DOWN-WEIGHTING UNCERTAIN SUPERVISION: FROZEN CHECKPOINT–LABEL COMPATIBILITY WEIGHTING FOR GRPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has emerged as an effective paradigm for fine-tuning vision-language models on classification tasks, yet the standard objective does not explicitly distinguish the reliability of the observed supervision associated with different training samples. We study **SCS-GRPO**, which assigns each sample a frozen checkpoint–observed-label compatibility score and applies it as a post-normalization multiplier on the GRPO advantage. We analyze why applying a prompt-level scalar before within-group normalization can largely remove its intended effect, and give a conditional gradient signal-to-noise analysis under explicit assumptions about corrupted supervision and the compatibility ranking. Experiments on multiple public benchmarks show consistent improvements in the evaluated classification and noise settings. An anonymous implementation is available.

1 INTRODUCTION

Reinforcement learning with verifiable rewards has become a standard recipe for eliciting reasoning capabilities in large language models (DeepSeek-AI et al., 2025; Team et al., 2025; Yu et al., 2025), and is increasingly studied in multimodal settings (Xu et al., 2025; Liu et al., 2025). GRPO (Shao et al., 2024) and its variants replace a learned reward model with verifiable programmatic signals, yielding strong performance at low computational cost (Hu, 2025; Zeng et al., 2025). Two common stability mechanisms are a KL divergence penalty that discourages distributional drift from the reference policy and importance-ratio clipping (Schulman et al., 2017) that limits per-step update magnitude. Neither mechanism explicitly uses the compatibility between a frozen checkpoint and each sample's observed label when determining that sample's contribution to the policy update.

This uniformity leaves an important axis unaddressed. In practice, a model's confidence on individual training samples varies substantially: when the predicted probability for an observed label is low, the resulting high cross-entropy may reflect either a *genuinely hard* example or an *incorrectly labeled* one. Noisy supervision can push the policy toward an incorrect target, while some hard but correctly labeled examples may also produce variable gradient estimates and remain useful for learning. KL regularization and clipping do not explicitly distinguish these cases. This motivates studying a complementary prompt-level weighting signal, while recognizing that model confidence alone cannot perfectly separate difficulty, label error, and distribution shift.

Prior work has approached uncertainty, difficulty, and noisy supervision from several related directions. In supervised noisy-label learning, small-loss methods (Han et al., 2018; Li et al., 2020) and training-dynamics analysis (Swayamdipta et al., 2020) exploit the observation that high-loss samples are disproportionately mislabeled, but their designs assume a cross-entropy training objective and do not account for RL-specific structures such as group normalization. Within the GRPO family, DAPO (Yu et al., 2025) filters homogeneous response groups, while noise-corrected GRPO (Mansouri et al., 2025) uses estimated class-conditional noise rates. Most closely, SEED-GRPO (Chen et al., 2025a) dynamically modulates policy updates using semantic entropy estimated from current-policy rollouts. Its multiplier measures label-agnostic rollout dispersion; SCS instead uses a frozen score conditioned on the observed label. DIET (Chen et al., 2025b) separately normalizes outcome and token-penalty advantages before applying a difficulty-dependent coefficient to the penalty term, targeting the

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

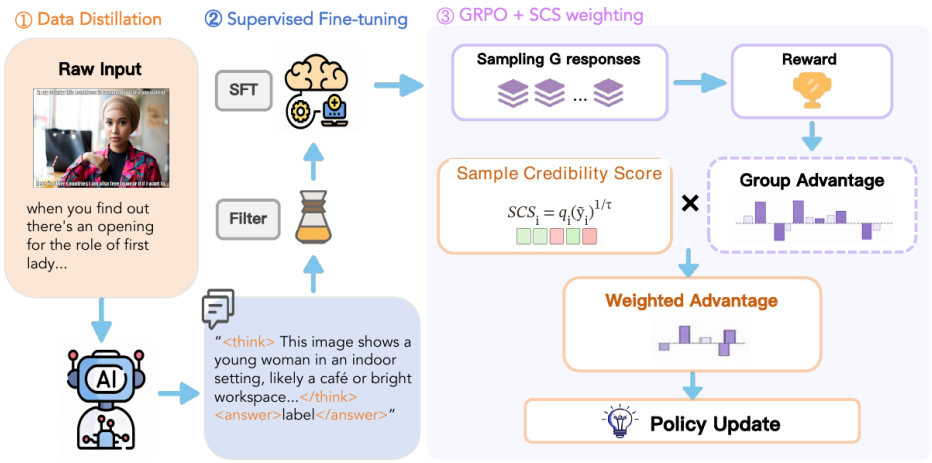


Figure 1: Overview of the SCS-GRPO training pipeline. **Step 1**: A teacher model generates chain-of-thought rationales paired with labeled answers to construct the cold-start training set. **Step 2**: The backbone is supervised fine-tuned (SFT) on this data as a cold-start initialization. **Step 3**: RL fine-tuning with SCS weighting. For each sample, the frozen SFT policy computes a *Sample Credibility Score* that measures frozen checkpoint–observed-label compatibility. During training, G rollouts are sampled per prompt and their rewards are group-normalized to obtain standard GRPO advantages. SCS weights are then applied as a post-normalization multiplier on the advantage: samples with lower checkpoint–label compatibility receive smaller advantage-weighted contributions, while higher-compatibility samples retain more of the original update.

performance–length trade-off in reasoning. These studies make clear that prompt-level modulation of policy updates is not, by itself, unique to SCS. Our focus is the narrower setting of frozen, observed-label-conditional compatibility weighting under potentially imperfect verifiable supervision.

We study **SCS-GRPO**, where the *Sample Credibility Score* (SCS) is a scalar derived from a frozen pre-RL checkpoint's compatibility with the observed label and applied as a post-normalization multiplier on the GRPO advantage. The score is a model-based compatibility signal, not a calibrated posterior probability that an annotation is correct. We analyze how within-group normalization can largely remove a prompt-level scalar applied at the reward level (§3.3), and provide a conditional gradient signal-to-noise analysis under explicit assumptions linking the compatibility ranking to corrupted supervision (§3.2). The method requires no explicit noise-rate estimate or separately trained noise detector; it reuses the frozen pre-RL checkpoint in one offline scoring pass and adds a scalar multiplication to the GRPO update. Experiments on Hateful Memes (Kiela et al., 2020), CHASM (Zheng et al., 2025), and Civil Comments (Borkan et al., 2019) demonstrate that SCS-GRPO achieves consistent and substantial Macro-F1 improvements: up to **+10.7 pp** over Vanilla GRPO and up to **+4.8 pp** over the strongest baseline DAPO in Qwen3-VL-2B-Thinking. Improvements across the evaluated benchmarks, including a clean-data condition, suggest that the weighting may also be useful beyond explicitly corrupted examples, without establishing a general regularization guarantee. Code is available at `https://anonymous.4open.science/r/SCS-E184/`.

Contributions.

- We formulate sample weighting in terms of frozen checkpoint–observed-label compatibility and analyze how the placement of a prompt-level scalar interacts with GRPO's within-group normalization (§3.3).
- We instantiate this idea as **SCS-GRPO**, applying a cached compatibility score after advantage normalization without an explicit noise-rate estimate or a separately trained noise detector.

- We provide a conditional gradient signal-to-noise analysis that states the assumptions under which suppressing corrupted contributions is beneficial, and empirically examine both noisy and clean settings (§3.2).

- We demonstrate consistent gains across three evaluated benchmarks spanning different modalities, controlled corruption, and unmodified benchmark annotations, and characterize how the weighting changes the observed advantage distribution (§4.4).

2 Related Work

Noise-robust RL fine-tuning. Label noise in RL fine-tuning corrupts the reward signal and drives the policy toward suboptimal fixed points (Plesner et al., 2026; Rad et al., 2026; Yang et al., 2026). The most direct remedies estimate class-conditional flip rates and correct rewards accordingly (Mansouri et al., 2025; Cai et al., 2025), but they require an explicit rate-estimation or calibration procedure and corresponding assumptions or calibration labels. Entropy RLVR (Xu et al., 2025) mitigates noise through entropy scheduling at the training-dynamics level, without per-sample discrimination. NoisyRollout (Liu et al., 2025) injects perceptual noise into visual inputs during rollout, targeting a qualitatively different failure mode from annotation noise. SCS does not estimate a flip-rate matrix or modify rollout generation. Instead, it reuses a frozen pre-RL checkpoint to assign an observed-label-conditional compatibility weight. This design introduces a different model-based assumption: the frozen checkpoint's ranking must retain useful information about the observed supervision.

Difficulty-aware training and advantage design. Several works observe that GRPO's within-group normalization can suppress sensitivity to sample difficulty (Yu et al., 2026). SEED-GRPO (Chen et al., 2025a) uses semantic entropy of current-policy rollouts to dynamically modulate policy updates, treating dispersion as prompt-level uncertainty. Its multiplier is independent of which observed label is attached to a fixed prompt and rollout set, whereas SCS is frozen and explicitly conditioned on that label. DIET (Chen et al., 2025b) estimates problem difficulty from pass rates, normalizes outcome and token-penalty advantages separately, and applies a difficulty-dependent coefficient to the penalty advantage. It is designed to allocate response length efficiently; its analysis concerns distortion from jointly normalizing multiple reward components rather than the reliability of an observed classification label. Hard Examples (Pikus et al., 2025) prioritizes the hardest examples under an annotation budget, using base-model success rates to estimate difficulty. VinePPO (Kazemnejad et al., 2025) refines credit assignment via Monte Carlo estimates of expected returns. Group DRO (Panaganti et al., 2026) applies distributionally robust optimization to protect worst-group performance. Difficulty and supervision reliability are related but distinct: a high-loss sample may be genuinely hard and useful, mislabeled, or shifted relative to the anchor. SCS does not perfectly separate these cases. Its narrower distinction is to use frozen checkpoint–observed-label compatibility as a soft weight, rather than current-policy rollout dispersion, a token-length objective, or a binary curriculum decision.

Small-loss weighting in supervised learning. The observation that noisy samples tend to incur higher loss than clean ones underpins a family of methods in supervised noisy-label learning (Natarajan et al., 2013). Co-Teaching (Han et al., 2018) selects the small-loss subset of each mini-batch for peer-network training; DivideMix (Li et al., 2020) fits a Gaussian mixture model to the loss distribution and uses posterior clean-sample probabilities as soft labels. Related training-dynamics analyses (Swayamdipta et al., 2020) likewise use confidence and variability to characterize training examples. Translating the small-loss principle to GRPO is non-trivial: the optimization target differs, and as we show in Section 3.3, within-group normalization can attenuate the intended effect of a prompt-level reward scalar. SCS therefore applies its frozen compatibility weight after normalization. We do not claim that post-normalization advantage weighting itself is new; the contribution studied here is the conditioning signal and its use for noisy verifiable supervision.

3 METHOD

3.1 PRELIMINARIES AND PROBLEM SETUP

Task formulation. We consider verifiable prediction tasks with a finite answer set $\mathcal{Y} = \{0, \ldots, K-1\}$; an output may contain a rationale, but its final answer can be parsed and checked by a verifier. Let (x_i^*, y_i^*) denote a latent clean input–label pair and $(x_i, \tilde{y}_i)$ the pair observed during RL. Corruption may affect the input ($x_i \neq x_i^*$), the attached label ($\tilde{y}_i \neq y_i^*$), or both; x_i may be text-only or multimodal. The method does not require knowing which examples were corrupted.

Let π_{θ_0} be the cold-start checkpoint available immediately before RL, and let π_θ be the trainable policy initialized from it. The construction of θ_0 and, in controlled experiments, the timing of synthetic corruption are specified in Section 4.1. For rollout $o_{i,j}$, we write $r_{i,j} = R(o_{i,j}; x_i, \tilde{y}_i)$ for the scalar verifiable reward used by the base GRPO implementation. This notation covers exact-match rewards and fixed composite rewards. When a base method instead normalizes reward components separately, $A_{i,j}$ below denotes their declared aggregate after that base normalization.

GRPO base update. We adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as our RL backbone. For each prompt x_i, GRPO samples a group of G rollouts. In the single-scalar case, it computes the normalized advantage

$$A_{i,j} = \frac{r_{i,j} - \mu_i}{\sigma_i + \varepsilon}, \tag{1}$$

where μ_i and σ_i are the within-group mean and standard deviation. For exposition, define the clipped per-rollout surrogate

$$\mathcal{L}_{i,j}^{\text{clip}}(A) = \min(\rho_{i,j} A, \text{clip}(\rho_{i,j}, 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}) A)\,, \tag{2}$$

where $\rho_{i,j} = \pi_\theta(o_{i,j} \mid x_i)/\pi_{\text{old}}(o_{i,j} \mid x_i)$. A standard GRPO objective maximizes the expectation of this surrogate, optionally together with a KL penalty to the reference policy. Clipping controls the surrogate importance ratio and the KL term discourages global policy drift; neither explicitly conditions its strength on the compatibility between a prompt and its observed label.

Prompt-level compatibility weighting. In practice, a model's confidence on individual training samples varies substantially. When the model assigns low probability to the observed label $\tilde{y}_i$, the resulting high cross-entropy reflects either a *genuinely hard* example—or an *incorrectly labeled* one. Hard examples may remain useful, while corrupted input–label pairs can direct the policy toward an incompatible target. Thus, low checkpoint probability is an imperfect compatibility signal rather than proof that an example is incorrect. We use this signal to modulate prompt-level contributions without requiring a known corruption rate or per-example clean/noisy indicators.

3.2 SAMPLE CREDIBILITY SCORE

Definition. We derive a scalar **Sample Credibility Score** (SCS) from the frozen checkpoint π_{θ_0}. Let $v(c)$ be the fixed answer verbalizer for class c, and let $z_{i,v(c)}$ be its next-token logit after a fixed answer prefix. For the binary tasks in our main experiments, the class verbalizers are single tokenizer tokens. We renormalize only over the valid classes:

$$q_i(c) = \frac{\exp z_{i,v(c)}}{\sum_{c' \in \mathcal{Y}} \exp z_{i,v(c')}}, \qquad \ell_i = -\log q_i(\tilde{y}_i), \qquad \text{SCS}_i = \exp(-\ell_i/\tau) = q_i(\tilde{y}_i)^{1/\tau}, \tag{3}$$

where $\tau > 0$ is a temperature. Unrelated vocabulary tokens do not enter the normalization, and the score uses the observed label $\tilde{y}_i$, not the most likely class. For a multi-token verbalizer, we replace the single-token logit with its length-normalized, teacher-forced sequence log score; the exact construction is given in Appendix A.

We use "credibility" operationally to mean *frozen checkpoint–label compatibility*. Equation (3) is not a calibrated posterior probability that the annotation is correct: low compatibility can also result from genuine difficulty, anchor error, bias, or distribution shift. SCS only reuses π_{θ_0} and the observed training label; it requires no additional clean label, corruption indicator, or separately trained noise estimator at RL time.

Temperature τ and its effect. The temperature τ controls the selectivity of the scoring: small τ sharpens the distribution, aggressively suppressing high-CE samples but risking the loss of genuinely hard examples; large τ flattens the distribution, recovering near-uniform weighting and standard GRPO. We use $\tau = 5.0$ and report a sensitivity study in Section 4.5. The reported results show stable performance over $\tau \in [3, 5]$ in the tested settings.

Theoretical justification. Weighting changes both the aligned gradient signal and its variance. Appendix B derives a conditional SNR decomposition: weighting improves SNR only when the reduction in noisy or high-variance contributions outweighs the attenuation of useful aligned signal. Compatibility that is monotonically associated with reliability can help satisfy this condition, but that association alone is not a universal guarantee. In particular, anchor errors and difficult clean examples can violate it. We therefore use the analysis to explain a conditional mechanism and treat performance in both corrupted and clean conditions as an empirical question.

3.3 SCS-GRPO Framework

Cancellation of prompt-constant reward weighting. A natural implementation would multiply every reward for prompt i by a positive prompt-constant scalar c_i before group normalization. Because c_i is shared by all G rollouts, the weighted group statistics satisfy $\mu_i^c = c_i\mu_i$ and $\sigma_i^c = c_i\sigma_i$, giving

$$A_{i,j}^{\text{rwd}} = \frac{c_i(r_{i,j} - \mu_i)}{c_i\sigma_i + \varepsilon} = \frac{r_{i,j} - \mu_i}{\sigma_i + \varepsilon/c_i}. \tag{4}$$

For $\varepsilon = 0$ and a non-degenerate group ($\sigma_i > 0$), the scalar cancels exactly and $A_{i,j}^{\text{rwd}} = A_{i,j}$. If $\sigma_i = 0$, both normalized advantages are zero. For $\varepsilon > 0$, cancellation is only approximate and depends on the realized c_i and σ_i; the perturbation is negligible when $\varepsilon \ll c_i\sigma_i$ and $\varepsilon \ll \sigma_i$. This observation applies specifically to a positive prompt-constant scalar shared by all rollouts. It does not cover completion-specific weighting or reward components normalized separately.

Post-normalization advantage weighting. We instead apply SCS_i *after* computing the advantage from unweighted rewards:

$$\tilde{A}_{i,j} = \text{SCS}_i \cdot A_{i,j}, \tag{5}$$

where $A_{i,j}$ is computed via Eq. (1) from the original rewards $\{r_{i,j}\}$. Because (μ_i, σ_i) are computed before the weight is applied, they remain invariant to the weighting. Since $\text{SCS}_i > 0$, it can equivalently multiply the clipped surrogate in Eq. (2). The resulting objective is

$$\mathcal{J}_{\text{SCS}} = \mathbb{E}\left[\sum_{i,j} \mathcal{L}_{i,j}^{\text{clip}}(\text{SCS}_i A_{i,j})\right] - \beta_{\text{KL}} \cdot D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{6}$$

where $\text{SCS}_i = q_i(\tilde{y}_i)^{1/\tau}$ is precomputed with π_{θ_0} and frozen throughout RL. SCS therefore changes only the prompt's contribution to the declared, already-normalized base advantage; it does not alter reward construction or group statistics.

Relation to other advantage-modulation methods. Multiplying an advantage-like policy-update term by a prompt-level scalar is not unique to SCS. Table 1 separates the methods by the signal that defines the scalar, when that signal is updated, and which advantage component it modulates. SCS is distinguished by a frozen, observed-label-conditional compatibility signal and a supervision-corruption motivation; it is not presented as a new generic weighting operator.

Offline computation and implementation. SCS weights are computed once before RL in a cached, batched scoring sweep with the frozen checkpoint π_{θ_0}. The proposed frozen variant makes no further scorer calls during RL. In practice, SCS_i is stored as a scalar per training sample and retrieved at each gradient step. Within the online training loop, the modification is a scalar lookup and multiplication after the base advantage has been formed. Reward construction, group sampling, and the base normalization procedure remain unchanged.

The full procedure is described in Algorithm 1.

Table 1: Conceptual comparison of prompt-level advantage-modulation methods. The methods share a broad scalar-weighting form but use different signals, update schedules, and target components.

Method	Weighting signal	Timing	Target of modulation
SCS (ours)	Observed-label compatibility under π_{θ_0}	Frozen; cached before RL	Declared normalized verifiable advantage
SEED-GRPO (Chen et al., 2025a)	Semantic dispersion of current-policy rollouts	Updated online	Mean-centered outcome advantage
DIET (Chen et al., 2025b)	Current-policy pass rate and training schedule	Updated online	Separately normalized token-penalty component

Algorithm 1 SCS-GRPO

Require: Dataset $\mathcal{D} = \{(x_i, \tilde{y}_i)\}$, frozen checkpoint π_{θ_0}, trainable policy $\pi_\theta \leftarrow \pi_{\theta_0}$, verifier reward R, answer template and verbalizers $\{v(c)\}_{c\in\mathcal{Y}}$, temperature τ
Ensure: Fine-tuned policy π_θ
// Phase 1: cached compatibility scoring
for each sample $x_i \in \mathcal{D}$ **do**
  Score every $c \in \mathcal{Y}$ with π_{θ_0} and the fixed answer template
  Renormalize over valid classes to obtain $q_i(c)$
  $\mathrm{SCS}_i \leftarrow q_i(\tilde{y}_i)^{1/\tau}$
end for
// Phase 2: RL training with compatibility-weighted advantage
for each training step **do**
  Sample mini-batch $\mathcal{B}$ from $\mathcal{D}$
  for each $(x_i, \tilde{y}_i) \in \mathcal{B}$ **do**
    Sample G rollouts $\{o_{i,j}\}_{j=1}^{G}$ from π_θ
    $r_{i,j} \leftarrow R(o_{i,j}; x_i, \tilde{y}_i)$ *% base verifiable reward; no SCS applied here*
    $A_{i,j} \leftarrow \mathrm{BaseGRPOAdvantage}(\{r_{i,k}\}_{k=1}^{G}, j)$ *% including the declared base normalization*
    $\tilde{A}_{i,j} \leftarrow \mathrm{SCS}_i \cdot A_{i,j}$ *% compatibility weighting post-normalization*
  end for
  Update π_θ to maximize $\mathcal{J}_{\mathrm{SCS}}$ using $\{\tilde{A}_{i,j}\}$
end for

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two multimodal classification benchmarks and one text-only benchmark, spanning different modalities, domains, and annotation conditions.

Hateful Memes (HM) (Kiela et al., 2020) is a binary classification benchmark requiring genuine cross-modal understanding. Each sample pairs an image with a caption; the label is *hateful* only when the two modalities combine to convey harmful content, making unimodal shortcuts ineffective. We use the standard splits (8,500 train / 500 dev) and report Macro-F1. For the main comparison (Table 2), we inject *modality-swap* noise at $\eta = 20\%$: each noisy sample has its image replaced with a randomly selected image from another sample, breaking cross-modal coherence while preserving text. For the robustness analysis (§4.3), we inject *symmetric label-flip* noise at $\eta \in \{0\%, 10\%, 20\%, 30\%, 40\%\}$. In both controlled-noise protocols, the shared cold-start checkpoint is trained on the original training pairs; corruption is introduced only into the RL training data after cold-start SFT, and the evaluation split remains unchanged.

CHASM (Zheng et al., 2025) is a multimodal advertisement detection benchmark comprising 4,992 social media posts (3,493 train / 499 dev / 1,000 test), each containing one or more images paired with text. The task is binary classification (advertisement vs. non-advertisement); we report Macro-F1 over both classes.

Table 2: Main results. HM is evaluated under 20% modality-swap corruption; CHASM and CC use their unmodified original annotations. For RL methods, all entries are means over three seeds. Sample standard deviation is shown in parentheses for Macro-F1 and omitted for per-class F1. SFT is the shared pre-RL reference. **Bold**: best per column.

	HM		CHASM		CC	
Method	Macro-F1	Hateful-F1	Macro-F1	Ad-F1	Macro-F1	Toxic-F1
Qwen3-VL-2B-Thinking						
SFT	53.5	37.2	53.4	36.8	53.7	44.5
Vanilla GRPO	53.3 (0.7)	38.7	56.0 (0.3)	37.1	57.2 (0.7)	45.3
DAPO	59.9 (0.4)	52.5	60.3 (0.8)	41.8	63.5 (0.4)	53.0
Dr.GRPO	58.9 (0.6)	48.1	59.5 (0.4)	40.5	59.0 (0.7)	49.2
SCS (ours)	**64.0 (0.2)**	**61.6**	**65.1 (0.6)**	**44.0**	**67.2 (0.2)**	**65.0**
Qwen3-VL-4B-Thinking						
SFT	55.1	39.0	54.8	38.1	55.3	46.2
Vanilla GRPO	55.9 (0.7)	40.5	57.3 (0.6)	43.3	64.2 (0.4)	47.0
DAPO	61.8 (0.3)	54.3	62.2 (0.3)	43.6	62.3 (0.6)	55.1
Dr.GRPO	60.8 (0.6)	50.0	61.3 (0.2)	42.3	60.7 (0.7)	52.4
SCS (ours)	**66.2 (0.3)**	**63.5**	**67.2 (0.3)**	**48.2**	**68.1 (0.5)**	**66.3**

Civil Comments (CC) (Borkan et al., 2019) is a large-scale text-only toxicity detection benchmark. Comments are annotated via crowd-sourcing and assigned binary labels (toxic / non-toxic) based on majority vote. We use the original annotations without injecting additional corruption. We sample a class-balanced subset of 20,000 training examples and evaluate on the official test split, reporting Macro-F1.

Baselines. We compare SCS-GRPO against four baselines using the same backbone; all RL baselines use the same training budget:

- **SFT**: the cold-start supervised fine-tuned model before any RL training, serving as the shared pre-RL reference.
- **Vanilla GRPO**: standard GRPO with the same verifier reward and no sample weighting.
- **DAPO** (Yu et al., 2025): dynamic group sampling that discards homogeneous rollout groups to prevent zero-gradient collapse.
- **Dr.GRPO (noise-corrected)** (Mansouri et al., 2025): the authors' noise-corrected Dr.GRPO variant, using estimated class-conditional reward-flip rates under a Bernoulli channel model; it is abbreviated as Dr.GRPO in figures and tables.

Implementation details. Our primary backbone is Qwen3-VL-2B-Thinking (Bai et al., 2025), fine-tuned with ms-swift (Zhao et al., 2025). We additionally conduct experiments with Qwen3-VL-4B-Thinking to verify that results are not specific to a single model scale. To construct the shared cold start, Qwen3-VL-32B-Instruct generates one rationale per original training example; we retain generations whose parsed final answer agrees with the training annotation. Cold-start SFT uses AdamW for three epochs with learning rate 2×10^{-5}, batch size 128, 3% warm-up, cosine decay, and seed 41; both backbone scales use this schedule. We set group size $G = 8$, learning rate 3×10^{-6}, and train for one epoch. SCS weights are computed offline in a cached, batched scoring sweep with the pre-RL (SFT) policy and stored as scalars. We select temperature $\tau = 5.0$ from $\{1, 3, 5, 10\}$ on an uncorrupted held-out validation split, disjoint from RL training and final evaluation, and report its sensitivity in §4.5. Macro-F1 results for RL methods are averaged over three random seeds (42, 123, 456); reporting details for the accompanying per-class values are given in the table captions. Full hyperparameter details are in Appendix A.

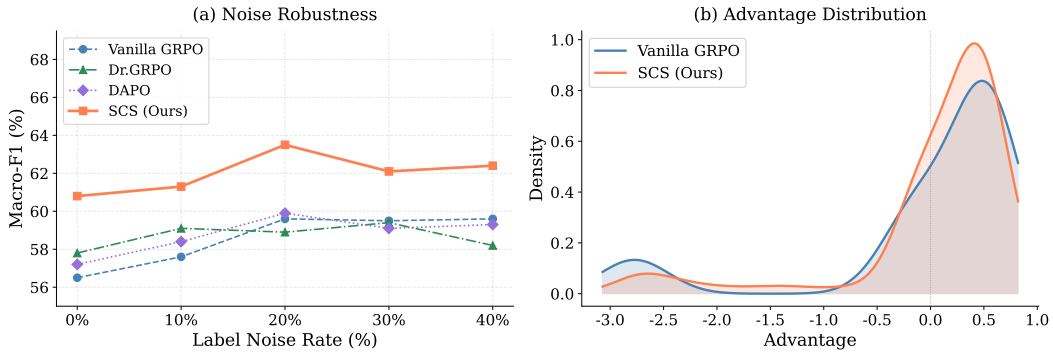


Figure 2: **(a) Noise robustness.** Macro-F1 on Hateful Memes as a function of symmetric label-flip noise rate η. SCS-GRPO consistently outperforms Vanilla GRPO at every noise level, including $\eta = 0\%$ (clean data). The clean-setting result suggests that the weighting can be useful beyond explicitly flipped labels in this experiment, without implying a general clean-data guarantee. **(b) Advantage distribution.** Kernel density estimate of effective advantage values under Vanilla GRPO ($A_{i,j}$) and SCS-GRPO ($\tilde{A}_{i,j} = \mathrm{SCS}_i \cdot A_{i,j}$). SCS contracts the distribution (σ: 1.000 $\rightarrow$ 0.805), as expected when normalized advantages are multiplied by weights in $(0, 1]$.

4.2 Main Results

Results. Across all three benchmarks and both backbone sizes, SCS attains the highest Macro-F1 (Table 2). With the 2B model, it improves over Vanilla GRPO by **+10.7**, **+9.1**, and **+10.0 pp** on HM, CHASM, and CC, respectively, and over DAPO by **+4.1**, **+4.8**, and **+3.7 pp**. These gains are accompanied by minority-class improvements of +22.9 Hateful-F1, +6.9 Ad-F1, and +19.7 Toxic-F1 over Vanilla GRPO. On HM, Vanilla GRPO (53.3%) does not improve over the SFT reference (53.5%), whereas SCS reaches 64.0%. With the 4B backbone, SCS remains best at 66.2%, 67.2%, and 68.1% Macro-F1, showing that the trend is not restricted to one model scale.

4.3 Robustness to Label Noise

To assess sensitivity to noise intensity, we apply symmetric label-flip noise to the HM RL training data at five rates $\eta \in \{0\%, 10\%, 20\%, 30\%, 40\%\}$ and evaluate on the unchanged evaluation split. This complements the modality-swap main experiment by stress-testing the method under varying proportions of corrupted labels. Figure 2(a) plots Macro-F1 for SCS-GRPO and the three RL baselines across all noise levels; we focus on the matched comparison with Vanilla GRPO.

SCS maintains a consistent advantage at every noise level. On clean data ($\eta = 0\%$), SCS already outperforms Vanilla GRPO by +4.3 pp (60.8% vs. 56.5%). Across the five tested rates, the reported gain over Vanilla GRPO ranges from +2.6 to +4.3 pp and is largest in the clean condition; there is no monotonic increase with the corruption rate. This result is consistent with, but does not by itself establish, a regularization effect or a particular causal mechanism.

4.4 Advantage Distribution Analysis

To understand *how* SCS reshapes the learning dynamics, we examine the distribution of effective advantage values before and after applying SCS weights. Under Vanilla GRPO, all advantages are passed to the policy gradient update with unit weight; under SCS-GRPO, each advantage is scaled by the corresponding credibility score, selectively suppressing low-credibility samples. Figure 2(b) shows kernel density estimates of these distributions computed on the HM training set under 20% modality-swap noise.

SCS reduces the standard deviation of the advantage distribution by approximately 20% (1.000 $\rightarrow$ 0.805). Because every SCS weight lies in $(0, 1]$, some contraction is a direct consequence of the multiplication rather than evidence that the down-weighted samples are mislabeled. The plot

is therefore a descriptive implementation check showing that SCS changes the scale and shape of prompt contributions; it is not used as a causal explanation of the gains in Table 2.

4.5 Ablation Study

We conduct ablations under the same data conditions as the main experiment. Values in Tables 3 and 4 are means over seeds 42, 123, and 456, with sample standard deviations in parentheses.

Temperature τ. Table 3 reports HM Macro-F1 as τ varies from 1.0 to 10.0, with all other settings fixed.

Table 3: Sensitivity to temperature τ. Macro-F1 (%) is averaged over three seeds; sample standard deviation is shown in parentheses.

τ	1.0	3.0	**5.0**	10.0
HM Macro-F1	63.1 (0.3)	63.8 (0.5)	**64.0 (0.2)**	62.7 (0.8)
CHASM Macro-F1	64.5 (0.5)	64.9 (0.6)	**65.1 (0.6)**	64.3 (0.7)
CC Macro-F1	66.6 (0.4)	67.0 (0.3)	**67.2 (0.2)**	66.4 (0.5)

Small τ produces more selective weighting, while large τ approaches uniform weighting. Among the tested values, $\tau = 5.0$ gives the highest reported Macro-F1 on all three datasets, and the values at $\tau \in \{3, 5\}$ are comparatively close.

Group size G. Table 4 reports all six metrics for $G \in \{4, 8, 16\}$.

Table 4: Effect of group size G on SCS-GRPO (Qwen3-VL-2B). Entries are mean F1 (%) over three seeds with sample standard deviation in parentheses. **Bold**: best per column.

	HM		**CHASM**		**CC**	
G	Macro-F1	Hateful-F1	Macro-F1	Ad-F1	Macro-F1	Toxic-F1
4	62.5 (0.6)	60.1 (0.9)	63.8 (0.4)	42.5 (0.4)	66.7 (0.3)	64.2 (0.5)
8	**64.0 (0.2)**	**61.6 (0.7)**	**65.1 (0.6)**	**44.0 (0.6)**	**67.2 (0.2)**	**65.0 (0.4)**
16	63.3 (0.9)	60.8 (0.5)	64.5 (0.3)	43.2 (0.8)	66.4 (0.4)	64.1 (0.8)

$G = 8$ gives the highest mean F1 in every reported column. Larger G increases rollout diversity within each group, providing a richer basis for advantage normalization, but incurs higher per-step compute.

5 Conclusion

We introduced **SCS-GRPO**, which uses a frozen pre-RL checkpoint's compatibility with the observed label to reweight GRPO advantages after within-group normalization. This placement preserves prompt-level weighting that reward-level scaling would largely cancel; implementation requires one cached scoring pass and no noise-rate estimate or separate detector. Across three benchmarks and two backbone sizes, SCS consistently outperforms the evaluated baselines. Its gains in the tested clean setting may reflect a regularization benefit, but do not imply a general guarantee.

Limitations and future directions. SCS depends on a frozen pre-RL checkpoint whose compatibility ranking may become stale or reflect anchor errors, bias, and distribution shift. Low observed-label compatibility can also indicate a genuinely difficult clean example rather than corrupted supervision, so the score is not a calibrated posterior probability of annotation correctness. The current evaluation focuses on binary classification with finite answer spaces; whether compatibility signals generalize to open-ended generation settings remains an open question. We also do not report a matched head-to-head implementation of SEED-GRPO, so the conceptual comparison in Table 1 should not be read as an empirical ranking. Slowly updated or independent anchors and compatibility signals beyond cross-entropy are useful directions for future work.

AI USE STATEMENT

Generative AI tools were used for writing assistance, including grammar checking and language polishing. All AI-assisted text was reviewed and edited by the authors, who take responsibility for the final content, scientific claims, and reported results. Model use that forms part of the experimental pipeline is described in the paper and is distinct from writing assistance.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *ArXiv*, abs/2511.21631, 2025. URL https://api.semanticscholar.org/CorpusID:283262018.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The Web Conference*, pp. 491–500, 2019.

Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers. *ArXiv*, abs/2510.00915, 2025. URL https://api.semanticscholar.org/CorpusID:281706047.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. SEED-GRPO: Semantic entropy enhanced GRPO for uncertainty-aware policy optimization, 2025a. URL https://arxiv.org/abs/2505.12346.

Weize Chen, Jiarui Yuan, Tailin Jin, Ning Ding, Huimin Chen, Zhiyuan Liu, and Maosong Sun. The overthinker's DIET: Cutting token calories with difficulty-aware training. *ArXiv*, abs/2505.19217, 2025b. URL https://api.semanticscholar.org/CorpusID:278905337.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, JingChang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, Ruiqi Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shao-Kang Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xia Yu, Wentao Zhang, Wangding Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao

Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Yukun Zha, Yuting Yan, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-Long Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638, 2025. URL https://api.semanticscholar.org/CorpusID:275789950.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Wai-Hung Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:52065462.

Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *ArXiv*, abs/2501.03262, 2025. URL https://api.semanticscholar.org/CorpusID:277509836.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. In *International Conference on Machine Learning*, 2025. URL https://api.semanticscholar.org/CorpusID:273026145.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *ArXiv*, abs/2005.04790, 2020. URL https://api.semanticscholar.org/CorpusID:218581273.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *ArXiv*, abs/2002.07394, 2020. URL https://api.semanticscholar.org/CorpusID:211146562.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *ArXiv*, abs/2504.13055, 2025. URL https://api.semanticscholar.org/CorpusID:277857224.

Omar El Mansouri, Fathinah Asma Izzati, Mohamed El Amine Seddik, and Salem Lahlou. Noise-corrected GRPO: From noisy rewards to unbiased gradients, 2025. URL https://arxiv.org/abs/2510.18924.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Neural Information Processing Systems*, 2013. URL https://api.semanticscholar.org/CorpusID:423350.

Kishan Panaganti, Zhenwen Liang, Wenhao Yu, Haitao Mi, and Dong Yu. Group distributionally robust optimization-driven reinforcement learning for llm reasoning. *ArXiv*, abs/2601.19280, 2026. URL https://api.semanticscholar.org/CorpusID:285071187.

Benjamin Pikus, Pratyush Ranjan Tiwari, and Burton Ye. Hard examples are all you need: Maximizing grpo post-training under annotation budgets. *ArXiv*, abs/2508.14094, 2025. URL https://api.semanticscholar.org/CorpusID:280692329.

Andreas Lindhardt Plesner, Francisco Guzmán, and Anish Athalye. An imperfect verifier is good enough: Learning with noisy rewards, 2026. URL https://api.semanticscholar.org/CorpusID:287256014.

Ali Rad, Khashayar Filom, Darioush Keivan, Peyman Mohajerin Esfahani, and Ehsan Kamalinejad. Rate or fate? rlvεr: Reinforcement learning with verifiable noisy rewards. *ArXiv*, abs/2601.04411, 2026. URL https://api.semanticscholar.org/CorpusID:285472873.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL https://api.semanticscholar.org/CorpusID:28695052.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL https://api.semanticscholar.org/CorpusID:267412607.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://api.semanticscholar.org/CorpusID:221856637.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Feng Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Haochen Ding, Hao-Xing Hu, Haoming Yang, Hao Zhang, Haotian Yao, Hao-Dong Zhao, Haoyu Lu, Haoze Li, Hao Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jia-Xing Guo, Jianling Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Li-Na Shi, Li tao Ye, Long Yu, Meng xiao Dong, Neo Y. Zhang, Ning Ma, Qi Pan, Qucheng Gong, Shaowei Liu, Shen Ma, Shu-Yan Wei, Sihan Cao, Si-Da Huang, Tao Jiang, Wei-Wei Gao, Weiming Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wen He, Xian sen Wei, Xian-Xian Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xue biao Pan, Y. Charles, Yang Li, Yan-Ling Hu, Yangyang Liu, Yanru Chen, Ye-Jia Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Yingbo Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhengxin Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziya Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms. *ArXiv*, abs/2501.12599, 2025. URL https://api.semanticscholar.org/CorpusID:275789974.

Donglai Xu, Hongzheng Yang, Yuzhi Zhao, Pingping Zhang, Jinpeng Chen, Wenao Ma, Zhijian Hou, Mengyang Wu, Xiaolei Li, Senkang Hu, Ziyi Guan, Jason Chun Lok Li, and Lai-Man Po. From exploration to exploitation: A two-stage entropy rlvr approach for noise-tolerant mllm training. *ArXiv*, abs/2511.07738, 2025. URL https://api.semanticscholar.org/CorpusID:282921502.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Sharon Li, Haobo Wang, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, and Gang Chen. Can llms learn to reason robustly under noisy supervision?, 2026. URL https://api.semanticscholar.org/CorpusID:287201557.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Honglin Yu, Weinan Dai, Yuxuan Song, Xiang Wei, Haodong Zhou, Jingjing Liu, Wei Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yong-Xu Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476, 2025. URL https://api.semanticscholar.org/CorpusID:277104124.

Zhiqi Yu, Zhangquan Chen, Mengting Liu, Heye Zhang, and Liangqiong Qu. Unveiling implicit advantage symmetry: Why grpo struggles with exploration and difficulty adaptation. *ArXiv*, abs/2602.05548, 2026. URL https://api.semanticscholar.org/CorpusID:285304028.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025. URL https://arxiv.org/abs/2503.18892.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2025. URL https://arxiv.org/abs/2408.05517.

Jingyi Zheng, Tianyi Hu, Yule Liu, Zhen Sun, Zongmin Zhang, Wenhan Dong, Zifan Peng, and Xinlei He. Chasm: Unveiling covert advertisements on chinese social media. https://huggingface.co/datasets/Jingyi77/CHASM-Covert_Advertisement_on_RedNote, 2025.

A Implementation Details

Table 5: RL hyperparameter settings used in all experiments unless otherwise noted.

Hyperparameter	**Value**	**Notes**
Backbone	Qwen3-VL-2B/4B-Thinking	frozen during SCS scoring
Training framework	ms-swift	
Optimizer	AdamW	
Learning rate	3×10^{-6}	
LR scheduler	cosine	
Warmup steps	10	
Training epochs	1	
Batch size	16	
Group size G	8	rollouts per prompt
KL coefficient β_{KL}	0.01	
Clip ratio ϵ_{clip}	0.2	
SCS temperature τ	5.0	sensitivity in Table 3
Max sequence length	2048	
Random seeds	42, 123, 456	
Reward function		
Accuracy reward	$1.0 \times \mathbf{1}[o_{i,j} = \tilde{y}_i]$	
Format reward	$2.0 \times \mathbf{1}$[valid format]	presence of `<answer>` tag
Length reward	$0.5 \times \min(1, L/L_{\mathrm{target}})$	$L_{\mathrm{target}} = 512$ tokens

Full hyperparameter table.

SCS offline computation. No classification head or sampled confidence is used. We append a fixed answer prefix T and score every class verbalizer with the frozen checkpoint π_{θ_0}. In the main binary tasks, "0" and "1" are single tokenizer tokens. If $z_{i,v(c)}$ is the next-token logit of class c, then the class distribution is obtained by renormalizing only over valid answers:

$$q_i(c) = \frac{\exp z_{i,v(c)}}{\sum_{c' \in \mathcal{Y}} \exp z_{i,v(c')}}. \tag{7}$$

The stored weight is $\mathrm{SCS}_i = q_i(\tilde{y}_i)^{1/\tau}$. Thus unrelated vocabulary tokens do not enter the normalization, and the score uses the observed label rather than the most likely class.

For a multi-token verbalizer $v(c) = (v_1(c), \ldots, v_{|v(c)|}(c))$, we teacher-force the complete string and length-normalize its log score:

$$s_i(c) = \frac{1}{|v(c)|} \sum_{t=1}^{|v(c)|} \log p_{\theta_0}(v_t(c) \mid x_i, T, v_{<t}(c)), \tag{8}$$

$$q_i(c) = \frac{\exp s_i(c)}{\sum_{c' \in \mathcal{Y}} \exp s_i(c')}. \tag{9}$$

All alternatives are evaluated in one cached, batched training-set scoring sweep. The resulting float32 scalar is indexed by sample ID, and the frozen variant makes no further scorer calls during RL. The score is a checkpoint–label compatibility measure, not a calibrated estimate of the probability that an annotation is correct. It requires white-box logits or equivalent log-probability access.

Compute resources. All experiments are conducted on 4× NVIDIA H800 80 GB GPUs with 4-way data parallelism. SCS scoring is performed once before RL; training thereafter uses the same rollout and update schedule as the corresponding base GRPO run.

B CONDITIONAL ANALYSIS

We state the normalization identity used in Section 3.3 and an explicit gradient signal-to-noise decomposition for the mechanism discussed in Section 3.2. The latter is conditional and does not claim that checkpoint compatibility identifies corrupted supervision in every setting.

B.1 PROMPT-CONSTANT REWARD WEIGHTING

Proposition 1 (Prompt-constant scaling) *Let $c_i > 0$ be shared by all rollouts of prompt i, and define μ_i^c and σ_i^c as the mean and standard deviation of $\{c_i r_{i,j}\}_{j=1}^G$. Then*

$$A_{i,j}^c = \frac{c_i r_{i,j} - \mu_i^c}{\sigma_i^c + \varepsilon} = \frac{r_{i,j} - \mu_i}{\sigma_i + \varepsilon/c_i}. \tag{10}$$

Consequently, for $\varepsilon = 0$ and $\sigma_i > 0$, the scaled and unscaled advantages are identical. For $\varepsilon > 0$, the identity is only approximate and depends on c_i and σ_i.

Proof. Because c_i is positive and constant across the group,

$$\mu_i^c = c_i \mu_i, \qquad \sigma_i^c = c_i \sigma_i. \tag{11}$$

Substitution gives Equation (10). Relative to the unscaled $A_{i,j} = (r_{i,j} - \mu_i)/(\sigma_i + \varepsilon)$, the exact difference is

$$A_{i,j}^c - A_{i,j} = \frac{(r_{i,j} - \mu_i)\varepsilon(1 - 1/c_i)}{(\sigma_i + \varepsilon/c_i)(\sigma_i + \varepsilon)}. \tag{12}$$

It is zero at $\varepsilon = 0$ for a non-degenerate group and small when the stabilizer is small relative to both σ_i and $c_i\sigma_i$. If $\sigma_i = 0$, the numerator is zero for every rollout and both conventional normalized advantages are zero. □

B.2 CONDITIONAL GRADIENT SNR DECOMPOSITION

Let the stochastic contribution associated with sample i be $g_i = s_i + \xi_i$, where $s_i = \mathbb{E}[g_i]$ is its mean signal, $\mathbb{E}[\xi_i] = 0$, and $v_i = \mathbb{E}[\|\xi_i\|_2^2]$. We make the explicit idealization that the noise terms are uncorrelated across samples. For deterministic nonnegative weights w_i, define $G(w) = \sum_i w_i g_i$ and

$$\mathrm{SNR}(w) = \frac{\|\mathbb{E}[G(w)]\|_2^2}{\mathbb{E}[\|G(w) - \mathbb{E}[G(w)]\|_2^2]}. \tag{13}$$

Dividing $G(w)$ by a common positive normalization, such as $\sum_i w_i$, leaves this ratio unchanged.

Proposition 2 (Conditional comparison with uniform weighting) *Under the assumptions above,*

$$\mathrm{SNR}(w) = \frac{\|\sum_i w_i s_i\|_2^2}{\sum_i w_i^2 v_i}. \tag{14}$$

Whenever the denominators and the uniform-weight signal are nonzero, $\mathrm{SNR}(w) > \mathrm{SNR}(\mathbf{1})$ if and only if

$$\frac{\|\sum_i w_i s_i\|_2^2}{\|\sum_i s_i\|_2^2} > \frac{\sum_i w_i^2 v_i}{\sum_i v_i}. \tag{15}$$

Thus weighting helps precisely when the relative loss of squared aligned signal is smaller than the relative reduction in stochastic variation.

Proof. Linearity gives $\mathbb{E}[G(w)] = \sum_i w_i s_i$. The zero-mean, uncorrelated noise assumption gives $\mathbb{E}[\|\sum_i w_i \xi_i\|_2^2] = \sum_i w_i^2 v_i$, proving Equation (14). Comparing this expression with $w_i = 1$ and cross-multiplying positive denominators yields Equation (15). □

Corollary 1 (Oracle attenuation case) *Partition the samples into $\mathcal{C}$ and $\mathcal{N}$. Suppose $s_i = 0$ for every $i \in \mathcal{N}$, retain $w_i = 1$ for every $i \in \mathcal{C}$, and choose $0 \le w_i \le 1$ for $i \in \mathcal{N}$. If the aggregate clean signal is nonzero and at least one noisy component with $v_i > 0$ is strictly down-weighted, then $\mathrm{SNR}(w) > \mathrm{SNR}(\mathbf{1})$.*

Proof. The mean signal is unchanged, while $\sum_i w_i^2 v_i < \sum_i v_i$. The claim follows from Equation (14). □

The corollary describes an oracle case, not a guarantee for SCS. SCS does not observe the partition and can attenuate useful signal from difficult clean examples; the frozen checkpoint can also be biased, inaccurate, or shifted. Equation (15) explains the condition that a compatibility ranking must satisfy for an SNR improvement. The clean-setting results in Section 4.3 are empirical observations rather than a formal consequence of this analysis.

C Additional Experimental Results

C.1 Per-Class F1 Breakdown

Table 6: Per-class F1 on Hateful Memes (20% modality-swap corruption), averaged over three seeds. These values complement the three-seed Macro-F1 values in Table 2; standard deviations are omitted for consistency with that table.

Method	Hateful-F1	Not-Hateful-F1
Vanilla GRPO	38.7	67.9
DAPO	52.5	67.2
Dr.GRPO	48.1	69.7
SCS (ours)	**61.6**	**66.5**

SCS achieves the highest mean Hateful-F1 for the minority class. Competing methods achieve higher Not-Hateful-F1 at the cost of significantly lower Hateful-F1, resulting in lower Macro-F1 overall. The smaller gap between the two per-class F1 values for SCS is a descriptive property of these runs; it is not, by itself, evidence about the learned decision boundary or the cause of the aggregate improvement.

D Prompt Template and Verbalizers

GRPO training prompt (joint).

```
System: You are a multimodal classification assistant.
        Analyze the image and text, then output your answer
        in the format: <answer>LABEL</answer>

User:   [IMAGE]
        Caption: {text}
        Question: Is this content hateful? Answer with 0 (no) or 1 (yes).
```

The dataset-specific question and label semantics replace the final line for CHASM and Civil Comments, while the output schema remains fixed. For offline scoring, the frozen checkpoint is evaluated immediately after the answer prefix `<answer>`; the valid class verbalizers are `0` and `1`. The tokenizer check confirms that each is a single token for the reported checkpoints. No decoding or sampling is used in this scoring pass.