STRUCTURED CATEGORY-WISE SELECTIVE ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) aligns teacher and student predictions on student-generated trajectories, but conventional token-level objectives entangle semantic-category choices with preferences among alternative realizations within the same category. We propose *Structured Category-wise Selective On-Policy Distillation* (SCS-OPD), which separates category-mass alignment from intra-category supervision. Category-level alignment drives the main rollout update; an optimizer-matched finite-step selective gating mechanism then compares category-plus-residual and category-only updates, retaining intra-category residuals only when the gate indicates a lower category-level loss. Across OPD experiments covering multiple methods, the Qwen and DeepSeek model families, and multiple random seeds, SCS-OPD achieves the best performance on most mathematical benchmarks. In the settings where it leads, SCS-OPD improves over methods including Reverse-KL OPD, EOPD, and TSD by 0.3–3.4 percentage points, while maintaining stable performance on out-of-domain benchmarks. Mechanistic analysis shows that, after semantic aggregation, intra-category supervision is sparse, highly concentrated, and heterogeneous. Selective gating over fine-grained realizations therefore provides substantial optimization benefits for OPD.

1 INTRODUCTION

On-policy distillation (OPD) reduces the mismatch between training and deployment by evaluating teacher supervision on prefixes generated by the student itself (Agarwal et al., 2024). This addresses a central limitation of distillation on fixed trajectories: at inference time, a student conditions on its own history, so supervision obtained only on externally supplied prefixes may not reflect the states the student actually visits. Existing OPD methods have therefore focused on trajectory construction, divergence choice, uncertainty, and token-level weighting (Gu et al., 2024; Ko et al., 2024; Jin et al., 2026).

A separate issue is the *granularity* of the supervision. Standard token-level divergence treats distinct vocabulary items as distinct outcomes even when several candidates represent closely related realizations of the same local semantic choice. Consequently, token-level alignment mixes two effects: matching how much probability mass the teacher and student assign to a semantic category, and matching how that mass is distributed among realizations inside the category. The latter can contain useful fine-grained teacher information, but it can also redirect the local optimization signal even when the category-level choice is preserved.

We study this interaction directly. At each student-generated prefix, we construct local semantic categories over the teacher candidate support and decompose normalized candidate-level reverse KL into a category-level term F and an intra-category conditional term. We use a detached residual R to represent within-category supervision while removing the explicit derivative through category weights. The chain-rule decomposition also makes the coupling explicit: candidate-level reverse KL combines category allocation with conditional fitting difficulty, motivating separate treatment of the two levels. SCS-OPD therefore uses category alignment for the main rollout update and handles intra-category supervision in a separate auxiliary stage.

The key selection mechanism is an optimizer-matched finite-step comparison. From the same model parameters and persistent auxiliary AdamW state, SCS-OPD compares an $F + \beta R$ trial with an F-only trial and retains R only when its inclusion yields a lower post-update category loss. This

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.

measures the incremental effect of intra-category supervision after the actual optimizer step rather than relying only on a gradient-level proxy.

Contributions. Our contributions are threefold:

- We formulate semantic-category supervision for OPD, separating category-mass alignment from conditional within-category fitting. We analyze their coupling under reverse KL and use category-preserving interventions to show how fine-grained allocation changes token-level objectives and gradients.
- We introduce an optimizer-matched finite-step selector that measures the incremental effect of intra-category supervision on category alignment. Matched category-only and category-plus-residual trials determine which residuals enter the auxiliary update, while persistent optimizer states keep the comparison faithful to the implemented update rule.
- We validate the method through complementary benchmark, ablation, and mechanistic evidence. SCS-OPD leads AIME24 and MATH500 at all three reported Qwen3 seeds and leads three of four benchmarks in a second teacher–student configuration. Controlled training isolates the benefit of selection over both removing and uniformly retaining residuals; local probes explain why these residuals need not support category alignment uniformly.

2 Related Work

OPD methods adapt student-generated trajectories and divergence objectives (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024), or allocate supervision through uncertainty (Jin et al., 2026). SCS-OPD instead controls the within-category component of supervision at a given context.

Semantic Entropy, CCP, and STC aggregate semantically related outcomes for uncertainty estimation (Kuhn et al., 2023; Fadeeva et al., 2024; Cao et al., 2026); DKD separates components of a distillation objective (Zhao et al., 2022). Our focus is the optimization interaction between category masses and within-category distributions.

Gradient-similarity weighting, PCGrad, and ATTITTUD assess or modify auxiliary updates to reduce interference (Du et al., 2018; Yu et al., 2020; Dery et al., 2021). We compare category-only and category-plus-residual updates after matched optimizer steps. Appendix G.1 gives the broader literature discussion.

3 Preliminaries

Let π_θ be the student and π_T the frozen teacher, sharing a tokenizer and vocabulary. OPD minimizes $\mathbb{E}_{s\sim d_\theta}[\mathcal{D}(\pi_T(\cdot \mid s), \pi_\theta(\cdot \mid s))]$ on student-generated prefixes, holding the rollout distribution fixed during each update. At a fixed prefix, p_v and q_v denote student and teacher probabilities. Reverse KL is $\sum_v p_v \log(p_v/q_v)$ and forward KL is $\sum_v q_v \log(q_v/p_v)$. Neither encodes semantic equivalence between vocabulary items. Both are baselines; Appendix G.2 provides the full definitions.

4 Method

We introduce *Structured Category-wise Selective On-Policy Distillation* (SCS-OPD). The method separates two decisions: *which semantic category should receive probability mass*, and *how probability should be distributed among candidate realizations within that category*. Category-level alignment drives the main on-policy update. Intra-category supervision enters a separate auxiliary stage only when an optimizer-matched finite-step trial indicates a better category-level outcome than an otherwise identical update without that supervision.

4.1 Constructing Semantic Categories

Let π_θ and π_T denote the student and frozen teacher, which share a tokenizer and vocabulary. Following on-policy distillation (Agarwal et al., 2024), we evaluate both models at prefixes generated

by the student. At a prefix s, let $\mathcal{S}_s = \{k_1, \ldots, k_K\}$ be the teacher's top-K token support. Define

$$p_i = \pi_\theta(k_i \mid s), \quad Z_p = \sum_{i=1}^{K} p_i, \qquad q_i = \pi_T(k_i \mid s), \quad Z_q = \sum_{i=1}^{K} q_i. \tag{1}$$

These are vocabulary probabilities restricted to a shared support; they need not sum to one. We retain Z_p and Z_q explicitly to distinguish probability masses from normalized distributions. We construct a local partition $\mathcal{C}_s = \{C_1, \ldots, C_{M_s}\}$ using decoded token forms and WordNet synsets (Miller, 1995). The grouping is fixed during differentiation and is treated as a rule-based approximation to semantic equivalence rather than a perfect contextual equivalence relation. Exact merge rules, exclusions, and component-size constraints are given in Appendix A.1.

4.2 Category Alignment and Intra-Category Supervision

For $i \in C_c$, define category masses, normalized category probabilities, and within-category conditionals:

$$\begin{aligned} P_c &= \sum_{i \in C_c} p_i, \quad Q_c = \sum_{i \in C_c} q_i, \\ \widetilde{P}_c &= P_c / Z_p, \quad \widetilde{Q}_c = Q_c / Z_q, \\ u_{i|c} &= p_i / P_c, \quad t_{i|c} = q_i / Q_c. \end{aligned} \tag{2}$$

The category-level loss and conditional discrepancy are

$$F = \sum_c \widetilde{P}_c \log \frac{\widetilde{P}_c}{\widetilde{Q}_c}, \qquad d_c = \sum_{i \in C_c} u_{i|c} \log \frac{u_{i|c}}{t_{i|c}}. \tag{3}$$

Thus, F is insensitive to redistributions within a category that preserve its total mass, whereas d_c retains the teacher's fine-grained preferences within that category.

Why separate the two levels? Writing $\bar{p}_i = p_i / Z_p$ and $\bar{q}_i = q_i / Z_q$, the chain rule for relative entropy (Cover & Thomas, 2006) yields

$$D_{\mathrm{KL}}(\bar{p} \| \bar{q}) = F + \underbrace{\sum_c \widetilde{P}_c d_c}_{W}. \tag{4}$$

Even with category probabilities held fixed, candidate-level KL remains sensitive to within-category disagreement, and its weighting couples conditional fitting difficulty to category allocation. Under a simple fixed-d_c relaxation, larger within-category discrepancy reduces the optimal relative mass of that category. This motivates separating the two objectives; the derivation and closed-form optimum are given in Appendix A.2.

Detached intra-category residual. We use the following auxiliary signal:

$$R = \sum_c \mathrm{sg}(P_c) d_c, \qquad \nabla_\theta R = \sum_c P_c \nabla_\theta d_c, \tag{5}$$

where sg denotes stop-gradient. Detaching category weights removes the explicit derivative through their dependence on θ. Category probabilities can nevertheless change, since shared model parameters also govern the within-category conditionals. We therefore select residuals by measuring their effect on F.

Further details on the support normalizer, the scalar relation between R and W, and the effect of stop-gradient are given in Appendix A.2. Singleton categories contribute zero to R, so intra-category supervision is restricted to multi-candidate categories.

4.3 Selecting Residuals with Matched Optimizer Trials

Residual selection uses auxiliary training packs $\mathcal{B} = \{b_1, \ldots, b_B\}$. A pack contains a fixed prefix, candidate continuations, cached teacher probabilities, and a category partition. Student continuation log-likelihoods are recomputed and normalized within each pack, giving losses F_b and R_b by Equations equation 2–equation 5 with $Z_p = 1$. These packs participate in optimization and are distinct from held-out evaluation data. Their category labels are supplied with the packs; the rollout token-grouping procedure need not be used to construct them.

Let θ_0 be the student parameters after the main rollout update, and let ω be the persistent auxiliary optimizer state. We write $\mathsf{U}_\alpha(\theta, \omega; L)$ for one AdamW step (Loshchilov & Hutter, 2019) on L, including its state update. For each pack, we form two temporary branches from the same snapshot:

$$\begin{aligned}(\theta_b^{FR}, \omega_b^{FR}) &= \mathsf{U}_\alpha(\theta_0, \omega; F_b + \beta R_b),\\ (\theta_b^{F}, \omega_b^{F}) &= \mathsf{U}_\alpha(\theta_0, \omega; F_b),\end{aligned} \tag{6}$$

where $\beta \geq 0$ controls residual strength. Both branches use the same parameter subset, learning rate, optimizer settings, pack, and evaluation conditions. The FR branch is evaluated first; all model parameters and optimizer state, including moments and step counters, are restored before the F branch and again after it. Trial updates are never retained as training updates.

The residual's incremental benefit is

$$\begin{aligned}\Gamma_b &= \left[F_b(\theta_0) - F_b(\theta_b^{FR})\right] - \left[F_b(\theta_0) - F_b(\theta_b^{F})\right]\\ &= F_b(\theta_b^{F}) - F_b(\theta_b^{FR}), \qquad A_b = \mathbf{1}[\Gamma_b > 0].\end{aligned} \tag{7}$$

We retain R_b only when its inclusion yields a strictly lower post-update category loss than the F-only reference.

4.4 Full Training Procedure

At iteration n, the student first performs the rollout update using

$$\mathcal{L}_{\text{main}}(\theta; \mathcal{D}_n) = \frac{\lambda_{\text{distill}}}{|\mathcal{D}_n|} \sum_{s \in \mathcal{D}_n} F(\theta; s), \tag{8}$$

so R does not enter the main actor backward pass. After residual selection, the auxiliary stage commits

$$\mathcal{L}_{\text{aux}}(\theta; A) = \frac{1}{B} \sum_{b \in \mathcal{B}} \left[F_b(\theta) + \beta\, \text{sg}(A_b) R_b(\theta)\right]. \tag{9}$$

Thus, training is sequential: category-level alignment drives the rollout update, and only accepted intra-category residuals enter the auxiliary update. The complete optimizer-state composition, rollback procedure, and Algorithm 1 are provided in Appendix A.4.

5 Experiments

5.1 Experimental Setup

Models and baselines. Our primary configuration uses Qwen3-1.7B as the student and Qwen3-4B-Non-Thinking-RL-Math as the teacher. We compare SCS-OPD with Reverse-KL, Forward-KL, Entropy-aware On-Policy Distillation (EOPD), TSD, and GRPO. SCS-OPD, EOPD, Reverse-KL, and GRPO are evaluated with seeds 17, 21, and 42; Forward-KL and TSD have results for seeds 21 and 42. The main table uses seed 21, where all six main methods are available, and the appendix retains every reported seed. A second configuration uses DeepSeek-R1-Distill-Qwen-1.5B with a 7B teacher and compares SCS-OPD, EOPD, TSD, and Reverse-KL at seed 17.

Benchmarks and evaluation. We evaluate mathematical reasoning on AIME24, AIME25, AMC23, and MATH500. AIME24/25 use Avg@8, while AMC23/MATH500 use Avg@5. Mathematical generation uses temperature $T = 1.0$, top-$p = 0.95$, top-$k = -1$, and MathVerify regrading without regeneration. MMLU 5-shot and HumanEval pass@1 provide additional capability diagnostics. Scores are percentages, and reported gains are percentage-point differences. Full settings and per-run results appear in Appendices C and B.

Table 1: Mathematical reasoning results in the Qwen3-4B → Qwen3-1.7B configuration using common seed 21. Scores are percentages. AIME24/25 use Avg@8 and AMC23/MATH500 use Avg@5.

Method	AIME24	AIME25	AMC23	MATH500
GRPO	15.8	10.0	46.5	71.8
TSD	19.6	16.7	51.5	77.1
Reverse-KL	54.2	40.0	86.5	90.8
Forward-KL	52.2	40.0	**91.0**	91.3
EOPD	54.3	39.6	89.0	91.0
SCS-OPD	**54.6**	**41.7**	88.5	**92.4**

Table 2: Controlled Qwen3 seed-17 ablation of intra-category supervision. Scores are percentages. AIME24/25 use Avg@8 and AMC23/MATH500 use Avg@5; bold indicates the best value, including ties.

Variant	AIME24	AIME25	AMC23	MATH500
F-only	54.6	42.1	87.5	91.2
F + All-R	**56.7**	42.1	90.5	91.6
SCS-OPD	**56.7**	**44.2**	**91.0**	**92.7**

5.2 Main Results

Primary comparison. Table 1 compares all six main methods at seed 21. SCS-OPD leads AIME24, AIME25, and MATH500, demonstrating gains across both AIME evaluations and the broader MATH500 benchmark.

Relative to Reverse-KL, SCS-OPD improves all four benchmarks by 0.4, 1.7, 2.0, and 1.6 points, respectively. It also exceeds EOPD on AIME24, AIME25, and MATH500.

Consistency across random seeds. Across seeds 17, 21, and 42, SCS-OPD has the highest reported AIME24 and MATH500 score at each seed; complete per-seed comparisons are in Appendix B.1.

Controlled ablation of intra-category supervision. The central design question is whether intra-category supervision should be removed, used uniformly, or selected. Table 2 compares these three choices in the Qwen3 seed-17 setting.

The ablation separates two questions: whether intra-category information is useful, and whether it should be applied uniformly. Adding all residuals to F-only improves AIME24, AMC23, and MATH500 by 2.1, 3.0, and 0.4 points and leaves AIME25 unchanged. Thus, discarding intra-category information loses useful supervision. Selection then adds a further 2.1 points on AIME25, 0.5 on AMC23, and 1.1 on MATH500 without reducing AIME24. Relative to F-only, the full method gains 2.1/2.1/3.5/1.5 points. This ordered comparison supports both retaining fine-grained knowledge and controlling which residuals contribute. The evidence is a controlled seed-17 training comparison, distinct from the local intervention diagnostics below; full results appear in Appendix B.2.

Residual strength. We use $\beta = 1.0$; the complete five-value sensitivity study and its benchmark-specific trade-offs are reported in Appendix B.3.

Transfer to a second teacher–student configuration. With DeepSeek-R1-Distill-Qwen-1.5B, SCS-OPD leads AIME25, AMC23, and MATH500 among the four compared methods. It improves over Reverse-KL and EOPD by diffirent points. These results extend the observed benefit to a second teacher–student configuration, even beyond the original Qwen3 pairing. Appendix B.5 reports all mathematical and additional capability scores.

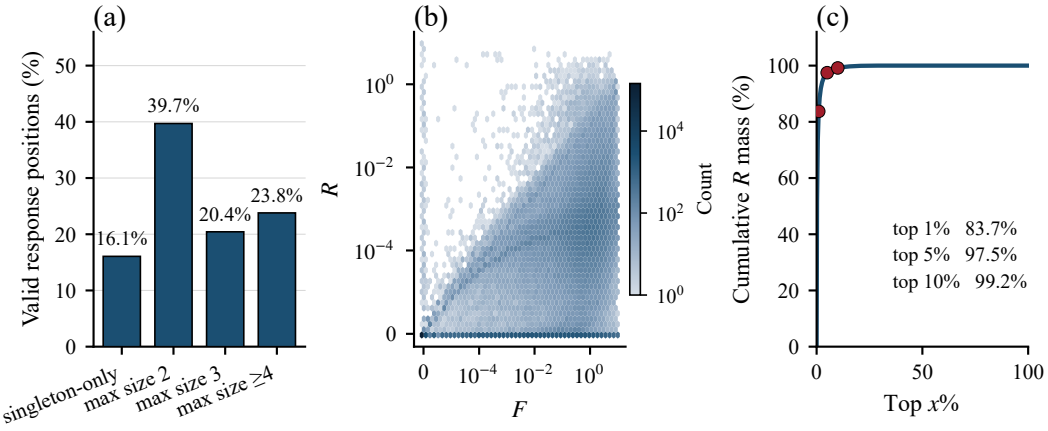


Figure 1: Characterizing intra-category supervision in token-level OPD. (a) Category-size coverage. (b) Distribution of inter-category and intra-category objectives. (c) Cumulative intra-category mass ranked by R.

Additional capabilities. MMLU and HumanEval are reported as capability diagnostics rather than primary targets; complete results are provided in Appendix B.4.

5.3 Mechanistic Analysis

We ask whether within-category supervision is uniformly compatible with category-level alignment. Candidate-level reverse KL decomposes into category alignment F and a conditional term, while the implemented residual R uses detached category weights. We study three questions: where intra-category supervision is concentrated, whether category-preserving reallocations alter token-level objectives and gradients, and whether residual-reducing updates are consistently compatible with F.

5.3.1 Intra-category supervision is sparse but highly concentrated

We first characterize the structure of intra-category supervision.

Figure 1(a) shows the distribution of semantic category sizes over valid response positions. At 16.1% of positions, all categories are singletons; at 39.7%, the largest category has size two. Thus, a substantial portion of response positions contains multiple candidate realizations within a semantic category, providing the structural basis for intra-category supervision.

However, the existence of such structure does not imply that all positions contribute meaningful intra-category loss. As shown in Figure 1(b), only 34.1% of eligible positions have non-zero R, and much of the distribution is concentrated near $R = 0$.

The non-zero intra-category signal is also extremely uneven. Figure 1(c) shows that the top 1%, 5%, and 10% of eligible positions ranked by R account for 83.7%, 97.5%, and 99.2% of the total intra-category loss contribution, respectively.

The fine-grained signal is therefore concentrated at a small subset of positions. This finding identifies where residual supervision is active; its usefulness is a separate question addressed by the compatibility probes and the training ablation.

5.3.2 Token-level OPD is sensitive to within-category token allocation

We next test whether token-level OPD reacts to token realization preferences even when teacher category masses are held fixed.

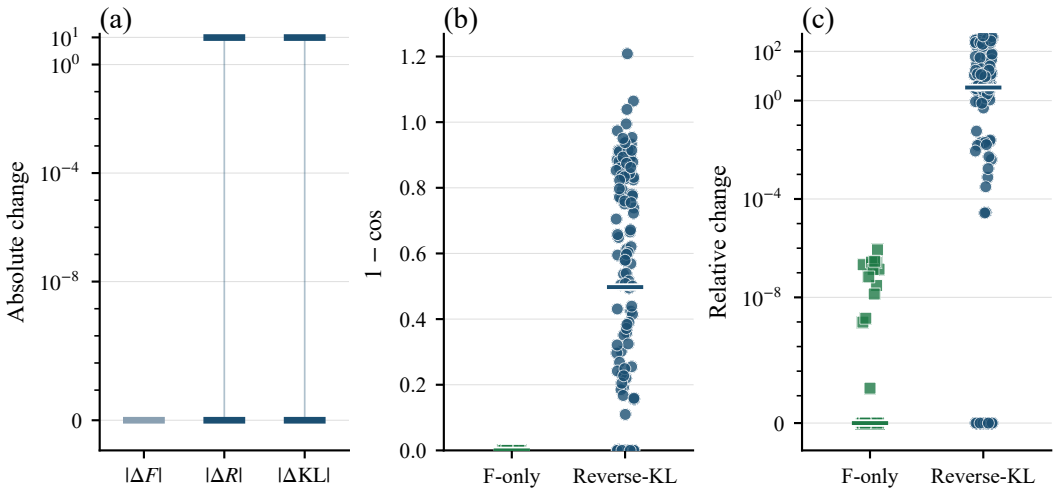


Figure 2: Semantic-preserving perturbations reveal the sensitivity of token-level OPD under fixed category-level distributions. (a) Objective changes under fixed category distribution. (b) Gradient direction changes. (c) Relative gradient variation.

We construct a category-mass-preserving perturbation that fixes the category-level probability Q_c while modifying only the allocation of probability among tokens belonging to the same category. The maximum category-mass deviation after perturbation is approximately 10^{-15}, confirming that category-level probabilities are preserved numerically.

Figure 2(a) shows that the inter-category objective F is effectively unchanged under this intervention, whereas both R and the token-level reverse-KL objective change substantially. The mean absolute change in F is approximately 5×10^{-19}, while the mean absolute changes in R and reverse-KL are both approximately 2.83.

Here, reverse-KL denotes the teacher-top-64 conditional reverse-KL computed on the normalized teacher support rather than a full-vocabulary KL.

This controlled intervention demonstrates that standard token-level OPD remains sensitive to within-category token allocation even when the chosen category masses are preserved. Token-level alignment therefore contains optimization pressure beyond matching semantic category mass.

5.3.3 Within-category variation redirects token-level gradients

The previous experiment establishes an objective-level effect. We next test whether the same category-mass-preserving perturbation changes the actual update direction received by the model.

We compare gradients of the category-level objective F with gradients of the token-level reverse-KL objective, measuring changes with respect to the final attention output projection.

As shown in Figure 2(b), the average gradient-direction change for F is only

$$1 - \cos(g, g') = 3.3 \times 10^{-8},$$

whereas reverse-KL exhibits a substantially larger change,

$$1 - \cos(g, g') = 0.458.$$

Figure 2(c) shows the same pattern in relative gradient magnitude. The F gradient remains nearly invariant, whereas the reverse-KL gradient changes substantially; its median relative change is approximately 3.4.

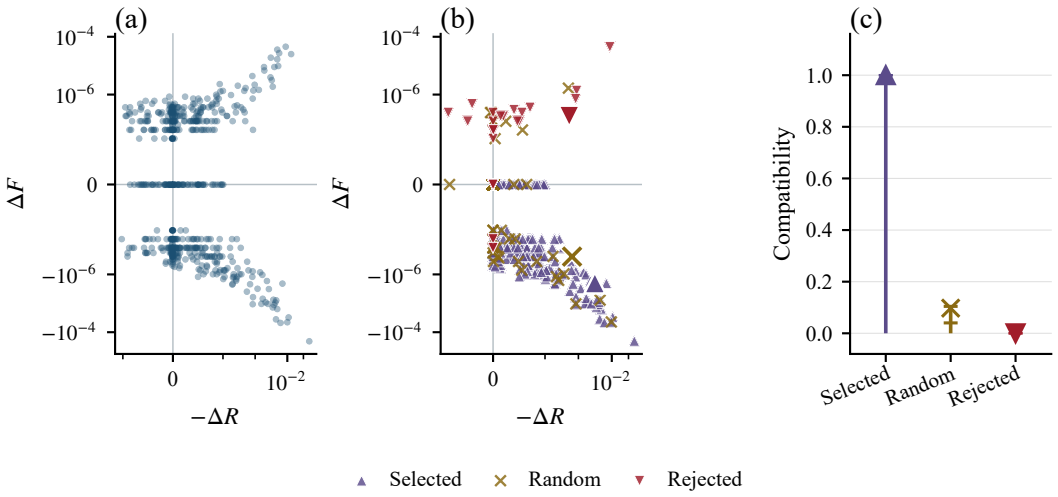


Figure 3: Selective gating preferentially selects compatible intra-category updates. (a) Update landscape. (b) Selected/Random/Rejected updates. (c) Compatibility rate with 95% confidence interval.

Thus, within-category realization preferences do not merely alter the numerical value of the token-level objective. They can substantially redirect the local token-level gradient while leaving the category-level semantic objective essentially unchanged.

5.3.4 Intra-category updates are heterogeneous

Sensitivity alone does not imply that intra-category supervision should be removed. Fine-grained teacher information may still be useful when its update direction is compatible with category-level alignment. We therefore directly study the interaction between changes in R and F.

For local intra-category interventions, Figure 3(a) plots the update landscape in terms of $-\Delta R$ and ΔF. Among updates that successfully reduce intra-category discrepancy ($\Delta R < 0$), 38.4% increase the category-level objective ($\Delta F > 0$), while 44.4% reduce both R and F.

Therefore, intra-category updates are heterogeneous: reducing within-category discrepancy can either support or interfere with category-level alignment. At the local objective level, this shows why neither uniform strategy is universally justified: preserving every intra-category update retains directions that interfere with F, whereas removing all intra-category supervision also discards directions that improve both objectives. This motivates selective retention.

5.3.5 Selective gating preferentially retains compatible updates

We finally characterize the compatibility of the Selected, Random, and Rejected groups in the observed local candidate landscape.

We define an update as compatible when it reduces intra-category discrepancy without increasing the category-level objective:

$$C = \mathbb{1}[\Delta R < 0 \wedge \Delta F \leq 0].$$

Figure 3(b) visualizes Selected, Random, and Rejected candidates in the same local update landscape. Figure 3(c) reports their compatibility rates. Among the evaluated candidates, Selected updates achieve a compatibility rate of 100%, compared with 9.7% for Random updates and 0% for Rejected updates.

These results support the central design principle of SCS-OPD: fine-grained intra-category supervision should not be uniformly preserved or removed, but selectively retained according to its in-

teraction with the category-level objective.Selected updates also differ from Random and Rejected updates in their initial R values and update magnitudes. The compatibility rates characterize selection on this local candidate landscape; they are not an independent downstream test of the production AdamW selector. Selection and compatibility are assessed on the observed objective changes, and the groups are not magnitude-matched. The controlled training ablation supplies the separate evidence for downstream benefit.

The controlled training ablation in Table 2 provides complementary downstream evidence. Under the reported seed-17 setting, SCS-OPD is higher than F-only on all four mathematical benchmarks and higher than $F + \text{All-}R$ on three while tying AIME24. This training-level result does not remove the magnitude confound in the local Selected/Random/Rejected diagnostic, but it directly tests the two uniform alternatives targeted by the method design.

6 Conclusion

We presented Structured Category-wise Selective On-Policy Distillation, which separates semantic-category alignment from fine-grained intra-category supervision and selects residuals through matched finite-step optimizer trials. The method addresses an optimization interaction: category alignment should not require uniform imitation of every within-category preference, but removing all such preferences also discards useful teacher information.

The experiments support this distinction at both training and local-update levels. SCS-OPD leads AIME24 and MATH500 at each of three Qwen3 seeds, exceeds Reverse-KL in 10 of 12 matched mathematical comparisons, and leads three of four benchmarks in a second teacher–student configuration. The controlled ablation improves on F-only across all four mathematical tasks and on uniform residual supervision on three while tying the fourth. Mechanistic probes explain why selection is relevant: residual supervision is sparse and concentrated, within-category allocation can redirect token-level gradients without changing category masses, and residual-reducing updates can either help or hinder category alignment. Together, these findings support *selective fine-grained transfer* as an effective complement to semantic-category supervision.

The scope remains mathematical reasoning with the reported model configurations. Grouping is a lexical approximation, trial updates add training cost, and packwise acceptance does not guarantee aggregate-update or held-out improvement. The residual ablation and weight study each use one seed, and the local Selected/Random/Rejected groups are not magnitude-matched. Improving contextual grouping and reducing selector cost are natural next steps.

AI use statement

Generative AI tools were used for drafting and editing prose, restructuring outlines, and checking internal consistency. Experimental values and citations were not accepted without author verification. The authors take responsibility for the final manuscript.

Ethics statement

This work studies on-policy distillation for mathematical reasoning using public model weights and mathematical prompt corpora. It does not involve human subjects or sensitive personal data. We do not claim deployment readiness beyond the reported academic setting.

Reproducibility statement

The optimization, semantic-grouping, and evaluation settings are provided in Appendix C. Method details and the finite-step selector appear in Section 4 and Appendix A. Appendix B reports individual-seed benchmarks, both residual ablations, the complete five-value residual-weight sweep, and second-configuration results. Appendix F specifies their reporting scope.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-

generated mistakes. In *International Conference on Learning Representations*, 2024.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for llms. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2402.12030.

Qi Cao, Andrew Gambardella, Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. Semantic token clustering for efficient uncertainty quantification in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 682–696. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-short.49. URL https://aclanthology.org/2026.eacl-short.49/.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018. URL https://proceedings.mlr.press/v80/chen18a.html.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2 edition, 2006. doi: 10.1002/047174882X.

Lucio M. Dery, Yann Dauphin, and David Grangier. Auxiliary task update decomposition: The good, the bad and the neutral. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2108.11346.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018. URL https://arxiv.org/abs/1812.02224.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9367–9385. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.558. URL https://aclanthology.org/2024.findings-acl.558/.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773. PMLR, 2020. URL https://proceedings.mlr.press/v108/farajtabar20a.html.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.08543v3.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level. *arXiv preprint arXiv:2605.06387*, 2026. URL https://arxiv.org/abs/2605.06387.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *ICLR 2026 Workshop on Scaling Post-training for LLMs (SPOT)*, 2026. URL https://arxiv.org/abs/2603.07079.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31044–31062. PMLR, 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

Chia-Hsuan Lee, Zelei Cheng, Yu Wang, Renkun Ni, Sambit Sahu, Shi-Xiong Zhang, and William Campbell. SEAD: Competence-aware on-policy distillation via entropy-guided supervision. *arXiv preprint arXiv:2606.28562*, 2026. URL https://arxiv.org/abs/2606.28562.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

George Michalopoulos, Ian McKillop, Alexander Wong, and Helen Chen. LexSubCon: Integrating knowledge from lexical resources into contextual embeddings for lexical substitution. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1226–1236. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.87. URL https://aclanthology.org/2022.acl-long.87/.

George A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the position bias of on-policy distillation. *arXiv preprint arXiv:2606.22600*, 2026. URL https://arxiv.org/abs/2606.22600.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026. URL https://arxiv.org/abs/2606.01249.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.11325.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation, 2026. URL https://arxiv.org/abs/2604.14084.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2001.06782.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2203.08679.

A Additional Method Details

A.1 WordNet-Based Category Construction

Categories are local to the candidate support at a prefix. We decode candidate token IDs and use rule-based edges to construct an undirected graph. The constrained connected components are remapped to contiguous category IDs. Table 3 summarizes the admissible merges. Lexical component construction is independent of candidate ordering; prohibited or oversized lexical merges do not delete candidates or their probability mass.

Table 3: Grouping rules used by the reference implementation. Exact-form and numeric equivalence are exempt from the lexical size caps.

Edge type	Eligibility	Size constraint
Exact form	Identical decoded surface strings	Uncapped
Complete numeral	Matching confirmed complete-number value keys	Uncapped
Language casefold	Nonmathematical, nonnumeric forms with the same casefolded form	At most 4 tokens and 1 canonical form
WordNet synset	Eligible natural-language forms sharing a synset; compatible POS/morphology; not linked solely by a common morphological root	At most 2 tokens and 2 forms

The lexical minimum alphanumeric length is three. Mathematical variables, operators, and expression fragments are excluded from WordNet edges. Case normalization is a separate language rule, not an exact string match, and does not identify case-sensitive variables. Numeric normalization is restricted to complete numeric forms; it must not reinterpret a token fragment as a full mathematical value.

The implementation calls this backend `lexsubcon`. The method uses lexical-resource-based candidate relations associated with lexical substitution research, including LexSubCon (Michalopoulos et al., 2022), but does not execute LexSubCon's contextual substitution and ranking pipeline. We therefore describe it as WordNet-based grouping. WordNet membership and component size restrictions provide a reproducible lexical approximation, not a proof of contextual semantic equivalence.

Additional grouping details moved from the main text. Identical decoded forms and complete numerals with the same verified numeric value may be grouped. For eligible natural-language tokens, we additionally allow case variants and shared-synset pairs with compatible lexical part-of-speech and morphological information; inflection alone is insufficient. Mathematical variables, operators, and expression fragments are excluded from lexical synonym matching. In particular, distinct numeric values and case-sensitive mathematical variables remain separate. Size constraints on lexical components limit transitive over-grouping without removing candidates from the support. The partition is fixed during differentiation and is not assumed to be error-free.

A.2 Normalization, Coupling, and Detached Gradients

The support normalizer matters. At a fixed support and partition, $\bar{p}_i = \widetilde{P}_c u_{i|c}$ and $\bar{q}_i = \widetilde{Q}_c t_{i|c}$ for $i \in C_c$. Substituting into candidate-level reverse KL gives

$$\begin{aligned} D_{\mathrm{KL}}(\bar{p}\|\bar{q}) &= \sum_c \sum_{i \in C_c} \widetilde{P}_c u_{i|c} \left[\log \frac{\widetilde{P}_c}{\widetilde{Q}_c} + \log \frac{u_{i|c}}{t_{i|c}} \right] \\ &= F + \sum_c \widetilde{P}_c d_c = F + W. \end{aligned} \tag{10}$$

This identity refers to normalized candidate distributions and does not equal full-vocabulary KL unless the support covers the vocabulary. The forward value of the implemented residual is $\sum_c P_c d_c = Z_p W$. Consequently, reporting $F + R$ as normalized candidate KL would require $Z_p = 1$, even before considering gradients. Within normalized control packs this value condition holds, while stop-gradient still changes the backward computation.

A conditional mass-allocation calculation. Assume $\widetilde{Q}_c > 0$, hold the finite discrepancies d_c fixed, and allow $\widetilde{P}$ to vary over the probability simplex. The objective is

$$J(\widetilde{P}) = \sum_c \widetilde{P}_c \log \frac{\widetilde{P}_c}{\widetilde{Q}_c} + \sum_c \widetilde{P}_c d_c. \tag{11}$$

For a Lagrange multiplier ξ, stationarity gives

$$\log \frac{\widetilde{P}_c}{\widetilde{Q}_c} + 1 + d_c + \xi = 0. \tag{12}$$

Normalization yields

$$\widetilde{P}_c^\star = \frac{\widetilde{Q}_c \exp(-d_c)}{\sum_j \widetilde{Q}_j \exp(-d_j)}, \tag{13}$$

and strict convexity on the simplex interior establishes uniqueness. For two categories c, j,

$$\frac{\widetilde{P}_c^\star}{\widetilde{P}_j^\star} = \frac{\widetilde{Q}_c}{\widetilde{Q}_j} \exp[-(d_c - d_j)]. \tag{14}$$

Thus a larger fixed conditional discrepancy lowers a category's mass relative to categories with smaller discrepancies. The assumption is deliberately local: in a neural student, both conditional distributions and masses normally change during training.

What stop-gradient does and does not remove. Let $R_{\mathrm{raw}} = \sum_c P_c d_c$ without detachment. Then

$$\nabla_\theta R_{\mathrm{raw}} = \sum_c d_c \nabla_\theta P_c + \sum_c P_c \nabla_\theta d_c, \qquad \nabla_\theta R = \sum_c P_c \nabla_\theta d_c. \tag{15}$$

Removing the first term does not constrain the actual update to preserve category probabilities. Setting $\beta = 1$ therefore does not recover the ordinary candidate-KL gradient, even on a normalized support.

A.3 Interpreting the Finite-Step Selector

Let $\delta_F = \theta_b^F - \theta_0$ and $\delta_{FR} = \theta_b^{FR} - \theta_0$ be the actual AdamW trial updates. If F_b has a locally Lipschitz Hessian, Taylor expansion gives

$$\begin{aligned} \Gamma_b = & -\nabla F_b(\theta_0)^\top (\delta_{FR} - \delta_F) \\ & - \frac{1}{2} \left[\delta_{FR}^\top H_b \delta_{FR} - \delta_F^\top H_b \delta_F \right] \\ & + O\big(\|\delta_{FR}\|^3 + \|\delta_F\|^3 \big), \end{aligned} \tag{16}$$

where $H_b = \nabla^2 F_b(\theta_0)$. The implemented selector measures Γ_b directly; it neither evaluates this expansion nor isolates its second-order term. For AdamW, the difference $\delta_{FR} - \delta_F$ depends on the existing first and second moments and need not be proportional to $-\beta \nabla R_b$.

The strict gate has zero comparison margin. A positive Γ_b is not a noise-floor certificate; numerical stability and transfer beyond the probed pack require empirical evaluation.

A.4 Complete Sequential Training Procedure

Section 4.4 gives the two committed training objectives. Here we provide the complete optimizer-state composition and pseudocode.

Let ν_n and ω_n denote the independent actor and auxiliary optimizer states. The complete iteration is

$$\begin{aligned}(\theta_n^+, \nu_{n+1}) &= \mathsf{U}_{\mathrm{actor}}(\theta_n, \nu_n; \mathcal{L}_{\mathrm{main}}),\\ A_n &= \mathsf{Select}(\theta_n^+, \omega_n; \mathcal{B}),\\ (\theta_{n+1}, \omega_{n+1}) &= \mathsf{U}_{\alpha}(\theta_n^+, \omega_n; \mathcal{L}_{\mathrm{aux}}(\,\cdot\,; A_n)).\end{aligned} \tag{17}$$

Here $\mathsf{U}_{\mathrm{actor}}$ includes the configured mini-batch optimization schedule. Both committed stages update all trainable student parameters. Auxiliary moments and step counters persist across iterations and checkpoints, including iterations in which all residuals are rejected. The two stages form a sequential optimization procedure, rather than one backward pass on a combined rollout–control loss.

The trials and commit share the auxiliary update rule. Grouping details and the reference configuration are provided in Appendix A.

Algorithm 1 SCS-OPD with packwise residual selection

Require: Student θ, frozen teacher π_T, control packs $\mathcal{B}$, residual weight β, actor and auxiliary optimizer states ν, ω

for each training iteration **do**
  Sample student rollouts and collect prefixes $\mathcal{D}$
  Gather teacher top-K support and both models' probabilities
  Build local categories; form $\mathcal{L}_{\mathrm{main}}$ by Equation equation 8
  $(\theta, \nu) \leftarrow \mathsf{U}_{\mathrm{actor}}(\theta, \nu; \mathcal{L}_{\mathrm{main}})$
  Snapshot $(\theta_0, \omega_0) \leftarrow (\theta, \omega)$
  for each usable pack $b \in \mathcal{B}$ **do**
    $(\theta, \omega) \leftarrow \mathsf{U}_{\alpha}(\theta_0, \omega_0; F_b + \beta R_b)$
    $f_b^{FR} \leftarrow F_b(\theta)$; restore $(\theta, \omega) \leftarrow (\theta_0, \omega_0)$
    $(\theta, \omega) \leftarrow \mathsf{U}_{\alpha}(\theta_0, \omega_0; F_b)$
    $f_b^{F} \leftarrow F_b(\theta)$; restore $(\theta, \omega) \leftarrow (\theta_0, \omega_0)$
    $A_b \leftarrow \mathbf{1}[f_b^{FR} < f_b^{F}]$
  end for
  Recompute $\mathcal{L}_{\mathrm{aux}}$ using the fixed gates A (Equation equation 9)
  $(\theta, \omega) \leftarrow \mathsf{U}_{\alpha}(\theta_0, \omega_0; \mathcal{L}_{\mathrm{aux}})$
  Retain committed parameters and both optimizer states
end for

A.5 Reference Training Configuration

Table 4 records the reference configuration. The main and auxiliary optimizers have separate persistent states; only the committed auxiliary step advances the stored auxiliary moments and step counters. Each trial starts from the same saved state and is fully rolled back. Auxiliary state must also be restored when training resumes from a checkpoint.

If B packs are probed, an auxiliary call performs two temporary forward/backward/optimizer steps and two post-update evaluations per pack, followed by one backward/optimizer step on the aggregate gated objective. Trial parameter and optimizer snapshots can be reused sequentially across packs. No additional network is introduced, and the grouping and selection procedure is used during training only.

B Full Experimental Results

We report every benchmark entry separately. The main Qwen3 comparison covers seeds 17, 21, and 42, with per-method coverage shown explicitly; seed 21 is used for the compact main-text table.

Table 4: Reference SCS-OPD configuration.

Setting	Value
Rollout main objective	$0.5F$
Residual coefficient β	1.0
Teacher candidate support	Top-64
Auxiliary optimizer / learning rate	AdamW / 10^{-6} (persistent state)
Auxiliary update frequency	Every training iteration
Residual gate	$\Gamma_b > 0$; ties reject
Class-loss / entropy prefilters	Disabled / disabled
Policy-gradient / reference-KL loss	Disabled / disabled
Lexical backend / minimum alphanumeric length	WordNet-based / 3
Casefold / synonym component cap	4 / 2 tokens
Training batch / mini-batch / rollouts per prompt	24 / 12 / 2

Table 5: Complete mathematical reasoning results for the primary Qwen3 configuration. Scores are percentages. AIME24/25 use Avg@8; AMC23/MATH500 use Avg@5. Bold indicates the best reported result within each seed; missing method–seed combinations are not inferred.

Method	Seed	AIME24	AIME25	AMC23	MATH500
GRPO	17	14.6	10.0	41.5	71.7
Reverse-KL	17	54.2	40.8	**91.5**	91.5
EOPD	17	50.8	38.3	85.0	90.0
SEM	17	54.2	39.2	90.8	91.6
SCS-OPD	17	**56.7**	**44.2**	91.0	**92.7**
GRPO	21	15.8	10.0	46.5	71.8
TSD	21	19.6	16.7	51.5	77.1
Reverse-KL	21	54.2	40.0	86.5	90.8
Forward-KL	21	52.2	40.0	**91.0**	91.3
EOPD	21	54.3	39.6	89.0	91.0
SCS-OPD	21	**54.6**	**41.7**	88.5	**92.4**
GRPO	42	15.0	6.7	46.5	72.2
TSD	42	22.5	18.8	56.0	76.8
Reverse-KL	42	50.0	42.9	88.0	91.1
Forward-KL	42	54.3	41.0	86.0	91.4
EOPD	42	52.1	**45.0**	89.5	91.7
SCS-OPD	42	**55.4**	39.2	**90.5**	**93.0**

The residual ablation, residual-weight sweep, and second teacher–student comparison use seed 17. Tables report scores in percent without averaging across seeds or tasks.

B.1 Complete Qwen3 Mathematical Reasoning Results

SCS-OPD is highest on AIME24 and MATH500 within each reported seed. Against Reverse-KL, its AIME24 gains are 2.5, 0.4, and 5.4 points for seeds 17, 21, and 42, and its MATH500 gains are 1.2, 1.6, and 1.9 points. Against EOPD, the corresponding gains are 5.9/0.3/3.3 on AIME24 and 2.7/1.4/1.3 on MATH500. The pattern supports repeatability of these two benchmark advantages across the three reported training seeds.

The exceptions also matter. At seed 42, EOPD, Reverse-KL, and Forward-KL exceed SCS-OPD on AIME25; at seed 17, Reverse-KL is 0.5 points higher on AMC23; at seed 21, Forward-KL and EOPD are higher on AMC23. SCS-OPD wins 10 of 12 matched benchmark–seed comparisons against each of Reverse-KL and EOPD. This is a descriptive count of the reported scores, not a statistical-significance test. SEM is retained under its experiment label and is not pooled with the separately reported controlled F-only variant.

B.2 Controlled F/R Ablation

Table 6 reports the full seed-17 comparison used in Table 2, including capability-retention diagnostics where available.

Variant	AIME24	AIME25	AMC23	MATH500	MMLU	HE
F-only	54.6	42.1	87.5	91.2	**60.2**	**35.4**
F + All-R	**56.7**	42.1	90.5	91.6	60.1	32.3
SCS-OPD	**56.7**	**44.2**	**91.0**	**92.7**	**60.2**	**35.4**

Table 6: Controlled Qwen3 seed-17 ablation of intra-category supervision. Scores are percentages; MMLU uses 5-shot evaluation and HE denotes HumanEval pass@1.

β	AIME24	AIME25	AMC23	MATH500	MMLU	HE
0.1	53.3	37.9	86.5	**93.3**	60.2	35.0
0.5	55.8	41.2	90.0	92.6	**60.3**	35.4
0.8	51.2	42.9	88.5	92.7	**60.3**	32.3
1.0	**56.7**	**44.2**	**92.0**	92.9	60.2	35.4
1.2	54.2	40.8	91.0	92.6	60.2	**36.0**

Table 7: Sensitivity of SCS-OPD to the intra-category residual coefficient β in the Qwen3 seed-17 configuration. All entries are actual evaluated scores (%); no averaging across β values is used.

The controlled variants retain the category-level objective and differ in residual use: F-only removes intra-category supervision, F + All-R retains it unconditionally, and SCS-OPD selects it. Uniform residual supervision already improves three mathematical benchmarks over F-only, demonstrating that the fine-grained signal is useful. Selection further improves AIME25, AMC23, and MATH500 while preserving AIME24. The full method exceeds F-only by 2.1/2.1/3.5/1.5 points and the all-residual variant by 0.0/2.1/0.5/1.1 points.

On additional capabilities, SCS-OPD matches F-only at 60.2 MMLU and 35.4 HumanEval, while the all-residual variant records 60.1 and 32.3. Thus, the measured mathematical gain from selection is not accompanied by a lower score on either of these two diagnostics relative to F-only. The ablation supports the selection mechanism at the evaluated residual weight and seed; it does not substitute for tuning the uniform-residual baseline across multiple weights.

B.3 Sensitivity to the Residual Weight β

Table 7 reports the Qwen3 seed-17 SCS-OPD sensitivity study. All rows use the same configuration family and differ in β.

Why use $\beta = 1.0$? The sweep supports $\beta = 1.0$ as a balanced setting across the mathematical benchmarks: it achieves the best AIME24, AIME25, and AMC23 scores, while its MATH500 score of 92.9 is 0.4 points below the maximum of 93.3 at $\beta = 0.1$. Relative to $\beta = 0.5$, setting $\beta = 1.0$ improves all four mathematical scores by 0.9/3.0/2.0/0.3 points. Increasing further to 1.2 reduces them by 2.5/3.4/1.0/0.3 points. The smaller or larger weight therefore does not improve the method monotonically: the coefficient affects both the trial comparison and the strength of an admitted residual.

MMLU varies only between 60.2 and 60.3, while the highest HumanEval score is 36.0 at $\beta = 1.2$. We select a default for the target mathematical tasks, not a universal optimum over every capability. This is a sensitivity study of selective supervision, not a weight sweep for the unfiltered-residual baseline.

Reporting scope. The $\beta = 1.0$ entry in this sweep reports AMC23/MATH500 as 92.0/92.9, whereas the separately reported main benchmark and controlled residual ablation give 91.0/92.7. We retain the values within their respective evaluation blocks and do not substitute the higher sweep

Table 8: Additional capability evaluations for the primary Qwen3 configuration. Scores are percentages. HumanEval uses pass@1.

Method	Seed	MMLU 5-shot	HumanEval
GRPO	17	60.3	36.1
Reverse-KL	17	60.2	32.3
EOPD	17	60.2	33.5
SEM	17	60.2	37.2
SCS-OPD	17	60.2	35.4
GRPO	21	60.2	35.6
TSD	21	60.1	42.1
Reverse-KL	21	60.2	34.8
Forward-KL	21	60.3	31.1
EOPD	21	60.2	34.1
SCS-OPD	21	60.2	37.5
GRPO	42	60.2	33.6
TSD	42	60.1	34.3
Reverse-KL	42	60.1	32.3
Forward-KL	42	60.2	31.1
EOPD	42	60.2	34.1
SCS-OPD	42	60.2	33.5
F-only	17	60.2	35.4
$F +$ All-R	17	60.1	32.3

Method	AIME24	AIME25	AMC23	MATH500	MMLU	HE
EOPD	27.1	22.1	71.0	83.4	**37.5**	**3.0**
TSD	**30.4**	22.1	73.0	83.7	**37.5**	1.8
Reverse-KL	28.9	25.6	73.5	83.8	37.4	1.8
SCS-OPD	29.6	**27.1**	**75.0**	**85.8**	**37.5**	1.8

Table 9: Complete mathematical and additional capability results for the second teacher–student configuration, seed 17. Scores are percentages; bold indicates the best value, including ties.

entry into the main comparison. No claim here depends on treating those entries as the same evaluation.

B.4 Additional Capability Evaluation

SCS-OPD scores 60.2 on MMLU at all three seeds, within the narrow 60.1–60.3 range across the reported Qwen3 methods. HumanEval is more sensitive to the training objective.

B.5 Second Teacher–Student Configuration

Table 9 reports the four methods evaluated with seed 17 in the second teacher–student configuration. The student is DeepSeek-R1-Distill-Qwen-1.5B and the teacher is a 7B model. This comparison changes the teacher–student configuration and student training provenance; it is not a comparison across unrelated backbone architectures.

SCS-OPD leads three of the four mathematical benchmarks. These results provide an additional model-configuration check of the mathematical gains observed with Qwen3.

The additional capability scores remain separate from that conclusion. SCS-OPD matches EOPD and TSD on MMLU at 37.5 and matches TSD and Reverse-KL on HumanEval at 1.8; EOPD is higher on HumanEval at 3.0. With a single reported seed, this second configuration establishes an additional empirical instance rather than the same level of seed coverage as the primary comparison.

Table 10: Training configuration for the primary Qwen3 SCS-OPD experiments.

Configuration	Value
Student	Qwen3-1.7B
Teacher	Qwen3-4B-Non-Thinking-RL-Math
Reported seeds	17 / 21 / 42
Hardware	8 Ascend NPUs
Distillation loss	`s2sm_token_fr_topk`
Rollout main objective	F
Main distillation coefficient	0.5
Auxiliary objective	$F + \beta AR$
Residual coefficient β	1.0
Residual selector	optimizer-matched FR vs. F trial
Gate behavior	strict selection; entropy gate disabled
Actor optimizer	AdamW
Actor learning rate	1×10^{-6}
Auxiliary optimizer	AdamW (persistent)
Auxiliary learning rate	1×10^{-6}
Weight decay	0.01
Training batch size	24
Mini-batch size	12
Micro-batch per NPU	1
Rollouts per prompt	2
Teacher candidate support	Top-64
Token grouping	`lexsubcon`
Casefold / synonym component cap	4 / 2 tokens
Minimum alphanumeric length	3
Additional KL / policy-gradient loss	disabled / disabled

C Training and Evaluation Details

C.1 Training Framework

All experiments are implemented using MindSpeed-MM. On-policy distillation and GRPO are implemented on top of verl. Unless otherwise specified, compared methods use the same student initialization, teacher model, training prompts, rollout configuration, and evaluation definitions. The primary experiments use eight Ascend NPUs.

C.2 Primary Qwen3 Configuration

The actor and auxiliary AdamW optimizers maintain separate persistent states. Trial branches begin from the same saved parameters and auxiliary optimizer state and are fully restored after evaluation. Only the committed auxiliary update advances the stored auxiliary moments and counters. No learned auxiliary selector is used.

C.3 Second Teacher–Student Configuration

The second configuration uses DeepSeek-R1-Distill-Qwen-1.5B as the student and a 7B teacher, with seed 17, $\beta = 1.0$, and eight Ascend NPUs. It uses the same selective SCS-OPD formulation as the primary configuration. Configuration-specific implementation details that do not alter the objective are kept separate from the scientific comparison.

Table 11: Generation configuration for mathematical reasoning benchmarks.

Configuration	Value
Temperature	1.0
Top-p	0.95
Top-k	-1
Maximum generation length	40,960 tokens
Generation seed	rollout ID
Math grader	MathVerify
Regrading	without regeneration
Hardware	8 Ascend NPUs per job

Table 12: Evaluation protocols used throughout the paper.

Benchmark	Protocol
AIME24	Avg@8
AIME25	Avg@8
AMC23	Avg@5
MATH500	Avg@5
MMLU	5-shot log-likelihood
HumanEval	0-shot pass@1

C.4 Generation Configuration

C.5 Benchmark Evaluation Protocol

For a benchmark with N problems and k sampled responses per problem,

$$\text{Avg@}k = \frac{1}{Nk} \sum_{i=1}^{N} \sum_{j=1}^{k} \mathbb{I}[y_{ij} \text{ is correct}]. \tag{18}$$

AIME24/25 use eight sampled responses per problem; AMC23/MATH500 use five. Mathematical responses are regraded using MathVerify without regeneration. MMLU follows the lm-evaluation-harness 5-shot log-likelihood protocol. HumanEval is evaluated zero-shot using pass@1 with code execution enabled through `--confirm_run_unsafe_code`.

D Mechanistic Probe Details

D.1 Semantic Grouping Audit

We conduct an AI-assisted automated audit of 500 sampled merge events from the semantic grouping procedure.

D.2 Structure of Intra-Category Supervision

The structural analysis contains 363,361 valid response positions, of which 304,980 satisfy the eligibility criterion used for the F/R analysis.

Among valid positions, 16.07% contain only singleton categories. Positions whose maximum category size equals two account for 39.70%, those with maximum size three account for 20.44%, and those with maximum size at least four account for 23.80%.

On eligible positions, the mean values of the decomposed objectives are

$$\mathbb{E}[F] = 0.4900, \qquad \mathbb{E}[R] = 0.004447. \tag{19}$$

Only 34.1% of eligible positions have non-zero R. Despite this sparsity, the non-zero intra-category signal is highly concentrated: the top 1%, 5%, and 10% of eligible positions ranked by R account for 83.7%, 97.5%, and 99.2% of the total intra-category loss contribution, respectively.

D.3 SEMANTIC-PRESERVING PERTURBATION

To isolate sensitivity to within-category token allocation, we perturb token probabilities while preserving category-level masses.

For the reverse-rank intervention, 304,980 eligible positions are perturbable. The maximum absolute deviation in category mass after perturbation is

$$\max_c |Q'_c - Q_c| = 2.331 \times 10^{-15}. \tag{20}$$

Singleton categories remain unchanged.

Under this intervention,

$$\mathbb{E}[|\Delta F|] = 4.968 \times 10^{-19}, \tag{21}$$

whereas

$$\mathbb{E}[|\Delta R|] = \mathbb{E}[|\Delta \mathrm{KL}|] = 2.833. \tag{22}$$

Here, KL refers to the teacher-top-64 conditional reverse-KL on the normalized teacher support rather than a full-vocabulary KL.

The same qualitative pattern holds under random within-category permutations with seeds 17, 29, and 43. These are perturbation seeds, not the training seeds used in the benchmark tables.

D.4 GRADIENT SENSITIVITY PROBE

The gradient probe uses 128 prompts, with up to four positions sampled from each prompt. Statistics are aggregated at the prompt level. Gradients are measured with respect to the final attention output projection.

For the category-level objective F, the mean gradient-direction change is

$$1 - \cos(g, g') = 3.306 \times 10^{-8}. \tag{23}$$

For token-level Reverse-KL, the corresponding mean is

$$1 - \cos(g, g') = 0.4581. \tag{24}$$

The relative gradient change

$$\frac{\|g - g'\|_2}{\|g\|_2 + \epsilon}, \qquad \epsilon = 10^{-12}, \tag{25}$$

has median approximately 3.40 for Reverse-KL, whereas the category-level objective remains effectively invariant.

D.5 LOCAL UPDATE LANDSCAPE

We evaluate 7,556 local candidate interventions using auxiliary update scale $\eta = 3 \times 10^{-4}$.

Among the 450 candidates satisfying $\Delta R < 0$,

- 44.4% also satisfy $\Delta F < 0$;
- 17.1% have $|\Delta F| \leq 10^{-12}$;
- 38.4% satisfy $\Delta F > 0$.

Table 13: Compatibility of Selected, Random, and Rejected intra-category updates.

Update group	Compatibility rate
Selected	100.0%
Random	9.7%
Rejected	0.0%

Thus, reducing intra-category discrepancy does not imply improvement of the category-level objective.

D.6 Selected, Random, and Rejected Updates

For the equal-count Selected/Random/Rejected analysis, each group contains 277 candidate updates drawn from 163 prompts.

Compatibility is defined as

$$\Delta R < 0 \quad \wedge \quad \Delta F \leq 0. \tag{26}$$

The observed compatibility rates are

Selected updates also exhibit larger initial intra-category losses and larger update magnitudes than Random and Rejected updates. Accordingly, this diagnostic supports the conclusion that the gate *preferentially selects compatible updates* from the observed candidate landscape. It does not establish superiority under perfectly magnitude-matched interventions.

E Baseline Comparison Protocol

To make the comparison controlled, the distillation baselines use the same primary student initialization, teacher model, training prompts, and evaluation definitions whenever applicable.

The primary comparison includes Reverse-KL, Forward-KL, EOPD, TSD, GRPO, and SCS-OPD.

The Reverse-KL objective is

$$D_{\mathrm{KL}}\left(p_\theta \| q\right), \tag{27}$$

where p_θ is the student distribution and q is the teacher distribution.

The Forward-KL objective is

$$D_{\mathrm{KL}}\left(q \| p_\theta\right). \tag{28}$$

EOPD, TSD, and GRPO are evaluated under the same benchmark definitions and the common-seed reporting protocol used for SCS-OPD. Method-specific algorithmic differences are retained rather than being replaced by a shared distillation loss.

F Result Reporting Protocol

The compact Qwen3 baseline table uses seed 21, where all six main methods have reported results. Complete per-seed tables preserve every reported evaluation, including exceptions to the main ranking. SCS-OPD, EOPD, Reverse-KL, and GRPO have three seeds; Forward-KL and TSD have two; SEM is an additional seed-17 reference. Unavailable method–seed combinations are not reconstructed. Statements about a method being highest at a seed refer only to the methods reported at that seed.

The residual ablation compares F-only, F + All-R, and SCS-OPD at seed 17. It tests the two uniform residual policies at the reported settings. The weight sweep records each of five tested β

values separately; its $\beta = 1.0$ row is not substituted for the main-run or residual-ablation row. Their reported AMC23/MATH500 values differ, as stated in Appendix B.3.

The second configuration compares the reported EOPD, TSD, Reverse-KL, and SCS-OPD results at seed 17. All six benchmark metrics are retained in Appendix B.5.

Scores are reported in percent. Differences such as +2.1 are percentage points, not relative percentage improvements. Descriptive win counts enumerate the matched benchmark–seed entries; they do not treat tasks or sampled responses as independent experimental replications. We do not report an unweighted cross-benchmark average, substitute results from another run, or claim statistical significance from these aggregate tables alone.

G Extended Background and Related Work

G.1 Additional Related Work

Knowledge distillation and student-generated trajectories. Knowledge distillation transfers predictive information from a teacher to a student through soft supervision (Hinton et al., 2015). Sequence-level distillation trains students on teacher-generated outputs (Kim & Rush, 2016), while learner-induced states have long been studied in imitation learning through methods such as DAgger (Ross et al., 2011). In language modeling, Generalized Knowledge Distillation (GKD) incorporates student-generated trajectories and supports multiple teacher–student divergence objectives (Agarwal et al., 2024). MiniLLM develops reverse-KL-based optimization for autoregressive language-model distillation (Gu et al., 2024), and Speculative Knowledge Distillation interleaves student generation with teacher corrections (Xu et al., 2025). These approaches establish the importance of *where* supervision is applied; our work asks how supervision should be structured among semantically related candidates at a fixed student-generated context.

Adaptive supervision in OPD. Recent work adapts the divergence or the allocation of teacher supervision according to uncertainty, reliability, or trajectory-level signals. DistiLLM combines skew KL objectives with adaptive reuse of student-generated responses (Ko et al., 2024), while DistiLLM-2 uses a contrastive formulation over teacher- and student-generated responses (Ko et al., 2025). EOPD supplements reverse KL with forward KL according to teacher entropy (Jin et al., 2026); SEAD adapts supervision using teacher–student entropy (Lee et al., 2026); AOPD distinguishes positive- and non-positive-advantage tokens (Jia et al., 2026); and Trust Region OPD, Importance-Weighted OPD, and TIP adapt supervision using reliability, accumulated discrepancy, or token importance (Xing et al., 2026; Xie et al., 2026; Xu et al., 2026). These methods motivate selective supervision, but their selection signals operate primarily at the token or trajectory level. SCS-OPD instead decomposes the supervision associated with a candidate set into category-level and within-category components.

Semantic aggregation and structured distillation. Semantic aggregation has been used to distinguish uncertainty about meaning from variation in surface realization. Semantic Entropy groups equivalent generations before uncertainty estimation (Kuhn et al., 2023); Claim Conditioned Probability reduces sensitivity to claim selection and surface-form variation (Fadeeva et al., 2024); and Semantic Token Clustering constructs token groups for efficient uncertainty quantification (Cao et al., 2026). Structured distillation has also considered alternative partitions of supervision: Decoupled Knowledge Distillation separates target- and non-target-class components (Zhao et al., 2022), while Universal Logit Distillation addresses distribution matching across different tokenizers (Boizard et al., 2025). Our partition instead concerns semantic relations among candidates in a shared support and is used to control optimization, not only to estimate uncertainty.

Auxiliary supervision and gradient compatibility. The usefulness of an auxiliary signal depends on how its update interacts with the primary objective. GradNorm balances objectives through gradient magnitudes (Chen et al., 2018); gradient-similarity weighting adapts auxiliary supervision according to first-order compatibility (Du et al., 2018); Learning to Reweight Examples uses a meta-objective to select example weights (Ren et al., 2018); and PCGrad, ATTITTUD, and Orthogonal Gradient Descent modify or decompose update directions to reduce interference (Yu et al., 2020; Dery et al., 2021; Farajtabar et al., 2020). SCS-OPD follows the same broad principle—auxiliary

information should be judged by its interaction with a primary objective—but uses an optimizer-matched finite-step comparison between category-only and category-plus-residual updates.

G.2 OPD and Token-Level KL Background

G.2.1 On-Policy Distillation

Let p_θ be the student and q a frozen teacher sharing vocabulary $\mathcal{V}$. Given prompt x, the student generates $y = (y_1, \ldots, y_T)$. At position t, both models are evaluated on the student-generated prefix $s_t = (x, y_{<t})$. Let $d_{\bar{\theta}}$ denote the prefix distribution induced by the parameters used to collect the current rollouts. A direct distribution-matching OPD objective is

$$\mathcal{L}_{\mathrm{OPD}}(\theta; \bar{\theta}) = \mathbb{E}_{s \sim d_{\bar{\theta}}} \left[\mathcal{D}\big(q(\cdot \mid s), p_\theta(\cdot \mid s)\big) \right], \tag{29}$$

where the rollout distribution is fixed during an optimization update and refreshed as training proceeds (Agarwal et al., 2024).

G.2.2 Token-Level KL Objectives

At a fixed prefix, reverse and forward KL are

$$D_{\mathrm{KL}}(p_\theta \| q) = \sum_{v \in \mathcal{V}} p_v \log \frac{p_v}{q_v}, \qquad D_{\mathrm{KL}}(q \| p_\theta) = \sum_{v \in \mathcal{V}} q_v \log \frac{q_v}{p_v}. \tag{30}$$

Reverse KL weights discrepancies by the student distribution and can favor mode concentration under restricted model families, whereas forward KL weights them by the teacher distribution and encourages coverage of teacher-supported alternatives (Gu et al., 2024). The two divergences differ in optimization behavior but neither represents semantic equivalence between distinct vocabulary items. We therefore include both directions as baselines and treat supervision granularity as a separate design dimension.