CONTEXT-CONSISTENT TEACHER SIGNALS FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides dense teacher supervision on student-generated trajectories, but the teacher signal is typically evaluated under a single context. Such single-context supervision can be sensitive to contextual changes, making it difficult to distinguish stable learning signals from context-specific teacher preferences. *Can we identify teacher supervision that remains consistent across task-relevant contexts and use it to improve on-policy distillation?* We introduce Factorial On-Policy Distillation (FOPD), which evaluates identical student actions across teacher checkpoints and contexts to identify cross-context consistent supervision. FOPD centers the resulting teacher signals over the student's action distribution, retains directions consistently supported across contexts with a conservative shared magnitude, and anchors the update to the standard OPD direction. On Qwen3-1.7B in non-thinking mode across 14 mathematics, code, and science benchmarks, FOPD achieves a Macro score of 43.35, corresponding to a 5.9% relative improvement over the strongest baseline, and improves performance on 13 of the 14 tasks. Ablation studies show that removing or replacing any defining component reduces Macro performance by 1.82–5.85 points. Controls matched for reward sparsity or magnitude remain 2.35–2.65 points below FOPD, while token-level evaluation shows that FOPD achieves the highest direction precision throughout the evaluated 10–70% coverage range. These results demonstrate that cross-context consistency provides an effective principle for identifying reliable token-level supervision in on-policy distillation.

1 INTRODUCTION

On-policy distillation (OPD) is an effective framework for training reasoning models from their own generated trajectories, providing teacher supervision directly on states visited by the student policy (Song & Zheng, 2026; Zhao et al., 2026). Existing OPD methods typically evaluate the teacher under a single context. Since teacher predictions are inherently conditioned on their input, the preference assigned to the same student action may change when the teacher is given another task-relevant context. As a result, single-context supervision makes it difficult to distinguish a stable learning signal from a context-specific teacher preference. This raises the question: *Can we identify teacher supervision that remains consistent across task-relevant contexts and use it to improve on-policy distillation?*

Recent work improves OPD from several complementary perspectives. Generalized OPD extends reward construction beyond direct teacher matching (Yang et al., 2026c), OPRD introduces representation-level supervision (Yang et al., 2026b), and Uni-OPD and DOPD explore complementary formulations of on-policy distillation (Hou et al., 2026; Yu et al., 2026). On-Policy Delta Distillation further compares a reasoning-tuned teacher with its base checkpoint to isolate preferences introduced during post-training (Heo et al., 2026). These methods improve how teacher supervision is constructed or represented, yet they largely determine that supervision from a single teacher context. Consequently, they do not directly address whether a signal assigned to the *same student action* remains stable when the teacher's task-relevant context changes. If it does not, context-specific teacher behavior may be transferred together with useful supervision. The reliability of teacher feedback should therefore reflect both the strength of the teacher's preference for an action and its consistency across contexts.

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.

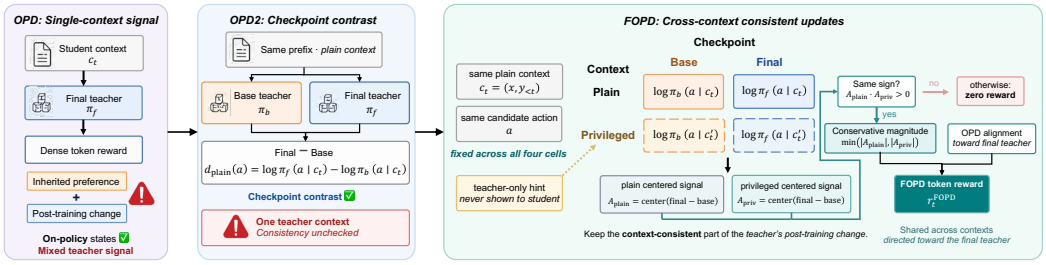


Figure 1: **From single-context supervision to cross-context consistent OPD.** FOPD evaluates identical student actions with base and final teachers under plain and teacher-only privileged contexts, retaining only directions consistently supported across contexts and aligned with standard OPD. Privileged information is used only for teacher scoring.

We propose **Factorial On-Policy Distillation (FOPD)**, which evaluates identical student actions across two dimensions: teacher checkpoint and teacher context. For each student-generated prefix, the base and final teachers score the same candidate actions under both the original task context and a teacher-only privileged context containing a high-level task hint, as illustrated in Figure 1. FOPD compares the checkpoint-induced preference in each context, centers both signals with respect to the student's current action distribution, and retains only directions consistently supported across contexts with a conservative shared magnitude. The retained supervision is also aligned with the standard OPD direction. The privileged context is used only for frozen teacher scoring, while student rollout, optimization, and inference remain entirely under the original context.

We evaluate FOPD on Qwen3-1.7B in non-thinking mode (Yang et al., 2025) over 14 mathematics, code, and science benchmarks. FOPD achieves a Macro score of **43.35**, outperforming the strongest baseline by **2.42 points** (**5.9%** relative), and improves performance on **13 of 14** tasks. The gains remain positive across all seven reported configurations spanning model scales, reasoning modes, and model families. Ablations show that removing or replacing any defining component lowers Macro performance by **1.82–5.85 points**. Reward-sparsity and reward-magnitude controls also remain below the full method, while token-level analysis shows that FOPD consistently provides more precise update directions at matched coverage. These results demonstrate that cross-context consistency is an effective criterion for selecting reliable token-level supervision in on-policy distillation.

We make three contributions:

1. We identify a limitation of single-context supervision in OPD: teacher preferences for the same student action can vary with task-relevant context. This motivates cross-context consistency as a criterion for identifying reliable teacher supervision.
2. We introduce FOPD, which evaluates identical student actions across teacher checkpoints and contexts and selects token-level supervision through action centering, cross-context agreement, conservative magnitude selection, and alignment with the standard OPD direction.
3. We demonstrate the effectiveness and generality of FOPD across 14 reasoning benchmarks and multiple model configurations. Component ablations, matched reward controls, and token-level analyses consistently support the proposed cross-context supervision principle.

2 Related Work

On-policy distillation for reasoning models. Knowledge distillation transfers behavior from a stronger teacher to a student and is widely used in reasoning post-training (Hinton et al., 2015; Shao et al., 2024; Yang et al., 2025). On-policy distillation (OPD) provides dense supervision on student-sampled trajectories, reducing train–test state mismatch (Song & Zheng, 2026; Zhao et al., 2026). Recent extensions include self-distillation (Yang et al., 2026a), flexible references and reward scaling (Yang et al., 2026c), representation-level supervision (Yang et al., 2026b), alternative supervision construction and routing in Uni-OPD and DOPD (Hou et al., 2026; Yu et al., 2026), and

checkpoint contrasts in On-Policy Delta Distillation (Heo et al., 2026). FOPD instead asks whether the signal for the same student action remains consistent across teacher contexts.

On-policy learning in interactive and agentic settings. Reliable on-policy feedback is also central to interactive language agents. Policy-gradient methods provide the optimization foundation (Schulman et al., 2017), while ReAct, ALFWorld, WebShop, and Search-R1 combine reasoning with interaction, tool use, or retrieval (Yao et al., 2022b; Shridhar et al., 2020; Yao et al., 2022a; Jin et al., 2025). Recent benchmarks study longer real-world, scientific, and dataset-discovery workflows (Li et al., 2026c; Wu et al., 2026b; Li et al., 2025), alongside work on efficient demonstrations, long-horizon supervision, selective guidance, and credit assignment (Xiao et al., 2025; Jiang et al., 2026; Fu et al., 2026; Feng et al., 2026). Multi-agent and resource-aware evaluations further highlight interaction structure and computation (Li et al., 2026a;b). FOPD targets context sensitivity at the token level by fixing the student prefix and action across evaluations.

Context-conditioned supervision and reliability. Privileged or hindsight information has been used to enrich on-policy supervision. StepOPSD uses hindsight-enriched teacher contexts (Zhang et al., 2026); Self-Distilled Agentic Reinforcement Learning uses gated privileged-context self-distillation (Lu et al., 2026); SEED distills hindsight skills (Wu et al., 2026a); and UCOB studies bidirectional self-distillation between skill-conditioned and ordinary policies (Tu et al., 2026). Yet privileged supervision can be unreliable: Self-Distilled RLVR separates privileged guidance from verifiable update directions (Yang et al., 2026a), while DOPD studies privileged-only information versus transferable capability (Yu et al., 2026). FOPD uses privileged context only as a second view of the *same student action*, retaining directions that are consistent across contexts and aligned with standard OPD.

3 Factorial On-Policy Distillation

3.1 Motivation and cross-context checkpoint contrasts

FOPD tests whether teacher supervision for the same student action remains consistent across task-relevant contexts. For a problem x, the student samples $y \sim \pi_s(\cdot \mid x)$ and visits the plain context $c_t = (x, y_{<t})$. We construct a second, teacher-only privileged context $c'_t = (x, h(x), y_{<t})$, where $h(x)$ is a high-level task hint describing a relevant strategy or fact without providing a complete answer. The hint is used only for frozen teacher scoring. Student rollout, optimization, and evaluation remain conditioned on the original plain context.

To separate preferences introduced during reasoning post-training from those already present in the base model, we compare a reasoning-tuned final teacher π_f with its base checkpoint π_b within each context, following checkpoint-difference supervision (Heo et al., 2026). For the same candidate action a, we define

$$d_{\text{plain}}(a) = \log \pi_f(a \mid c_t) - \log \pi_b(a \mid c_t),$$
$$d_{\text{priv}}(a) = \log \pi_f(a \mid c'_t) - \log \pi_b(a \mid c'_t). \tag{1}$$

Their difference

$$I_t(a) = d_{\text{priv}}(a) - d_{\text{plain}}(a) \tag{2}$$

measures how the checkpoint-induced preference changes with teacher context. Because the problem, sampled prefix, token position, and candidate action are fixed, the two contrasts differ only in the context presented to the teacher. This gives the 2×2 design in table 1: teacher checkpoint (base vs. final) crossed with teacher context (plain vs. privileged).

Table 1: **Teacher checkpoint $\times$ context design.** Base and final teachers score identical student-selected actions under plain and teacher-only privileged contexts.

Teacher context	Base score	Final score	Checkpoint contrast
Plain c_t	$\log \pi_b(a \mid c_t)$	$\log \pi_f(a \mid c_t)$	$d_{\text{plain}}(a)$
Privileged c'_t	$\log \pi_b(a \mid c'_t)$	$\log \pi_f(a \mid c'_t)$	$d_{\text{priv}}(a)$

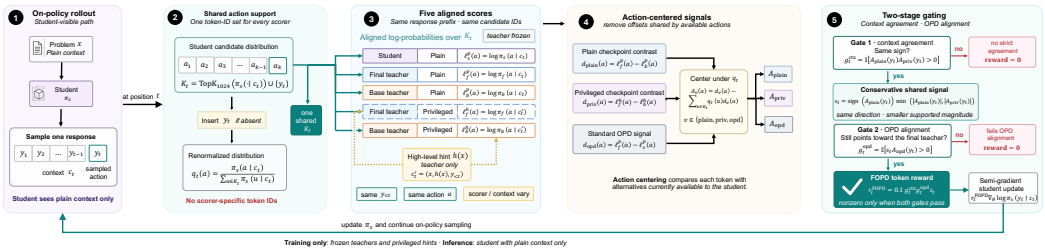


Figure 2: **Per-token computation in FOPD.** All scoring branches share the same student prefix and candidate actions. Base and final teachers evaluate them under plain and teacher-only privileged contexts. FOPD retains an update only when the centered checkpoint contrasts agree across contexts and the resulting direction agrees with standard OPD.

Before policy training, we verify that this context dimension carries meaningful variation. Using 256 balanced prompts and 2,048 fixed Qwen3-1.7B trajectories (Yang et al., 2025), the plain-context checkpoint contrast reaches a correctness AUROC of 0.52255, whereas a cross-context shared signal reaches 0.58938, with a prompt-cluster bootstrap interval of $[0.04366, 0.09096]$ for the gain. The negative interaction reaches 0.61162 AUROC, and the privileged-only contrast reaches 0.62489.

These observations motivate the role of the second context in FOPD. The privileged context provides a second evaluation of the same action without serving as an alternative target for the student. If the checkpoint-induced preference changes direction across contexts, the supervision is context-sensitive. FOPD therefore uses cross-context consistency to determine which part of the teacher signal should be transferred.

3.2 Shared action support and action centering

Cross-context comparison is meaningful only when every scoring branch evaluates the same action. Independently selecting top-k candidates from different scorers can silently compare different token IDs. We therefore construct a shared candidate set K_t from the student's plain-context top-k actions and explicitly insert the sampled token y_t when it is absent. The student, base teacher, and final teacher are then queried on these identical token IDs in every context.

The student distribution restricted to this shared support is

$$q_t(a) = \frac{\pi_s(a \mid c_t)}{\sum_{u \in K_t} \pi_s(u \mid c_t)}, \qquad a \in K_t. \tag{3}$$

Across all scoring branches, only the teacher checkpoint or teacher context changes, while the action being compared remains fixed.

Raw log-probability differences may also contain offsets shared across the student's available actions. We therefore compare each action relative to the alternatives currently available to the student. Alongside the two checkpoint contrasts, define the standard OPD signal

$$d_{\mathrm{opd}}(a) = \log \pi_f(a \mid c_t) - \log \pi_s(a \mid c_t). \tag{4}$$

For $v \in \{\mathrm{plain}, \mathrm{priv}, \mathrm{opd}\}$, we action-center all three signals under the same student distribution:

$$A_v(a) = d_v(a) - \mathbb{E}_{u \sim q_t}[d_v(u)]. \tag{5}$$

The centered quantity $A_v(a)$ measures whether action a is preferred relative to alternatives that the current student can select. A context-wide additive shift therefore does not affect the resulting direction.

3.3 Cross-context consistency and OPD alignment

FOPD next asks whether the centered checkpoint contrasts support the same direction. For the sampled action y_t, we retain only strict sign agreement:

$$g_t^{\mathrm{ctx}} = \mathbb{I}[A_{\mathrm{plain}}(y_t) A_{\mathrm{priv}}(y_t) > 0]\,. \tag{6}$$

Table 2: **Three-domain reasoning performance.** FOPD improves all three domain averages and achieves the highest overall Macro score.

Method	Math	Code	Science	Macro
Qwen3-1.7B	34.80	10.50	30.80	25.37
+ OPD	51.00	21.00	36.50	36.17
+ ExOPD	51.40	24.60	36.50	37.50
+ OPD^2	54.60	29.40	38.80	40.93
+ FOPD	**58.19**	**30.53**	**41.33**	**43.35**
FOPD − OPD^2	**+3.59**	**+1.13**	**+2.53**	**+2.42**

If the two contexts disagree, the token receives no distillation update. If they agree, FOPD retains the smaller magnitude supported by both contexts:

$$s_t = \operatorname{sign}(A_{\text{plain}}(y_t)) \min\{|A_{\text{plain}}(y_t)|, |A_{\text{priv}}(y_t)|\}\,. \tag{7}$$

This conservative aggregation transfers only the portion of the direction supported in both contexts.

The connection to context dependence can be written explicitly. At y_t, let

$$M_t = \tfrac{1}{2}(A_{\text{plain}} + A_{\text{priv}}), \qquad \widetilde{I}_t = A_{\text{priv}} - A_{\text{plain}}.$$

Then

$$g_t^{\text{ctx}} s_t = \operatorname{sign}(M_t) \left[|M_t| - \tfrac{1}{2}|\widetilde{I}_t|\right]_+ . \tag{8}$$

Here $\widetilde{I}_t$ is the action-centered counterpart of the context interaction in eq. (2). The shared component provides the common direction, while disagreement between contexts reduces the magnitude that is transferred. Thus, the consistency rule directly suppresses the context-sensitive portion of the checkpoint contrast.

Cross-context consistency alone is not sufficient because a final-minus-base checkpoint contrast does not depend on the current student and may remain nonzero even after the student matches the final teacher. We therefore require the retained direction to agree with the action-centered standard OPD signal:

$$g_t^{\text{opd}} = \mathbb{I}[s_t A_{\text{opd}}(y_t) > 0]\,, \tag{9}$$

$$r_t^{\text{FOPD}} = \lambda\, g_t^{\text{ctx}} g_t^{\text{opd}} s_t, \qquad \lambda = 0.1. \tag{10}$$

The two conditions play complementary roles. Cross-context consistency removes checkpoint-induced preferences whose direction changes with context, while OPD alignment ensures that the retained update still points from the current student toward the final teacher. When the student matches the final teacher on the shared support, $A_{\text{opd}} = 0$ and the FOPD reward vanishes.

3.4 Optimization and implementation

We update the student with the on-policy semi-gradient estimator

$$\nabla_\theta J_{\text{FOPD}} = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta}\left[\sum_{t=1}^{T} r_t^{\text{FOPD}} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\right]. \tag{11}$$

The scorer log probabilities, q_t, centered signals, gates, and token rewards are treated as detached batch quantities. Gradients therefore flow only through the sampled student log probability. Reference-model KL is disabled in the main experiments.

Figure 2 summarizes the complete computation for one sampled token. The four teacher scores share the same response prefix and candidate IDs and can be packed into the same scoring pipeline. The experiments separately test the contribution of cross-context consistency, action centering, shared candidate support, conservative magnitude selection, and OPD alignment.

Table 3: **Performance across 14 reasoning benchmarks.** FOPD achieves higher scores on 13 of 14 tasks, with improvements spanning mathematics, code, and science.

	Mathematics						
Method	AIME24	AIME25	AMC23	HMMT25	MATH500	Olympiad	RGMath
Qwen3-1.7B	14.2	9.4	41.2	5.0	68.6	26.0	79.5
+ OPD	36.7	23.8	70.8	17.0	81.0	37.1	90.9
+ ExOPD	35.4	23.3	75.5	14.7	81.9	37.8	90.9
+ OPD^2	**41.0**	28.8	79.5	15.0	83.9	40.1	93.9
+ FOPD	40.0	**31.04**	**80.5**	**18.0**	**87.28**	**55.98**	**94.5**

	Code				Science		
Method	CodeCont.	CodeForces	LCBv5	RGAlgo	GPQA	SuperGPQA	SciBench
Qwen3-1.7B	4.8	10.8	7.0	19.4	39.4	26.1	26.9
+ OPD	9.9	16.0	27.9	30.0	48.4	26.1	35.0
+ ExOPD	16.8	18.6	29.2	33.6	48.0	26.2	35.4
+ OPD^2	24.6	21.3	34.1	37.5	49.8	27.3	39.2
+ FOPD	**25.7**	**22.0**	**35.8**	**38.6**	**50.8**	**32.6**	**40.6**

4 EXPERIMENTS

Our experiments evaluate the central premise of FOPD: teacher supervision that remains consistent across task-relevant contexts should provide a more reliable learning signal for on-policy distillation. We examine this premise from four perspectives. We first evaluate downstream reasoning performance, then study generalization across model scales, reasoning modes, and model families. We next intervene on the key components and teacher contexts used by FOPD, and finally analyze whether the selected token-level directions provide more reliable supervision than alternative signals.

4.1 EXPERIMENTAL SETUP

The main experiment follows the model, training, and evaluation setting of OPD^2 (Heo et al., 2026) to enable a direct comparison. We use Qwen3-1.7B in non-thinking mode as the student and Qwen3-4B-Instruct-2507 / Qwen3-4B-Base as the final and base teachers, respectively (Yang et al., 2025). Training uses 25,600 problems balanced across mathematics, code, and science, drawn from OpenMathReasoning (Moshkov et al., 2025), OpenCodeReasoning, and OpenScienceReasoning-2. For each problem, a high-level task hint is constructed from the reference solution and used only in the teacher-only privileged context; the student always receives the original problem.

We train for 100 steps with a global batch size of 256. FOPD uses a shared top-1,024 candidate support, reward scale $\lambda = 0.1$, and no reference-model KL. Evaluation uses repeated mean pass@1 over 14 reasoning benchmarks. The mathematics suite contains AIME24, AIME25, AMC23, HMMT25, MATH500 (Hendrycks et al., 2021), OlympiadBench, and RGMath. The code suite contains CodeContests, CodeForces, LiveCodeBench v5 (Jain et al., 2025), and RGAlgo. The science suite contains GPQA (Rein et al., 2023), SuperGPQA (Du et al., 2026), and SciBench (Wang et al., 2023). We compare FOPD with standard OPD, ExOPD (Yang et al., 2026c), and OPD^2 (Heo et al., 2026) under the aligned evaluation setting.

Table 4: **Generalization across model configurations.** FOPD improves Macro performance across all seven model-scale, reasoning-mode, and model-family settings.

Student	Mode	OPD^2	FOPD	Gain
Qwen3-1.7B	Non-thinking	40.93	**43.35**	+2.42
Qwen3-1.7B	Thinking	48.83	**49.73**	+0.90
Qwen3-4B	Non-thinking	53.63	**54.77**	+1.14
Qwen3-4B	Thinking	59.03	**59.97**	+0.94
Qwen3-8B	Non-thinking	54.37	**55.43**	+1.06
Qwen3-8B	Thinking	62.77	**63.80**	+1.03
Gemma4-E4B-it	Native	55.37	**58.33**	**+2.96**

Table 5: **Ablation of FOPD design choices.** Full FOPD achieves the highest performance, while removing or replacing each defining component lowers the Macro score.

Variant	Math	Code	Science	Macro
Full FOPD	**58.19**	**30.53**	**41.33**	**43.35**
OPD[2]	54.6	29.4	38.8	40.93
No cross-context consistency	54.7	29.5	38.9	41.03
No OPD alignment	54.9	29.8	39.0	41.23
No action centering	54.4	28.9	38.5	40.60
Mean shared magnitude	55.2	30.1	39.3	41.53
Maximum shared magnitude	54.6	29.2	38.6	40.80
Sign only, no magnitude	55.0	29.9	39.1	41.33
Independent scorer top-k sets	50.8	25.7	36.0	37.50

4.2 Main Results and Generalization

FOPD consistently improves reasoning performance across mathematics, code, and science. As shown in table 2, FOPD reaches a Macro score of **43.35**, compared with **40.93** for the strongest baseline. This corresponds to an absolute improvement of **2.42 points** and a **5.9% relative gain**. The improvement appears in every domain: FOPD gains **3.59 points** in mathematics, **1.13 points** in code, and **2.53 points** in science. The gains extend to individual tasks. FOPD achieves higher performance on **13 of the 14 benchmarks**, spanning all three reasoning domains. In mathematics, the largest improvement is **+15.88 points** on OlympiadBench. In science, SuperGPQA improves by **+5.30 points**. The complete task-level results in Table 3 show that these gains are distributed across a diverse benchmark suite.

We next ask whether the benefit of cross-context consistency extends beyond the main Qwen3-1.7B setting. We evaluate FOPD across Qwen3-1.7B, Qwen3-4B, and Qwen3-8B in both non-thinking and thinking modes (Yang et al., 2025), and on Gemma4-E4B-it (Team et al., 2026). FOPD improves Macro performance in **all seven configurations**. The Qwen3 gains range from **+0.90 to +2.42 points**, while the cross-family Gemma4 setting achieves a **+2.96-point** improvement. These results, together with the task-level gains and training trajectories, are summarized in Figure 3. The corresponding configuration-level comparison appears in Table 4. The checkpoint trajectories provide a complementary view of the same behavior. FOPD continues to improve through later stages of training, particularly in mathematics and science, while maintaining its gains in code. Together with the consistent improvements across scales, reasoning modes, and model families, these results indicate that cross-context consistent supervision transfers beyond a single task or model configuration.

Table 6: **Teacher-context interventions.** A task-relevant high-level context yields the strongest FOPD performance, outperforming plain-context copies, random or incorrect contexts, and direct answer disclosure.

Teacher-only context	Math	Code	Science	Macro
Task-relevant high-level hint	**58.19**	**30.53**	**41.33**	**43.35**
Copy of plain context	54.7	29.5	38.9	41.03
In-domain random hint	53.8	28.4	38.0	40.07
Cross-domain random hint	53.1	27.6	37.4	39.37
Incorrect but fluent hint	51.8	26.4	36.9	38.37
Oracle answer disclosure	55.1	30.0	39.0	41.37

4.3 Ablations and Context Interventions

We next examine whether the performance gains depend on FOPD's design choices. The complete method combines shared candidate support, action centering, cross-context consistency, conservative shared-magnitude selection, and OPD alignment. Each component serves a distinct role in ensuring that identical student actions are compared consistently and that only teacher directions supported across contexts are transferred. The corresponding ablations are summarized in Table 5.

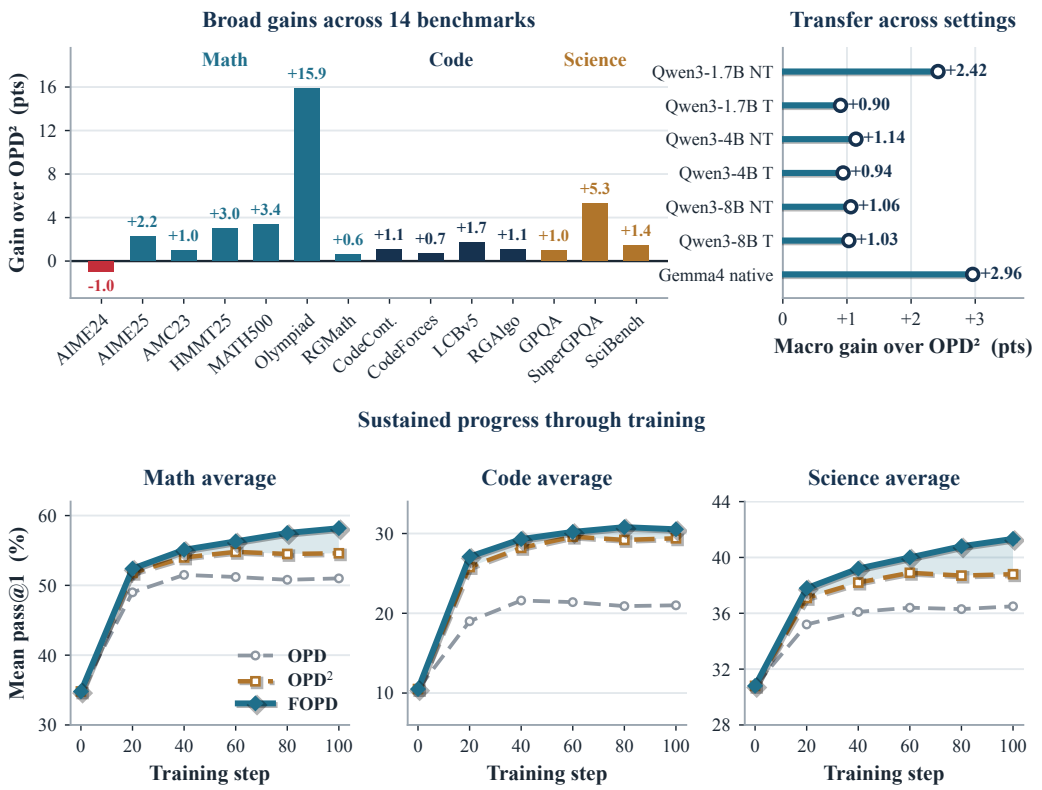


Figure 3: **Performance breadth and generalization of FOPD.** (a) Per-benchmark gains over OPD^2 across 14 reasoning tasks. (b) Macro gains across seven model configurations. (c) Mathematics, code, and science performance across training checkpoints, where FOPD continues to improve through later stages of training.

The ablations show a consistent dependence on these design choices. Removing cross-context consistency lowers Macro from 43.35 to 41.03, a **2.32-point drop**. Removing action centering and OPD alignment reduces performance by **2.75** and **2.12 points**, respectively. Replacing the conservative shared magnitude with mean, maximum, or sign-only aggregation reduces Macro by **1.82–2.55 points**. The largest decrease occurs when different scorers use independent top-k candidate sets, which lowers Macro by **5.85 points**. This result directly supports the shared-action construction in FOPD: cross-context consistency is most meaningful when every teacher condition evaluates the same student actions.

Table 7: **Matched reward controls.** Matching FOPD's reward sparsity or magnitude does not reproduce its downstream performance.

Method	Nonzero rate	Rel. reward norm	Macro
OPD^2	0.58	1.00	40.93
OPD^2, sparsity matched	0.42	0.73	40.70
OPD^2, magnitude matched	0.58	0.73	41.00
FOPD	0.42	0.73	**43.35**

We further isolate the role of the second teacher context by changing only the teacher-only privileged input. The resulting teacher-context interventions are reported in Table 6. A task-relevant high-level hint achieves the strongest performance in all three domains and reaches **43.35 Macro**. Copying the plain context lowers Macro to 41.03, while random or incorrect contexts produce 38.37–40.07. Direct answer disclosure reaches 41.37. The strongest result therefore comes from a task-relevant alternative context that provides a useful second view of the same student action. More information alone does not yield the same performance.

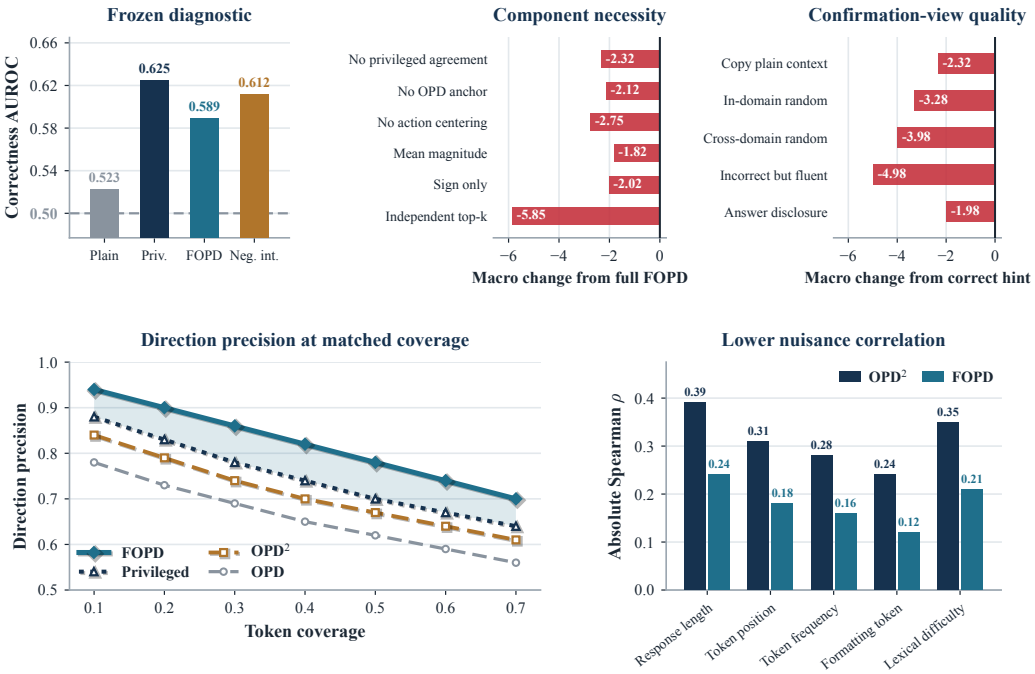


Figure 4: **Ablation and token-level evidence for FOPD.** (a) Trajectory-level comparison of plain, privileged, cross-context shared, and interaction signals. (b) Performance changes under FOPD component ablations. (c) Performance changes under different teacher-only contexts. (d) FOPD achieves the highest direction precision across matched token-coverage levels. (e) FOPD exhibits lower absolute correlation with five measured nuisance variables.

4.4 Matched Controls and Token-Level Analysis

FOPD retains only token directions consistent across teacher contexts, yielding a more selective reward signal. Compared with OPD^2, the nonzero reward rate decreases from 0.58 to 0.42 and the relative reward norm from 1.00 to 0.73. To separate better direction selection from simply fewer or smaller updates, we construct matched controls in Table 7. FOPD remains **2.65 points** above the sparsity-matched control and **2.35 points** above the magnitude-matched control, indicating that the gain comes from which token directions are retained rather than reward thinning alone.

Token-level measurements further support this interpretation. Before OPD alignment, the cross-context shared signal achieves **0.76 positive precision** and **0.59 first-error localization accuracy**. With the full FOPD selection rule, FOPD achieves the highest direction precision across all evaluated coverage levels from **0.10 to 0.70**. Its token rewards also show lower absolute correlation with all five measured nuisance variables, with **38.5%–50.0%** relative reductions versus OPD^2. These results, summarized in Figure 4, show that FOPD selects more precise update directions, while matched reward sparsity or magnitude alone does not reproduce the gain.

5 Discussion

Conclusion. We identify context sensitivity as a limitation of single-context supervision in on-policy distillation and propose **Factorial On-Policy Distillation (FOPD)** to retain token-level signals consistent across teacher checkpoints and contexts. FOPD achieves 43.35 Macro on 14 benchmarks, improving over the strongest baseline by 2.42 points (5.9% relative) and outperforming it on 13 tasks. Ablations support cross-context consistency as a reliable supervision criterion.

Limitations and Future Work. FOPD currently uses two teacher contexts and a fixed top-1,024 candidate support. Future work can explore multiple contexts, adaptive candidate selection, and applications to long-horizon reasoning, tool use, and interactive agents.

REFERENCES

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Advances in Neural Information Processing Systems*, 38, 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Dayuan Fu, Yunze Wu, Xiaojie Cai, Lyumanshan Ye, Shijie Xia, Zhen Huang, Weiye Si, Tianze Xu, Jie Sun, Keyu Li, et al. Argo: Asynchronous rollout with human guidance for research agent optimization. 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-policy delta distillation. *arXiv preprint arXiv:2607.15161*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, et al. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Mohan Jiang, Dayuan Fu, Junhao Shi, Ji Zeng, Weiye Si, Keyu Li, Xuefeng Li, Yang Xiao, Wenjie Li, Dequan Wang, et al. davinci-agency: Unlocking long-horizon agency data-efficiently. *arXiv preprint arXiv:2602.02619*, 2026.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Keyu Li, Mohan Jiang, Dayuan Fu, Yunze Wu, Xiangkun Hu, Dequan Wang, and Pengfei Liu. Datasetresearch: Benchmarking agent systems for demand-driven dataset discovery. *arXiv preprint arXiv:2508.06960*, 2025.

Keyu Li, Jin Gao, and Dequan Wang. Aligned agents, biased swarm: Measuring bias amplification in multi-agent systems. *arXiv preprint arXiv:2604.08963*, 2026a.

Keyu Li, Jin Gao, and Dequan Wang. The cost of knowing: A resource-aware protocol for benchmarking hallucination beyond static leaderboards. *arXiv preprint arXiv:2607.24063*, 2026b.

Keyu Li, Junhao Shi, Yang Xiao, Mohan Jiang, Jie Sun, Yunze Wu, Dayuan Fu, Shijie Xia, Xiaojie Cai, Tianze Xu, et al. Agencybench: Benchmarking the frontiers of autonomous agents in 1m-token real-world contexts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7422–7440, 2026c.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, et al. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 2026.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

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

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

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Songjun Tu, Chengdong Xu, Qichao Zhang, Yiwen Ma, Yaocheng Zhang, Linjing Li, Dong Li, Xiangyuan Lan, and Dongbin Zhao. Ucob: Learning to utilize and evolve agentic skills via credit-aware on-policy bidirectional self-distillation. *arXiv preprint arXiv:2606.29502*, 2026.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, et al. Seed: Self-evolving on-policy distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.14777*, 2026a.

Yunze Wu, Dayuan Fu, Weiye Si, Zhen Huang, Mohan Jiang, Keyu Li, Shijie Xia, Jie Sun, Tianze Xu, Yang Xiao, et al. Innovatorbench: Evaluating agents' ability to conduct innovative ai research. In *International Conference on Learning Representations*, volume 2026, pp. 36144–36185, 2026b.

Yang Xiao, Mohan Jiang, Jie Sun, Keyu Li, Jifan Lin, Yumin Zhuang, Ji Zeng, Shijie Xia, Qishuo Hua, Xuefeng Li, et al. Limi: Less is more for agency. *arXiv preprint arXiv:2509.17567*, 2025.

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

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, et al. Oprd: On-policy representation distillation. *arXiv preprint arXiv:2606.06021*, 2026b.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026c.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, et al. Dopd: Dual on-policy distillation. *arXiv preprint arXiv:2606.30626*, 2026.

Yanfei Zhang, Xu Lin, and Chenglin Wu. Stepopsd: Step-aware online preference distillation for agent reinforcement learning. *arXiv preprint arXiv:2605.27140*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

AI USE STATEMENT

Large language models assist with editorial polishing of the manuscript. The authors determine the scientific claims, supply the experimental results, check the mathematical content, and remain responsible for the final text.

A SUPPLEMENTARY EXPERIMENTS AND ANALYSIS

This appendix provides additional experimental details and analyses that complement the main paper. We report the full frozen diagnostic, training and evaluation settings, detailed transfer results, training dynamics, sensitivity analyses, token-level measurements, and FOPD pseudocode.

A.1 FROZEN DIAGNOSTIC OF CROSS-CONTEXT CONSISTENCY

We conduct a frozen diagnostic before policy training to examine whether teacher supervision for the same student action changes with task-relevant context. We sample 256 prompts balanced across mathematics, code, and science. Qwen3-1.7B generates eight responses per prompt, producing 2,048 fixed student trajectories (Yang et al., 2025). Qwen3-4B-Instruct-2507 and Qwen3-4B-Base then score the same response prefixes and actions under both the plain context and a teacher-only privileged context. The student trajectory, token position, candidate action, and teacher checkpoints are held fixed across contexts.

We aggregate four token-level signals at the trajectory level and use final-answer correctness as the diagnostic label. Prompt-macro AUROC is computed over the 152 prompts containing both correct and incorrect trajectories.

Table 8: **Frozen diagnostic of cross-context consistency.** Cross-context signals provide stronger correctness discrimination than the plain checkpoint contrast.

Signal	Correctness AUROC	Gain vs. plain
Plain checkpoint contrast	0.52255	–
Privileged checkpoint contrast	**0.62489**	+0.10234
FOPD shared signal	0.58938	+0.06683
Negative context interaction	0.61162	+0.08907

The plain checkpoint contrast reaches 0.52255 AUROC, whereas the FOPD shared signal reaches 0.58938. A 5,000-sample prompt-cluster bootstrap gives a 95% interval of $[0.04366, 0.09096]$ for this gain, with all bootstrap replicates remaining positive. The negative context interaction reaches 0.61162 AUROC, showing that disagreement between teacher contexts is itself informative. The privileged checkpoint contrast reaches 0.62489, showing that the additional teacher context reveals variation beyond the plain context alone. These results motivate FOPD's use of cross-context consistency to select token-level supervision.

A.2 TRAINING AND EVALUATION DETAILS

The main experiment uses Qwen3-1.7B in non-thinking mode as the student, with Qwen3-4B-Instruct-2507 and Qwen3-4B-Base as the final and base teachers (Yang et al., 2025). Training uses 25,600 problems balanced across mathematics, code, and science, drawn from OpenMathReasoning (Moshkov et al., 2025), OpenCodeReasoning, and OpenScienceReasoning-2. The student always receives only the original problem statement.

Table 9: **Main FOPD training configuration.** All teacher branches share the same student trajectory and candidate token support.

Item	Setting
Student	Qwen3-1.7B, non-thinking
Final / base teacher	Qwen3-4B-Instruct-2507 / Qwen3-4B-Base
Training examples	25,600
Math / code / science	8,534 / 8,533 / 8,533
Completions per prompt	1
Global batch size	256
Training steps	100
Prompt / response cap	4,096 / 8,192 tokens
Rollout sampling	temperature 0.7, top-p 1.0
Shared candidate support	student top-1,024 plus sampled token
Learning rate	5×10^{-6}
LR schedule	cosine; 0.1 warmup; 0.1 min-LR ratio
Gradient clipping	1.0
Reward scale	0.1
Reference KL	disabled
Hardware	8 NVIDIA H200 GPUs

Teacher-only context construction. For each training problem, Qwen3-4B-Instruct-2507 generates a high-level task hint from the reference solution before training. Deterministic decoding uses temperature 0 and at most 128 tokens, requesting one to three high-level cues: a method or conceptual step for mathematics, a governing principle for science, or an algorithm and complexity target for code. The generation prompt excludes final answers, answer choices, answer-bearing arithmetic, executable code, and test-specific constants. An automated filtering and sanitization pipeline checks hint length, answer markers, code blocks, exact answers, and answer choices. All 25,600 hints pass the final automated leakage audit.

Evaluation protocol. We follow the benchmark set, prompts, parsers, sampling configuration, and repeated mean pass@1 protocol of OPD[2] (Heo et al., 2026). The mathematics suite contains AIME24, AIME25, AMC23, HMMT25, MATH500 (Hendrycks et al., 2021), OlympiadBench, and RGMath. The code suite contains CodeContests, CodeForces, LiveCodeBench v5 (Jain et al., 2025), and RGAlgo. The science suite contains GPQA (Rein et al., 2023), SuperGPQA (Du et al., 2026), and SciBench (Wang et al., 2023).

At evaluation time, we use temperature 0.7 and top-p 0.95 and report the step-100 checkpoint. Domain scores average the benchmarks within mathematics, code, and science, and Macro is the arithmetic mean of the three domain scores.

Table 10: **Evaluation repetitions for each fixed checkpoint.**

Repetitions	Benchmarks
16	AIME24, AIME25
10	AMC23, HMMT25, MATH500, GPQA
3	OlympiadBench, RGMath, CodeContests, CodeForces, LiveCodeBench v5, RGAlgo, SuperGPQA, SciBench

A.3 Detailed Transfer Results

The main paper reports Macro improvements across seven model configurations. Here we provide the corresponding domain-level results beyond the main Qwen3-1.7B non-thinking setting. Across Qwen3 scales and reasoning modes (Yang et al., 2025), FOPD improves mathematics, code, science, and Macro performance in every listed configuration. The same pattern extends to Gemma4-E4B-it (Team et al., 2026).

Table 11: **Detailed transfer results beyond the main setting.** FOPD improves all three domain averages and the Macro score across Qwen3 scales, reasoning modes, and Gemma4.

Student / Mode	Method	Math	Code	Science	Macro
Qwen3-1.7B / Thinking	OPD^2	62.7	40.4	43.4	48.83
	FOPD	**63.6**	**41.4**	**44.2**	**49.73**
Qwen3-4B / Non-thinking	OPD^2	70.3	40.1	50.5	53.63
	FOPD	**71.2**	**41.4**	**51.7**	**54.77**
Qwen3-4B / Thinking	OPD^2	74.8	50.8	51.5	59.03
	FOPD	**75.6**	**52.0**	**52.3**	**59.97**
Qwen3-8B / Non-thinking	OPD^2	71.6	39.9	51.6	54.37
	FOPD	**72.5**	**41.2**	**52.6**	**55.43**
Qwen3-8B / Thinking	OPD^2	75.9	57.8	54.6	62.77
	FOPD	**76.8**	**59.1**	**55.5**	**63.80**
Gemma4-E4B-it / Native	OPD^2	67.8	49.5	48.8	55.37
	FOPD	**69.0**	**56.0**	**50.0**	**58.33**

Table 12: **Domain performance across training checkpoints.** FOPD continues to improve through the later stages of training.

	Math			Code			Science		
Step	OPD	OPD^2	FOPD	OPD	OPD^2	FOPD	OPD	OPD^2	FOPD
0	34.8	34.8	34.8	10.5	10.5	10.5	30.8	30.8	30.8
20	49.0	51.8	52.4	19.0	25.8	27.1	35.2	37.1	37.8
40	51.5	54.0	55.1	21.6	28.2	29.3	36.1	38.2	39.2
60	51.2	54.8	56.3	21.4	29.6	30.2	36.4	38.9	40.0
80	50.8	54.5	57.5	20.9	29.2	**30.8**	36.3	38.7	40.8
100	51.0	54.6	**58.19**	21.0	29.4	30.53	36.5	38.8	**41.33**

The detailed results follow the same trend as the aggregate results in the main paper. The Qwen3 configurations improve across all three domains, while the Gemma4 setting increases from 55.37 to 58.33 Macro, including a code improvement from 49.5 to 56.0.

A.4 Training Dynamics

We record domain performance throughout training to examine how FOPD performance changes during optimization. FOPD continues to improve through later checkpoints, particularly in mathematics and science.

The selection behavior also changes systematically during optimization. The fraction of tokens receiving nonzero FOPD rewards decreases as the student adapts, while the shared candidate support continues to cover most of the student probability mass.

The nonzero rate settles around 0.42–0.43 in later training, while the top-k support continues to cover about 0.91 of the student probability mass. The gradient norm decreases to approximately 0.18–0.20 and the mean absolute reward remains near 0.019, showing stable optimization as cross-context selection retains fewer token-level directions.

A.5 Token-Level Signal Analysis

We provide the detailed token-level measurements underlying the mechanism analysis in the main paper. Here, *FOPD shared signal* denotes the sign-consistent cross-context signal before OPD alignment, while *FOPD* denotes the final selected reward after both cross-context consistency and OPD alignment are applied.

The FOPD shared signal reaches 0.76 positive precision and 0.59 first-error localization accuracy. This complements the trajectory-level diagnostic by showing that cross-context consistency is particularly effective at identifying reliable local update directions.

Table 13: **Selected FOPD gate and optimization statistics.** FOPD becomes more selective while maintaining high candidate-support mass and stable optimization.

Step	Factorial agreement	OPD alignment	Nonzero rate	Top-k mass	Grad. norm	Reward abs. mean
1	.7920	.6594	.6344	.9217	1.6762	.02880
20	.7733	.5354	.5188	.9019	.5539	.02681
40	.7154	.4594	.4407	.9143	.2458	.01994
60	.7127	.4475	.4286	.9142	.2006	.01923
80	.7052	.4368	.4186	.9152	.1971	.01828
99	.7203	.4478	.4290	.9124	.1841	.01894

Table 14: **Token-level signal quality.** The FOPD shared signal achieves the highest positive precision and first-error localization accuracy.

Signal	Correctness AUROC	First-error acc.	Positive precision	Negative recall	Length $\lvert\rho\rvert$
OPD	0.560	0.41	0.61	0.48	0.44
OPD^2 plain checkpoint contrast	0.523	0.45	0.64	0.55	0.41
Privileged checkpoint contrast	**0.625**	0.52	0.69	0.61	0.36
FOPD shared signal	0.589	**0.59**	**0.76**	0.66	**0.24**
Negative context interaction	0.612	0.56	0.72	**0.69**	0.27

A.5.1 Direction Precision at Matched Coverage

Table 15: **Direction precision at matched token coverage.** FOPD achieves the highest precision throughout the evaluated coverage range.

Coverage	OPD	OPD^2	Privileged-only	FOPD
0.10	0.78	0.84	0.88	**0.94**
0.20	0.73	0.79	0.83	**0.90**
0.30	0.69	0.74	0.78	**0.86**
0.40	0.65	0.70	0.74	**0.82**
0.50	0.62	0.67	0.70	**0.78**
0.60	0.59	0.64	0.67	**0.74**
0.70	0.56	0.61	0.64	**0.70**

FOPD maintains the highest measured direction precision from 0.10 to 0.70 token coverage. Its advantage persists near the training operating region around 0.42 nonzero coverage, showing that it also holds when supervision density is controlled.

A.5.2 Correlation with Nuisance Variables

We measure the absolute Spearman correlation between the final token reward and five response properties: response length, token position, token frequency, formatting-token status, and prompt lexical difficulty.

Table 16: **Absolute Spearman correlation with nuisance variables.** FOPD exhibits lower correlation with all five measured response properties than OPD^2.

Nuisance variable	OPD^2	FOPD	Relative reduction
Response length	0.39	**0.24**	38.5%
Token position	0.31	**0.18**	41.9%
Token frequency	0.28	**0.16**	42.9%
Formatting-token indicator	0.24	**0.12**	50.0%
Prompt lexical difficulty	0.35	**0.21**	40.0%

Across all five properties, FOPD reduces the absolute correlation by 38.5%–50.0% relative to OPD^2. Together with the matched-coverage results, these measurements show that cross-context consistency selects more precise token-level supervision with lower correlation to the measured nuisance variables.

A.6 FOPD PSEUDOCODE

Algorithm 1 summarizes the per-token computation of FOPD. All scoring branches evaluate identical student candidate IDs, while the teacher-only privileged context remains outside the student policy.

Algorithm 1 Factorial On-Policy Distillation (FOPD)

Require: Dataset $\mathcal{D}$, student π_s, final/base teachers π_f, π_b, hints $h(x)$, top-k, scale λ
Ensure: Updated student policy π_s

for each problem x in an on-policy batch **do**
  Sample $y \sim \pi_s(\cdot \mid x)$ under the plain context
  for each position t **do**
    Construct $c_t = (x, y_{<t})$ and $c'_t = (x, h(x), y_{<t})$
    Build the shared student top-k candidate set K_t and insert y_t if needed
    Score identical candidate IDs with the student and final/base teachers under both teacher contexts
    Compute A_{plain}, A_{priv}, and A_{opd} under q_t
    if $A_{\text{plain}}(y_t)A_{\text{priv}}(y_t) \leq 0$ **then**
      $r_t^{\text{FOPD}} \leftarrow 0$
    else
      Compute s_t using Equation 7
      if $s_t A_{\text{opd}}(y_t) > 0$ **then**
        $r_t^{\text{FOPD}} \leftarrow \lambda s_t$
      else
        $r_t^{\text{FOPD}} \leftarrow 0$
  Update π_s using Equation 11