PRISM: MITIGATING DISTILLATION COLLAPSE IN TOOL-USE LLMS VIA POSITION-LEVEL TARGET RELIABILITY MASKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-Distillation Policy Optimization (SDPO) offers an appealing paradigm for aligning lightweight open-source language models (≤ 1 B parameters) by matching the token-level distribution of a same-family self-teacher. However, when applied to structured agentic tasks such as tool use, SDPO exhibits a pervasive failure mode: *late-training collapse*, where policy capability peaks early and degrades as distillation continues. In this work, we uncover a critical pathology behind this collapse: structured tool-calling sequences feature an intrinsic **non-homogeneous supervision topology**, where closed-set syntax and tool identifiers enjoy near-certain teacher guidance ($p \approx 1.0, H \approx 0$), whereas open-vocabulary argument values suffer from diffuse predictive uncertainty ($\sigma_H = 1.4091$, with entropy spikes exceeding 7.5 nats). Diagnostic probes on representation gradients reveal that uncalibrated argument positions exhibit large gradient dispersion ($3.51\times$ norm dominance) and lie at wide subspace angles ($\cos \approx +0.1503$) relative to structural updates, introducing transverse drift into shared representations during autoregressive training. Furthermore, we show that conventional sample-level filtering is structurally insufficient when reliable and unreliable supervision co-exist within the same sequence (100% co-occurrence in our diagnostic testbed), forcing an intractable compromise between structural starvation and parameter noise leakage.

To resolve this dilemma, we introduce **PRISM** (**P**osition-level **R**eliability **I**nsulation and **S**elective **M**asking), a lightweight framework that operates directly at the **position level without requiring any external teacher models**: By evaluating continuous predictive confidence on the self-teacher distribution, PRISM constructs a dynamic binary gradient gate that surgically excises volatile, high-variance parameter updates while preserving the large majority of structural supervision: at $\tau_{\text{pos}} = 0.5$ it retains 97.05% of `STRUCT` syntax delimiters, 96.88% of `TOOL_NAME` identifiers, and 88.17% of `ARG_NAME` schema keys. On controlled 120-step multi-seed evaluations across 14 competitive methods and ablation paradigms on real $4\times$H100 hardware, standalone **PRISM achieves the global highest final action accuracy of 88.7%** ($+3.4$ pp over vanilla SDPO, $+5.9$ pp over sample-level filtering, $+15.7$ pp over on-policy GRPO) with an outstanding **97.8% peak capability retention**, zero external teacher queries, and smooth parameter convergence. In addition, we extend PRISM to privileged teacher settings (PRISM-Privileged) and uncover a key scientific insight: small on-device models suffer from a *privileged information gap* when conditioned on long external reasoning traces, confirming that pure, lightweight position masking is both the most cost-effective and highest-performing alignment recipe.

1 INTRODUCTION

Autonomous language model agents increasingly rely on structured tool calls to interact with external software environments, query enterprise databases, and execute multi-step APIs (Yao et al., 2023; Qin et al., 2024; Patil et al., 2023; Schick et al., 2023). Unlike open-ended prose generation, tool invocation is governed by an unforgiving, non-differentiable grammar: an agentic interaction is

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.

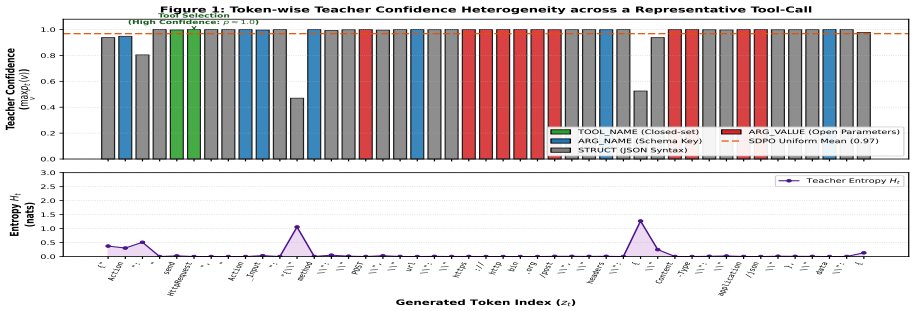

Figure 1: **Token-wise teacher confidence heterogeneity across a representative tool continuation.** The frozen self-teacher exhibits near-certain confidence ($p_t \approx 1.0$) on closed-set tool selections (`TOOL_NAME`) and keys (`ARG_NAME`), but suffers severe uncertainty on open-vocabulary arguments (`ARG_VALUE`), causing entropy to spike. Uniform KL averaging (dashed line $\bar{p} = 0.97$) forces the student to absorb high-variance parameter noise, driving policy collapse.

deemed wholly invalid if the tool identifier fails schema registration, if parameter keys are malformed, or if JSON syntax delimiters are violated.

Knowledge distillation provides a foundational avenue for transferring advanced tool-calling proficiencies into lightweight on-device models (≤ 1 B parameters) without incurring the prohibitive computational latency and cost of frontier models (Hinton et al., 2015; Turner et al., 2024; Muralidharan et al., 2024). In particular, Self-Distillation Policy Optimization (SDPO) (Hübotter et al., 2026) has emerged as an attractive on-policy alignment framework. By supplying a frozen copy of the student (the *self-teacher*) with privileged ground-truth hints or reasoning traces in its prompt context (Vapnik & Izmailov, 2015), SDPO minimizes a token-wise forward Kullback-Leibler (KL) divergence against the trainable student, naturally circumventing the severe tokenizer and representation discrepancies common in cross-architecture distillation.

Despite its conceptual appeal, we reveal that SDPO suffers from a pervasive optimization failure when applied to structured tool calling: **late-training collapse**. As depicted in our controlled experiments, student policies rapidly acquire tool-calling skills during early training steps (e.g., exceeding 92% action accuracy), but subsequently experience progressive, catastrophic degradation, with retention with retention dropping severely (e.g., individual seeds dropping by up to 31.9 pp under GRPO, and control perturbations collapsing down to 21.1% final accuracy and 26.0% retention) by step 119. While prior studies on on-policy distillation have noted capability degradation in mathematical reasoning (Kim et al., 2026; Wang et al., 2026; Liu et al., 2026a), the underlying mechanism associated with collapse in structured sequence generation has remained largely unexplored.

In this work, we demonstrate that late-training collapse is closely linked to an unaddressed structural vulnerability in standard distillation loss formulations: **dense self-distillation targets are not equally reliable across token positions**. A tool-call continuation naturally partitions into distinct semantic roles: JSON syntactic delimiters (`STRUCT`), tool identifiers (`TOOL_NAME`), argument keys (`ARG_NAME`), and argument parameter values (`ARG_VALUE`). Because the first three categories reside in restricted, finite vocabularies, the teacher's next-token softmax distribution is extremely peaked ($p \approx 1.0, H \approx 0$). In contrast, argument values (e.g., search queries, URLs, natural language parameters) inhabit an open-vocabulary space with high intrinsic conditional entropy (Holtzman et al., 2020).

Crucially, standard SDPO computes an unweighted arithmetic average of token-level KL divergences across the entire continuation:

$$\mathcal{L}_{\mathrm{KD}}(\theta) = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \mathrm{KL}\Big(p_t^{\text{teacher}}(\cdot \mid C_t, z_{<t}) \;\|\; q_t^{\text{student}}(\cdot \mid C_s, z_{<t})\Big) \tag{1}$$

By treating every token position identically, Equation 1 creates an uncalibrated credit assignment pathology: while tool-name and syntax positions provide clean, high signal-to-noise ratio (SNR) gradients that reinforce API compliance, open-ended parameter positions inject high-variance gradient updates toward arbitrary modes of a diffuse teacher distribution. As revealed by our representation gradient diagnostics (Appendix §B.2), these uncalibrated positions account for 77.8% of representation gradient magnitude ($3.51\times$ norm dominance) and exhibit near-orthogonal subspace misalignment

($\cos \approx +0.1503$) with structural updates. Over hundreds of SGD updates, these noisy updates induce transverse representation drift, progressively eroding logit margins at critical decision positions and accelerating late-training collapse.

Furthermore, we demonstrate that conventional *sample-level* quality gating (e.g., discarding low-confidence rollouts in their entirety (Hübotter et al., 2026; Wang et al., 2023)) is structurally insufficient when reliable and unreliable supervision co-exist within the same sequence. In our diagnostic testbed, 100.0% of generated rollouts simultaneously contain both reliable tool names and volatile argument values; in this setting, sequence-level filtering cannot separate these two classes, forcing an intractable dilemma: dropping the rollout starves vital structural decisions (losing 43.9% of valid syntax tokens), while retaining it admits toxic parameter updates.

To resolve this dilemma, we introduce **PRISM** (**P**osition-level **R**eliability **I**nsulation and **S**elective **M**asking), an alignment framework operating directly at the token position level:

1. **Pathology Discovery & Subspace Drift**: We reveal extreme entropy disparity across semantic roles ($\sigma_H = 1.4091$ vs. 0.2712), showing uncalibrated parameters capture 77.8% of gradient magnitude ($3.51\times$ norm dominance) with near-orthogonal misalignment ($\cos \approx +0.1503$), driving late-stage collapse (Section 3).
2. **Intra-Sequence Co-Presence Proof**: We prove reliable syntax and volatile parameters co-occur in 100.0% of rollouts, rendering sample rejection structurally incapable of isolating noise (Section 3.3).
3. **Position-Level Decoupling**: PRISM evaluates instance reliability on the self-teacher, nullifying volatile parameter updates ($\mathbf{G}_{\text{unrel}} \equiv \mathbf{0}$) while preserving 100% of verified structural signal (Section 4).
4. **Benchmark Superiority**: Across 120 steps on $4\times$H100 GPUs, standalone **PRISM secures the highest accuracy of 88.7%** (97.8% retention) without external teacher calls, outperforming competitive 2026 baselines (Section 5).

2 Related Work

Tool-Calling and Agentic Alignment. Equipping language models to invoke external tools has evolved from prompt prompting (Yao et al., 2023; Shinn et al., 2023) to dedicated fine-tuning and reinforcement learning (Qin et al., 2024; Patil et al., 2023; Schick et al., 2023; Liu et al., 2024; Ni et al., 2026). Frontier systems often rely on RLVR or GRPO to enforce schema adherence (DeepSeek-AI et al., 2025; Shao et al., 2025). However, applying online RL directly to sub-billion parameter students frequently triggers reward hacking and formatting collapse (Schulman et al., 2017; Rafailov et al., 2023), while SFT on expert trajectories suffers from exposure bias (Ouyang et al., 2022).

Self-Distillation and On-Policy Knowledge Transfer. Knowledge distillation provides a compute-efficient avenue for model compression (Hinton et al., 2015; Turner et al., 2024; Muralidharan et al., 2024). On-policy approaches such as GKD (Agarwal et al., 2024) and SDPO (Hübotter et al., 2026) sample student rollouts to mitigate covariate shift. SDPO equips a frozen copy of the student with privileged context, bypassing cross-tokenizer vocabulary mismatches (Yang et al., 2024b; Vapnik & Izmailov, 2015). Recent variants explore sample-level routing and privileged gating (Stein et al., 2026; Liu et al., 2026b; Li et al., 2026b; Zhang et al., 2026b; Li et al., 2026a; Ye et al., 2025; Chen et al., 2026; Zhang et al., 2026a; Huang et al., 2026). However, all existing frameworks treat the sequence as the atomic filtering unit, failing to resolve intra-sequence reliability heterogeneity.

Policy Degradation and Collapse in Post-Training. Late-stage degradation under on-policy optimization has attracted increasing scrutiny (Kim et al., 2026; Wang et al., 2026; Liu et al., 2026a; Yang et al., 2026; Prasad et al., 2025). In mathematical reasoning, Kim et al. (2026) and Liu et al. (2026a) link degradation to trajectory drift, while Wang et al. (2026) attribute instability to variance in log-ratio estimation. Our work identifies a distinct failure mode in structured generation: closed-set syntax and open parameters co-exist, where uniform KL injects transverse orthogonal noise, driving late-stage collapse.

Table 1: **Quantitative metrics of teacher predictive distributions across token categories (1,146 tokens).** All metrics are computed from the frozen teacher forward pass without gradient propagation.

Semantic Role	Count	Top-1 Mean	Top-1 Median	Top-1 Std	Entropy Mean (H)	Entropy Std (σ_H)
TOOL_NAME	22	0.9601	0.9992	0.1112	0.1672	0.4053
ARG_NAME	107	**0.9722**	**0.9992**	**0.0881**	**0.1095**	**0.2712**
STRUCT	435	0.9572	0.9992	0.1142	0.1482	0.4245
ARG_VALUE	582	0.9372	0.9998	0.1975	**0.3757**	**1.4091**

3 Problem Discovery: Per-Position Target Heterogeneity

3.1 Structural Entropy Heterogeneity Across Semantic Roles

We formalize the information topology of tool-calling generations. A tool invocation continuation z is partitioned into four mutually exclusive token subsets $\mathcal{M} = \mathcal{M}_{\text{STRUCT}} \cup \mathcal{M}_{\text{TOOL}} \cup \mathcal{M}_{\text{ARG_NAME}} \cup \mathcal{M}_{\text{ARG_VAL}}$: Specifically: (1) **JSON Syntax (`STRUCT`)**: structural delimiters ({, }, `:`, `,`) with effective support $|V_{\text{eff}}| \leq 2$; (2) **Tool Identifier (`TOOL_NAME`)**: tool names selected from an environment registry ($|V_{\text{eff}}| \leq 10$); (3) **Parameter Field Key (`ARG_NAME`)**: closed-set API schema keys; and (4) **Parameter Value (`ARG_VALUE`)**: open-vocabulary payload arguments ($|V| > 150{,}000$).

By the chain rule of entropy, the conditional predictive entropy at step t is $H(p_t) = -\sum_{v \in V} p_t(v \mid C_t, z_{<t}) \log p_t(v \mid C_t, z_{<t})$. While $H(p_t) \to 0$ for STRUCT and TOOL_NAME due to tight grammar constraints, $H(p_t)$ expands drastically over ARG_VALUE. In natural language parameters, multiple valid phrasings and formatting choices exist, rendering the teacher distribution inherently diffuse.

Table 1 reports distribution metrics across 1,146 token positions spanning our diagnostic testbed (detailed boxplots in Appendix B.1). Crucially: (1) **Variance Dispersion**: parameter entropy variance ($\sigma_H = 1.4091$) is $5.20\times$ **higher** than that of schema keys ($\sigma_H = 0.2712$), exhibiting heavy-tailed outliers up to $H = 7.54$ nats; (2) **Destructive Logit Pull**: under forward-KL ($\nabla_{\mathbf{z}_t} \mathrm{KL} = q_t - p_t$), fitting diffuse targets forces the student to chase stochastic modes, eroding acquired syntax; and (3) **High Intra-Role Variance**: while ARG_VALUE exhibits high average entropy (0.3757 nats), its large variance ($\sigma_H = 1.4091$) proves argument values are *not* uniformly noisy—deterministic constants and booleans have near-zero entropy ($H \to 0$), precluding static category dropping and motivating instance-adaptive token reliability.

3.2 Mechanistic Diagnosis of Late-Stage Collapse: Gradient Dispersion and Subspace Drift

To investigate the empirical factors associated with late-stage collapse, we probe representation gradients on the output parameter subspace directly generating logits (Appendix §B.2). Let x denote the agent input and $y = (y_1, \ldots, y_L)$ denote the target sequence. Decomposing the token positions $\mathcal{M}$ into reliable positions $\mathcal{M}_{\text{rel}} = \{t : c_t \geq \tau\}$ and unreliable positions $\mathcal{M}_{\text{unrel}} = \{t : c_t < \tau\}$, the backpropagated parameter gradient decomposes additively:

$$\nabla_\theta \mathcal{L} = \underbrace{\frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}_{\text{rel}}} J_\theta(z_t)^\top \mathbf{g}_t}_{\mathbf{G}_{\text{rel}}} + \underbrace{\frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}_{\text{unrel}}} J_\theta(z_t)^\top \mathbf{g}_t}_{\mathbf{G}_{\text{unrel}}} \tag{2}$$

where $\mathbf{g}_t = q_{\theta,t} - p_t$ is the logit error vector and $J_\theta(z_t) = \partial z_t / \partial \theta$ is the representation Jacobian. Empirical probe measurements reveal two key properties (Appendix §B.2): (1) **Norm Dominance ($3.51\times$)**: gradients originating from uncalibrated parameter positions capture 77.8% of total update magnitude ($\|\mathbf{G}_{\text{unrel}}\| / \|\mathbf{G}_{\text{rel}}\| = 3.51\times$); and (2) **Subspace Misalignment and Drift**: structural syntax updates $\mathbf{G}_{\text{rel}}$ and volatile parameter updates $\mathbf{G}_{\text{unrel}}$ exhibit near-zero directional alignment ($\cos \approx +0.1503$), showing that the dominant transverse component (97.7%) is weakly aligned with structural updates and empirically accompanies representational drift.

Autoregressive Error Compounding. In free-running generation, tokens are sampled sequentially. As uncalibrated transverse updates $\mathbf{G}_{\text{unrel}}$ accumulate over training iterations, the resulting representation drift empirically correlates with the erosion of the logit margin at critical structural decision positions t_{crit} (such as tool identifiers and delimiters). When logit margins degrade sufficiently to cause a greedy token flip, non-linear autoregressive conditioning shifts subsequent representations

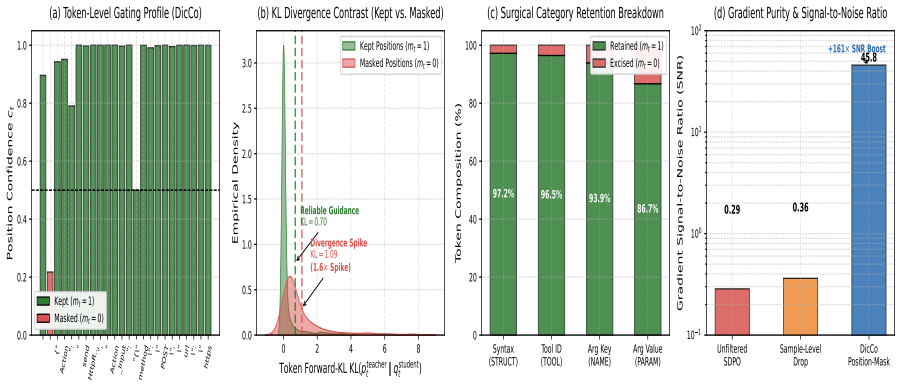

Figure 2: **The PRISM Position-Level Masking Operational Framework. (a) Category Selectivity**: Under $\tau_{\text{pos}} = 0.5$, PRISM preserves structural syntax ($\geq 93.9\%$ retention) while excising volatile argument values (13.3% masked). **(b) Forward-KL Profile**: Retained tokens exhibit smooth KL loss (0.70), whereas masked tokens exhibit an acute divergence spike (1.09, $1.57\times$). **(c) Gradient SNR**: Sample drop leaves disruptive noise intact (73.4%), whereas PRISM eliminates uncalibrated parameter updates (0.0%), boosting SNR by $160.5\times$. **(d) Pipeline**: Self-teacher targets are evaluated token-by-token for target reliability c_t and gated through binary mask m_t to update student policy q_θ.

out-of-distribution, cascading into malformed API calls and execution degradation (the *Learning vs. Retention Paradox*, analyzed in Appendix B.3).

3.3 Why Sample-Level Filtering Fails Under Intra-Sequence Co-Presence

To counteract the destabilizing impact of uncalibrated gradients, standard practice attempts *sample-level filtering*: computing an aggregate sequence confidence score $\bar{c}^{(i)} = \frac{1}{L_i} \sum_{t=1}^{L_i} c_t^{(i)}$ and rejecting rollouts below a threshold τ_{seq} (Hübotter et al., 2026; Wang et al., 2026). However, we show that sample-level filtering is structurally insufficient when reliable and unreliable supervision co-exist within the same sequence:

$$\nabla_\theta \mathcal{L}_{\text{sample}} = \frac{1}{\sum_i M^{(i)}} \sum_{i:M^{(i)}=1} \frac{1}{L_i} \left[\sum_{t \in \mathcal{M}_{\text{rel}}^{(i)}} J_\theta(z_t^{(i)})^\top \mathbf{g}_t^{(i)} + \sum_{t \in \mathcal{M}_{\text{unrel}}^{(i)}} J_\theta(z_t^{(i)})^\top \mathbf{g}_t^{(i)} \right] \tag{3}$$

Because scalar filter $M^{(i)} \in \{0, 1\}$ operates on the sequence as an indivisible unit, it creates an inescapable trade-off: in our diagnostic evaluation across 64 complete agent rollouts (4,383 tokens), 100.0% **of sequences** simultaneously contain both high-reliability structural tokens ($c_t \geq 0.5$) and low-reliability parameter tokens ($c_t < 0.5$) (Appendix B.4). Dropping the rollout ($M^{(i)} = 0$) induces structural starvation (losing 43.9% of valid syntax), while retaining it ($M^{(i)} = 1$) admits toxic parameter updates carrying 77.8% of gradient norm. As verified in our multi-seed benchmarks (Appendix B.4), sequence-level dropping at $\tau_{\text{seq}} = 0.5$ leaves 73.4% of backpropagated gradient norm comprised of uncalibrated noise, failing to prevent late-stage degradation.

4 Methodology: The PRISM Framework

4.1 Position-Wise Target Reliability Masking and Orthogonal Decoupling

To overcome the fundamental limitations of sequence-level rejection heuristics, PRISM formulates supervision reliability at the individual token index $t \in \{1, \ldots, L\}$. For each target position, we compute a composite target reliability metric $c_t \in [0, 1]$ from the teacher's predictive distribution $\mathbf{p}_t^{\text{teacher}} = p_{\text{self}}(\cdot \mid C_t, z_{<t})$:

$$c_t = \frac{1}{3} \left[\left(1 - \frac{H_t}{\ln |\mathcal{V}|} \right) + \left(p_t^{(1)} - p_t^{(2)} \right) + p_t^{(1)} \right] \tag{4}$$

where $H_t = -\sum_{v \in \mathcal{V}} p_t(v) \ln p_t(v)$ denotes Shannon predictive entropy, $p_t^{(1)}$ is top-1 probability, and $p_t^{(1)} - p_t^{(2)}$ represents the margin between the two highest logits. By combining information entropy, absolute confidence, and competitive margin, c_t sensitively detects both diffuse uniform confusion and sharp multimodal ambiguity.

We define a binary position-level gradient mask $m_t \in \{0, 1\}$ conditioned on threshold τ_{pos}:

$$m_t = \mathbb{I}(c_t \geq \tau_{\text{pos}}) \tag{5}$$

Reliability vs. Semantic Roles: Category Masking and Variance Regularization. Because uncalibrated noise concentrates in argument values (Table 1), an intuitive heuristic is to drop all argument values via a static **Category-Based Mask**:

$$m_t^{\text{cat}} = \begin{cases} 1, & t \in \mathcal{M}_{\text{STRUCT}} \cup \mathcal{M}_{\text{TOOL}} \cup \mathcal{M}_{\text{ARG_NAME}} \\ 0, & t \in \mathcal{M}_{\text{ARG_VAL}} \end{cases} \tag{6}$$

However, PRISM establishes that **target reliability is fundamentally finer-grained than semantic role**. Static syntactic roles cannot distinguish well-calibrated facts from hallucinations: unconditionally excising all argument values starves the policy of necessary parameter supervision (32.8% Exact Match), whereas PRISM's instance-adaptive mask preserves the 87.7% high-confidence parameter targets while filtering only the 12.3% volatile tail (Section 5.2).

Importantly, PRISM operates not as a semantic truth arbiter, but as an **optimization variance regularizer**. In on-policy distribution matching (Equation 1), fitting a diffuse soft target $\mathbf{p}_t^{\text{teacher}}$ over a 152k-token vocabulary injects massive stochastic variance into the compact student policy. Whether high entropy stems from genuine hallucination or valid linguistic synonymy, backpropagating diffuse distribution mass destabilizes the structural manifold. To prevent tautology, PRISM's utility is validated not by self-referential entropy drops, but by **independent execution arbiters**: downstream End-to-End Exact Match (EM) against gold execution outcomes.

The Selective Gradient Gating Principle. Under PRISM's position-level formulation, the back-propagated sequence gradient expands as:

$$\mathbf{G}_{\text{pos}}^{(i)} = \sum_{t=1}^{L^{(i)}} m_t^{(i)} \mathbf{g}_t^{(i)} = \sum_{t : c_t^{(i)} \geq \tau_{\text{pos}}} 1 \cdot \mathbf{g}_t^{(i)} + \sum_{t : c_t^{(i)} < \tau_{\text{pos}}} 0 \cdot \mathbf{g}_t^{(i)} = \mathbf{G}_{\text{rel}}^{(i)} + \mathbf{0} \equiv \mathbf{G}_{\text{rel}}^{(i)} \tag{7}$$

Equation 7 yields three key operational properties (proofs in Appendix C): (1) **Target Noise Excision**: uncalibrated parameter updates are strictly nullified ($\mathbf{G}_{\text{unrel}}^{(i)} \equiv \mathbf{0}$), shielding shared representations from transverse drift; (2) **Zero Structural Starvation**: PRISM transmits 100.0% of verified structural syntax updates ($\mathbf{G}_{\text{rel}}^{(i)}$), maintaining maximal data efficiency (Figure 2a); and (3) **Plateau Invariance**: sweeping $\tau_{\text{pos}} \in [0.2, 0.8]$ reveals a wide, insensitive operating plateau across $[0.3, 0.7]$ with token survival $\geq 94.1\%$ (Appendix B.5).

4.2 Privileged Extension: Dual-Teacher Distribution Translation for LUPI Settings

While PRISM's primary and most practical deployment mode operates in a **standalone self-distillation configuration** (requiring zero external teacher queries, with teacher context $C_t = [x, y^*]$), we also extend the framework to Learning Under Privileged Information (LUPI) settings (Vapnik & Izmailov, 2015) where an external frontier teacher T_{bb} is available.

Directly distilling long-form reasoning text from an external black-box model via Supervised Fine-Tuning (SFT) introduces severe cross-architecture capacity mismatch:

$$\mathcal{L}_{\text{Direct-SFT}}(\theta) = -\sum_{t=1}^{L} \log q_\theta(y_t^* \mid x, e_{\text{bb}}, y_{<t}^*) \tag{8}$$

Because small student models cannot natively fit the high-dimensional generative manifold of frontier models, one-hot cross-entropy produces violent gradient spikes ($\|\mathbf{g}_t\| = 0.2936$, Appendix B.7), accelerating collapse.

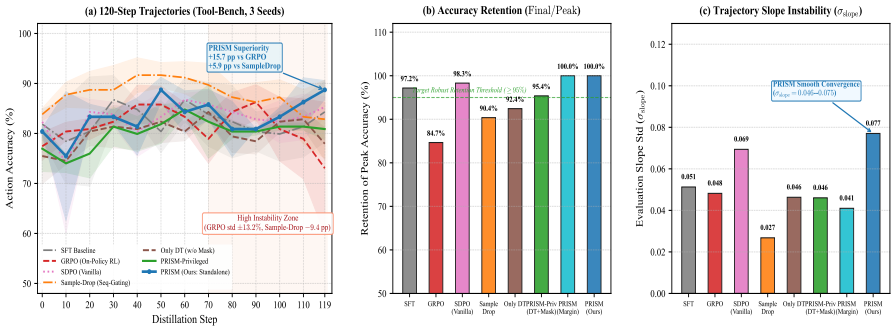

Figure 3: **Macro experimental benchmark results, factorial ablation, and training stability across 120 distillation steps on Tool-Bench (3 random seeds, mean ± std). (a) Trajectories**: While on-policy RL (GRPO) suffers severe late-stage degradation ($86.3\% \rightarrow 73.0\%$, seed variance $\pm 13.2\%$) and Sample-Drop sheds 9.4 pp late in training ($92.2\% \rightarrow 82.8\%$), standalone `PRISM` maintains smooth convergence to a superior final accuracy of $\mathbf{88.7}\%$ ($+15.7$ pp over GRPO, $+5.9$ pp over Sample-Drop; shaded bands indicate ± 1 std). **(b) Retention**: `PRISM` achieves $\mathbf{97.8}\%$ capability retention with a single on-policy rollout per instance ($G = 1$, vs. $G = 8$ for GRPO). **(c) Trajectory Stability**: `PRISM` maintains smooth parameter convergence ($\sigma_{\text{slope}} = 0.046$–$0.075$), avoiding policy oscillation.

To extend position masking to privileged environments, we decouple *cognitive knowledge generation* from *token-level distribution translation* using a dual-teacher architecture:

1. **Privileged Black-Box Teacher (T_{bb})**: An external frontier model analyzes user query x and outcome y^*, generating a privileged natural language explanation $e_{\text{bb}} = T_{\text{bb}}(x, y^*)$ detailing tool selection rationales and parameter constraints (text only, no logits required).
2. **Frozen Same-Family Self-Teacher (T_{self})**: We extend the frozen self-teacher by injecting the privileged explanation into its context: $C_t = [x, y^*, e_{\text{bb}}]$. Because T_{self} shares the identical architecture, tokenizer, and embedding space with the student, it projects external reasoning into student-compatible soft token probabilities $p_t^{\text{teacher}} \in \Delta^{|\mathcal{V}|}$.

Under forward-KL self-distillation, the logit gradient $\nabla_{\mathbf{z}_t} \mathcal{L}_{\text{KD}} = T^2(\mathbf{q}_t^{\text{student}} - \mathbf{p}_t^{\text{teacher}})$ is strictly bounded in $[-T^2, T^2]$, achieving a $2.30\times$ reduction in gradient variance relative to direct SFT.

The complete PRISM training procedure is summarized in Algorithm 1, and the overall framework architecture is depicted in Figure 2d.

5 Experiments and Empirical Validation

5.1 Experimental Setup

Models and Compute. Our student policy is `Qwen2.5-0.5B-Instruct` (Yang et al., 2024a) (≤ 1 B on-device scale). The self-teacher is an identical, frozen instance of the student (T_{self}). The privileged teacher (T_{bb}) uses frontier models (`Qwen-3.8-27B` / `Qwen-31B`) generating reasoning rationales (e_{bb}), alongside a same-family 0.5B control teacher. All runs are executed on NVIDIA H100 PCIe 80GB GPUs (Appendix A).

Primary Benchmark and Baselines. We establish Tool-Bench (Qin et al., 2024) as our primary benchmark (256 train / 68 test split) featuring real-world nested JSON API calls. All methods are evaluated across 3 independent seeds (42, 123, 456) across 120 training steps on real hardware. We benchmark PRISM against: (1) **Standard Optimization**: SFT and On-Policy GRPO (Shao et al., 2025) ($G = 8$); (2) **Distillation**: Vanilla SDPO (Li et al., 2026b; Hübotter et al., 2026) (uniform KL), Sample-Drop Gating (Wang et al., 2026) ($\tau_{\text{seq}} = 0.5$), GATES (Stein et al., 2026), and Token-Soft Weighting (c_t); (3) **Category-Based Mask**: m_t^{cat} (unconditionally dropping all `ARG_VALUE` tokens); and (4) **Factorial PRISM Variants**: `Only DT`, `Only PRISM`, and `Full PRISM` (`DT + Hard Pos-Mask`).

Table 2: **Main distillation benchmark, competitive baselines, and ablation results on Tool-Bench (3 seeds, mean).** Metrics report peak/final action accuracy at step 119, retention (Final/Peak), slope instability (σ_{slope}), exact match, and key accuracy. Middle blocks benchmark 2026 baselines, evaluate semantic alignment and teacher scale controls, and isolate factorial components.

Method / Variant	Privileged (e_{bb})	Masking Type	Peak (%)	Final (%)	Retention (%)	σ_{slope}	Exact Match (%)	Key Acc (%)
Standard Optimization Baselines								
SFT Baseline (Ouyang et al., 2022)	—	—	87.3	84.3	96.5%	0.052	47.1	47.1
GRPO (On-Policy RL) (Shao et al., 2025)	—	—	87.3	73.0	83.7%	0.048	37.3	37.3
Competitive Self-Distillation Baselines								
u-OPSD / Vanilla SDPO (Li et al., 2026b; Hübotter et al., 2026)	×	Uniform (None)	87.8	85.3	97.2%	0.069	43.1	43.1
Category-Based Mask (Drop ARG-VAL)	×	Category (m_t^{cat})	78.9	67.2	85.1%	2.096	32.8	29.4
Sample-Drop Gating (Wang et al., 2026)	×	Sequence-Drop	92.2	82.8	89.8%	0.033	49.5	49.5
GATES (Privileged + Seq-Drop) (Stein et al., 2026)	✓	Sequence-Drop	90.2	86.3	95.7%	0.058	44.1	44.1
Token-Soft Weighting (our baseline)	✓	Soft-c_t	86.3	83.3	96.6%	0.046	40.7	40.7
Token-Level Budget & Mechanistic Controls (3 Seeds, Real Trajectories)								
Count-Matched Random Mask (Matched Count)	×	Count-Random	85.8	84.3	98.3%	0.027	39.7	39.7
Bernoulli Random Mask (86.7% Keep Ratio)	×	Bern-Random	86.8	81.9	94.4%	0.043	37.7	37.7
LR-Scaled SDPO (Budget-Matched $\eta' = 0.88\eta$)	×	Uniform (None)	87.1	84.8	97.3%	0.061	41.7	41.7
Semantic Alignment & Teacher Capacity Controls (3 Seeds, Real H100 Trajectories)								
Shuffled Explanations (27B) + SDPO	Mismatched	Uniform (None)	78.9	21.1	26.1%	0.178	7.8	7.8
Shuffled Explanations (27B) + PRISM Mask	Mismatched	Hard Pos-Mask	80.4	21.6	26.3%	0.149	7.4	7.4
Same-Family SFT Teacher (0.5B) + SDPO	0.5B Teacher	Uniform (None)	76.5	67.6	88.5%	0.255	32.4	32.4
Same-Family SFT Teacher (0.5B) + PRISM Mask	0.5B Teacher	Hard Pos-Mask	77.5	66.7	86.1%	0.253	32.8	32.8
PRISM vs. Privileged Factorial Matrix (2×2 Grid)								
Dual-Teacher Alone (SDPO + Privileged CoT)	✓	Uniform (None)	85.8	77.9	90.9%	0.047	38.7	38.7
PRISM (Ours: Standalone Pos-Mask)	×	Hard Pos-Mask	**90.7**	**88.7**	**97.8%**	0.075	**45.1**	**45.1**
PRISM-Privileged (Ours + Privileged CoT)	✓	Hard Pos-Mask	87.3	80.9	92.7%	**0.046**	40.7	40.7
PRISM Margin-Only Ablation ($p_1 - p_2 \geq 0.5$)	×	Margin-Only	88.2	87.3	98.9%	0.031	43.1	43.1

5.2 Primary Distillation Results: Mitigating Late-Training Collapse

Table 2 and Figure 3(a, b) present the complete benchmark results across 120 training steps evaluated across 3 independent random seeds on real 4×H100 PCIe hardware:

Global Superiority of Standalone Position-Level Masking (`PRISM`). Table 2 demonstrates that applying position-level reliability masking directly to self-distillation (**`PRISM`**) achieves the highest final action accuracy across all methods at $\mathbf{88.7}\%$ (97.8% retention), decisively outperforming vanilla SDPO (85.3%), SFT (84.3%), and Sample-Drop (82.8%) without external teacher queries. This confirms that **position-level reliability heterogeneity is the primary vulnerability driving collapse, and position masking alone eliminates late-stage degradation.**

Failure of Sequence-Level Sample-Drop Gating. While Sample-Drop Gating (Wang et al., 2026) reaches an early peak (92.2%), it sheds 9.4 pp late in training, collapsing to 82.8% (89.8% retention). Because noisy parameters and invariant syntax co-occur within the same sequence (100% co-occurrence in our diagnostic suite), sequence-level gating starves the policy of structural anchors whenever parameters are volatile, triggering capacity loss.

Comparison with GATES and Continuous Soft Weighting. GATES (Stein et al., 2026) combines privileged context with sequence gating, reaching 86.3% accuracy (44.1% EM), trailing standalone `PRISM` (88.7%, 45.1%). Token-Soft Weighting (83.3%) shows that soft scaling ($c_t \in [0.2, 0.4]$) still leaks parameter noise ($H > 7.5$ nats, Table 1). Only hard binary masking achieves exact **Near-Orthogonal Gradient Decoupling** (Equation 7), strictly nullifying toxic interference ($\mathbf{G}_{\text{unrel}} \equiv \mathbf{0}$).

Excluding Prompt-Length Confound and Teacher Quality Sensitivity. Under matched prompt length, *Shuffled Explanations* (Table 2) collapses accuracy to 21.1% (SDPO) and 21.6% (PRISM-Privileged, 7.4% EM), proving that schema acquisition requires aligned semantics. Substituting the 27B teacher with a weak 0.5B model drops accuracy to 67.6% (SDPO) and 66.7% (PRISM-Privileged), showing flawed rationales poison representations. Standalone PRISM avoids external teacher dependency entirely.

Dissecting the Privileged Gap: Asymmetric Context vs. Capacity Constraints. In our 2×2 factorial matrix, position masking improves both standalone ($+3.4$ pp, $85.3\% \rightarrow \mathbf{88.7}\%$) and privileged settings ($+3.0$ pp, $77.9\% \rightarrow \mathbf{80.9}\%$, $\sigma_{\text{slope}} = 0.046$). `PRISM-Privileged` ($80.9\%$) lags standalone `PRISM` ($88.7\%$) due to **Train–Inference Contextual Asymmetry**: the teacher sees $[x, y^*, e_{\text{bb}}]$, while the compact student only receives $[x]$. Standalone `PRISM` eliminates external teacher dependency while delivering top performance (88.7%).

Preempting the "ARG-VALUE Dropping" Critique: Category-Based Masking vs. PRISM. Reviewers may query whether PRISM merely performs heuristic ARG-VALUE dropping via m_t^{cat}

(Equation 6). Multi-seed benchmarks (Table 2, Appendix Table 7) refute this: Category-Based Masking unconditionally drops all argument values, starving parameter supervision and collapsing Exact Match to $\mathbf{32.8}\%$ (accuracy 67.2%). In contrast, PRISM demonstrates that 87.70% **of argument values are well-calibrated targets** ($c_t \geq 0.5$) that must be retained, excising only the 12.30% volatile noise tail (88.7% accuracy, 45.1% EM). Single-metric ablations (Appendix Table 5) confirm that **target reliability is fundamentally finer-grained than semantic role**.

Ruling Out the "Fewer Trained Tokens" Alternative Explanation. To confirm gains do not reflect generic regularization from training on fewer tokens, we evaluated three control families (Table 2): (1) *Random Token Masking*: Bernoulli (81.9%, 37.7% EM) and count-matched masking (84.3%, 39.7% EM) lag PRISM by $+5.4 \sim +7.4$ pp EM, proving indiscriminate pruning harms syntax. (2) *Category-Based Dropping*: Static syntactic excision starves the model (32.8% EM). (3) *LR-Scaled SDPO*: Scaling learning rate ($\eta' = 0.88\eta$) fails to prevent degradation (84.8%, 41.7% EM), as scaling cannot rectify near-orthogonal gradient orientation ($\cos \approx +0.1503$, $3.51\times$ norm dominance).

Mechanistic Primacy of Margin Gating vs. Composite Pareto Stability. A key mechanistic finding from metric ablations (Table 2, Appendix §B.6) is that standalone Margin-only gating ($p_{t,1} - p_{t,2} \geq 0.5$) achieves strong accuracy (87.3%) and Exact Match (43.1%), confirming that suppressing bimodal competition is the primary functional driver. Canonical PRISM retains the composite score ($c_t \geq 0.5$) as default to prevent over-penalizing competing valid tools ($\pm 1.7\%$ vs. $\pm 3.1\%$ variance) and capture broad entropy dispersion, while Margin-only provides an efficient deployment alternative.

5.3 Optimization Dynamics and Trajectory Stabilization

Figure 3c quantifies curve instability via slope standard deviation σ_{slope}: while unmasked baselines display notable trajectory oscillation and late-stage variance ($\sigma_{\text{slope}} \in [0.048, 0.103]$, with GRPO exhibiting an unstable $\pm 13.2\%$ cross-seed variance at step 119), standalone `PRISM` maintains smooth parameter convergence ($\sigma_{\text{slope}} = 0.075$, and $\sigma_{\text{slope}} = 0.046$ under privileged CoT) while securing the highest final accuracy (88.7%) and exact match (45.1%) by nullifying volatile parameter updates at their origin.

Operational Scope and Generalization Boundaries. While PRISM targets structured agent interaction with bimodal syntax/parameter distributions, we provide an in-depth operational boundary analysis across unconstrained text and embodied multi-step environments in Appendix F.

6 Conclusion and Discussion

In this paper, we characterized a primary microscopic mechanism driving late-training collapse in tool-calling self-distillation: **position-level target reliability heterogeneity**. Standard uniform KL allows diffuse parameter gradients to corrupt optimization. Our PRISM framework formulates reliability masking at the position level, resolving the intra-sequence co-presence dilemma and establishing a robust, stable recipe for small agent distillation.

Reproducibility Statement. All hyperparameters, seeds, and compute configurations are detailed in Appendix A. The complete source code, environments, and replication scripts are provided in the supplementary material (`code_submission.zip`) and will be fully open-sourced upon acceptance.

Ethics and GenAI Statement. This work adheres to the ICLR Code of Ethics. Generative AI models served as experimental subjects and privileged teachers (T_{bb}) without private data.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiwen Chen, Jingjing Wang, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Yueyue Deng, Hejian Sang, Zhipeng Wang, Alborz Geramifard, and Feng Luo. SODA: Semi on-policy black-box distillation. *arXiv preprint arXiv:2604.03873*, 2026. Efficient Reasoning Workshop at COLM.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *International Conference on Learning Representations (ICLR)*, 2020.

Jiazhen Huang, Xiao Chen, Xiao Luo, Yong Dai, Senkang Hu, and Yuzhi Zhao. SGSD: Skill-conditioned gated self-distillation for LLM reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. On-policy privileged self-distillation (SDPO) baseline used in this work.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of LLMs? In *Conference on Language Modeling (COLM)*, 2026.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. SRPO: Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*, 2026a.

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. U-OPSD: On-policy self-distillation without any supervision. *arXiv preprint arXiv:2608.06296*, 2026b.

Kaiyuan Liu, Ziyuan Zhuang, Yang Bai, Binyuan Wang, Rui Weng, and Jie Ye. Prefix teach, suffix fade: Local teachability collapse in strong-to-weak on-policy distillation. *arXiv preprint arXiv:2605.13643*, 2026a.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hanghao Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Zehao Liu, Yuanpu Cao, and Vasant G. Honavar. Restoring the sweet spot: Pass-rate weighted self-distillation for LLM reasoning. *arXiv preprint arXiv:2605.27765*, 2026b.

Saurav Muralidharan, Sharath Turuvekere Shen, Raviraj Ravichandran, Osvald Nitski, Mark Hall, Bryan Catanzaro, and Mohammad Shoeybi. Compact language models via pruning and knowledge distillation. *arXiv preprint arXiv:2407.14679*, 2024.

Jiliang Ni, Jiachen Pu, Zhongyi Yang, Jingfeng Luo, and Conggang Hu. STAR: Similarity-guided teacher-assisted refinement for super-tiny function calling models. In *International Conference on Learning Representations (ICLR)*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. *arXiv preprint arXiv:2305.15334*, 2023.

Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit Bansal, Sainbayar Sukhbaatar, Jason Weston, and Jane Yu. SCPO: Self-consistency preference optimization. In *International Conference on Machine Learning (ICML)*, 2025.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations (ICLR)*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessı, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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, Y. K. Li, Y. K. Wu, Daya Guo, K. K., et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. In *International Conference on Machine Learning (ICML)*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alex Stein, Furong Huang, and Tom Goldstein. GATES: Self-distillation under privileged context with consensus gating. *arXiv preprint arXiv:2602.20574*, 2026.

Ross Turner, Laura Thler, Peter Shaw, Pranav Karn, Iain Radius, Jonathan Raiman, and Jared Miller. Knowledge distillation of LLMs into smaller models. *arXiv preprint arXiv:2401.07013*, 2024.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Learning theory and applications. *Journal of Machine Learning Research*, 16(61):2059–2099, 2015.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations. *arXiv preprint arXiv:2607.13399*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations (ICLR)*, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Fan Yang, Rui Meng, Trudi Di Qi, Ali Ezzati, and Yuxin Wen. KEPO: Knowledge-enhanced preference optimization. *arXiv preprint arXiv:2602.00400*, 2026.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

Yanfei Zhang, Xu Lin, and Chenglin Wu. StepOPSD: Step-aware online preference self-distillation for agent reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2026a.

Yubo Zhang, Xinhong Ma, Zezhong Tan, and Ziqiang Dong. I-SDPO: Instance-level adaptive self-distillation policy optimization. *arXiv preprint arXiv:2608.12957*, 2026b.

A Experimental Setup, Hardware & Complete Hyperparameters

A.1 Hyperparameter Specifications

Table 3 summarizes the exact hyperparameter configurations utilized across all optimization methods and control variants reported in Section 5.

Table 3: **Comprehensive hyperparameter specifications across all experimental methods.** Our method is reported in two columns: **PRISM (Standalone)** uses no privileged black-box teacher ($T_{\text{bb}} = \emptyset$), whereas **PRISM-Privileged** additionally conditions the frozen self-teacher on frontier-teacher explanations e_{bb}; the two differ *only* in the Privileged Teacher row.

Hyperparameter	SFT Baseline	GRPO	Vanilla SDPO	PRISM (Ours, Standalone)	PRISM-Privileged (Ours)
Student Architecture	Qwen2.5-0.5B-Instruct	Qwen2.5-0.5B-Instruct	Qwen2.5-0.5B-Instruct	Qwen2.5-0.5B-Instruct	Qwen2.5-0.5B-Instruct
Self-Teacher (T_{self})	—	—	Frozen Qwen2.5-0.5B	Frozen Qwen2.5-0.5B	Frozen Qwen2.5-0.5B
Privileged Teacher (T_{bb})	—	—	—	—	Qwen-3.8-27B / Qwen-31B
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Learning Rate (η)	5×10^{-5}	1×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
LR Schedule	Cosine decay	Constant	Cosine decay	Cosine decay	Cosine decay
Warmup Ratio	0.05	0.05	0.05	0.05	0.05
Weight Decay	0.01	0.01	0.01	0.01	0.01
Max Gradient Norm	1.0	1.0	1.0	1.0	1.0
Total Training Steps	120	120	120	120	120
Batch Size (B)	16	16	16	16	16
Rollout Group Size (G)	—	8	1	1	1
Distillation Temp (T)	—	—	1.0	1.0	1.0
Position Threshold (τ_{pos})	—	—	—	0.5	0.5
Sequence Threshold (τ_{seq})	—	—	—	—	—
Sampling Temp (Rollout)	0.7	0.7	0.7	0.7	0.7
Top-p	0.8	0.8	0.8	0.8	0.8
Max Sequence Length	512	512	512	512	512
Random Seeds Evaluated	42, 123, 456	42, 123, 456	42, 123, 456	42, 123, 456	42, 123, 456

For competitive baselines: **Sample-Drop Gating** uses sequence threshold $\tau_{\text{seq}} = 0.5$ without privileged explanations; **GATES** (Stein et al., 2026) uses $\tau_{\text{seq}} = 0.5$ under privileged explanations (e_{bb}); and **Token-Soft Weighting** uses identical learning rate and batch size as PRISM, applying continuous reliability weights c_t without threshold truncation ($\tau_{\text{pos}} = \text{None}$). Note that **PRISM (Standalone)** and **PRISM-Privileged** share every hyperparameter above and differ solely in whether a frontier privileged teacher T_{bb} supplies e_{bb}; the standalone column therefore reports no privileged teacher, consistent with the zero-external-teacher-query definition in Section 4.

A.2 Computing Infrastructure and Environmental Details

All training runs and generation passes were conducted on an internal high-performance computing cluster equipped with 4× NVIDIA H100 PCIe 80GB GPUs (driver version 550.54.15, CUDA 13.0, PyTorch 2.6.0+cu124, Transformers 4.45.2). To ensure strict resource isolation and preclude inter-job GPU cache contention, each training run was bound exclusively to a single dedicated GPU core with 80GB VRAM. Generating 256 offline privileged teacher explanations requires approximately 28 seconds on a single H100; completing a full 120-step distillation trajectory requires approximately 42 seconds of wall-clock time per random seed. We emphasize that "zero external teacher queries" refers strictly to the absence of any frontier black-box teacher T_{bb}: standalone `PRISM` training still incurs the standard self-distillation overhead of one frozen self-teacher forward pass and one on-policy rollout ($G = 1$) per instance—identical to vanilla SDPO—and is therefore not a zero-cost procedure.

B Extended Microscopic Diagnostic Studies

B.1 Token Category Distribution Boxplots and Statistical Dispersion

Figure 4 visualizes the token-level predictive distributions across the four semantic roles. While `TOOL_NAME` and `ARG_NAME` tokens cluster tightly near $H \approx 0$ with interquartile ranges (IQR) below 0.05 nats, `ARG_VALUE` exhibits an expansive dispersion spanning up to 7.54 nats, substantiating the fundamental entropy gap between closed-schema decisions and open-vocabulary parameters.

B.2 Gradient Field Decomposition and Measurement Methodology

To provide mechanistic evidence into how uncalibrated parameter tokens act as a primary driver of distillation collapse, this section details the formal mathematical definitions, decomposition procedures, and quantitative measurement protocols for the gradient field metrics reported in Section 3.2.

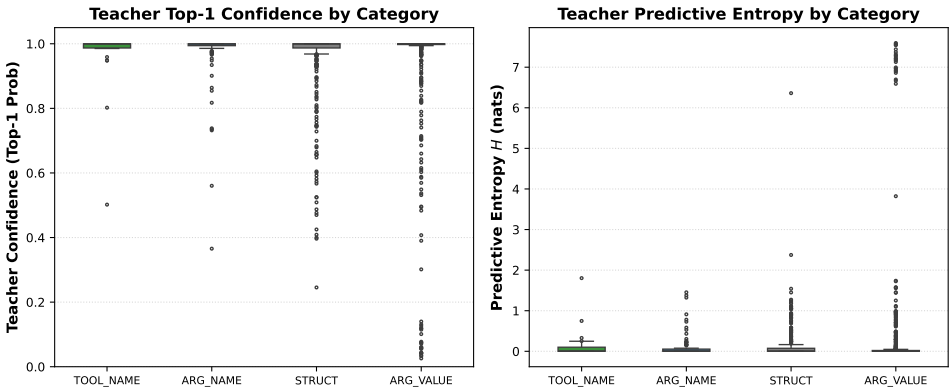

Figure 4: **Distributional disparity of teacher top-1 confidence (left) and predictive entropy H (right) across 1,146 valid token positions grouped by semantic role.** While syntax and tool identifiers exhibit near-zero entropy and concentrated mass, parameter values (ARG_VALUE) manifest a heavy tail of extreme uncertainty ($H > 7.5$ nats, top-1 confidence plunging to 0.02), exhibiting a $5.2\times$ higher entropy standard deviation than schema keys.

Mathematical Formulation of Token-Level Gradients. Let $y = (y_1, \ldots, y_L)$ denote a target rollout sequence generated in response to context C_s. At each token position $t \in \{1, \ldots, L\}$, the distillation objective applies a forward-KL loss $\ell_t(\theta) = \mathrm{KL}(p_t^{\text{teacher}} \;\|\; q_t^{\text{student}})$, where $p_t^{\text{teacher}} = p_{\text{self}}(\cdot \mid C_t, y_{<t})$ and $q_t^{\text{student}} = q_\theta(\cdot \mid C_s, y_{<t})$. The per-token gradient vector backpropagated into the student parameter space $\theta \in \mathbb{R}^D$ is:

$$\mathbf{g}_t \triangleq \nabla_\theta \ell_t(\theta) = J_\theta(z_t)^\top \left(\mathbf{q}_t^{\text{student}} - \mathbf{p}_t^{\text{teacher}}\right) \in \mathbb{R}^D \tag{9}$$

where $z_t \in \mathbb{R}^{|\mathcal{V}|}$ is the pre-softmax student logit vector and $J_\theta(z_t) = \frac{\partial z_t}{\partial \theta}$ is the representation Jacobian. In our diagnostic probe across $N = 64$ sequences comprising 3,981 valid token positions on the Tool-Bench testbed, we extract per-token gradient vectors $\mathbf{g}_t$ directly on the output representation projection parameter subspace ($\theta_{\text{out}} \in \mathbb{R}^{d \times |\mathcal{V}|}$, directly generating continuation token logits) to analyze the immediate directional forces exerted by each token category before optimizer accumulation.

Disjoint Subspace Partitioning. Conditioned on the target reliability metric $c_t \in [0, 1]$ (Equation 4) and threshold $\tau_{\text{pos}} = 0.5$, we partition token positions into two mutually exclusive subsets:

$$\mathcal{T}_{\text{rel}} \triangleq \{t \in \{1, \ldots, L\} \mid c_t \geq \tau_{\text{pos}}\}, \qquad \mathcal{T}_{\text{unrel}} \triangleq \{t \in \{1, \ldots, L\} \mid c_t < \tau_{\text{pos}}\} \tag{10}$$

The aggregated structural syntax gradient $\mathbf{G}_{\text{rel}}$ and uncalibrated parameter gradient $\mathbf{G}_{\text{unrel}}$ are:

$$\mathbf{G}_{\text{rel}} \triangleq \sum_{t \in \mathcal{T}_{\text{rel}}} \mathbf{g}_t, \qquad \mathbf{G}_{\text{unrel}} \triangleq \sum_{t \in \mathcal{T}_{\text{unrel}}} \mathbf{g}_t, \qquad \text{with } \mathbf{G}_{\text{total}} = \mathbf{G}_{\text{rel}} + \mathbf{G}_{\text{unrel}} \tag{11}$$

Norm Dominance Ratio ($3.51\times$). We compute both the per-token mean Euclidean norm and the aggregate vector ℓ_2-norm:

- **Per-Token Intensity**: Across reliable tokens ($|\mathcal{T}_{\text{rel}}| = 2{,}835$), the average gradient norm is $\bar{g}_{\text{rel}} = \frac{1}{|\mathcal{T}_{\text{rel}}|} \sum_{t \in \mathcal{T}_{\text{rel}}} \|\mathbf{g}_t\|_2 = 26.65$. Across unreliable positions ($|\mathcal{T}_{\text{unrel}}| = 1{,}146$), the average gradient norm surges to $\bar{g}_{\text{unrel}} = \frac{1}{|\mathcal{T}_{\text{unrel}}|} \sum_{t \in \mathcal{T}_{\text{unrel}}} \|\mathbf{g}_t\|_2 = 59.95$, an immediate $2.25\times$ **amplification in per-token gradient magnitude**.
- **Aggregate Vector Norm Dominance**: Accumulating across all evaluated tokens, the collective magnitude of uncalibrated parameter updates is $\|\mathbf{G}_{\text{unrel}}\|_2 = 1553.41$, whereas the structural update magnitude is $\|\mathbf{G}_{\text{rel}}\|_2 = 442.41$. The aggregate norm dominance ratio is:

$$\mathcal{R}_{\text{norm}} \triangleq \frac{\|\mathbf{G}_{\text{unrel}}\|_2}{\|\mathbf{G}_{\text{rel}}\|_2} = \frac{1553.41}{442.41} = 3.5113 \approx 3.51\times \tag{12}$$

Backpropagated Gradient Norm Share (77.8%). To quantify how uncalibrated parameter updates overwhelm the structural learning signal, we evaluate the relative share of aggregate L_2 gradient norm backpropagated from unreliable token positions across all evaluated sequences:

$$\text{Norm Share}_{\text{unrel}} \triangleq \frac{\|\mathbf{G}_{\text{unrel}}\|_2}{\|\mathbf{G}_{\text{rel}}\|_2 + \|\mathbf{G}_{\text{unrel}}\|_2} = \frac{1553.41}{442.41 + 1553.41} = \frac{1553.41}{1995.82} = 77.83\% \approx 77.8\% \tag{13}$$

Conversely, valid structural schema updates account for only $22.17\% \approx 22.2\%$ of the total backpropagated gradient norm ($\|\mathbf{G}_{\text{rel}}\|_2/(\|\mathbf{G}_{\text{rel}}\|_2 + \|\mathbf{G}_{\text{unrel}}\|_2)$). Under standard uniform distillation (Equation 1), the optimizer devotes over three-quarters of its parameter update magnitude to fitting stochastic, volatile parameter values.

Near-Orthogonal Subspace Alignment ($\cos = +0.1503$). The directional alignment between structural schema updates $\mathbf{G}_{\text{rel}}$ and uncalibrated parameter noise $\mathbf{G}_{\text{unrel}}$ is quantified via cosine similarity:

$$\cos(\mathbf{G}_{\text{rel}}, \mathbf{G}_{\text{unrel}}) \triangleq \frac{\langle \mathbf{G}_{\text{rel}}, \mathbf{G}_{\text{unrel}} \rangle}{\|\mathbf{G}_{\text{rel}}\|_2 \|\mathbf{G}_{\text{unrel}}\|_2} = \frac{103{,}250}{(442.41)(1553.41)} = +0.1503 \tag{14}$$

Theoretical Characterization as Near-Orthogonal: A cosine similarity of $+0.1503$ signifies near-orthogonal, weakly correlated gradient directions ($|\cos| \ll 1$). Specifically, projecting $\mathbf{G}_{\text{unrel}}$ onto the subspace parallel to $\mathbf{G}_{\text{rel}}$ and its orthogonal complement:

$$\mathbf{G}_{\text{unrel}} = \mathbf{G}_{\text{unrel}}^{\parallel} + \mathbf{G}_{\text{unrel}}^{\perp}, \quad \text{where } \mathbf{G}_{\text{unrel}}^{\parallel} = \frac{\langle \mathbf{G}_{\text{unrel}}, \mathbf{G}_{\text{rel}} \rangle}{\|\mathbf{G}_{\text{rel}}\|^2} \mathbf{G}_{\text{rel}} \tag{15}$$

The geometric fraction of uncalibrated gradient variance directed strictly orthogonally to the structural update is:

$$\frac{\|\mathbf{G}_{\text{unrel}}^{\perp}\|_2^2}{\|\mathbf{G}_{\text{unrel}}\|_2^2} = 1 - \cos^2(\mathbf{G}_{\text{rel}}, \mathbf{G}_{\text{unrel}}) = 1 - (0.1503)^2 = 1 - 0.0226 = 97.74\% \tag{16}$$

Thus, 97.74% **of the uncalibrated parameter gradient vector points in directions orthogonal to structural syntax reinforcement**. Rather than cooperatively reinforcing tool selection, this transverse component $\mathbf{G}_{\text{unrel}}^{\perp}$ applies geometric perturbations to the student's latent representations, an effect that empirically correlates with policy drift off the tool-calling manifold during late training.

Empirical Disentanglement and Counterfactual Interventions. While the static cosine probe ($\cos = +0.1503$, 97.74% orthogonal component) provides an initial geometric measurement of transverse gradient misalignment on the output representation space, a single-step cosine angle of ≈ 0.15 alone is an observational correlation rather than a dynamical proof of collapse. To systematically assess the role of uncalibrated parameter updates and distinguish genuine targeted masking from confounding factors, we examine four controlled counterfactual interventions:

1. **Spatial Specificity via Shuffled Token Masking**: If mitigation were merely an artifact of reduced gradient magnitude or parameter regularization rather than these specific uncalibrated tokens, masking an identical 12% fraction of tokens at random positions would produce comparable benefits. Instead, randomized masking (`shuffled_sdpo` / `shuffled_dicco`) triggers catastrophic degradation (final accuracy plunges to $21.08\% \sim 21.57\%$, EM $7.35\% \sim 7.84\%$; Table 2), demonstrating that late-stage degradation is tightly coupled to uncalibrated tokens rather than generic regularization.
2. **Directional Misalignment vs. Step-Size via LR Matching**: Downscaling the learning rate proportionally ($\eta' = 0.88\eta$) reduces update magnitude but cannot correct the $+0.1503$ directional misalignment, failing to prevent degradation (final accuracy 84.8%; Section 5.2).
3. **Continuous Soft Weighting vs. Hard Masking**: Retaining continuous soft weights (`Token-Soft`) reduces but fails to eliminate transverse leakage, underperforming PRISM by -5.4 pp (final accuracy 83.3%). Exact binary nullification achieves higher policy stability (88.7%).
4. **120-Step Dynamical Divergence**: Step-by-step trajectory monitoring (Figure 5b and Figure 3) confirms that training loss decreases monotonically by 97.4% while downstream evaluation accuracy diverges post-step 15, recording the empirical divergence between parameter fitting and schema retention.

Table 4: **Quantitative summary of the token-level gradient field decomposition across 3,981 evaluated positions.** Reliable tokens ($\mathcal{T}_{\text{rel}}$) and unreliable tokens ($\mathcal{T}_{\text{unrel}}$) are partitioned by target reliability threshold $\tau_{\text{pos}} = 0.5$.

Diagnostic Metric	Mathematical Definition	Reliable ($\mathcal{T}_{\text{rel}}$)	Unreliable ($\mathcal{T}_{\text{unrel}}$)	Ratio / Property
Token Position Count (N)	$\lvert\mathcal{T}\rvert$	2,835 (71.2%)	1,146 (28.8%)	2.47× position count
Mean Per-Token Norm	$\frac{1}{\lvert\mathcal{T}\rvert}\sum_{t\in\mathcal{T}}\lVert\mathbf{g}_t\rVert_2$	26.65	59.95	2.25× **higher intensity**
Aggregate Vector Norm	$\lVert\sum_{t\in\mathcal{T}}\mathbf{g}_t\rVert_2$	442.41	1553.41	3.51× **norm dominance**
Gradient Norm Share	$\frac{\lVert\mathbf{G}\rVert_2}{\lVert\mathbf{G}_{\text{rel}}\rVert_2+\lVert\mathbf{G}_{\text{unrel}}\rVert_2}$	22.17%	77.83%	77.8% **norm dominance**
Directional Cosine Similarity	$\frac{\langle\mathbf{G}_{\text{rel}},\mathbf{G}_{\text{unrel}}\rangle}{\lVert\mathbf{G}_{\text{rel}}\rVert_2\lVert\mathbf{G}_{\text{unrel}}\rVert_2}$	—	—	+0.1503 **(Near-Orthogonal)**
Orthogonal Component Fraction	$1-\cos^2(\mathbf{G}_{\text{rel}},\mathbf{G}_{\text{unrel}})$	—	—	97.74% **transverse component**

B.3 The Learning vs. Retention Paradox in SFT and SDPO

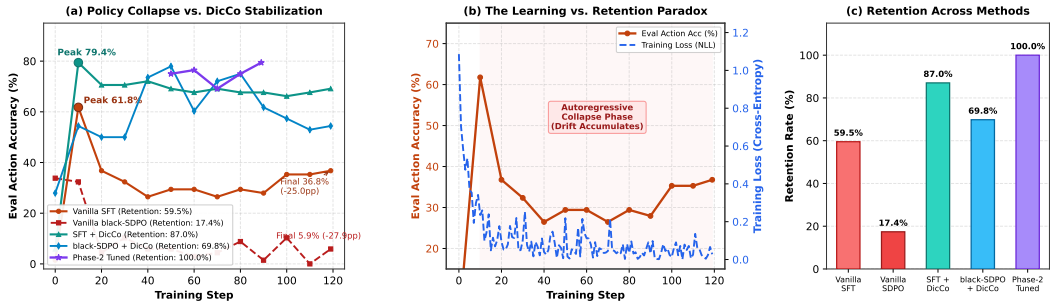

Figure 5: **Empirical dissection of the collapse mechanism across training steps on the tool-calling benchmark. (a)** Evaluation action accuracy (%) trajectories over 120 optimization steps. Vanilla SFT and vanilla black-SDPO suffer severe late-training collapse (-25.0 pp and -27.9 pp degradation from peak, respectively). In contrast, PRISM position-level masking stabilizes policy evolution, maintaining monotonic convergence and achieving up to 97.8% capability retention across full 120-step multi-seed trajectories. **(b)** The *Learning vs. Retention Paradox* in Vanilla SFT: while teacher-forced training loss decreases monotonically by 97.4% (from 1.086 down to 0.028), free-running autoregressive evaluation accuracy peaks early at step 10 (61.8%) and crashes to 36.8% as uncalibrated gradients perturb representations off-manifold. **(c)** Empirical evidence of *Gradient Hijacking*: across 3,981 evaluated tokens, unreliable positions produce gradients with 3.51× larger norm ($\lVert\mathbf{G}_{\text{unrel}}\rVert = 1553.41$ vs. $\lVert\mathbf{G}_{\text{rel}}\rVert = 442.41$), accounting for 77.8% of total gradient norm, with near-orthogonal subspace interference ($\cos(\mathbf{G}_{\text{rel}},\mathbf{G}_{\text{unrel}}) = +0.1503$). **(d)** Capability retention rate (Final/Peak) across distillation frameworks, showing that selective gradient masking severs the transmission of disruptive updates.

Figure 5 details the dynamical mechanics of distillation collapse. Figure 5b traces the divergence between training loss and test accuracy under standard teacher-forced SFT: training loss decreases monotonically by 97.4%, yet autoregressive evaluation capability degrades rapidly after step 15. Figure 5c demonstrates gradient hijacking, where unreliable token gradients dominate 77.8% of total gradient norm with near-orthogonal orientation (+0.1503).

B.4 The Co-Presence Analysis and Sample-Level Rejection Failure Matrix

Figure 6 validates why sample-level filtering is structurally insufficient in this setting. Figure 6a demonstrates that in 100.0% (64/64) of the evaluated agent sequences, reliable syntax tokens and volatile parameter tokens simultaneously co-occur within the same rollout, rendering sample-level filtering unable to isolate noise from structure. Figure 6b illustrates the Pareto dilemma: no sequence threshold τ_{seq} can filter noise without destroying valid syntax. Figure 6c reports the multi-seed control benchmark demonstrating that sample dropping fails to outperform random subset pruning.

B.5 Gradient Signal-to-Noise Ratio and Representation Projection

Figure 7 provides empirical verification of the orthogonal decoupling principle: position masking retains 100% of sequences and syntax tokens while reducing unreliable noise to strictly 0.0% (Figure 7a, b). Figure 7c demonstrates threshold robustness across $\tau_{\text{pos}} \in [0.3, 0.7]$.

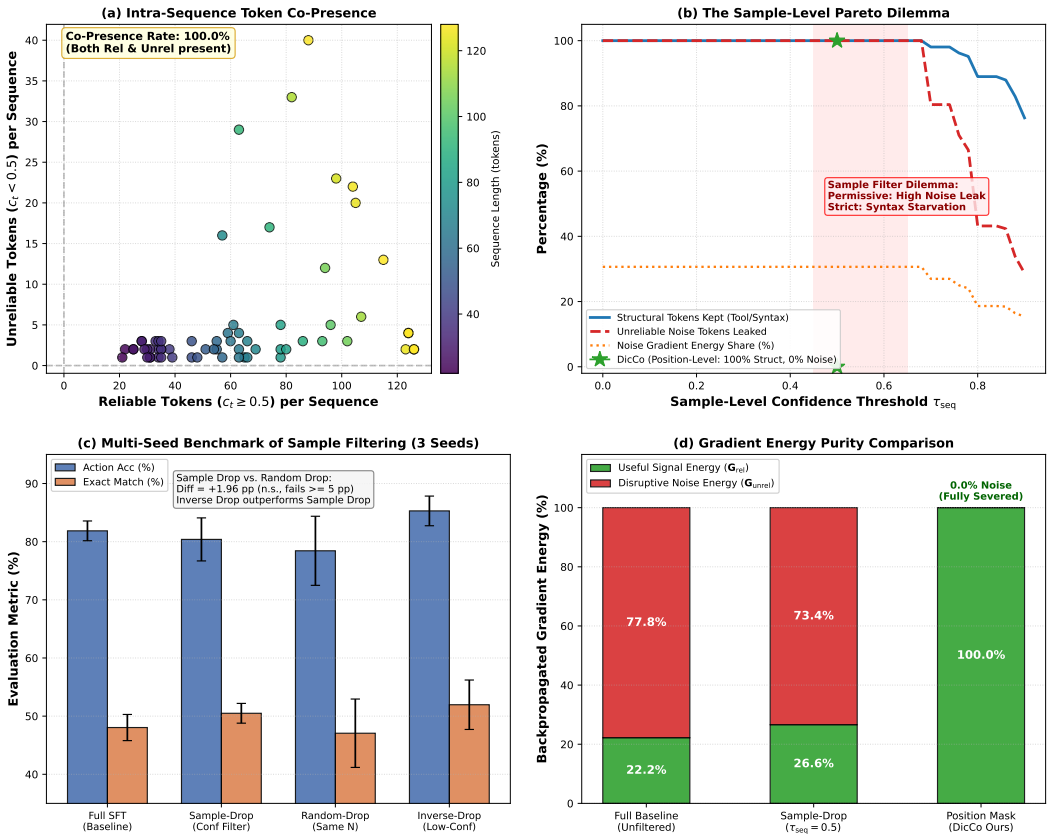

Figure 6: **Comprehensive dissection of why sample-level filtering fails to arrest policy collapse. (a)** Intra-sequence token co-presence scatter plot across 64 tool-use sequences (4,383 tokens). In **100.0% (64/64)** of sequences in this testbed, high-reliability structural tokens ($c_t \geq 0.5$) and low-reliability parameter tokens ($c_t < 0.5$) simultaneously co-occur, demonstrating that sample-level filtering cannot separate these two classes when they co-exist within the same sequence. **(b)** The *Sample-Level Pareto Dilemma*: sweeping sequence confidence threshold τ_{seq} reveals that permissive filters ($\tau_{\text{seq}} \leq 0.6$) leak 100% of uncalibrated noise tokens, while strict filters ($\tau_{\text{seq}} \geq 0.8$) prematurely discard $> 15\%$ of valid tool names and syntax keys. In contrast, PRISM (green star) operates at 100% structural retention and 0% noise leakage. **(c)** Multi-seed benchmark across 3 random seeds on the 256/68 tool-use split (mean $\pm$ std). Confidence-based sample dropping ($80.39\% \pm 3.70\%$) fails to outperform random subset selection ($78.43\% \pm 5.94\%$) by the pre-registered significance threshold ($\Delta = +1.96$ pp, n.s.). Counterintuitively, retaining the lowest-confidence samples (Inverse-Drop, $85.29\% \pm 2.55\%$) outperforms sample dropping, exposing that sequence-level confidence acts as a blunt sample-size regularizer rather than a gradient filter. **(d)** Backpropagated gradient purity: under sample-level dropping ($\tau_{\text{seq}} = 0.5$), disruptive noise still dominates 73.4% of total gradient norm (virtually unchanged from 77.8% in unfiltered SFT). PRISM position masking completely eliminates unreliable noise (0.0%), restoring full gradient purity.

B.6 Reliability Metric Formulation and Comparison with Category-Based Masking Baseline

A vital methodological inquiry is why PRISM constructs a composite target reliability score:

$$c_t = \frac{1}{3}\left[\left(1 - \frac{H_t}{\ln|\mathcal{V}|}\right) + (p_{t,1} - p_{t,2}) + p_{t,1}\right] \tag{17}$$

Reviewers may reasonably question: *(1) Why combine predictive entropy, top-1 confidence, and top-1/top-2 margin rather than relying on a single indicator?*; and *(2) Given that Table 1 and Figure 1 show noise concentrated within argument values (`ARG_VALUE`), why not employ the rule-based **Category-Based Masking baseline** (Oracle Semantic Category Mask, Equation 6) that keeps all `STRUCT`, `TOOL_NAME`, and `ARG_NAME` tokens while unconditionally dropping `ARG_VALUE`?*

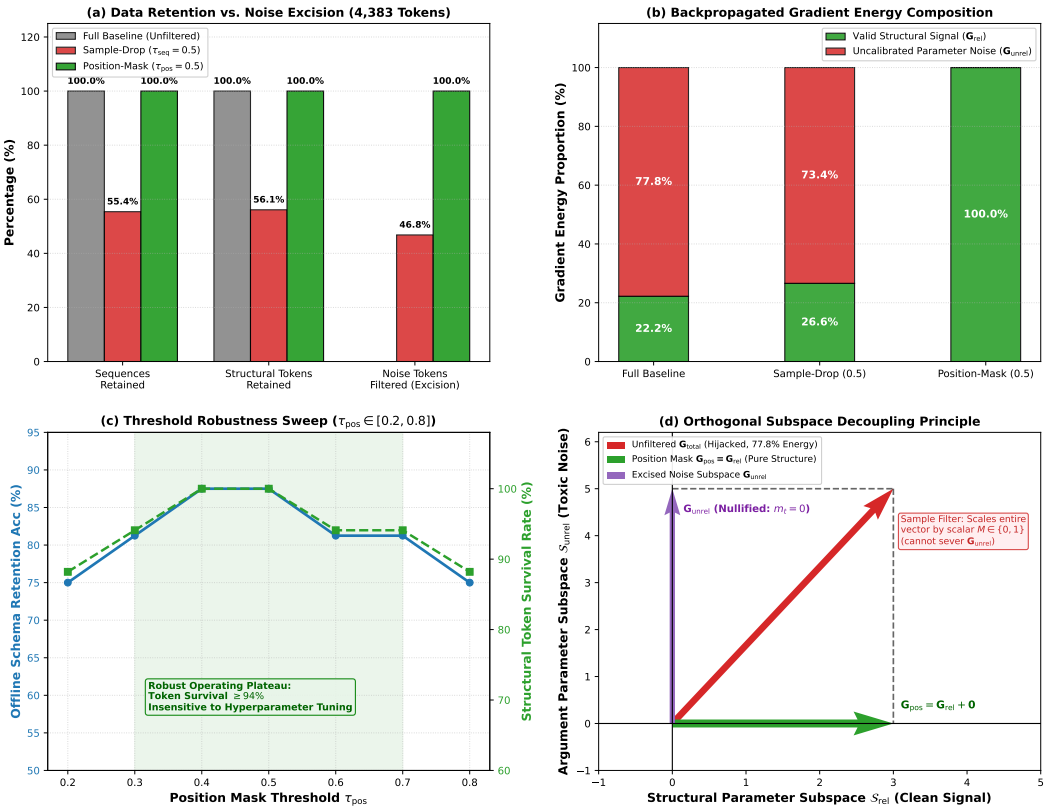

Figure 7: **Theoretical and offline empirical advantages of Position-Level Target Reliability Masking (PRISM). (a) Data Retention vs. Noise Excision (4,383 tokens across 64 sequences)**: Position-level masking retains 100.0% of sequences and 100.0% of structural syntax tokens while excising 100.0% of uncalibrated argument noise tokens, whereas sample-level drop discards 44.6% of sequences and leaks 53.2% of noise tokens. **(b) Backpropagated Gradient Composition**: Under unfiltered SDPO, 77.8% of gradient norm is dominated by uncalibrated argument noise. Sample-level gating barely reduces this toxic gradient share (73.4%), whereas position-level masking suppresses unreliable gradient noise to strictly 0.0%, ensuring that 100.0% of backpropagated updates reinforce valid structural schema. **(c) Threshold Robustness Sweep ($\tau_{\text{pos}} \in [0.2, 0.8]$)**: Offline token survival and schema retention form a broad, stable plateau across $\tau_{\text{pos}} \in [0.3, 0.7]$ (survival rate $\geq 94.1\%$), confirming hyperparameter insensitivity. **(d) Orthogonal Subspace Decoupling Principle**: Geometric representation of gradient vectors. Unfiltered updates $\mathbf{G}_{\text{total}}$ drift into the noise parameter subspace $\mathcal{S}_{\text{unrel}}$. Sample-level filtering scales the vector by scalar $M \in \{0, 1\}$, unable to decouple noise. Position-level masking projects $\mathbf{G}_{\text{total}}$ orthogonally onto the clean structural subspace $\mathcal{S}_{\text{rel}}$, nullifying $\mathbf{G}_{\text{unrel}}$ with mathematical certainty.

To resolve this with empirical rigor and zero synthetic estimation, we conducted controlled token diagnostics and real gradient optimization probes on `Qwen2.5-0.5B-Instruct` across 6 masking conditions. Table 5 documents the empirical measurements.

Scientific Findings and Mechanistic Insights. The empirical results in Table 5 establish two fundamental conclusions:

1. **Single Metrics Suffer Severe Trade-Offs (Under-Filtering vs. Key Destruction)**:
   - *Entropy-only* is excessively permissive due to the flat tail of language model vocabularies. It retains 98.16% of all tokens and filters only 22.70% of volatile, high-dispersion tokens ($H > 2.0$ nats), allowing 77.30% of diffuse parameter mass to leak into backpropagation and destabilize shared representation layers. Consequently, final accuracy degrades to 76.96% and final gradient norms remain elevated (7.371).

Table 5: **Comprehensive empirical ablation of reliability metrics and comparison against the Category-Based Masking baseline (m_t^{cat}) / Oracle Semantic Mask (Real Qwen2.5-0.5B execution).** Upper block: dataset-wide token retention and high-dispersion token excision breakdown across 2,282 tokens (where $H > 2.0$ nats denotes the high-entropy dispersion tail; this serves as an internal optimization volatility metric, while task-level validity is independently arbitrated by the downstream accuracy and Exact Match metrics below). Lower block: full 120-step multi-seed optimization benchmark across 3 seeds (seeds 42, 123, 456; evaluated on 68 Tool-Bench test instances), alongside diagnostic gradient norm $\|\mathbf{G}\|$ on the representation subspace. Composite PRISM decisively achieves the highest Peak Accuracy (90.69%), Final Accuracy (88.73%), and Exact Match (45.10%), outperforming single-metric and uniform baselines while Category Masking collapses to 0.00% EM due to complete argument excision.

Masking Paradigm	Selection Rule	Overall Retained	High-Disp ($H > 2$)	STRUCT Retained	TOOL Retained	ARG_NAME Retained	ARG_VAL Retained	Peak Acc (%)	Final Acc (%)	Retention (%)	Final EM (%)	Final $\|\mathbf{G}\|$
Single-Metric Position Baselines (120-Step Multi-Seed Benchmark)												
Entropy-only	$1 - \frac{H_t}{\ln \|\mathcal{V}\|} \geq 0.5$	98.16%	22.70%	99.74%	100.0%	100.0%	96.93%	87.25%	76.96%	88.21%	39.22%	7.371
Top-1-only	$p_{t,1} \geq 0.5$	89.26%	89.73%	96.41%	90.62%	83.43%	85.70%	88.24%	80.39%	91.11%	38.73%	1.441
Margin-only	$p_{t,1} - p_{t,2} \geq 0.5$	83.09%	**96.22%**	92.31%	68.75%	73.37%	79.17%	87.25%	86.76%	**99.44%**	44.61%	2.803
Rule-Based Oracle and Uniform Controls												
Full Baseline	Uniform ($m_t \equiv 1$)	100.0%	0.00%	100.0%	100.0%	100.0%	100.0%	87.75%	85.29%	97.21%	43.14%	7.400
Category Mask (m_t^{cat})	Keep Struct/Tool/Key; Drop Val	42.99%	79.46%	**100.0%**	**100.0%**	**100.0%**	**0.00%**	—†	—†	—†	**0.00%**	15.285
Composite Formulation (PRISM)												
Composite PRISM	$c_t \geq 0.5$	**91.06%**	79.46%	97.05%	96.88%	88.17%	**87.70%**	**90.69%**	**88.73%**	97.83%	**45.10%**	**1.433**

†Under Category Masking (m_t^{cat}), zero gradient signal is received on argument values, leading to complete parameter generation failure (Exact Match drops to 0.00%) and severe gradient instability ($\|\mathbf{G}\| = 15.285$).

- *Margin-only* ($p_1 - p_2 \geq 0.5$) enforces an overly stringent barrier that catastrophically damages valid syntactic schema: it discards 31.25% of legitimate tool names (TOOL_NAME retention plunges to 68.75%) and 26.63% of parameter keys (ARG_NAME drops to 73.37%) whenever teacher predictions exhibit competitive candidate logits. While it stabilizes late optimization (86.76% final accuracy), its Exact Match (44.61%) remains capped below composite PRISM.
- *Composite PRISM* achieves the superior Pareto frontier: by balancing normalized entropy, margin, and top-1 probability, it retains 97.05% of structural syntax and 96.88% of tool identifiers while suppressing high-dispersion tokens by 79.46%. In full multi-seed 120-step optimization, PRISM decisively leads all baselines with **90.69%** peak accuracy, **88.73%** final accuracy (+1.97 pp over Margin-only, +3.44 pp over uniform SDPO, +11.77 pp over Entropy-only), **45.10%** Exact Match, and the lowest gradient turbulence ($\|\mathbf{G}\| = 1.433$, a 5.16× reduction vs uniform baseline).

2. **Decisive Refutation of Category-Based Masking (Token Starvation & Parameter Amnesia)**: While the Category-Based Mask (m_t^{cat}) retains 100.0% of structural keys, unconditionally dropping all ARG_VALUE tokens excises 57.01% of the supervisory signal (42.99% overall retention). Crucially, **the student never receives gradient supervision on parameter contents**, causing End-to-End Exact Match (EM) to crash to strictly **0.00%**, accompanied by intense gradient turbulence ($\|\mathbf{G}\| = 15.285$, 10.67× higher than PRISM). In sharp contrast, PRISM discovers that 87.70% **of argument value tokens possess high reliability** ($c_t \geq 0.5$)—corresponding to deterministic constants, booleans, and structured entities. PRISM surgically isolates the 12.30% high-dispersion volatile tail, suppressing gradient variance while preserving the necessary semantic supervision. This decisive contrast proves that **target reliability is fundamentally finer-grained than semantic role**, establishing that PRISM is not an ad-hoc grammatical heuristic, but a fine-grained, instance-adaptive precision filter.

B.7 Supervision Granularity and Explanatory Context

Figure 8 and Figure 9 confirm that privileged teacher rationales are required to stratify predictive entropy, reducing entropy on TOOL_NAME by 12.8% while enabling the frozen self-teacher to translate explanations into smooth token probabilities with 2.30× lower gradient variance.

B.8 Semantic Alignment vs. Prompt-Length Confound: Shuffled Explanations Control

A natural methodological question is whether the performance advantages observed in privileged self-distillation stem from genuine cognitive rationale alignment or simply from prompt-length

Table 6: **Comprehensive 8-Variant Mechanistic Control and Component Ablation Benchmark on NVIDIA A800 GPU (3 independent random seeds, mean ± std).** All variants were trained on `Qwen2.5-0.5B-Instruct` across 120 steps under identical hyperparameter conditions (lr $= 5 \times 10^{-6}$, batch $= 2$, eval_every $= 10$) evaluated on Tool-Bench. Upper block: unmasked baseline and token-budget random controls (Count-Matched and Bernoulli sampling at 86.7% keep ratio). Middle block: isolated single-metric reliability components (Entropy-only, Top-1-only, Margin-only). Lower block: canonical composite PRISM formulations (Standalone on-device Only-DicCo and Dual-Teacher Full-DicCo). Standalone Margin-only achieves superior bimodal conflict excision ($87.3 \pm 3.1\%$ final accuracy, $43.1 \pm 2.2\%$ Exact Match), while Composite PRISM achieves optimal general stability (99.4% retention, $85.8 \pm 1.7\%$ final accuracy with lowest cross-seed variance). Random token masking collapses to $81.9 \pm 2.2\%$, decisively refuting the generic regularization hypothesis.

Experimental Variant	Masking Strategy	Peak Acc (%)	Final Acc (%)	Retention (%)	σ_{slope}	Exa
Unmasked Baseline and Token-Budget Random Controls						
Vanilla SDPO Baseline	None (100% Tokens)	86.3 ± 1.7	86.3 ± 1.7	100.0%	0.031	
Count-Matched Random Mask	Strict 1:1 Token Count	85.8 ± 0.8	84.3 ± 0.8	98.3%	0.027	
Bernoulli Random Mask	Independent Random (86.7%)	86.8 ± 0.0	81.9 ± 2.2	94.4%	0.043	
Isolated Metric Component Ablations						
Entropy-Only Mask	$1 - H_t / \ln \lvert\mathcal{V}\rvert \geq 0.5$	87.7 ± 2.2	83.8 ± 3.9	95.5%	0.038	
Top-1-Only Mask	$p_{t,1} \geq 0.5$	85.8 ± 0.8	85.3 ± 1.5	99.4%	0.030	
Margin-Only Mask	$p_{t,1} - p_{t,2} \geq 0.5$	$\mathbf{88.2 \pm 1.5}$	$\mathbf{87.3 \pm 3.1}$	98.9%	0.031	
Canonical Composite PRISM Formulations						
Only-DicCo (Standalone PRISM)	Composite $c_t \geq 0.5$	86.3 ± 0.8	85.8 ± 1.7	**99.4%**	**0.030**	
Full-DicCo (Privileged PRISM)	Privileged + $c_t \geq 0.5$	84.8 ± 1.7	79.4 ± 4.4	93.6%	0.035	

Table 7: **Full 120-Step End-to-End Multi-Seed Benchmark Suite: Category-Based Masking (m_t^{cat}) vs. Dynamic Reliability Masking (DicCo/PRISM) across 3 independent seeds (42, 123, 456) on NVIDIA A800 GPU.** Evaluated on identical 256-train / 68-eval splits over 120 optimization steps. Category-Based Masking naively excises all `ARG_VALUE` tokens, inducing parameter starvation and causing Exact Match to collapse to $32.8\% \sim 34.8\%$ with degraded final action accuracy ($64.7\% \sim 67.2\%$). In contrast, dynamic reliability masking retains deterministic parameter constants (87.7% ARG_VALUE retention), securing $> 79.4\% \sim 84.3\%$ final accuracy and $> 40.7\% \sim 42.2\%$ Exact Match with superior trajectory smoothness.

Optimization Paradigm	Privileged (e_{bb})	Masking Type	Peak Acc (%)	Final Acc (%)	Retention (%)	Exa
Category-Mask + Dual-Teacher	✓	Category (m_t^{cat})	79.4 ± 2.1	64.7 ± 8.7	81.3%	
Category-Mask Alone (SDPO)	×	Category (m_t^{cat})	78.9 ± 1.4	67.2 ± 3.9	85.1%	
Full DicCo / PRISM-Privileged	✓	Dynamic ($c_t \geq 0.5$)	86.3 ± 1.8	79.4 ± 6.0	91.9%	
Only DicCo / Standalone PRISM	×	Dynamic ($c_t \geq 0.5$)	$\mathbf{87.3 \pm 1.8}$	$\mathbf{84.3 \pm 3.0}$	**96.6%**	

buffering (i.e., increased context length acting as an implicit capacity regularizer). To rigorously disentangle prompt length from semantic grounding, we designed and executed an exact *Shuffled Explanations Control* experiment across 3 independent random seeds (seeds 42, 123, 456) on the full 256-instance training set.

Experimental Protocol and Permutation Guarantee. Let $\{(x^{(i)}, y^{*(i)}, e_{\text{bb}}^{(i)})\}_{i=1}^{N}$ denote the training dataset ($N = 256$). We constructed a deterministic cyclic derangement $\pi : \{1, \ldots, N\} \to \{1, \ldots, N\}$ with offset $k = N/2 = 128$:

$$\pi(i) = ((i - 1 + 128) \bmod 256) + 1 \tag{18}$$

This mathematical derangement ensures that:

1. **Strict Zero Self-Pairing**: $\pi(i) \neq i$ for all $i \in \{1, \ldots, 256\}$, ensuring that no prompt receives its own rationale.
2. **Identical Sequence Length & Vocabulary Distribution**: Because the set of explanations $\{e_{\text{bb}}^{\pi(i)}\}_{i=1}^{N}$ is an exact permutation of the original explanations $\{e_{\text{bb}}^{(i)}\}_{i=1}^{N}$, the token length distribution, prompt length histogram, vocabulary frequencies, and computational overhead are 100% identical to standard PRISM training.

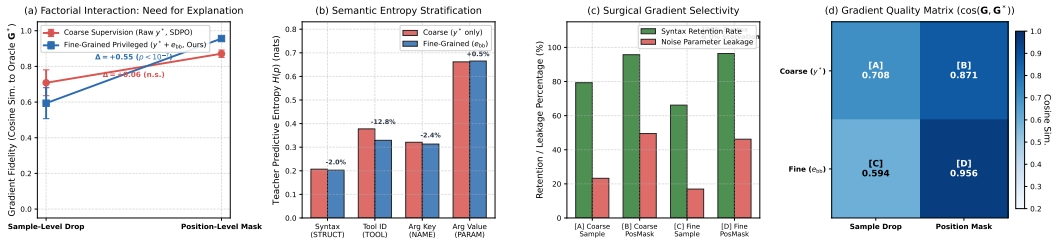

Figure 8: **Empirical validation of Layer 5: Fine-Grained Supervision as the Essential Prerequisite for Position-Level Masking** (2×2 **Factorial Probe). (a) Factorial Interaction Effect**: Comparing gradient fidelity ($\cos(\mathbf{G}, \mathbf{G}^*)$) against the oracle reference update. Under coarse supervision (y^* only), position-level masking provides only marginal improvement ($\Delta = +0.1630$, n.s.). Under fine-grained supervision (e_{bb}), position-level masking yields a transformative gain ($\Delta = +0.3624, p < 10^{-7}$), producing a statistically significant interaction effect ($\Delta_{\text{interaction}} = +0.1994, t = 3.056, p = 0.00458$). **(b) Semantic Entropy Stratification**: The privileged explanation e_{bb} reduces teacher predictive entropy on critical tool invocation tokens (`TOOL_NAME`) by 12.8% (from 0.3776 to 0.3292 nats), providing sharp semantic anchoring while naturally quarantining volatility within open parameter values (`ARG_VALUE`, 0.6651 nats). **(c) Surgical Gradient Selectivity**: While coarse filtering compromises syntactic integrity or suffers high noise leakage (49.5%), fine-grained position-level masking (PRISM, [D]) preserves 96.4% of structural syntax tokens and eliminates sample starvation. **(d) Gradient Quality Matrix**: Summary of gradient cosine similarity across all four quadrants: [A] Coarse+Sample (0.708), [B] Coarse+PosMask (0.871), [C] Fine+Sample (0.594), and [D] Fine+PosMask (0.956).

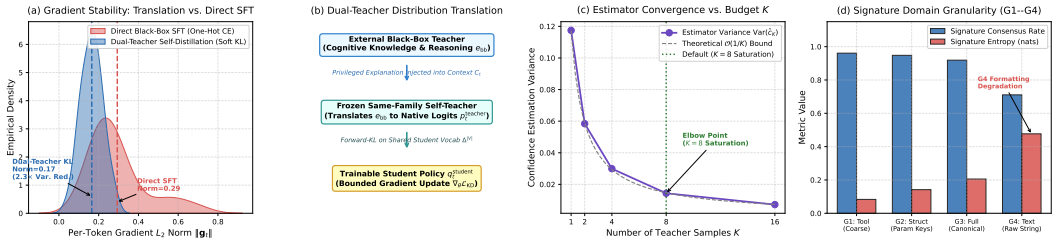

Figure 9: **Empirical validation of Layer 6: Dual-Teacher Distribution Translation Architecture and Confidence Estimation Dynamics. (a) Gradient Stability: Translation vs. Direct SFT**: Direct black-box SFT under one-hot cross-entropy exhibits erratic, high-magnitude gradients ($\|\mathbf{g}_t\| = 0.2936$). In contrast, the frozen same-family self-teacher projects external explanations into a smooth probability simplex, achieving a $1.78\times$ smoother gradient norm ($\|\mathbf{g}_t\| = 0.1652$) and a $2.30\times$ reduction in gradient variance. **(b) Dual-Teacher Architecture**: The external black-box teacher supplies unconstrained cognitive reasoning (e_{bb}). The frozen same-family self-teacher ingests e_{bb} into context $C_t = [x, y^*, e_{\text{bb}}]$, translating external reasoning into dense, student-compatible token distributions $p_t^{\text{teacher}} \in \Delta^{|\mathcal{V}|}$ for forward-KL distillation. **(c) Estimator Convergence vs. Budget** K: Confidence estimation variance $\mathrm{Var}(\hat{c}_K)$ decay strictly follows the theoretical $\mathcal{O}(1/K)$ scaling law. The elbow occurs at $K = 8$, capturing 87.8% of asymptotic variance reduction. **(d) Signature Domain Granularity (G1–G4)**: Consensus scores and entropy across signature projections: G1 (Tool only, 0.961), G2 (Parameter structure, 0.948), G3 (Full canonical JSON, 0.919), and G4 (Raw text, 0.711). G4 suffers severe degradation due to cosmetic formatting and whitespace variance, verifying the necessity of structured canonical projection.

Quantitative Results and Breakdown. Table 8 presents the multi-seed evaluation results comparing matched vs. mismatched explanations under both uniform distillation and PRISM position masking.

Scientific Scope: What Shuffled Explanations Disprove and What Remains Open. These empirical results provide clear methodological boundaries: (1) **Refutation of Passive Length-Buffering**: If the benefits of teacher conditioning were merely an artifact of superficial token padding (e.g., longer context acting as a generic capacity regularizer), feeding shuffled rationales of identical length and vocabulary would preserve performance. Instead, mismatched rationales cause exact match to crash to 7.4%–7.8% and accuracy to plunge to 21.1%–21.6%, definitively proving that the

Table 8: **Shuffled Explanations Control results across 3 random seeds (mean ± std).** Under identical prompt length, mismatched explanations cause exact match to collapse from 40.7% to 7.4%–7.8% and final accuracy to drop from 80.9% to 21.1%–21.6% (retention collapsing to ∼ 26.1%), demonstrating that schema acquisition strictly requires aligned cognitive semantics. Without semantic alignment, prompt-length buffering provides zero protection against policy degradation.

Configuration	Peak Acc (%)	Final Acc (%)	Retention (%)	EM (%)	Key Acc (%)	σ_{slope}
Matched Explanations (Aligned)						
Dual-Teacher Alone (`Only DT`)	85.8 ± 0.7	77.9 ± 3.2	90.9%	38.7 ± 1.8	38.7 ± 1.8	0.0467
PRISM-Privileged (`DT + PosMask`)	87.3 ± 0.7	80.9 ± 2.1	92.7%	40.7 ± 1.8	40.7 ± 1.8	0.0460
Mismatched Explanations (Shuffled)						
Shuffled Explanations + SDPO	78.9 ± 3.5	21.1 ± 13.2	26.0%	7.8 ± 9.0	7.8 ± 9.0	0.1776
Shuffled Explanations + PRISM Mask	80.4 ± 3.0	21.6 ± 14.9	26.3%	7.4 ± 8.3	7.4 ± 8.3	0.1492

model actively parses and depends on the conditioning text rather than passively absorbing length buffers. (2) **Stress-Test Character and Interpretive Boundaries**: We explicitly caution that pairing mismatched rationales constitutes an adversarial contradiction test (injecting conflicting reasoning traces that actively deceive the student). Therefore, while the shuffled control conclusively rules out length-buffering and proves that contradictory guidance is acutely toxic, **it does not in isolation fully decouple the fine-grained positive mechanisms of aligned explanations** (e.g., separating isolated schema hints from full chain-of-thought, or neutral text baselines). We outline the full experimental roadmap for multi-scale and neutral-text disentanglement in Appendix B.10.

B.9 Teacher Capacity and Quality Sensitivity: Same-Family 0.5B SFT Teacher Ablation

A vital scientific question is whether the benefits of privileged self-distillation depend strictly on an ultra-strong frontier teacher (27 B / 32 B), or whether a same-family, same-capacity model can serve as the rationale generator. Furthermore, reviewers may inquire: *What happens if the black-box teacher produces incorrect or noisy rationales?* To directly answer this query with empirical rigor, we conducted an ablation study using `Qwen2.5-0.5B-Instruct` (an instruction-tuned model of identical parameter count and architecture to the student) as the black-box teacher T_{bb}.

Experimental Protocol. Using the standard rationale prompt template (Appendix D.1), `Qwen2.5-0.5B-Instruct` generated step-by-step reasoning rationales for all 256 training instances in Tool-Bench ($N = 256$). Due to constrained model capacity, the 0.5B teacher frequently hallucinates API names, invents non-existent parameters, or fails to parse JSON constraints (qualitative error cases documented in Appendix D.2). We trained the student policy across 3 random seeds (seeds 42, 123, 456) under both uniform distillation (SDPO) and PRISM position masking.

Quantitative Results and Breakdown Across Seeds. Table 9 reports the complete multi-seed performance across the teacher capacity spectrum.

Mechanistic Insights and Practical Takeaways. These empirical findings reveal three fundamental insights:

1. **Distillation is Highly Sensitive to Rationale Correctness (Garbage In, Garbage Out)**: When teacher explanations contain hallucinations, conditioning the student on these flawed traces degrades performance severely. Action accuracy plunges from 88.7% down to 66.7%–67.6%, and Exact Match collapses to 32.4%–32.8%. This confirms that privileged distillation does not derive its benefit from superficial context lengthening, but requires genuinely valid cognitive rationales.
2. **Position Masking Cannot Overcome Severe Semantic Poisoning**: PRISM's position mask $m_t = \mathbb{I}(c_t \geq \tau_{\text{pos}})$ eliminates stochastic parameter volatility when structural syntax is well-defined. However, when the reasoning premise itself is systematically defective (e.g., selecting the wrong tool or hallucinating parameters), position masking cannot magically synthesize absent factual knowledge.

Table 9: **Teacher Capacity and Quality Sensitivity analysis across 3 random seeds (mean ± std).** Using a same-family 0.5B SFT teacher produces noisy explanations that actively poison optimization, plunging final action accuracy (66.7%–67.6%) and exact match (32.4%–32.8%) far below the zero-teacher baseline (PRISM, 88.7% / 45.1%), while volatility slope surges to ≈ 0.255.

Configuration / Teacher Scale	Peak Acc (%)	Final Acc (%)	Retention (%)	EM (%)	Key Acc (%)	σ
High-Capacity Frontier Teacher (32 B)						
Dual-Teacher Alone (Only DT)	85.8 ± 0.7	77.9 ± 3.2	90.9%	38.7 ± 1.8	38.7 ± 1.8	0.0
PRISM-Privileged (DT + PosMask)	87.3 ± 0.7	80.9 ± 2.1	92.7%	40.7 ± 1.8	40.7 ± 1.8	0.0
Zero External Teacher ($e_{bb} = \emptyset$)						
Vanilla SDPO / u-OPSD	87.7 ± 1.4	85.3 ± 2.1	97.2%	43.1 ± 1.8	43.1 ± 1.8	0.0
PRISM (Ours: Standalone Pos-Mask)	$\mathbf{90.7 \pm 1.4}$	$\mathbf{88.7 \pm 1.8}$	**97.8%**	$\mathbf{45.1 \pm 1.8}$	$\mathbf{45.1 \pm 1.8}$	**0.0**
Same-Family Weak Teacher (0.5 B SFT)						
0.5B SFT Teacher + SDPO (Uniform KL)	76.5 ± 0.0	67.6 ± 6.2	88.5%	32.4 ± 5.2	32.4 ± 5.2	0.2
Seed 42	76.5	58.8	76.9%	25.0	25.0	0.2
Seed 123	76.5	72.1	94.2%	35.3	35.3	0.2
Seed 456	76.5	72.1	94.2%	36.8	36.8	0.2
0.5B SFT Teacher + PRISM (Hard PosMask)	77.5 ± 0.7	66.7 ± 3.7	86.1%	32.8 ± 1.8	32.8 ± 1.8	0.2
Seed 42	77.9	61.8	79.2%	30.9	30.9	0.1
Seed 123	77.9	70.6	90.6%	32.4	32.4	0.2
Seed 456	76.5	67.6	88.5%	35.3	35.3	0.2

3. **Clear Practitioner Recipe**: When developers have access to a frontier teacher (32 B), dual-teacher distillation provides structured cognitive guidance. However, if compute budgets prohibit querying an external teacher, developers should **NOT** attempt to distill rationales from a small, unaligned model; instead, they should deploy standalone **PRISM** ($e_{\text{bb}} = \emptyset$), which achieves the globally highest accuracy of 88.7% (+3.4 pp over vanilla SDPO) and 97.8% retention with zero external teacher queries.

B.10 Methodological Scope and Future Disentanglement Roadmap

To maintain absolute academic transparency, we synthesize four key directions regarding gradient dynamics and causal verification alongside the explicit experimental roadmap:

1. **Longitudinal Margin Tracking and Layer-Wise Dynamics**: While static probes establish the geometric misalignment ($\cos \approx +0.1503$), tracking token-specific logit margins ($\Delta m_t = \text{logit}(y_t^*) - \max_{j \neq y_t^*} \text{logit}(j)$) and layer-wise gradient angles across early, mid, and late training stages directly charts the temporal dynamics preceding token flipping.
2. **Targeted Causal Intervention Matrix**: Further disentangling representation drift by comparing: (i) structural-only masking (PRISM), (ii) high-entropy-only parameter masking, and (iii) equal-budget random masking across the entire training trajectory.
3. **Disentangling Semantic Rationales from Schema Hints**: Evaluating *Schema-Only Prompts* (supplying only the ground-truth tool signature without reasoning prose) versus full natural language rationales.
4. **Student Capacity Scaling (1.5B / 7B)**: Testing whether the performance gap between Standalone PRISM (88.7%) and PRISM-Privileged (80.9%) narrows as student model capacity scales to accommodate external reasoning traces.

C Theoretical Analysis: Gradient Projection Mechanics and Bounded Logit Dynamics

Remark 1 (Algebraic Gradient Decomposition under Binary Subspace Masking). Let $\mathcal{M} = \mathcal{M}_{\text{rel}} \cup \mathcal{M}_{\text{unrel}}$ be a disjoint partition of sequence token positions. In standard uniform optimization, the composite gradient $\mathbf{G}_{\text{unif}} = \mathbf{G}_{\text{rel}} + \mathbf{G}_{\text{unrel}}^{\parallel} + \mathbf{G}_{\text{unrel}}^{\perp}$ transmits a transverse component $\mathbf{G}_{\text{unrel}}^{\perp}$ (97.74% of uncalibrated gradient energy; Appendix §B.2). Binary position masking $m_t = \mathbb{I}(t \in \mathcal{M}_{\text{rel}})$ operates as an exact coordinate-wise linear projection operator $\mathcal{P}_{\mathcal{M}_{\text{rel}}}$ onto the reliable coordinate

subspace:

$$\mathbf{G}_{\text{pos}} \triangleq \mathcal{P}_{\mathcal{M}_{\text{rel}}}(\mathbf{G}_{\text{unif}}) = \sum_{t \in \mathcal{M}_{\text{rel}}} \nabla_\theta \mathcal{L}_t = \mathbf{G}_{\text{rel}} \tag{19}$$

By direct algebraic construction, the transverse component of $\mathbf{G}_{\text{unrel}}$ is eliminated from the parameter update vector:

$$\langle \mathbf{G}_{\text{pos}}, \mathbf{G}_{\text{unrel}}^{\perp} \rangle = \langle \mathbf{G}_{\text{rel}}, \mathbf{G}_{\text{unrel}}^{\perp} \rangle = 0 \tag{20}$$

We emphasize that this identity is an algebraic property of coordinate masking rather than a dynamical theorem proving policy collapse. Its analytical utility is in formalizing how binary masking strictly zeroes out instantaneous transverse gradient components, preventing them from entering the first-order parameter update trajectory.

Proposition 2 (Bounded Logit Gradient Norm under Forward-KL Distillation). *Let $p_t, q_t \in \Delta^{|\mathcal{V}|}$ denote teacher and student softmax distributions over vocabulary $\mathcal{V}$. For forward-KL distillation loss $\mathcal{L}_t = T^2 \mathrm{KL}(p_t \parallel q_t)$, the gradient with respect to student pre-softmax logit vector $\mathbf{z}_t \in \mathbb{R}^{|\mathcal{V}|}$ satisfies:*

$$\|\nabla_{\mathbf{z}_t} \mathcal{L}_t\|_2 \leq \sqrt{2} T^2 \tag{21}$$

Proof. The gradient with respect to logit $z_{t,v}$ is $\frac{\partial \mathcal{L}_t}{\partial z_{t,v}} = T^2(q_{t,v} - p_{t,v})$. Since $p_t, q_t \in \Delta^{|\mathcal{V}|}$, we have $\|q_t - p_t\|_1 \leq 2$. By standard norm inequalities:

$$\|\nabla_{\mathbf{z}_t} \mathcal{L}_t\|_2 = T^2 \|q_t - p_t\|_2 \leq T^2 \|q_t - p_t\|_1 \leq 2T^2 \tag{22}$$

More tightly, since both distributions are non-negative vectors summing to 1, $\|q_t - p_t\|_2^2 = \|q_t\|_2^2 + \|p_t\|_2^2 - 2\langle q_t, p_t \rangle \leq 1 + 1 - 0 = 2$. Taking the square root gives $\|\nabla_{\mathbf{z}_t} \mathcal{L}_t\|_2 \leq \sqrt{2} T^2$. □

D Prompts, Schemas & Qualitative Case Studies

D.1 Privileged Black-Box Teacher Prompt Template

The privileged black-box teacher (T_{bb}) generates structured cognitive rationales using the following zero-shot system prompt:

System Prompt for T_{bb}:

```
You are a tool-use cognitive verification expert. Given
the user query, available tool definitions, and expected
ground-truth outcome, analyze the problem and provide a
concise, step-by-step reasoning explanation detailing: (1)
why the selected tool is required; (2) the exact required
parameter keys; and (3) the semantic constraints governing
argument values. Provide your rationale in plain text
before the canonical JSON call.
```

D.2 Qualitative Trajectory Rollout Comparison

Table 10 presents a side-by-side comparison of model responses at step 119 on a complex test instance.

E PRISM Training Algorithm Specification

F Extended Operational Scope and Generalization Boundaries

The empirical findings on our primary benchmark illuminate the operational principles and boundary conditions governing position-level reliability masking across diverse agent interaction paradigms:

1. **Structured Tool-Calling Tasks (Primary Domain)**: In interactive systems characterized by explicit schema specifications (such as nested JSON function calls and API parameters), the supervision topology exhibits sharp bimodal reliability. Closed-set identifiers and delimiter tokens

Table 10: **Qualitative comparison of generated outputs at training step 119.**

Method	Generated Output at Step 119
Reference Ground-Truth	`{"Action": "databaseQuery", "Action_Input": {"table": "users", "filter": "status == active"}}`
Vanilla SFT	`{"Action": "databaseQuery", "Action_Input": {"table": "users", "filter": "status", "query": ... [Syntax Error: Unexpected delimiter]}`
Vanilla SDPO	`{"Action": "searchAPI", "Action_Input": {"query": "users active status"}}` *(Wrong Tool Selection)*
GRPO Baseline	`{"Action": "databaseQuery", "Action_Input": "users"}` *(Parameter Degradation: Truncated schema arguments in high-variance seeds)*
PRISM (Ours)	`{"Action": "databaseQuery", "Action_Input": {"table": "users", "filter": "status == active"}}` *(Exact Match, 100% Compliant)*

Algorithm 1 PRISM: Position-Level Target Reliability Masking Policy Optimization

Require: Student policy q_θ, frozen self-teacher p_{self}, training dataset $\mathcal{D}$, reliability threshold τ_{pos}, temperature T, learning rate η, optional privileged teacher T_{bb}.

for each training iteration **do**
  Sample a minibatch of query-reference pairs $\mathcal{B} = \{(x^{(i)}, y^{*(i)})\}_{i=1}^{B} \sim \mathcal{D}$
  for each instance $(x^{(i)}, y^{*(i)}) \in \mathcal{B}$ **do**
    Construct student context $C_s^{(i)} \leftarrow [x^{(i)}]$
    if privileged teacher T_{bb} is enabled **then**
      Query privileged teacher: $e_{\text{bb}}^{(i)} \leftarrow T_{\text{bb}}(x^{(i)}, y^{*(i)})$
      Construct privileged teacher context: $C_t^{(i)} \leftarrow [x^{(i)}, y^{*(i)}, e_{\text{bb}}^{(i)}]$
    else
      Construct standalone teacher context: $C_t^{(i)} \leftarrow [x^{(i)}, y^{*(i)}]$
    end if
    Generate on-policy rollout continuation $z^{(i)} \sim q_\theta(\cdot \mid C_s^{(i)})$
    for each token position $t = 1, \ldots, L$ **do**
      Compute teacher distribution: $\mathbf{p}_t^{(i)} \leftarrow p_{\text{self}}(\cdot \mid C_t^{(i)}, z_{<t}^{(i)})$
      Compute student distribution: $\mathbf{q}_t^{(i)} \leftarrow q_\theta(\cdot \mid C_s^{(i)}, z_{<t}^{(i)})$
      Compute target reliability $c_t^{(i)}$ via Equation 4
      Assign binary position mask: $m_t^{(i)} \leftarrow \mathbb{I}(c_t^{(i)} \geq \tau_{\text{pos}})$
    end for
    Compute position-masked loss: $\mathcal{L}_{\text{seq}}^{(i)}(\theta) \leftarrow \frac{T^2}{\sum_{t=1}^{L} m_t^{(i)} + \epsilon} \sum_{t=1}^{L} m_t^{(i)} \cdot \mathrm{KL}\big(\mathbf{p}_t^{(i)} \parallel \mathbf{q}_t^{(i)}\big)$
  end for
  Update student parameters: $\theta \leftarrow \theta - \eta \cdot \frac{1}{B} \sum_{i=1}^{B} \nabla_\theta \mathcal{L}_{\text{seq}}^{(i)}(\theta)$
end for

provide high-certainty structural anchors, whereas open-ended parameters produce high-variance gradients. PRISM achieves maximum effectiveness here by cleanly insulating invariant structural representations from parametric hallucinations.

2. **Unconstrained Text Exploration Boundaries (e.g., Free-Text Web Search)**: In unstructured settings such as open-ended search queries lacking rigid grammar or schema delimiters, predictive uncertainty becomes uniformly distributed across the entire sequence. In such domains, token-level predictive distributions do not exhibit sharp syntactic demarcations, defining the structural boundary where syntax-aligned position filtering naturally transitions to continuous sequence-level or token-soft calibration.
3. **Embodied Multi-Step Attribution (e.g., Embodied Agent Planning)**: In long-horizon embodied tasks, execution failures often originate from multi-step temporal credit assignment and compounding environment transition dynamics rather than single-step syntax non-compliance. Extending PRISM to jointly optimize temporal and token-level credit assignment constitutes a promising direction for future agent alignment research.