7). This gradient expression differs from the OPSD gradient in two structural properties.

Property 1: Direction anchoring. The token-level advantage $\hat{A}_t$ strictly preserves the sign of the sequence-level advantage A:

$$\text{sign}(\hat{A}_t) = \text{sign}(A) \quad \forall\, t, \tag{32}$$

since $\text{clip}(w_t, 1-\epsilon_w, 1+\epsilon_w) > 0$. This ensures that the environment reward has *exclusive authority* over the direction of policy updates: all tokens in correct trajectories are reinforced, and all tokens in incorrect trajectories are penalized. The teacher's privileged information, channeled through w_t, modulates only the *relative intensity* of reinforcement or penalization across tokens within a trajectory, but can never flip the direction. In contrast, the OPSD gradient direction at each token position is determined by $P_T(v \mid r)$ or its difference from $P_S(v)$, so the teacher's privileged preference directly controls *where* the gradient points.

Property 2: Student-determined support. The expectation in the RLSD gradient is taken over trajectories $y \sim \pi_\theta(\cdot \mid x)$, sampled from the student's own policy without access to r. The log-derivative $\nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})$ is evaluated only at tokens the student itself generates. The gradient's support, defined as the set of token sequences receiving non-zero gradient signal, is therefore identical to the support of $\pi_\theta(\cdot \mid x)$.

In contrast, the OPSD per-sample gradient $g(\theta; r) = -\sum_{v \in \mathcal{V}} P_T(v \mid r) \nabla_\theta \log P_S(v)$ ranges over the entire vocabulary $\mathcal{V}$: tokens that the teacher strongly favors due to r receive $P_T(v \mid r)$-weighted gradient contributions, actively pulling unseen privileged patterns into the parameter updates. RLSD eliminates this channel entirely.

B.6.2 Zero-Leakage Guarantee

Combining the two structural properties yields the following guarantee.

Theorem 5 (Direction-Level Leakage-Free Training Under RLSD)**.** *In the RLSD policy gradient objective, the privileged information r enters the computation only through the stop-gradient scalar w_t, and w_t is isolated at three levels:*

- *(i)* ***Directional isolation.*** *By Property 1, $\text{sign}(\hat{A}_t) = \text{sign}(A)$, so r cannot influence the gradient sign of any token. The direction of parameter updates in θ-space is fully determined by the student-sampled token y_t (via $\nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})$) and the environment correctness judgment (via $\text{sign}(A)$), neither of which contains information about r.*
- *(ii)* ***Support isolation.*** *By Property 2, any token $y_{\text{leak}} \notin \text{supp}(\pi_\theta(\cdot \mid x, y_{<t}))$ that exists only in the privileged mode has strictly zero sampling probability and contributes zero expected gradient.*
- *(iii)* ***Magnitude boundedness.*** *The weight w_t is clipped to $[1-\epsilon_w, 1+\epsilon_w]$ and further attenuated by the mixing coefficient λ, strictly bounding its influence. As the student internalizes reasoning ability and $P_S \to P_T$, the weights satisfy $w_t \to 1$, and RLSD automatically degrades to vanilla GRPO.*

Consequently, there exists no mathematical pathway through which patterns from the privileged information r can be injected into the direction of parameter updates. The privileged information in RLSD influences only the magnitude of token-level credit, not which tokens are reinforced or penalized, nor whether reinforcement or penalization occurs.

B.6.3 Resolution of the Impossibility Trilemma

Appendix B.4 establishes that distribution matching under shared parameters cannot simultaneously satisfy objective stability (a), sustained improvement (b), and leakage-free training (c). We verify that RLSD satisfies all three.

(a) Objective stability. RLSD optimizes the environment reward $R(x, y) \in \{0, 1\}$, an external signal independent of θ. The teacher contributes only through the stop-gradient scalar w_t, which does not constitute part of the optimization objective. The teacher drift term Δ_T from Theorem 2 is therefore inapplicable: the optimization anchor is a fixed external reward, not a drifting internal distribution. Even as θ updates cause w_t to change at subsequent steps, this affects only the credit magnitude, not the gradient direction (Property 1) or support (Property 2).

(b) Sustained improvement. RLSD is architecturally designed as a *curriculum-aware* training strategy that seamlessly unifies two complementary learning phases. In the early stage, when the student's policy is far from optimal and token-level credit discrimination is most valuable, the self-distillation signal provides dense, fine-grained guidance that accelerates convergence well beyond what uniform advantages can achieve. As training progresses, the mixing coefficient λ is linearly decayed to zero, smoothly transitioning RLSD into vanilla GRPO. This scheduled transition reflects a principled design choice: the teacher's dense credit assignment is most beneficial when the student's policy is still coarse-grained and requires fine-grained steering, whereas in the later stage, the environment reward alone suffices to drive continued improvement. Crucially, the GRPO policy gradient remains non-zero as long as the student has not reached the optimal policy, ensuring that training does not stall after the transition. This two-phase structure, dense credit assignment for rapid early improvement followed by standard RLVR for sustained long-term optimization, allows RLSD to inherit the convergence ceiling of GRPO while substantially accelerating the path toward it. The explicit scheduling of λ also provides practitioners with a transparent and controllable mechanism to balance the intensity of teacher guidance against training stability, making RLSD a practical and fully compatible enhancement to existing RLVR pipelines.

(c) Leakage-free training. By Theorem 5, the gradient direction is determined entirely by the environment reward and the student's on-policy samples. The deviation component δ from Proposition 1 does not arise, because RLSD does not perform distribution matching and therefore does not generate the $[P_T(v \mid r) - \bar{P}_T(v)]$ bias that drives the gradient direction in OPSD. The self-reinforcing feedback loop of Proposition 3 is likewise absent: even if shared parameters cause $S(\theta)$ to change, this affects only the numerical value of w_t, which, after clipping, enters $\hat{A}_t$ as a bounded scalar multiplier that does not alter the gradient direction or the token-level support.

Table 4 summarizes the trilemma from the perspective of both OPSD strategies and RLSD.

Table 4: Comparison of OPSD parameter management strategies and RLSD across the three desiderata identified in the impossibility trilemma.

Property	**OPSD (Frozen Teacher)**	**OPSD (Online Teacher)**	**RLSD**
(a) Objective stability	✓	×	✓
(b) Sustained improvement	×	✓	✓
(c) Leakage-free training	×	×	✓

C A Unified Token-Level Advantage Perspective

To situate RLSD within the broader landscape, we observe that GRPO, on-policy self-distillation, and RLSD can all be expressed as instances of a single policy gradient template:

$$\Delta\theta \;\propto\; \mathbb{E}_{y\sim P_S(\cdot|x)}\left[\sum_{t=1}^{|y|} \hat{A}_t \;\nabla_\theta \log P_S(y_t \mid x, y_{<t})\right], \tag{33}$$

where the methods differ solely in how they define the token-level advantage $\hat{A}_t$.

GRPO assigns a uniform advantage across all tokens: $\hat{A}_t = A$, where A is the sequence-level advantage from the verifier. This ensures that the optimization direction is fully grounded in the environment reward, but provides no token-level discrimination: every token within a trajectory receives identical credit regardless of its contribution to the final answer.

On-policy self-distillation methods replace the environment reward with a dense teacher signal. The specific form of $\hat{A}_t$ depends on the divergence measure: minimizing the reverse KL $D_{\mathrm{KL}}(P_S \| P_T)$ via the log-derivative trick yields $\hat{A}_t = \Delta_t = \log P_T(y_t) - \log P_S(y_t)$. The environment reward $R(x, y)$ is entirely absent from $\hat{A}_t$: even when a trajectory produces a wrong answer ($A < 0$), tokens favored by the teacher ($\Delta_t > 0$) still receive positive advantage, decoupling the optimization direction from the verifiable correctness signal.

RLSD resolves this tension by combining both sources of information. Its advantage (Eq. 7) uses the environment reward to determine the *direction* (sign) of each token's update, while using the teacher's privileged assessment to determine the *magnitude* (relative credit) within a trajectory. We provide a formal proof that these properties render RLSD structurally immune to privileged information leakage in Appendix B.6, and verify that RLSD simultaneously satisfies all three desiderata of the impossibility trilemma (Appendix B.4).

D RLSD Algorithm

We present the complete RLSD training procedure in Algorithm 1, providing a self-contained reference for implementation. The algorithm shows that RLSD requires only minimal modifications to the standard GRPO pipeline: an additional teacher forward pass (line 13) to compute log-probabilities under the privileged context, followed by per-token credit reweighting (lines 15–17). All other components, including on-policy rollout, group-relative advantage estimation, and the clipped policy update, remain identical to GRPO. The mixing coefficient λ controls the strength of the self-distillation signal and is linearly decayed to 0 during training, smoothly transitioning RLSD to vanilla GRPO once the teacher's informational advantage diminishes (see Section 5.3 for empirical justification).

Algorithm 1 RLSD: Reinforcement Learning with Self-Distillation

Require: Policy π_θ, dataset $\mathcal{S} = \{(x_i, r_i)\}_{i=1}^{N}$, verifier $R(\cdot,\cdot)$, group size G, mixing coefficient λ, clip bounds ϵ_w

for each training iteration **do**
  Sample a batch of questions $\{x\}$ from $\mathcal{S}$
  for each question x with privileged information r **do**
    // Step 1: On-policy rollout
    Sample G responses $\{y^{(1)}, \ldots, y^{(G)}\} \sim \pi_\theta(\cdot \mid x)$
    // Step 2: Sequence-level advantage from environment
    for $i = 1, \ldots, G$ **do**
      Obtain reward $R(x, y^{(i)}) \in \{0, 1\}$ from the verifier
    end for
    Compute $A^{(i)} = \frac{R(x,y^{(i)}) - \mu_G}{\sigma_G}$ ▷ Group-relative advantage
    // Step 3: Token-level credit assignment via self-distillation
    for $i = 1, \ldots, G$ **do**
      Compute teacher logits via forward pass with $(x, r, y^{(i)})$ ▷ Single extra forward pass
      for $t = 1, \ldots, |y^{(i)}|$ **do**
        $\Delta_t \leftarrow \mathrm{sg}\Big(\log \pi_\theta(y_t^{(i)} \mid x, r, y_{<t}^{(i)}) - \log \pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)})\Big)$
        $w_t \leftarrow \exp\big(\mathrm{sign}(A^{(i)}) \cdot \Delta_t\big)$
        $\hat{A}_t^{(i)} \leftarrow A^{(i)} \cdot ((1-\lambda) + \lambda \cdot \mathrm{clip}(w_t,\, 1-\epsilon_w,\, 1+\epsilon_w))$
      end for
    end for
  end for
  // Step 4: Policy update
  Update θ by maximizing $\mathcal{L}_{\mathrm{RLSD}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \frac{1}{|y^{(i)}|} \sum_{t=1}^{|y^{(i)}|} \hat{A}_t^{(i)}$
end for

E More Dataset Details.

Training data. We use different training corpora corresponding to the three evaluation domains.

Multimodal reasoning. We train on **MMFineReason-123K** [12], a challenging subset of MMFineReason-1.8M obtained by filtering for problems on which the Qwen3-VL-4B-Thinking model fails across four independent rollouts. This conservative criterion discards trivial examples and concentrates the training signal on challenging problems, yielding faster convergence and more efficient use of computational resources.

Text-only reasoning. We train on **Polaris-Dataset-53K** [41], which is filtered from DeepScaleR-Preview-Dataset [42] and AReal-boba-Data [43] to retain high-quality mathematical reasoning problems.

Video reasoning. We collect training data from OneThinker [44], VideoChat-R1 [45], and Video-R1 [46], then filter the combined pool with Qwen3-VL-8B by removing samples with a pass rate of either 0% (always wrong) or 100% (always correct), retaining 40K samples of moderate difficulty for training.

Multimodal reasoning benchmarks. We evaluate on eight benchmarks covering diverse reasoning capabilities. **MMMU-Pro** [17] is a massive multi-discipline benchmark covering college-level subjects across science, engineering, and humanities. **MathVista** [13] assesses mathematical reasoning grounded in visual contexts. **MathVision** [14] extends mathematical evaluation to competition-level visual problems. **ZeroBench** [20] is specifically designed to be unsolvable by current frontier models, providing a stress test for reasoning robustness. **WeMath** [15] evaluates fine-grained mathematical problem solving with structured difficulty levels. **MMBench** [18] probes broad multimodal competence across perception and reasoning subskills. **MM-Star** [19] stresses fine-grained visual discrimination through carefully curated samples. **MathVerse** [16] measures whether multimodal models truly understand diagrams in visual math problems.

Video understanding benchmarks. We evaluate video understanding on eight benchmarks spanning diverse capabilities. **Video-MME** [47] is a comprehensive multiple-choice benchmark for general video understanding. **TempCompass** [48] evaluates fine-grained temporal reasoning in video LLMs. **MotionBench** [49] targets fine-grained motion-level perception and analysis. **LongVideoBench(LongVB)** [50] stresses long-context interleaved video understanding. **VCR-Bench** [51] provides a comprehensive evaluation framework for video causal reasoning, combining multiple-choice and open-ended formats. **LongVideo-Reason(LongVR)** [52] evaluates reasoning on long videos under a scaling-RL setting. **MMVU** [53] measures expert-level multi-discipline understanding across STEM domains. **VideoMathQA** [54] benchmarks video-grounded mathematical reasoning.

Text-only reasoning benchmarks. We evaluate text-only reasoning on eight benchmarks. **GSM8K** [55] contains 1,319 elementary multi-step arithmetic problems. **MATH-500** [56] consists of competition-level problems across high school mathematics domains, using the 500-problem subset curated by OpenAI. **AMC 2023** [57] contains 40 problems from the American Mathematics Competitions, covering algebra, geometry, number theory, and combinatorics. **AIME 2024** and **AIME 2025** [58] each contain 30 challenge problems from the American Invitational Mathematics Examination, where solutions are integers between 000 and 999. **Minerva Math** [59] evaluates quantitative reasoning across STEM topics. **OlympiadBench** [60] is an Olympiad-level scientific benchmark; we use the 675-sample subset adopted in LIMO. **GPQA Diamond** [61] is the highest-quality 198-question subset of a PhD-level benchmark spanning physics, chemistry, and biology, on which domain experts achieve only 69.7% accuracy.

F Implementation Details.

We implement our method based on the VERL [62] and EasyR1 [63] frameworks. During training, the model's maximum context length is set to 8192, with a maximum prompt length of 4096 and a maximum response length of 4096. The same prompt and response length settings are used consistently during both training and evaluation. For GRPO, GRPO+OPSD, and RLSD, the learning rate is fixed at 1×10^{-6}; for OPSD and SDPO, the learning rate is set to 1×10^{-5} following their original implementations. The training batch size is set to 256, and for each prompt, we sample 8 rollouts with a sampling temperature of 1.0. For all GRPO-based methods, we use clipping thresholds of $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$, and omit both the KL penalty loss and entropy regularization loss from the objective function. The hyperparameter λ in our RLSD is initialized at 0.5 and linearly decayed to 0 over the first 60 training steps, and ϵ_w is set to 0.2. To maintain a stable self-distillation signal, the teacher model parameters are synchronized with the student model every 20 training steps and kept frozen in between. Regarding the privileged information required by different methods: OPSD requires verified reasoning traces, for which we use reasoning trajectories distilled from Qwen3-VL-235B-A22B-Thinking and verified as correct within the MMFineReason-123K dataset; SDPO follows its original setting in the paper, using a successful previous rollout as privileged context; RLSD requires **only the final ground-truth answer** without any reasoning trace, making it the least demanding in terms of privileged information. For the GRPO+OPSD baseline, the linear interpolation coefficient between the GRPO loss and the OPSD distillation loss is a critical hyperparameter that significantly affects performance. To ensure a fair comparison, we conduct a grid search over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and report results under the best-performing setting of 0.9, which assigns dominant weight to the GRPO loss while incorporating moderate self-distillation guidance. All experiments are conducted on 4 compute nodes, each equipped with 8 NVIDIA H200 140GB GPUs. To reduce sampling variance during evaluation, we report averaged accuracy across multiple independent runs. For benchmarks with a small number of samples (e.g., AMC 2023 with 40 problems and AIME 2024/2025 with 30 problems each), we average over **16 independent runs** to ensure stable estimates. For larger benchmarks, we average over **4 independent runs**. This protocol is applied uniformly to all methods (Base LLM, GRPO, OPSD, SDPO, GRPO+OPSD, and RLSD) to ensure fair comparison.

G More Experimental Results

This section reports the full numerical results for the generalization experiments referenced in Section 5.2. Table 5 reports video understanding results on Qwen3-VL-8B-Instruct across eight

Table 5: Video understanding results on the Qwen3-VL-8B-Instruct model. **Bold** indicates the best result among all models. Avg. is computed over the eight selected benchmarks. LongVB denotes LongVideoBench and LongVR denotes LongVideo-Reason.

Method	VideoMME	Tempcompass	MotionBench	LongVB	VCRBench	LongVR	MMVU	VideoMathQA	Avg.
Base LLM	68.26	75.69	56.14	62.00	44.58	77.32	61.70	47.81	61.69
GRPO	**70.11**	75.98	57.01	64.85	44.00	78.61	60.40	50.33	62.66
RLSD *(Ours)*	69.44	**76.53**	**57.16**	**65.97**	**45.74**	**79.32**	**65.20**	**51.95**	**63.91**

Table 6: Text Reasoning results on the Qwen3-8B (Non-thinking) model. **Bold** indicates the best result among all models. Avg. is computed over the eight selected benchmarks.

Method	GSM8K	MATH-500	AMC	AIME24	AIME25	Minerva	OlympiadBench	GPQA	Avg.
Base LLM	84.87	79.35	68.59	23.54	17.71	24.45	46.89	44.07	48.68
GRPO	**96.00**	92.00	89.06	58.13	46.67	51.29	60.74	58.46	69.04
RLSD *(Ours)*	95.60	**93.20**	**92.66**	**61.46**	**47.50**	**53.13**	**65.93**	**58.71**	**71.02**

benchmarks covering general video QA, temporal reasoning, motion analysis, long-video understanding, and video-grounded mathematics. Table 6 reports text-only reasoning results on Qwen3-8B (Non-thinking) across eight mathematical and scientific benchmarks ranging from grade-school problems to competition-level olympiads. In both settings, RLSD consistently outperforms the GRPO baseline, confirming that the benefits of decoupled credit assignment generalize across modalities and model families.

H Ablation Study

We provide comprehensive ablation studies to validate RLSD's design choices. Table 7 reports the full results across four ablation dimensions: (a) the credit-weight clipping threshold ϵ_w, (b) the initial mixing coefficient λ, (c) the decay step at which λ is fully decayed to 0, and (d) the granularity of privileged information (final answer versus full chain-of-thought). The summary findings discussed in Section 5.3 are derived directly from this table; readers are referred there for the corresponding interpretation.

I Case Study

To complement the quantitative results in Section 5.2, Figure 6 visualizes RLSD's token-level credit assignment on two representative multimodal reasoning examples: one correct trajectory (cube-counting) and one incorrect trajectory (bar-model misinterpretation). The heatmaps illustrate how the privileged teacher modulates the relative magnitude of token-level credit while the environment reward determines the overall direction of reinforcement, providing a qualitative confirmation of the design principles formalized in Section 4.

J Limitations

While RLSD demonstrates consistent improvements over RLVR and self-distillation baselines, several limitations remain to be addressed in future work. Our experiments are conducted on models up to 8B parameters. We have not validated RLSD on substantially larger models such as Qwen3-235B-A22B or other frontier-scale architectures. Notably, we observe that the improvement of RLSD over GRPO grows substantially as model scale increases (from +0.84 points on Qwen3-VL-4B to +2.41 points on Qwen3-VL-8B; see Table 2), which we attribute to stronger instruction-following capabilities in larger models that enable more effective absorption and utilization of the privileged information injected through in-context conditioning. This trend suggests that RLSD's gains may further amplify at frontier scales, where models exhibit even stronger in-context reasoning capabilities, making validation at such scales an important direction for future work. The teacher in RLSD derives its informational advantage solely from privileged conditioning rather than from greater capacity, and

Table 7: Comprehensive ablation studies on RLSD's design choices, conducted on Qwen3-VL-Instruct across eight multimodal reasoning benchmarks. **(a)** Credit-weight clipping threshold ϵ_w, which bounds the per-token influence of the teacher's evidence ratio. **(b)** Initial mixing coefficient λ, which controls the strength of the self-distillation signal at the start of training. **(c)** Decay step, denoting the training step at which λ is fully linearly decayed to 0 (i.e., the point at which RLSD completely reverts to vanilla GRPO). **(d)** Granularity of privileged information: providing only the final ground-truth answer (*answer*) versus a full verified chain-of-thought (*cot*). Default settings used in our main experiments are $\epsilon_w = 0.2$, $\lambda = 0.5$, decay step $= 60$, and *answer*-only privileged information. **Bold** indicates the best result in each column within each panel.

Method	MMMU-Pro	MathVista	MathVision	ZeroBench	WeMath	MMBench	MM-Star	MathVerse	Avg.
(a) Effect of credit-weight clipping threshold ϵ_w									
RLSD ($\epsilon_w = 0.2$)	57.73	**78.50**	**56.12**	22.90	62.29	**90.41**	**73.47**	59.82	**62.66**
RLSD ($\epsilon_w = 0.3$)	57.22	77.70	54.80	**24.25**	62.86	90.34	73.00	**59.86**	62.50
RLSD ($\epsilon_w = 0.4$)	**57.92**	77.55	55.00	23.80	**63.33**	89.99	73.27	59.76	62.58
(b) Effect of initial mixing coefficient λ									
RLSD ($\lambda = 0.5$)	**57.73**	78.50	**56.12**	22.90	62.29	90.41	73.47	**59.82**	62.66
RLSD ($\lambda = 0.75$)	57.44	**78.50**	54.92	22.75	**64.10**	**90.53**	**73.53**	59.59	**62.67**
RLSD ($\lambda = 1.0$)	57.53	77.80	54.41	**23.80**	61.24	90.09	72.90	59.78	62.19
(c) Effect of decay step (training step at which λ is fully decayed to 0, reverting RLSD to GRPO)									
RLSD (decay@60)	**57.73**	78.50	**56.12**	22.90	62.29	90.41	73.47	**59.82**	**62.66**
RLSD (decay@80)	56.73	**78.85**	54.37	22.16	**65.14**	**90.52**	**73.77**	58.24	62.47
RLSD (decay@100)	55.89	77.05	53.78	23.20	61.14	90.31	73.40	58.94	61.71
RLSD (decay@200)	53.96	77.20	51.92	**23.65**	57.24	89.70	71.57	58.03	60.41
(d) Effect of privileged information granularity									
RLSD (answer)	**57.73**	**78.50**	**56.12**	22.90	62.29	**90.41**	73.47	**59.82**	62.66
RLSD (cot)	55.95	78.05	55.36	**24.40**	**64.95**	90.17	**74.00**	59.30	**62.77**

therefore does not benefit from the student's continued RLVR optimization. As the student's intrinsic reasoning capability improves, the teacher's relative advantage diminishes, eventually rendering the self-distillation signal stale. To preserve training stability, RLSD linearly decays the mixing coefficient λ to zero, smoothly reverting to vanilla GRPO in the late training phase. While RLSD consistently outperforms GRPO across all scales and modalities in our experiments by leveraging dense token-level credit assignment during the early- to mid-training phase, the magnitude of improvement is inherently bounded: once the teacher's signal becomes uninformative, additional training proceeds under standard GRPO and cannot extract further gains from self-distillation. Achieving sustained improvement substantially beyond GRPO would therefore require a teacher whose informational advantage *also* evolves through training, e.g., via an external stronger model or curriculum-style privileged information, which we leave to future work.

K Broader Impact

RLSD offers a token-efficient approach to post-training large reasoning models, achieving stronger performance than standard RLVR with negligible additional computational overhead (only a single extra forward pass per response). This efficiency has several positive societal implications. First, by improving the convergence speed of RLVR, RLSD reduces the total compute required to reach a given level of reasoning performance, helping lower the energy consumption and carbon footprint of training large reasoning models. Second, RLSD requires only the final ground-truth answer as privileged information, rather than expensive verified chain-of-thought traces or stronger teacher models, making advanced reasoning training more accessible to research teams with limited compute or curated datasets, promoting the democratization of AI capabilities. Third, our theoretical analysis of why distribution-matching under information asymmetry fails contributes to a deeper understanding of self-distillation, which may inform safer and more reliable training practices in the broader community.

However, as with any general-purpose post-training method that enhances reasoning capabilities, RLSD carries potential risks. Stronger reasoning models could be misused to generate more persuasive



Figure 6: Token-level credit heatmaps under RLSD on two multimodal reasoning examples. **Top (correct trajectory):** credit concentrates on decisive steps such as identifying the relevant yellow cube and executing the final subtraction, while generic narration like "Looking at the image, I see…" receives smaller weight. **Bottom (incorrect trajectory):** the strongest blame is assigned to the misread bar-model relation "3x = 28.5" and the derived wrong answer "x = 9.5", whereas neutral setup tokens receive comparatively smaller penalties. Green denotes larger token-level credit magnitude; red denotes smaller magnitude.

misleading content or to assist in tasks that benefit from sophisticated multi-step reasoning. Moreover, as a verifier-grounded training paradigm, RLSD could in principle be applied to consolidate harmful capabilities if paired with verifiers that reward unsafe behaviors. We emphasize, however, that these risks are not unique to RLSD but are shared by all reasoning-enhancement and post-training methods, including the RLVR baselines we build upon. Mitigating such risks requires complementary efforts in safety alignment, content filtering, and responsible deployment, which lie beyond the scope of the present work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1) clearly state our three main contributions: (1) identifying the structural cause of OPSD's failure through controlled experiments and formal analysis (Section 3), (2) proposing RLSD as a new training paradigm that decouples update direction from update magnitude (Section 4), and (3) demonstrating consistent gains across eight multimodal reasoning benchmarks and additional video and text-only reasoning settings (Section 5.2). All claims are substantiated by the corresponding theoretical proofs (Appendix B) and empirical results (Tables 2–6).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a dedicated discussion of limitations in Appendix J.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results (Theorems 1–5, Propositions 1–3) are formally stated with explicit assumptions in Sections 3 and 4, and proofs are provided in Appendix B.

4. **Experimental result reproducibility**

   Answer: [Yes]

   Justification: The complete training procedure is provided as Algorithm 1 (Appendix D). All hyperparameters, learning rates, training configurations, privileged information construction, and hardware setup are detailed in Appendix F. The training and evaluation datasets are publicly available.

5. **Open access to data and code**

   Answer: [No]

   Justification: To preserve anonymity during the review process, we do not release code at submission time. However, all training and evaluation datasets are publicly available, our method is built on the publicly released VERL [62] and EasyR1 [63] frameworks, and Algorithm 1 together with Appendix F provides sufficient detail to reproduce our results. We will release the complete code and configuration files upon acceptance.

6. **Experimental setting/details**

   Answer: [Yes]

   Justification: Section 5.1 summarizes the training data, models, baselines, and overall experimental setup. Full implementation details—including learning rates, batch size, context length, sampling temperature, GRPO clipping thresholds, RLSD-specific hyperparameters (λ, ϵ_w, decay schedule), teacher synchronization frequency, privileged information construction for each method, and hardware—are provided in Appendix F.

7. **Experiment statistical significance**

   Answer: [No]

   Justification: Due to the substantial computational cost of RL training on large reasoning models (each training run requires 4 nodes of 8 H200 GPUs over multiple hours), we do not report error bars from multiple random seeds. However, RLSD's improvements over baselines are consistent across eight diverse multimodal benchmarks, two model scales (4B and 8B), and three evaluation domains (multimodal, video, text), suggesting that the observed gains are robust rather than artifacts of seed variation.

8. **Experiments compute resources**

   Answer: [Yes]

   Justification: As described in Appendix F, all experiments are conducted on 4 compute nodes, each equipped with 8 NVIDIA H200 140GB GPUs. The additional computational overhead of RLSD relative to GRPO consists of a single teacher forward pass per response, which is negligible compared to rollout generation (Section 4).

9. **Code of ethics**

   Answer: [Yes]

   Justification: This research conforms to the NeurIPS Code of Ethics. We use only publicly available datasets and base models, do not involve human subjects or private data, and discuss potential broader societal impacts in Appendix K.

10. **Broader impacts**

    Answer: [Yes]

    Justification: We discuss both positive and negative societal impacts in Appendix K, including improvements in training efficiency and accessibility (positive) and potential misuse risks shared by all reasoning-enhancement methods (negative).

11. **Safeguards**

    Answer: [N/A]

    Justification: This paper proposes a general post-training algorithm and does not release new pre-trained models, datasets, or assets that pose specific high-risk misuse concerns beyond those already shared by the underlying base models (Qwen3-VL-Instruct and Qwen3) and standard public benchmarks.

12. **Licenses for existing assets**

    Answer: [Yes]

    Justification: All existing assets used in this work are properly cited. We use Qwen3-VL-Instruct and Qwen3-8B [21] as base models, MMFineReason-123K [12] for training, the VERL [62] and EasyR1 [63] frameworks for implementation, and standard public evaluation benchmarks listed in Appendix E, all of which are released under licenses permitting research use.

13. **New assets**

    Answer: [N/A]

    Justification: This paper does not release new datasets, pre-trained models, or other assets at submission time.

14. **Crowdsourcing and research with human subjects**

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects and therefore does not require IRB approval.

16. **Declaration of LLM usage**

    Answer: [N/A]

    Justification: LLM-based tools were only used for writing assistance (e.g., editing and formatting), which per the NeurIPS LLM policy does not require declaration.