TANGENT ON-POLICY DISTILLATION: LOCAL-DIRECTION SUPERVISION BEYOND OUTPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) reduces the train–test mismatch by training on student-generated trajectories, but most existing methods supervise output distributions and do not directly constrain how intermediate representations evolve along those trajectories. This raises the question: can local changes in hidden representations provide an effective supervision signal beyond output matching? We introduce **Tangent On-Policy Distillation (TOPD)**, which constructs teacher and student token-by-layer hidden-state grids on the same student-generated rollout and aligns first finite-difference directions along adjacent response tokens and aligned network layers. TOPD combines this local-direction supervision with pointwise state matching, using cosine distance and detached, capped loss-ratio balancing to calibrate their scales. Distilling JustRL-DeepSeek-1.5B into DeepSeek-R1-Distill-Qwen-1.5B on DAPO-Math-17K, TOPD achieves 58.66 Macro Avg@16 across AIME24, AIME25, and AIMO, outperforming the strongest prior distillation baseline by 4.16 percentage points (7.6% relative). Ablations support this design: removing temporal or depth supervision lowers Macro Avg@16 by 1.70 and 1.00 points, respectively, while the proposed cosine objective with detached, capped balancing outperforms fixed weighting, magnitude matching, differentiable balancing, and uncapped balancing. These results show that local representation directions provide an effective complement to output-level supervision in on-policy distillation.

1 INTRODUCTION

On-policy distillation (OPD) provides an effective way to train reasoning models on the trajectories they generate at inference time. By learning from student-generated rollouts, OPD reduces the distribution mismatch between distillation on fixed sequences and autoregressive deployment (Song & Zheng, 2026). Most existing OPD objectives define supervision at the output level: they primarily constrain what the student predicts without directly supervising how its intermediate representations evolve along its own trajectory. Long-form reasoning requires both appropriate token predictions and coherent internal computation as reasoning progresses. This leads to the question: *can local changes in hidden representations provide an effective supervision signal beyond output matching?*

Recent OPD methods strengthen output-level supervision through generalized objectives, dual-perspective training, dual distillation, and step-aware preference signals (Yang et al., 2026c; Hou et al., 2026; Yu et al., 2026b; Zhang et al., 2026). A complementary line of work moves supervision into the representation space. OPRD matches teacher and student hidden states on the same student-generated trajectories (Yang et al., 2026b). These approaches enrich the supervision available to the student, but local representation changes remain implicit. Output-level objectives do not directly constrain intermediate representations, while pointwise hidden-state matching aligns individual states without explicitly matching the local changes between neighboring states. In an autoregressive reasoning model, computation progresses both as new tokens are generated and as representations pass through successive network layers. Two models may therefore be well aligned at individual states while still changing in different directions between them. Explicit alignment of these local changes directly supervises how internal representations evolve and complements the alignment of individual states.

We introduce **Tangent On-Policy Distillation (TOPD)**, a simple on-policy distillation objective that jointly supervises hidden states and their local first-order changes. For each student-generated rollout,

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.

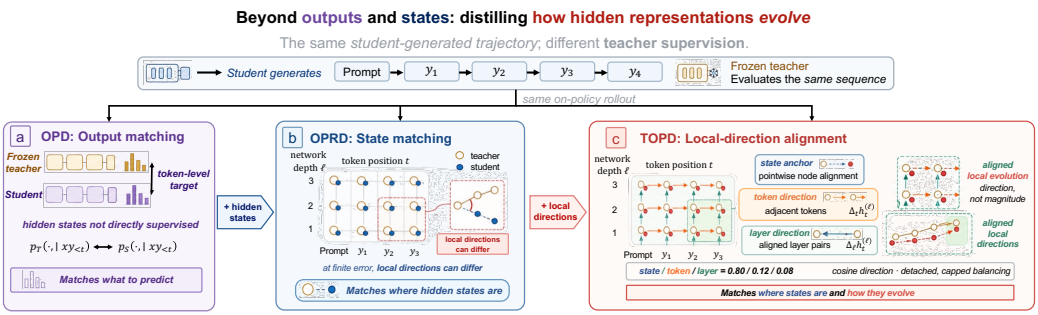

Figure 1: **From output supervision to local representation alignment.** OPD supervises output distributions, OPRD (Yang et al., 2026b) additionally matches hidden-state nodes, and TOPD aligns both hidden states and first finite-difference directions along the token and layer axes. In every panel, the frozen teacher evaluates the same student-generated sequence.

a frozen teacher evaluates the same sequence, producing paired token-by-layer hidden-state grids. TOPD retains pointwise state matching and additionally aligns the directions of first finite differences along two axes: adjacent response tokens and aligned network layers. The two directional objectives use cosine distance, while detached, capped loss-ratio balancing calibrates their scales relative to the state objective. The token-axis term captures how representations change as autoregressive reasoning advances, whereas the layer-axis term captures how the same prefix is transformed through network depth. Together, they complement absolute state alignment with explicit supervision of local representation changes, as illustrated in Figure 1.

We evaluate TOPD in a 1.5B-to-1.5B mathematical reasoning setting, distilling JustRL-DeepSeek-1.5B into DeepSeek-R1-Distill-Qwen-1.5B on DAPO-Math-17K. Across AIME24, AIME25, and AIMO, TOPD achieves a Macro Avg@16 of **58.663**, improving over the strongest prior distillation baseline by **4.163 percentage points** (**7.6% relative**). Ablations further show that both token- and layer-axis direction supervision contribute to the improvement, and that the proposed cosine-based objective with detached, capped balancing is important for stable and effective optimization. We also observe consistent gains in cross-architecture and cross-tokenizer transfer settings, suggesting that local representation-direction supervision provides a robust complement to pointwise state matching rather than relying on a particular teacher–student pairing.

Our contributions are:

- We identify local changes in intermediate representations as a direct supervision target for on-policy distillation, complementing conventional output-level objectives and pointwise hidden-state matching.
- We propose **Tangent On-Policy Distillation (TOPD)**, which aligns first finite-difference directions along both the token and layer axes using cosine distance and combines them with state matching through detached, capped loss-ratio balancing.
- We demonstrate the effectiveness of TOPD on mathematical reasoning, where it improves Macro Avg@16 by 4.163 percentage points over the strongest prior distillation baseline. Ablations, heterogeneous-model transfer experiments, and trajectory-level analyses further support the contribution of local representation direction supervision.

2 Related Work

On-policy distillation. Classical knowledge distillation transfers teacher knowledge by matching predictions on a fixed dataset (Hinton et al., 2015), whereas on-policy distillation (OPD) supervises student-generated trajectories to reduce train–test mismatch (Song & Zheng, 2026). Recent work extends OPD through reward extrapolation, dual-perspective objectives, dual distillation, step-aware preference supervision, and on-policy self-distillation (Yang et al., 2026c; Hou et al., 2026; Yu et al., 2026b; Zhang et al., 2026; Zhao et al., 2026; Yang et al., 2026a). Despite differing in how teacher signals, rewards, preferences, or self-generated targets are constructed, these methods primarily

supervise outputs or policy behavior. TOPD instead asks whether the evolution of intermediate representations on the same student-generated trajectory can provide an additional supervision signal.

Representation-level supervision and on-policy self-improvement. Representation-level distillation moves supervision inside the network. OPRD aligns teacher and student hidden states on student-generated trajectories and supports heterogeneous teacher–student pairs through a representation interface (Yang et al., 2026b). More broadly, PPO and recent group-based, self-distilled, credit-aware, and self-evolving methods study how on-policy interaction can be reused for iterative improvement (Schulman et al., 2017; Feng et al., 2026; Lu et al., 2026; Tu et al., 2026; Wu et al., 2026a). TOPD complements these approaches by aligning *local first finite-difference directions* between neighboring representations. Pointwise matching constrains where representations should be, while the token-axis and layer-axis objectives additionally constrain how they change along the generated trajectory and through network depth.

Long-horizon reasoning and agentic trajectories. Long-horizon reasoning and agentic systems further motivate trajectory-level supervision. ReAct, ALFWorld, WebShop, and search-augmented reasoning all require models to accumulate decisions over multi-step interactions (Yao et al., 2022b; Shridhar et al., 2020; Yao et al., 2022a; Jin et al., 2025). Recent work extends this setting to million-token evaluation, dataset discovery, data-efficient long-horizon agency, research agents, and asynchronous rollout with human guidance (Li et al., 2026c; 2025; Xiao et al., 2025; Jiang et al., 2026; Wu et al., 2026b; Fu et al., 2026), while other studies examine trajectory-level phenomena such as bias amplification and resource-aware hallucination evaluation (Li et al., 2026a;b). Model families such as Qwen3 and DeepSeekMath further strengthen the backbones used for reasoning-intensive training (Yang et al., 2025; Shao et al., 2024). TOPD builds on this trajectory perspective by extracting token- and layer-axis finite differences directly from teacher and student hidden states on the same rollout, providing dense local supervision without relying on a specific environment or annotation scheme.

3 Tangent On-Policy Distillation

For each student-generated response, the student and a frozen teacher process the same prompt–response sequence, producing paired hidden-state representations indexed by token position and network layer. For model $M \in \{S, T\}$, we write

$$\mathcal{H}_M = \left\{ h_{M,t}^{(l)} \,\middle|\, l = 1, \ldots, L,\ t = 1, \ldots, T \right\}. \tag{1}$$

Each hidden state is a point in the token-by-layer grid. In addition to matching corresponding states, we explicitly supervise how representations change between neighboring token positions and neighboring aligned layers. These two first finite differences define the token-axis and layer-axis local directions used by **Tangent On-Policy Distillation (TOPD)**, as illustrated in Figure 2. All equations below use a common representation space. When teacher and student architectures differ, we apply projection and layer alignment before evaluating the losses.

3.1 Local Representation Changes at Finite Error

Pointwise state matching constrains the absolute locations of teacher and student representations. Let

$$e_t^{(l)} = h_{S,t}^{(l)} - h_{T,t}^{(l)} \tag{2}$$

denote their residual mismatch. Exact state agreement also implies agreement between adjacent differences. At finite error, however, the mismatch between local changes depends on the variation of neighboring residuals:

$$\Delta_t h_{S,t}^{(l)} - \Delta_t h_{T,t}^{(l)} = e_t^{(l)} - e_{t-1}^{(l)}, \qquad \Delta_l h_{S,t}^{(l)} - \Delta_l h_{T,t}^{(l)} = e_t^{(l+1)} - e_t^{(l)}. \tag{3}$$

State matching therefore constrains local differences only indirectly. For example,

$$\left\| e_t^{(l)} - e_{t-1}^{(l)} \right\|_2^2 \le 2 \left\| e_t^{(l)} \right\|_2^2 + 2 \left\| e_{t-1}^{(l)} \right\|_2^2, \tag{4}$$

with an analogous bound along the layer axis. TOPD complements this pointwise constraint by directly aligning the directions of local representation changes.

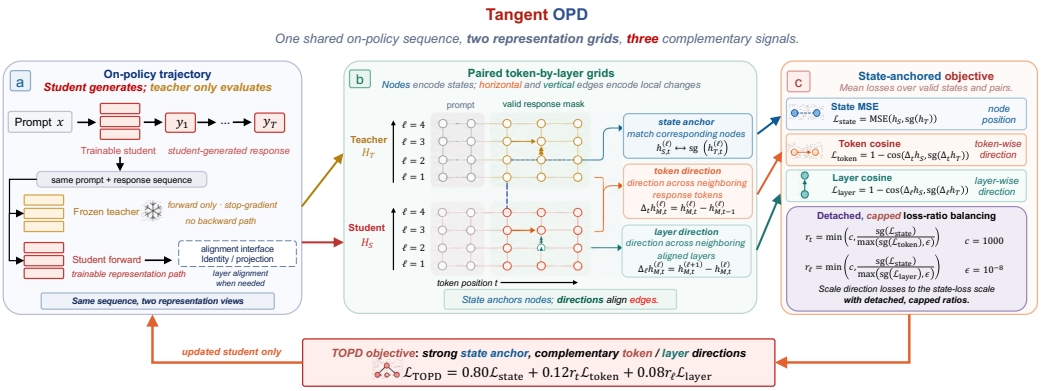

Figure 2: **Overview of Tangent On-Policy Distillation (TOPD).** The student generates a response and the frozen teacher processes the same sequence. Pointwise state MSE aligns corresponding hidden states, while cosine losses align first finite-difference directions along the token and aligned-layer axes. Detached, capped loss-ratio balancing calibrates the three objectives, and gradients update only the student.

3.2 Token- and Layer-Axis Direction Alignment

For student or teacher $M \in \{S, T\}$, we define first finite differences along the token and layer axes as

$$\Delta_t h^{(l)}_{M,t} = h^{(l)}_{M,t} - h^{(l)}_{M,t-1}, \qquad \Delta_l h^{(l)}_{M,t} = h^{(l+1)}_{M,t} - h^{(l)}_{M,t}. \tag{5}$$

The token-axis difference is defined for $t = 2, \ldots, T$ and describes how the representation changes after the next response token. The layer-axis difference is defined for $l = 1, \ldots, L-1$ and describes how the same prefix is transformed between aligned network layers.

TOPD retains pointwise hidden-state matching through

$$\mathcal{L}_{\text{state}} = \operatorname*{mean}_{l,t} \left[\frac{1}{d} \left\| h^{(l)}_{S,t} - \operatorname{sg}\left(h^{(l)}_{T,t} \right) \right\|_2^2 \right], \tag{6}$$

where the mean is taken over supervised layers and valid response positions.

To supervise local changes, we align the directions of the finite differences in eq. (5). Because hidden-state scale can vary across architectures and representation mappings, we use cosine distance rather than matching difference magnitudes:

$$\mathcal{L}_{\text{token}} = \operatorname*{mean}_{l,t} \left[1 - \cos\left(\Delta_t h^{(l)}_{S,t}, \operatorname{sg}\left(\Delta_t h^{(l)}_{T,t} \right) \right) \right], \tag{7}$$

$$\mathcal{L}_{\text{layer}} = \operatorname*{mean}_{l,t} \left[1 - \cos\left(\Delta_l h^{(l)}_{S,t}, \operatorname{sg}\left(\Delta_l h^{(l)}_{T,t} \right) \right) \right]. \tag{8}$$

Direction vectors are normalized using an epsilon of 10^{-6}. The token-axis loss follows local changes across successive response positions, whereas the layer-axis loss follows local transformations through network depth. State MSE in eq. (6) preserves absolute representation alignment, while the two cosine objectives focus specifically on local direction.

For masking, a token-axis difference is valid only when both adjacent token positions are valid. A layer-axis difference uses the valid-position mask at the corresponding response position for each aligned layer pair.

3.3 Loss Scale Balancing

The state MSE and cosine-direction losses operate at different numerical scales. We therefore calibrate each direction loss to the current state-loss scale using detached, capped ratios:

$$r_t = \min\left(c, \frac{\operatorname{sg}(\mathcal{L}_{\text{state}})}{\max(\operatorname{sg}(\mathcal{L}_{\text{token}}), \epsilon)} \right), \tag{9}$$

$$r_l = \min\left(c, \frac{\operatorname{sg}(\mathcal{L}_{\text{state}})}{\max(\operatorname{sg}(\mathcal{L}_{\text{layer}}), \epsilon)} \right), \tag{10}$$

where $c = 1000$ and $\epsilon = 10^{-8}$. The stop-gradient operator prevents the balancing ratios from introducing additional gradient paths, while the cap limits extreme rescaling when a direction loss becomes small.

The final TOPD objective is

$$\boxed{\mathcal{L}_{\text{TOPD}} = \alpha_s \mathcal{L}_{\text{state}} + \alpha_t r_t \mathcal{L}_{\text{token}} + \alpha_l r_l \mathcal{L}_{\text{layer}}}, \qquad (\alpha_s, \alpha_t, \alpha_l) = (0.80, 0.12, 0.08). \tag{11}$$

The coefficients in eq. (11) control the relative contributions of state, token-axis, and layer-axis supervision after scale calibration. In configurations with $\alpha_s = 0$, the detached state loss is still computed as the scale reference for the active direction terms. These configurations remove the state gradient while retaining the state-derived scale reference.

TOPD requires neither step-level annotations nor an additional teacher forward pass. It reuses the hidden states produced when the frozen teacher evaluates the student-generated sequence and adds only adjacent differences, normalization, and dot products during training. For heterogeneous teacher–student pairs, the same losses are evaluated after the chosen representation projection and layer alignment. At inference time, the student generates with its original architecture, without teacher access or representation matching, so TOPD introduces no additional computation.

4 Experiments

4.1 Experimental Setup

We evaluate TOPD primarily in a same-architecture mathematical reasoning setting, distilling JustRL-DeepSeek-1.5B (He et al., 2025) into DeepSeek-R1-Distill-Qwen-1.5B on DAPO-Math-17K (Yu et al., 2026a). Following the training and evaluation protocol used in OPRD (Yang et al., 2026b), training runs for 500 optimizer steps with 8 prompts and 2 student-generated responses per prompt at temperature 1.0. We supervise all 28 Transformer layers over the final 2,000 valid response tokens. Optimization uses AdamW with peak learning rate 10^{-5}, bf16 precision, and FSDP. Unless otherwise specified, TOPD uses the state/token/layer coefficients $(0.80, 0.12, 0.08)$ and the balancing cap $c = 1000$ introduced in section 3.3.

We evaluate mathematical reasoning on AIME24, AIME25, and AIMO. Each problem is sampled 16 times at temperature 0.7 and top-$p = 0.95$. Avg@16 measures average response accuracy, and Macro Avg@16 assigns equal weight to the three benchmarks. Baseline results follow the corresponding evaluation setting in Yang et al. (2026b).

In addition to the main setting, we evaluate cross-architecture transfer from Qwen3-4B to Qwen3-1.7B (Yang et al., 2025), as well as cross-tokenizer transfer from Phi-4-mini-reasoning to Qwen3-1.7B. These experiments test whether the same state and local-direction objectives can be applied through different representation interfaces.

4.2 Main Results

TOPD improves accuracy across all three mathematical reasoning benchmarks. As shown in table 1, it reaches 56.039, 35.389, and 84.561 Avg@16 on AIME24, AIME25, and AIMO, respectively,

Table 1: **Same-architecture mathematical reasoning accuracy (%).** TOPD achieves the highest distilled-student accuracy on AIME24, AIME25, AIMO, and Macro Avg@16. Baseline results follow Yang et al. (2026b). Avg@16 averages 16 sampled responses per problem, and Macro gives equal weight to the three benchmarks.

Method	AIME24	AIME25	AIMO	Macro
Student (R1-Distill-1.5B)	32.900	21.900	62.200	39.000
Teacher (JustRL-1.5B)	50.800	35.600	79.500	55.300
OPD top-1	42.300	33.500	77.000	50.900
OPD top-16	47.100	34.000	76.500	52.500
OPRD	49.800	34.600	79.100	54.500
TOPD (ours)	**56.039**	**35.389**	**84.561**	**58.663**

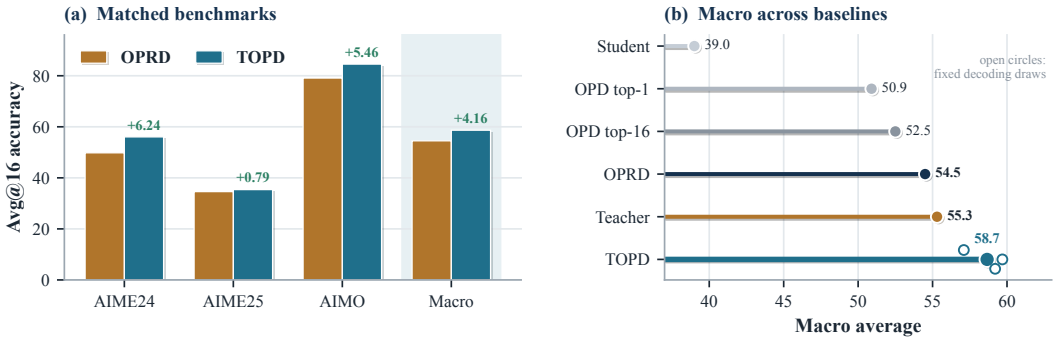

Figure 3: **Main results of TOPD.** (a) TOPD improves over OPRD by $+6.239$, $+0.789$, and $+5.461$ Avg@16 points on AIME24, AIME25, and AIMO, respectively. (b) Macro Avg@16 increases from 54.500 for the strongest prior distillation baseline to 58.663 for TOPD.

yielding **58.663 Macro Avg@16**. This improves over the strongest prior distillation baseline in this setting, which achieves 54.500 Macro Avg@16, by **4.163 percentage points**, corresponding to a **7.6% relative improvement**.

The gains span all three benchmarks: TOPD improves Avg@16 by $+6.239$ points on AIME24, $+0.789$ on AIME25, and $+5.461$ on AIMO. Figure 3 visualizes these benchmark-level gains and the resulting improvement in Macro Avg@16.

4.3 Ablation Studies

Contribution of token- and layer-axis supervision. We examine the contribution of each local-direction objective to final performance. Table 2 compares the component configurations. State matching alone achieves 54.500 Macro Avg@16. Adding token-axis supervision raises performance to 57.667, while adding layer-axis supervision reaches 56.967. Combining both directions produces the best result of 58.663.

Starting from the full TOPD objective, removing token-axis supervision decreases Macro Avg@16 by 1.696 points, while removing layer-axis supervision decreases it by 0.996 points. State + token and state + layer also outperform their corresponding direction-only configurations. These results support the central design of TOPD: pointwise state alignment provides an absolute representation reference, while token- and layer-axis directions contribute complementary local supervision.

Direction metric and loss balancing. We next compare the direction metric and loss-balancing choices. As shown in table 3, fixed-weight state-and-direction training reaches 56.900 Macro Avg@16, outperforming state-only matching. Replacing directional cosine alignment with L2 magnitude matching yields 57.100, whereas cosine alignment with the proposed detached, capped ratio balancing reaches 58.663.

Performance also depends on the balancing mechanism. A differentiable balance ratio obtains 56.200, while removing the cap gives 57.600. The full TOPD design improves over these alternatives by

Table 2: **Contribution of state, token-axis, and layer-axis supervision.** Combining state matching with both local-direction objectives gives the strongest performance. Checkmarks indicate active gradient objectives.

Method	State	Token	Layer	AIME24	AIME25	AIMO	Macro
State only (OPRD)	✓			49.800	34.600	79.100	54.500
Token direction only		✓		48.900	33.800	78.600	53.767
Layer direction only			✓	48.100	33.200	77.900	53.067
State + token	✓	✓		54.700	35.200	83.100	57.667
State + layer	✓		✓	53.800	34.900	82.200	56.967
TOPD	✓	✓	✓	**56.039**	**35.389**	**84.561**	**58.663**

Table 3: **Direction metric and loss-balancing ablations.** The final cosine direction objective with detached, capped ratio balancing achieves the highest Macro Avg@16. "TOPD gain" measures the improvement of the complete objective over each alternative.

Variant	Macro	TOPD gain
State only (OPRD)	54.500	+4.163
Fixed weights	56.900	+1.763
Differentiable balance ratio	56.200	+2.463
L2 finite-difference matching	57.100	+1.563
Cosine direction, no cap	57.600	+1.063
Cosine + detached ratio + cap	**58.663**	**0.000**

2.463 and 1.063 points, respectively. These results show that both direction-only comparison and stable scale calibration contribute to the final objective.

Objective weights and supervision positions. The same trend appears when varying the relative strength of local-direction supervision. Among the tested coefficient configurations, $(0.80, 0.12, 0.08)$ achieves the highest Macro Avg@16 of 58.663, compared with 57.400 for $(0.90, 0.06, 0.04)$ and 57.900 for $(0.70, 0.20, 0.10)$. The best tested configuration therefore combines a strong state anchor with both token- and layer-axis direction objectives.

The location of representation supervision also matters. Supervising the final 2,000 response tokens reaches 58.663 Macro Avg@16, compared with 55.300 for the first 2,000 tokens, 56.700 for uniformly sampled positions, and 57.500 when all valid positions are used. Figure 4 summarizes both trends.

4.4 Heterogeneous Distillation

We further evaluate whether the same local-direction objective transfers beyond the same-architecture setting. For Qwen3-4B $\rightarrow$ Qwen3-1.7B cross-architecture distillation (Yang et al., 2025), TOPD uses proportional layer alignment together with a learned representation projection. As shown in table 4, TOPD reaches **17.633 Macro Avg@16** and **36.200 Macro Best@16**. The strongest competing Avg@16 result is 15.300, giving TOPD a $+2.333$-point improvement. It also achieves the best result on each of AIME24, AIME25, and AIMO among the distilled students in this setting.

TOPD also transfers across tokenizers. For Phi-4-mini-reasoning $\rightarrow$ Qwen3-1.7B, the representation-space implementation reaches 7.767 Macro Avg@16 and 25.600 Macro Best@16, compared with 5.000 and 21.300 for OPRD-Bridge. This corresponds to improvements of $+2.767$ points in Avg@16 and $+4.300$ points in Best@16.

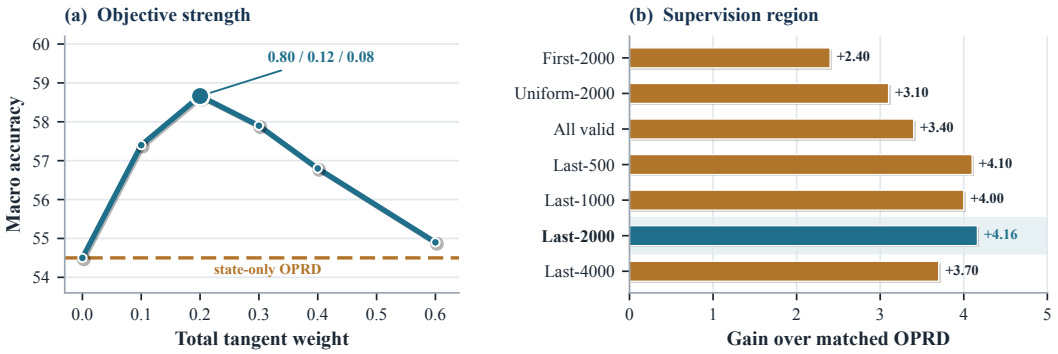

Figure 4: **Objective-weight and supervision-position ablations.** (a) A strong state anchor with complementary token- and layer-axis direction supervision achieves the best Macro Avg@16. (b) Supervising the final 2,000 response tokens gives the strongest result among the tested position configurations.

Table 4: **Cross-architecture distillation from Qwen3-4B to Qwen3-1.7B.** TOPD achieves the strongest distilled-student performance in both Macro Avg@16 and Macro Best@16.

Method	AIME24	AIME25	AIMO	Macro Avg@16	Macro Best@16
Teacher Qwen3-4B	23.100	25.200	59.400	35.900	56.900
Student Qwen3-1.7B	0.600	0.200	7.500	2.800	4.500
OPD top-1	8.500	6.300	31.000	15.300	30.800
OPD top-16	6.700	4.000	20.500	10.400	27.100
OPRD-Bridge (r=8)	5.000	3.800	24.800	11.200	30.800
OPRD-Bridge → OPD top-1	9.600	3.000	28.200	13.600	33.900
TOPD Cross-Arch	**10.800**	**7.200**	**34.900**	**17.633**	**36.200**

The generation statistics show a similar trend. In the cross-architecture experiment, TOPD obtains Dist-4g 23.800 with an average response length of 4,980 tokens, compared with 20.400 and 5,333 tokens for OPRD-Bridge. These results demonstrate that token- and layer-axis local-direction supervision remains effective when teacher and student representations require nontrivial alignment.

4.5 Training Dynamics and Response Behavior

The optimization dynamics reflect the three-part objective introduced in section 3. Across training, state MSE decreases from 4.269 to 0.212, token-axis direction loss decreases from 0.0391 to 0.00624, and layer-axis direction loss decreases from 0.0852 to 0.00993. At the same time, teacher–student representation cosine similarity increases from 0.9534 to 0.9960. These trajectories show that absolute state alignment and both forms of local-direction alignment improve together during training.

Validation performance also improves. Monitoring Macro Avg@16 reaches 59.078 at step 300 and remains strong through the end of training, while the final evaluation reaches 58.663. Mean response length also decreases over training, from 7,164 tokens in the first 20-step block to 5,336 tokens in the final block. Figure 5 summarizes these training dynamics.

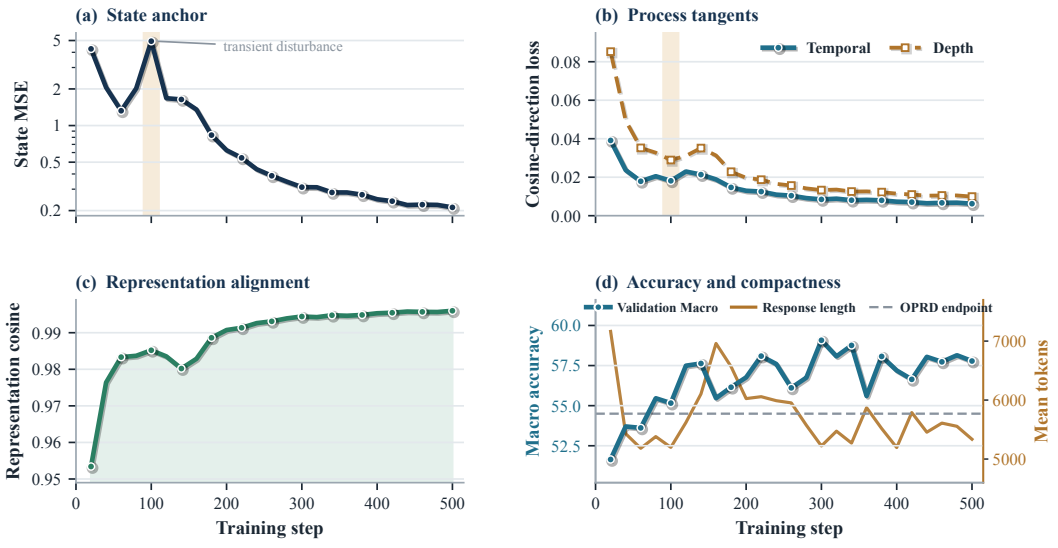

Figure 5: **Representation alignment and response behavior during training.** State, token-axis, and layer-axis losses decrease as teacher–student representation cosine increases. Validation accuracy improves while generated responses become shorter over the course of training.

The final generations are also shorter and less repetitive. TOPD responses average 5,180 tokens, compared with 5,700 for OPRD, while Dist-4g increases from 17.9 to 20.7 and self-consistency increases from 0.74 to 0.81. Thus, the accuracy gain is accompanied by shorter, less repetitive, and more consistent reasoning traces. The accuracy–length relationship is shown in panel (a) of fig. 6.

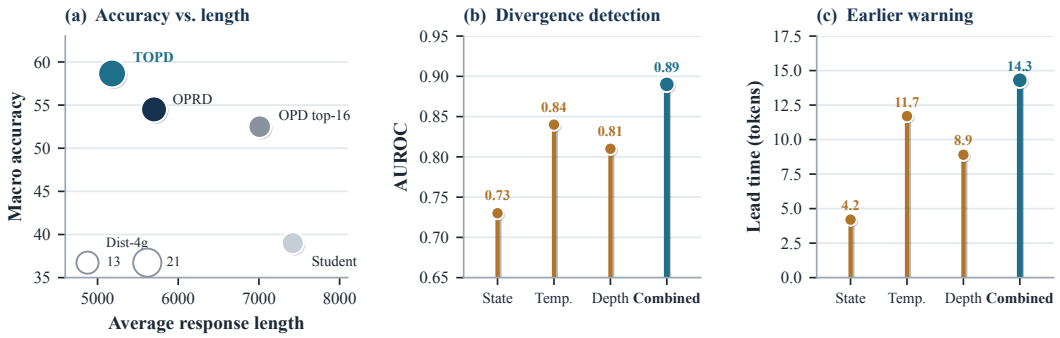

Figure 6: **Response behavior and trajectory-divergence detection.** (a) TOPD combines higher accuracy with shorter responses and stronger Dist-4g. (b) Token- and layer-axis direction gaps provide stronger divergence detection than state distance, with the combined direction signal reaching 0.89 AUROC. (c) The combined direction signal identifies divergence 14.3 tokens before visible failure, compared with 4.2 tokens for state distance.

4.6 Trajectory Divergence Detection

Finally, we examine whether the local-direction quantities optimized by TOPD are informative about reasoning divergence. State distance achieves 0.73 AUROC and identifies divergence 4.2 tokens before visible failure. In comparison, the token-axis direction gap reaches 0.84 AUROC with 11.7 lead tokens, and the layer-axis direction gap reaches 0.81 AUROC with 8.9 lead tokens. Combining the two direction signals further improves performance to **0.89 AUROC** with **14.3 lead tokens**, as summarized in Figure 6.

As shown in panels (b–c) of fig. 6, local representation directions provide both stronger discrimination and earlier warning than state distance alone. This result complements the accuracy ablations: the same token- and layer-axis changes that improve distillation performance also provide a sensitive signal for where a reasoning trajectory begins to diverge.

The experiments support local representation directions as complementary supervision beyond pointwise state matching. Both token- and layer-axis objectives improve reasoning accuracy, and the cosine direction objective with detached and capped balancing gives the strongest ablation result. The same formulation transfers across heterogeneous representation interfaces, while local-direction gaps provide earlier signals of reasoning divergence.

5 Discussion

Conclusion. We introduce **Tangent On-Policy Distillation (TOPD)**, which complements pointwise hidden-state matching with first finite-difference direction alignment along the token and layer axes. TOPD reaches 58.663 Macro Avg@16 in the main mathematical reasoning setting, outperforming the strongest prior distillation baseline by 4.163 percentage points. Ablations show that both local-direction objectives contribute to performance and that the proposed cosine objective with detached, capped loss balancing achieves the best result among the tested variants. TOPD also remains effective in heterogeneous teacher–student settings, indicating that local direction supervision is not limited to identical architectures. The trajectory analysis further suggests that local direction gaps can reveal reasoning divergence earlier than pointwise state distance alone. These results support local representation directions as an effective complement in on-policy distillation.

Limitations and Future Work. TOPD currently models representation dynamics using first finite differences between neighboring tokens and aligned layers. This formulation does not capture longer-range dependencies or higher-order structure in representation trajectories. Future work can extend TOPD to longer-range or higher-order relations and explore its effectiveness beyond mathematical reasoning, including code generation, scientific reasoning, and agentic tasks. It would also be useful to study whether direction supervision can be made adaptive across different reasoning stages or layers. More flexible representation alignment may further improve heterogeneous distillation.

REFERENCES

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.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, et al. Justrl: Scaling a 1.5 b llm with a simple rl recipe. *arXiv preprint arXiv:2512.16649*, 2025.

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.

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.

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.

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.

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.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

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*, 2026b.

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 ADDITIONAL EXPERIMENTAL AND IMPLEMENTATION DETAILS

This appendix provides supplementary details for the method and experiments in sections 3 and 4. We include the representation alignment used for heterogeneous teacher–student pairs, additional coefficient and supervision-position ablations, optimization diagnostics, and implementation details.

A.1 HETEROGENEOUS REPRESENTATION ALIGNMENT

The state and local-direction objectives in TOPD require teacher and student representations to lie in a common space. When their hidden dimensions differ, $d_S \neq d_T$, we project the student representation to the teacher dimension with a learned linear map,

$$\widetilde{h}_{S,t}^{(l)} = W_P h_{S,t}^{(l)}, \qquad W_P \in \mathbb{R}^{d_T \times d_S}. \tag{12}$$

When network depths differ, student layer i is aligned to teacher layer

$$\phi(i) = 1 + \mathrm{round}\left(\frac{(i-1)(L_T-1)}{L_S-1}\right), \qquad i = 1, \ldots, L_S. \tag{13}$$

The teacher-side layer difference corresponding to student layers $(i, i+1)$ is $h_T^{(\phi(i+1))} - h_T^{(\phi(i))}$. After projection and layer alignment, the state, token-axis, and layer-axis losses are computed exactly as defined in sections 3.2 and 3.3. For low-rank Bridge interfaces, teacher and student representations are instead mapped into the same r-dimensional space before applying the TOPD objective (Yang et al., 2026b).

Once representations are expressed in a common space, state matching over L' supervised layers, K positions, and hidden width d requires $O(L'Kd)$ operations. Each local-direction axis adds only adjacent differences, normalization, and dot products, preserving the same asymptotic dependence on sequence length and hidden width. In particular, the local-direction objectives introduce no vocabulary-size dimension.

A.2 ADDITIONAL ABLATIONS

Objective coefficients. Table 5 reports representative state/token/layer coefficient configurations. The final $(0.80, 0.12, 0.08)$ combination achieves the highest Macro Avg@16, supporting a strong state anchor together with complementary token- and layer-axis supervision.

Table 5: **Representative state/token/layer coefficient configurations.** The complete TOPD configuration achieves the highest Macro Avg@16. The state-only reference follows Yang et al. (2026b).

State / token / layer	Macro Avg@16
1.00 / 0.00 / 0.00	54.500
0.90 / 0.06 / 0.04	57.400
0.80 / 0.12 / 0.08	**58.663**
0.70 / 0.20 / 0.10	57.900
0.60 / 0.24 / 0.16	56.800

Supervision positions. TOPD is also sensitive to which response positions receive representation supervision. As shown in table 6, supervising the final 2,000 response tokens gives the strongest result among the tested regions. Supervision in later response regions consistently outperforms supervision at the beginning of the response or at uniformly sampled positions.

Table 6: **Supervision-position ablation for TOPD.** Supervising the final 2,000 valid response tokens achieves the best Macro Avg@16 among the tested configurations.

Position	Tokens	Macro Avg@16
First-k	2,000	55.300
Uniform sampled	2,000	56.700
All valid	all	57.500
Last-k	1,000	58.200
Last-k	**2,000**	**58.663**
Last-k	4,000	58.100

A.3 Optimization Diagnostics

The three objectives become increasingly compatible over training. On the shared student backbone, pairwise gradient cosine similarities increase from early to late training, while the measured gradient-conflict rate decreases from 28% to 11% (table 7). This trend complements the representation-loss dynamics reported in section 4.5: state alignment and both forms of local-direction alignment improve together throughout optimization.

Table 7: **Gradient alignment across training phases.** Pairwise gradient cosine similarities increase while gradient conflicts become less frequent.

Metric	Early	Mid	Late
$\cos(g_{\text{state}}, g_{\text{token}})$	0.31	0.44	0.52
$\cos(g_{\text{state}}, g_{\text{layer}})$	0.24	0.37	0.46
$\cos(g_{\text{token}}, g_{\text{layer}})$	0.18	0.29	0.41
Gradient conflict rate	28%	17%	11%

A.4 Implementation Details

For the main same-architecture experiment, training uses a maximum response length of 16,384 tokens. Each optimization step uses 8 prompts with 2 student rollouts per prompt, and the per-GPU micro-batch size is 1. Training uses AdamW, bf16 precision, FSDP, and gradient clipping at 1000. Evaluation uses a maximum response length of 15,360 tokens.

For the TOPD losses, direction vectors use L2 normalization with $\epsilon = 10^{-6}$. The detached loss ratios use a denominator floor of 10^{-8} and cap $c = 1000$, consistent with section 3.3.