BACKTRACKING WHEN IT STRAYS: MITIGATING DUAL EXPOSURE BIASES IN LLM REASONING DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have achieved remarkable success in complex reasoning tasks via long chain-of-thought (CoT), yet their immense computational overhead hinders real-world deployment. LLM reasoning distillation addresses this by transferring reasoning capabilities from formidable teacher models to compact student models. However, existing distillation paradigms face a fundamental dilemma. Typical off-policy distillation strictly utilizes teacher-generated golden trajectories, suffering from an *exposure bias* due to the mismatch between training distributions and student-generated inference contexts, which leads to error cascades in long CoT reasoning. To address this, on-policy distillation allows students to explore their own trajectories, but we demonstrate that it inherently introduces a reciprocal *reversed exposure bias*: the teacher model also struggles to provide positive guidance when conditioned on student-generated sub-optimal contexts. To resolve this dual exposure bias problem, we propose *Monitoring Trajectories and Backtracking when it strays* (MOTAB), a new LLM reasoning distillation pipeline. Specifically, MOTAB dynamically monitors the student's on-policy generation against an adaptive safety boundary. When the generation strays and exceeds this threshold, MOTAB backtracks to the last safe state and leverages teacher intervention to correct the course. This approach inherently tolerates minor student errors to mitigate exposure bias, while preventing sub-optimal contexts to circumvent reversed exposure bias. Extensive experiments on the *LIMO-v2* and *AceReason* datasets demonstrate that MOTAB effectively alleviates the dual exposure biases, yielding a roughly 3% average performance improvement in reasoning tasks.

1 INTRODUCTION

Recently, various *large language models* (LLMs) have demonstrated remarkable capabilities in complex reasoning tasks that require multi-step *chain-of-thought* (CoT) (Wei et al., 2022; Shao et al., 2024; Yang et al., 2025; Yuan et al., 2026). Despite their efficacy, deploying these state-of-the-art models in real-world applications remains computationally intensive, motivating the adoption of knowledge distillation to transfer their reasoning capabilities into smaller, faster student models, namely *LLM reasoning distillation* (Agarwal et al., 2024; Yang et al., 2025; Yan et al., 2026).

Typically, most previous works follow an *off-policy distillation* paradigm (Ho et al., 2023; Muennighoff et al., 2025): a teacher model generates an extensive corpus of reasoning trajectories for a given set of complex problems; these trajectories are subsequently filtered using heuristic pipelines (Li et al., 2025; Guha et al., 2025) and utilized for fine-tuning of the student model. However, this off-policy approach empirically suffers from an *exposure bias* issue (Agarwal et al., 2024; Xu et al., 2025). Specifically, during the training phase, the student model is exclusively optimized using the *golden* trajectories provided by the teacher model. Conversely, during inference, the student model is exposed to its own *sub-optimal* generated context. This mismatch inevitably leads to an error cascade, particularly within the context of long CoT reasoning trajectories (Yan et al., 2026).

To mitigate this issue, the community recently proposed an alternative paradigm, namely *on-policy distillation* (Agarwal et al., 2024; Ko et al., 2024). Specifically, the student model generates its own reasoning trajectories, which are then guided through dense feedback provided by the teacher model, *e.g.,* token probabilities (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2025). Although this method effectively eliminates exposure bias by exposing the student's own inference-time

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.

trajectories, in this work, we formally and empirically demonstrate that on-policy distillation also introduces a reciprocal *reversed exposure bias* challenge. Analogously, when the student model generates sub-optimal on-policy trajectories as context, the teacher model often suppresses these tokens and struggles to provide positive supervisory signals (Fu et al., 2026). As experimentally evidenced by Figs. 1 and 2, a consistent discrepancy exists between the teacher's and student's next-token probability distributions when conditioned on student-generated on-policy trajectories, and the divergence between the teacher's distributions under off-policy versus on-policy contexts consistently increases as the sequence length extends.

Therefore, to specifically address these *dual exposure biases*, the ideal distillation dataset should *cover* the student model's own sub-optimal on-policy data distribution to mitigate exposure bias, while the teacher model must provide supervision within a *relatively accurate* student-generated context to circumvent reversed exposure bias, which creates an apparent dilemma. To achieve these goals, we propose a new LLM reasoning distillation method, namely ***Mo**nitoring **T**rajectories **a**nd **B**acktracking when it strays* (**MOTAB**). Specifically, MOTAB consistently monitors the student model's on-policy trajectories against an adaptive safety boundary, *e.g.,* based on the entropy of teacher-generated tokens. Once a generation step exceeds this threshold, MOTAB backtracks to the last safe state through a credit assignment strategy, where the teacher model intervenes to steer the subsequent generation. Accordingly, this approach utilizes a safety boundary to tolerate sub-optimal on-policy trajectories, encouraging the model to expose its own errors and thereby mitigating the exposure bias. When the student model strays, MOTAB employs backtracking to identify a high-quality on-policy context, allowing the teacher model to provide guidance that effectively addresses the reversed exposure bias.

To evaluate the performance of MOTAB, we utilize Qwen3-32B (Yang et al., 2025) and gpt-oss-120b (OpenAI, 2025) as teacher models to generate distillation data on the questions from *LIMO-v2* (Ye et al., 2025) and *AceReason* (Chen et al., 2025c) datasets using MOTAB, and perform full-parameter fine-tuning on three distinct student models. Experimental results demonstrate that our approach yields an average improvement of approximately 3% in complex reasoning performance and effectively mitigates dual exposure biases. *Our anonymous code is released in the repository* https://anonymous.4open.science/r/MOTAB.

In summary, our contributions can be summarized as the follow three-folds:

- We formally and experimentally summarize a fundamental dilemma inherent in existing off-policy and on-policy distillation methods, *i.e.,* the dual exposure biases problem.
- To address the biases, we introduce MOTAB, a new reasoning distillation framework that monitors student-generated on-policy trajectories within an adaptive safety boundary, and backtracks to the last safe point when the student model strays.
- Extensive experimental results demonstrate that our approach consistently enhances LLM reasoning performance and alleviates the dual exposure biases.

2 INVESTIGATION ON DUAL EXPOSURE BIASES

In this section, we formally define and empirically investigate the dual exposure biases in LLM reasoning distillation. The detailed theoretical analysis of these biases is provided in Appendix A.

2.1 FORMAL DEFINITION OF DUAL EXPOSURE BIASES

Task descriptions of LLM reasoning and its distillation. We formalize the LLM reasoning process as an auto-regressive sequence generation task. Given an initial question $\mathbf{q}$, a policy LLM $\boldsymbol{\pi}$ generates a reasoning trajectory consisting of a sequence of reasoning steps $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L)$. At any step l, the reasoning context $\mathbf{s}_l$ is defined as the concatenation of the question and the previously generated prefix as $\mathbf{s}_l = [\mathbf{q}, \mathcal{Y}_{<l}]$. The policy $\boldsymbol{\pi}(\cdot \mid \mathbf{s}_l)$ defines the probability distribution over the vocabulary space for the next generation step. We denote $d^{\boldsymbol{\pi}}(\mathbf{s})$ as the marginal distribution of contexts induced by the policy $\boldsymbol{\pi}$ during free-form inference. The ultimate goal of reasoning distillation is to optimize a student policy $\boldsymbol{\pi}_S$ to accurately replicate the reasoning capabilities of a capable teacher policy $\boldsymbol{\pi}_T$.

Formulation of dual exposure biases. Existing reasoning distillation methods can be broadly categorized into off-policy and on-policy approaches. However, both paradigms inherently suffer from different forms of exposure biases:

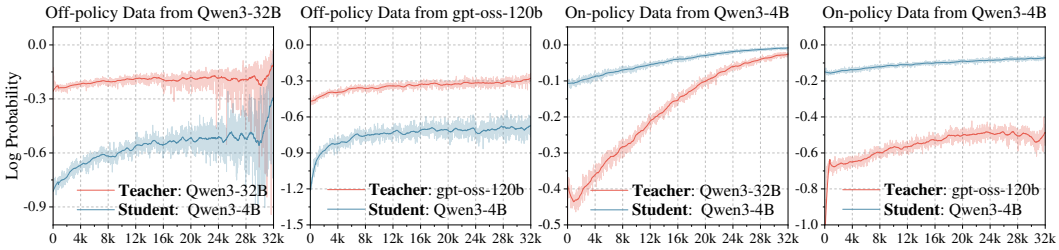

Figure 1: Next-token log probabilities of teacher and student under off-policy and on-policy contexts.

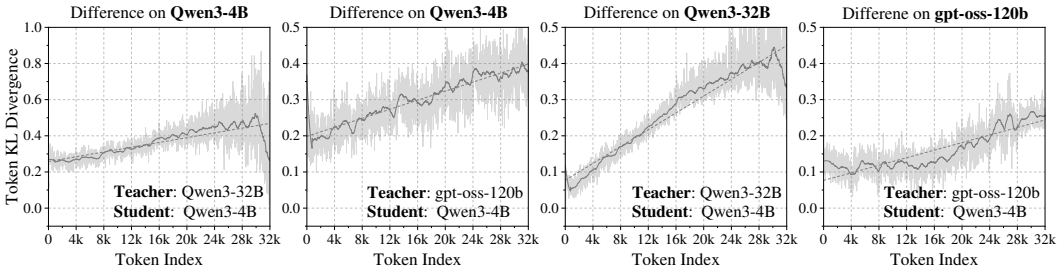

Figure 2: KL divergence between off-policy and on-policy trajectories across context lengths.

- **Exposure bias of off-policy distillation:** Typical off-policy methods follow the standard teacher-forcing objective, which minimizes the *Kullback-Leibler* (KL) divergence over the teacher's context distribution $d^{\boldsymbol{\pi}_T}$:

$$\mathcal{J}_{\text{off}}(\boldsymbol{\pi}_S) = \mathbb{E}_{\mathbf{s}\sim d^{\boldsymbol{\pi}_T}}\left[D_{\text{KL}}(\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}) \,\|\, \boldsymbol{\pi}_S(\cdot \mid \mathbf{s}))\right]. \tag{1}$$

  However, during test-time inference, the student auto-regressively generates trajectories from its own distribution $d^{\boldsymbol{\pi}_S}$. Because the training distribution mismatches the inference distribution ($d^{\boldsymbol{\pi}_T} \neq d^{\boldsymbol{\pi}_S}$), a distribution shift occurs. Once the student makes an initial error, it encounters a context $\mathbf{s} \notin \text{supp}(d^{\boldsymbol{\pi}_T})$, inevitably leading to cascading errors.

- **Reversed exposure bias of on-policy distillation:** To mitigate this distribution shift, on-policy methods directly sample contexts from the student's own distribution $d^{\boldsymbol{\pi}_S}$:

$$\mathcal{J}_{\text{on}}(\boldsymbol{\pi}_S) = \mathbb{E}_{\mathbf{s}\sim d^{\boldsymbol{\pi}_S}}\left[D_{\text{KL}}(\boldsymbol{\pi}_S(\cdot \mid \mathbf{s}) \,\|\, \boldsymbol{\pi}_T(\cdot \mid \mathbf{s}))\right]. \tag{2}$$

  However, when $\mathbf{s} \sim d^{\boldsymbol{\pi}_S}$ contains severe logical flaws due to the student's sub-optimal capabilities, this trajectory acts as an *out-of-distribution* input for the teacher. Consequently, the supervision signal $\boldsymbol{\pi}_T(\cdot \mid \mathbf{s})$ always suppresses student tokens and fails to provide the positive feedback, ultimately degrading into uninformative noise.

Accordingly, existing methods struggle to simultaneously align the training with the inference-time distribution while maintaining valid, high-quality supervision, which forms a fundamental *dilemma*, resulting in the **dual exposure biases**. More formulations can be seen in Appendices A.1 and A.2.

2.2 Empirical Analysis of Dual Exposure Biases

To substantiate the existence of the dual exposure biases, we conduct preliminary experiments using Qwen3-32B and gpt-oss-120b as teacher models, and Qwen3-4B-Instruct, which is initialized via an off-policy distillation phase on datasets corresponding to their respective teachers, as the student.

We first evaluate the models on off-policy data, where the context comprises golden trajectories generated exclusively by the teacher. As illustrated in Fig. 1, the teacher's predictive confidence is consistently high, given that the context perfectly aligns with its native distribution. Conversely, the student model struggles to replicate these exact reasoning steps and exhibits significantly lower confidence. Furthermore, as shown in Fig. 2, we illustrate the KL divergence of the next-token probability distributions produced by the student model when conditioned on its own auto-regressively

generated on-policy contexts versus teacher-generated off-policy contexts. Generally, this divergence consistently increases as the reasoning sequence lengthens. Therefore, these results clearly illustrate the exposure bias: training solely on golden contexts leaves the student increasingly vulnerable to its own stylistic deviations, causing errors to accumulate rapidly during inference-time generation.

To mitigate exposure bias, on-policy methods rely on the teacher to supervise student-generated trajectories. However, our experiments also uncover a critical *reversed exposure bias*. As shown in Fig. 1, when conditioning on sub-optimal on-policy trajectories generated by the student, the student model's step-wise confidence for its self-generated token is consistently higher than the teacher's confidence for the same token. Therefore, the guidance provided by the teacher model, based on token probabilities (Agarwal et al., 2024), consistently suppresses these student-generated tokens, thereby hindering the provision of effective positive supervision. Similarly, in Fig. 2, the KL divergence of the next-token probability distributions estimated by the teacher model across on-policy and off-policy contexts increases with sequence length, thereby also exhibiting an error accumulation phenomenon.

2.3 What is the Ideal Unbiased State for Reasoning Distillation?

Given the dual exposure biases formulated in Eqs. 1 and 2 and empirically demonstrated above, an *ideal LLM reasoning distillation process* must satisfy the following two theoretical conditions:

- ***Coverage* for exposure bias**: The training data must be sampled from the student's intrinsic context distribution $d^{\boldsymbol{\pi}_S}(\mathbf{s})$, encompassing the sub-optimal prefixes the student is likely to encounter during real-world inference to eliminate distribution shift.
- ***Validity* for reversed exposure bias**: The teacher model should be conditioned on a valid context $\mathbf{s}^*$, enabling it to provide informative positive supervision signals.

Formally, building upon these two conditions, the ideal reasoning distillation objective aims to minimize the expected negative log-likelihood over this optimal configuration:

$$\mathcal{J}_{\text{ideal}}(\boldsymbol{\pi}_S) = \mathbb{E}_{\mathbf{s}\sim d^{\boldsymbol{\pi}_S}}\left[\mathbb{E}_{\mathbf{y}^*\sim\boldsymbol{\pi}_T(\cdot|\mathbf{s}^*)}\left[-\log \boldsymbol{\pi}_S(\mathbf{y}^* \mid \mathbf{s})\right]\right]. \tag{3}$$

In practice, however, this remains challenging, as they exists an inherent dilemma, *i.e.,* $\mathbf{s}^* \neq \mathbf{s} \sim d^{\boldsymbol{\pi}_S}$.

2.4 Can Speculative KD Resolve Dual Exposure Biases?

In the community, to balance off-policy and on-policy data, some recent works propose *speculative knowledge distillation* (SKD) (Xu et al., 2025; Kim et al., 2025), where the teacher continuously monitors and corrects the student's on-policy generation. The core intuition is that if an erroneous token $\hat{y}_i$ is produced by the student, it is immediately replaced by a teacher-generated token sampled from $\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}^S)$. The training objective for this pipeline is as follows:

$$\mathcal{J}_{\text{skd}}(\boldsymbol{\pi}_S) = \mathbb{E}_{\mathbf{s}\sim d_{\text{mix}}}\left[\mathbb{E}_{\hat{y}\sim\boldsymbol{\pi}_T(\cdot|\mathbf{s})}[-\log \boldsymbol{\pi}_S(\hat{y} \mid \mathbf{s})]\right], \tag{4}$$

where $d_{\text{mix}} = \beta d^{\boldsymbol{\pi}_S} + (1-\beta)d^{\boldsymbol{\pi}_T}$ denotes a mixture distribution between the teacher's and the student's context distributions. However, by comparing $\mathcal{J}_{\text{skd}}$ in Eq. 4 with the ideal objective $\mathcal{J}_{\text{ideal}}$ in Eq. 3, we identify two critical gaps in SKD:

- **Coverage gap** ($d_{\text{mix}} \neq d^{\boldsymbol{\pi}_S}$): By monitoring the student's generation process step-by-step, SKD strictly corrects the steps where the student deviates from the teacher's distribution, thereby failing to expose the student's sub-optimal trajectories.
- **Validity gap** ($\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}) \neq \boldsymbol{\pi}_T(\cdot \mid \mathbf{s}^*)$): Forced token replacements break the natural auto-regressive coherence of the trajectory, so the teacher $\boldsymbol{\pi}_T$ still suffers from out-of-distribution collapse. Lacking the global error-correction, it will produce high-entropy noise rather than meaningful guidance, violating the validity condition.

Therefore, to bridge these two gaps and more accurately approximate the ideal objective $\mathcal{J}_{\text{ideal}}$, we propose a more robust data synthesis framework MOTAB in this work.

3 The Proposed MOTAB

Method overview. To address the dilemma of the dual exposure biases, the ideal conditions should satisfy both coverage of the student's sub-optimal distribution and validity, ensuring the teacher

provides guidance within a relatively correct context. To achieve this, the basic idea of MOTAB involves consistently monitoring the on-policy trajectories generated by the student, *tolerating sub-optimal paths within an adaptive safety boundary, and backtracking to a relatively safe state for teacher intervention once that boundary is exceeded.*

Specifically, MOTAB is an iterative data synthesis framework comprised of three critical steps: (1) ***on-policy trajectory monitoring***: As the student generates each reasoning step, its likelihood is consistently monitored by the teacher model. Specifically, the entropy of the step generated by the teacher within the given context serves as the adaptive safety boundary. Higher entropy signifies the existence of multiple viable paths toward the correct solution (Wang et al., 2025), thereby expanding the safety margin. As long as the student's trajectory remains within this boundary, its sub-optimal trajectories are tolerated; (2) ***State backtracking***: Once the student violates this safety boundary, a credit assignment mechanism (Schulman et al., 2016; Xi et al., 2026) is employed to trace back along the on-policy trajectory. This process identifies the specific step characterized by a precipitous decline in likelihood scores, facilitating a backtrack to the last safe state preceding the error; (3) ***off-policy trajectory stitching***: Within this relatively correct context following the backtrack, the teacher model performs a continuation to generate a corrected sequence. Instead of directly using this teacher-generated trajectory as training data, we append the teacher's suffix to the on-policy trajectory precisely at the point where the safety boundary is violated, therefore to explicitly teach the student model how to self-correct its previous deviations. For clarity, the overall framework of MOTAB is depicted in Fig. 3.

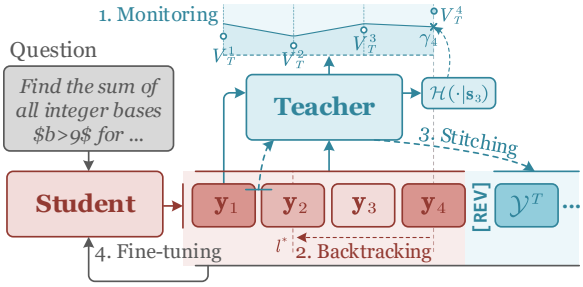

Figure 3: Overall framework of the proposed MOTAB.

3.1 ON-POLICY TRAJECTORY MONITORING VIA ADAPTIVE BOUNDARY

To satisfy the *coverage* condition, the student model must be allowed to explore its intrinsic distribution d^{π_S}. However, while unconstrained exploration inevitably leads to severe logical hallucinations, excessively restrictive student exploration, *e.g.,* SKD (Xu et al., 2025; Yang et al., 2026), fails to expose the student's sub-optimal distribution. Therefore, MOTAB monitors the student's step-wise generation using a teacher-guided *adaptive safety boundary*. Given a reasoning context $\mathbf{s}_{l-1}$ at step l, the student policy $\boldsymbol{\pi}_S$ generates a candidate reasoning step $\mathbf{y}_l^S$. To evaluate the validity of this step without aggressively truncating acceptable sub-optimal logic, we define a step-level value function $V_T \in [0, 1]$ using the teacher's length-normalized likelihood (Xu et al., 2025; Liu et al., 2026):

$$V_T\left(\mathbf{s}_{l-1}, \mathbf{y}_l^S\right) = \exp\left(\frac{1}{|\mathbf{y}_l^S|}\sum_{i=1}^{|\mathbf{y}_l^S|}\log \boldsymbol{\pi}_T\left(y_{l,i}^S \mid \mathbf{s}_{l-1}, y_{l,<i}^S\right)\right), \tag{5}$$

where $y_{l,i}^S$ denotes the i-th token of the reasoning step $\mathbf{y}_l^S$. Instead of utilizing a rigid, static threshold which artificially restricts the student's exploration, we introduce a state-adaptive boundary $\gamma(\mathbf{s}_{l-1})$. This boundary is inversely proportional to the teacher model's predictive sequence entropy $\mathcal{H}(\boldsymbol{\pi}_T \mid \mathbf{s}_{l-1})$. A high entropy indicates a reasoning manifold with multiple acceptable paths (Wang et al., 2025; Cheng et al., 2026), warranting a lower threshold, which allows the student to explore more freely; conversely, deterministic axiomatic steps require strict adherence. The adaptive boundary is formulated as

$$\gamma(\mathbf{s}_{l-1}) = \gamma_0 \cdot \exp\left(-\alpha \cdot \mathcal{H}(\boldsymbol{\pi}_T \mid \mathbf{s}_{l-1})\right), \tag{6}$$

where $\gamma_0 \in (0, 1]$ is the base threshold and $\alpha > 0$ is a scaling factor. If $V_T(\mathbf{s}_{l-1}, \mathbf{y}_l^S) \geq \gamma(\mathbf{s}_{l-1})$, the step is deemed safe, and the state transitions normally to $\mathbf{s}_l = [\mathbf{s}_{l-1}, \mathbf{y}_l^S]$. This mechanism guarantees that the student's logically sound, albeit stylistic, explorations are preserved.

3.2 STATE BACKTRACKING VIA CREDIT ASSIGNMENT

When the student's generation breaches the safety boundary, *i.e.,* $V_T < \gamma(\mathbf{s}_{l-1})$, a critical logical collapse is detected. However, directly replacing the immediate step $\mathbf{y}_l^S$ is myopic, as logical errors

are often delayed, *e.g.,* a flawed premise established at step $l-2$ only manifests as a mathematical contradiction at step l. Furthermore, forcing the teacher to correct from the flawed state $\mathbf{s}_{l-1}$ violates the *validity* condition, triggering reversed exposure bias. To address this, MOTAB employs a *credit assignment* strategy (Lowe et al., 2017; Pignatelli et al., 2024) to backtrack to the safe bifurcation step. Specifically, we compute the step-wise *temporal difference error* δ_k (Schulman et al., 2016; Xi et al., 2026) along the historical trajectory to identify the precise moment the logic deviated from the teacher's safe manifold:

$$\delta_k = V_T\left(\mathbf{s}_{k-1}, \mathbf{y}_k^S\right) - V_T\left(\mathbf{s}_{k-2}, \mathbf{y}_{k-1}^S\right), \quad \forall k \in \{1, \dots, l\}. \tag{7}$$

The safe step l^* is identified as the node exhibiting the most severe degradation in oracle value, strictly constrained such that the preceding state remains safe:

$$l^* = \arg\min_{k \le l}(\delta_k) \quad \text{s.t.} \quad V_T\left(\mathbf{s}_{l^*-2}, \mathbf{y}_{l^*-1}^S\right) \ge \gamma(\mathbf{s}_{l^*-2}). \tag{8}$$

By isolating l^*, MOTAB rewinds the reasoning process strictly to the last pristine context $\mathbf{s}_{l^*-1}$. This structurally isolates the teacher from any out-of-distribution sub-optimal steps, ensuring that the subsequent teacher supervision remains uncorrupted by reversed exposure bias.

3.3 OFF-POLICY TRAJECTORY STITCHING

Having identified the safe prefix $\mathbf{s}_{l^*-1}$, the teacher policy $\boldsymbol{\pi}_T$ is queried to generate a valid, off-policy corrective suffix $\mathcal{Y}_{l^*:T}^T$ as

$$\mathcal{Y}_{l^*:T}^T \sim \boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l^*-1}). \tag{9}$$

Because $\mathbf{s}_{l^*-1}$ belongs to the valid support manifold of the teacher, $\boldsymbol{\pi}_T$ provides highly confident and accurate guidance. To construct the final trajectory for fine-tuning, instead of directly use $[\mathbf{s}_{l^*-1}, \mathcal{Y}_{l^*:T}^T]$, we seamlessly stitch the student's sub-optimal exploration, an explicit semantic transition, and the teacher's correction. Specifically, we retain the student's fully flawed trajectory up to unsafe step l to act as negative exposure. A discrete semantic revising token, denoted as `[REV]`, *e.g., "However"*, is appended to asynchronously stitch the pristine correction to the flawed context:

$$\boldsymbol{\tau}_{\text{SFT}} = \left[\mathbf{q}, \mathcal{Y}_{1:l}^S\right] \oplus \texttt{[REV]} \oplus \mathcal{Y}_{l^*:T}^T. \tag{10}$$

During fine-tuning, by conditioning the student's loss on the sub-optimal on-policy state to predict `[REV]` and then output the correct steps from l^*, the student explicitly learns to: (1) operate robustly within its own generation distribution (mitigating exposure bias), (2) recognize its own logical drift, and (3) perform self-correction using valid oracle targets (bypassing reversed exposure bias).

3.4 THEORETICAL ALIGNMENT TO THE IDEAL OBJECTIVE

To theoretically validate MOTAB, we demonstrate its alignment with the ideal objective $\mathcal{J}_{\text{ideal}}$ in Eq. 3 and illustrate how it resolves the *coverage* and *validity* gaps inherent in SKD in Eq. 4.

In MOTAB, the student policy is optimized via maximum likelihood estimation over the dynamically stitched trajectories. The expected empirical loss $\mathcal{J}_{\text{MOTAB}}(\boldsymbol{\pi}_S)$ can be decomposed into a piecewise formulation conditionally dependent on the adaptive boundary:

$$\begin{aligned}\mathcal{J}_{\text{MOTAB}}(\boldsymbol{\pi}_S) =& \mathbb{E}_{\mathbf{s}_{l-1}\sim d^{\boldsymbol{\pi}_S}} \mathbb{E}_{\mathbf{y}_l^S \sim \boldsymbol{\pi}_S} \Big[\mathbb{I}_{[V_T \ge \gamma(\mathbf{s})]} \left[-\log \boldsymbol{\pi}_S\left(\mathbf{y}_l^S \mid \mathbf{s}_{l-1}\right)\right] \\ &+ \mathbb{I}_{[V_T < \gamma(\mathbf{s})]} \mathbb{E}_{\mathcal{Y}^T \sim \boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l^*-1})} \left[-\log \boldsymbol{\pi}_S\left(\texttt{[REV]} \oplus \mathcal{Y}^T \mid \mathbf{s}_{l-1} \oplus \mathbf{y}_t^S\right)\right] \Big].\end{aligned} \tag{11}$$

By comparing Eq. 11 with $\mathcal{J}_{\text{ideal}}$, we show that MOTAB strictly satisfies both prerequisite conditions: Unlike SKD which restricts states to a narrow mixture d_{mix}, Eq. 11 operates strictly over $\mathbf{s}_{l-1} \sim d^{\boldsymbol{\pi}_S}$ and $\mathbf{y}_t^S \sim \boldsymbol{\pi}_S$. Retaining the flawed context $(\mathbf{s}_{l-1} \oplus \mathbf{y}_t^S)$ as the prefix ensures full coverage of the student's authentic inference manifold, effectively eliminating the distribution shift. Forcing the teacher to predict upon a flawed context $\mathbf{s}_t$ induces out-of-distribution noise ($\boldsymbol{\pi}_T \neq \boldsymbol{\pi}_T^*$). MOTAB approximates the ideal oracle $\boldsymbol{\pi}_T^*$ by sampling the correction $\mathcal{Y}^T$ exclusively from the pristine state $\mathbf{s}_{l^*-1} \in \text{supp}(d^{\boldsymbol{\pi}_T})$. Stitching via the `[REV]` token guarantees the teacher's supervision remains valid and highly confident. More rigorous theoretical analysis of MOTAB can be seen in Appendix A.4.

Accordingly, $\mathcal{J}_{\text{MOTAB}}$ serves as a robust estimator for $\mathcal{J}_{\text{ideal}}$, simultaneously mitigating the dual exposure biases through structurally sound data synthesis.

Table 1: Experimental results using Qwen3-4B-Instruct as the student model.

Teach.	Method	AIME24			AIME25			MATH500					Avg.
		P@8	P@4	P@1	P@8	P@4	P@1	L1	L2	L3	L4	L5	
Qwen3-32B	SFT	50.00	37.81	21.25	40.00	35.43	24.17	90.70	82.22	77.14	70.31	51.49	52.77
	RFT	46.67	38.38	21.67	40.00	30.95	21.67	83.72	82.22	85.71	65.62	51.49	51.65
	ImitKD	46.67	40.71	24.17	40.00	33.05	20.00	93.02	82.22	87.62	77.34	60.45	55.02
	SKD ($\beta = 0.6$)	53.33	45.00	25.00	46.67	43.33	25.00	88.37	85.55	89.52	75.78	60.45	58.09
	SKD ($\beta = 0.8$)	53.33	48.33	24.17	46.67	45.00	27.50	88.37	86.67	87.61	82.03	63.43	59.37
	SKD ($\beta = 1.0$)	56.67	43.33	27.50	46.67	40.00	25.83	90.07	86.67	88.57	81.25	56.72	58.47
	MOTAB (ours)	**66.67**	**57.57**	**34.17**	**56.67**	**50.71**	**31.67**	**93.02**	**86.67**	**95.24**	**86.72**	**73.88**	**66.64**
gpt-oss-120b	SFT	76.67	73.14	43.33	73.33	61.48	37.50	90.70	90.00	80.00	74.22	50.75	68.28
	ImitKD	76.67	72.29	44.17	63.33	57.29	37.92	81.40	67.78	71.43	69.53	61.19	63.91
	SKD ($\beta = 0.6$)	80.00	74.57	45.83	66.67	61.48	39.58	69.77	74.44	77.14	71.88	72.39	66.70
	SKD ($\beta = 0.8$)	83.33	75.29	46.66	80.00	72.52	45.00	90.70	92.22	84.76	81.25	74.63	75.12
	SKD ($\beta = 1.0$)	80.00	73.14	36.25	80.00	64.33	33.75	74.42	72.22	71.43	75.00	73.13	66.70
	MOTAB (ours)	**86.67**	**78.33**	**47.08**	**83.33**	**74.14**	**47.50**	**93.02**	**95.56**	**88.57**	**82.81**	**76.87**	**77.62**

4 EXPERIMENTAL EVALUATION

Experimental settings. Our problem sets are derived from the *LIMO-v2* repository (Ye et al., 2025) and a randomly sampled subset of 10,000 questions from *AceReason* (Chen et al., 2025c). We evaluate our trained student models across three mathematical reasoning benchmarks: AIME24, AIME25, and MATH500 (Lightman et al., 2024), and two cross-domain benchmarks: IFEval (Zhou et al., 2023) and GPQA (Rein et al., 2023). For AIME24, AIME25, and GPQA, we sample 8 times per problem and report the pass@8, pass@4, and pass@1 metrics. We utilize Qwen3-32B and gpt-oss-120b as teacher models, while Qwen3-4B-Instruct, Qwen2.5-7B-Instruct, and Llama3.2-3B serve as the student models. Our selected baselines include *supervised fine-tuning* (SFT), *rejection sampling fine-tuning* (RFT), ImitKD (Lin et al., 2020), and SKD (Xu et al., 2025). More detailed settings, *e.g.,* LLM model cards and implementation details, can be seen in Appendix D.

4.1 MAIN RESULTS

Table 1 summarizes the results of our proposed MOTAB against various baseline methods. Overall, MOTAB consistently and significantly outperforms all baselines across different teacher models and metrics. Notably, when utilizing Qwen3-32B as the teacher, MOTAB achieves an average accuracy of 66.64%, yielding a substantial improvement of over 7% compared to the strongest baseline SKD. This superiority is strictly maintained when scaling up to the gpt-oss-120b teacher. Furthermore, our method exhibits remarkable resilience on highly complex tasks requiring long CoT trajectories. On challenging benchmarks, *e.g.,* AIME24 and AIME25, especially under the strict Pass@1 metric, and the most difficult subsets of MATH500, standard off-policy methods, *e.g.,* SFT and RFT, struggle severely due to error cascades caused by exposure bias. Similarly, aggressive on-policy interventions like ImitKD and SKD yield limited gains, as they force the teacher to generate from sub-optimal contexts, exacerbating reversed exposure bias. In contrast, MOTAB drastically boosts performance by adaptively tolerating sub-optimal explorations and backtracking to the safe point for intervention.

Beyond the primary evaluations, we extensively validate the robustness and generalizability of our framework across varied data sources, student architectures, and downstream domains. As shown in Table 3, training on the distinct *AceReason* dataset yields identically superior trends, confirming that the efficacy of MOTAB is independent of specific data distributions. Moreover, Table 6 demonstrates that our method seamlessly transfers to diverse student architectures, consistently elevating the reasoning capabilities of both Qwen2.5-7B and Llama3.2-3B. Finally, to assess out-of-domain transferability, we evaluate the math-trained models on instruction following, *i.e.,* IFEval, and scientific reasoning, *i.e.,* GPQA. As detailed in Table 7, MOTAB achieves state-of-the-art cross-domain performance regardless of the teacher model used. This verifies that our trajectory monitoring and backtracking mechanisms not only prevent mathematical error cascades but also instill a generalized, domain-agnostic self-reflection ability that broadly enhances complex, multi-step problem-solving.

Table 4 compares the performance of MOTAB against several on-policy distillation approaches, *e.g.,* top-k OPD and DASD. Since on-policy methods are highly memory-intensive and require vocabulary

Table 2: Ablation study of MOTAB.

Method	AIME24		AIME25		MATH
	P@8	P@1	P@8	P@1	
MOTAB	**66.67**	**34.17**	**56.67**	**31.67**	**85.60**
w/o adapt	63.33	32.50	50.00	29.16	84.00
w/o back	56.67	30.83	50.00	27.91	82.80
w/o stitch	63.33	32.08	53.33	29.58	83.60

Table 3: Results on 10k questions from *AceReason*.

Method	AIME24		AIME25		MATH
	P@8	P@1	P@8	P@1	
SFT	56.67	25.41	53.33	23.33	81.00
ImitKD	63.33	31.25	53.33	28.33	81.20
SKD	63.33	32.92	53.33	28.75	82.60
MOTAB	**70.00**	**35.00**	**56.67**	**30.42**	**85.80**

Table 4: Results of on-policy methods.

Method	AIME24		AIME25		MATH
	P@8	P@1	P@8	P@1	
SFT	50.00	18.75	**60.00**	21.25	69.00
RFT	13.33	5.83	33.33	12.08	64.20
SKD	53.33	20.41	56.67	22.08	71.40
top-k OPD	**60.00**	32.91	53.33	30.41	**79.20**
DASD	53.33	26.67	53.33	28.75	74.60
MOTAB	**60.00**	**33.75**	**60.00**	**31.67**	79.00

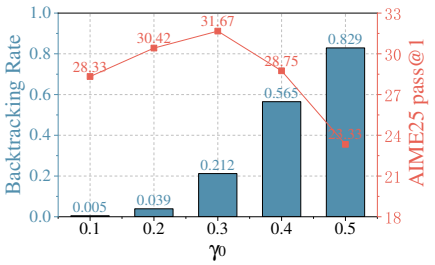

Table 5: Sensitivity analysis of the parameter γ_0.

alignment between the teacher and student models, we employ Qwen3-8B as the teacher and Qwen3-1.7B as the student. The results demonstrate that our method not only significantly outperforms off-policy baselines but also remains comparable to on-policy methods. Furthermore, given that on-policy methods entail substantially higher time complexity and GPU memory consumption than our approach, these results further validate the effectiveness and efficiency of our method.

4.2 EVALUATING DUAL EXPOSURE BIASES

To explicitly visualize the mitigation of dual exposure biases, we track the token-level log probabilities of both the student and teacher models along the trajectories synthesized by MOTAB. As illustrated in Fig. 4, our state backtracking mechanism ensures that the teacher's confidence remains stable, successfully averting the severe probability collapse typically associated with reversed exposure bias.

Under a boundary $\gamma_0 = 0.3$, a distributional gap persists, preserving the student's exploratory diversity to mitigate exposure bias while maintaining valid teacher supervision. Remarkably, when the boundary is tightened to $\gamma_0 = 0.5$, the probability distributions of the student and teacher align almost perfectly. This empirical evidence demonstrates that MOTAB effectively harmonizes the student's authentic generation manifold with the teacher's valid supervision manifold, fundamentally resolving the distributional discrepancies in dual exposure biases.

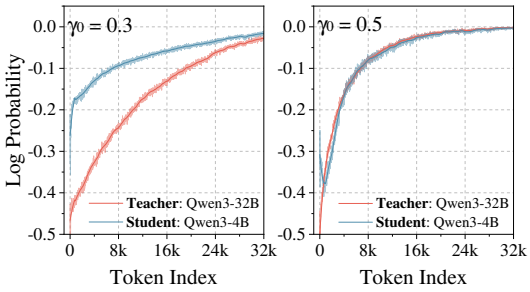

Figure 4: Evaluation of dual exposure biases.

4.3 ABLATION AND SENSITIVITY ANALYSIS

Ablation study. To validate our framework's core components, we conduct an ablation study on the AIME24, AIME25, and MATH benchmarks. As shown in Table 2, removing any module consistently degrades performance. Crucially, disabling state backtracking, *i.e., w/o back*, incurs the most severe decline, *e.g.,* dropping from 34.17% to 30.83% on AIME24 Pass@1. This highlights that forcing immediate teacher intervention without tracing the logical root cause pollutes the context, thereby triggering reversed exposure bias. Furthermore, omitting sub-optimal trajectory stitching, *i.e., w/o stitch*, via direct sequence truncation noticeably impairs accuracy. This confirms that explicitly exposing the student to its own sub-optimal prefixes is vital for learning self-correction and mitigating exposure bias. Finally, substituting the entropy-based adaptive boundary with a static threshold, *i.e., w/o adapt*, also harms performance, demonstrating that rigid constraints cannot effectively balance student exploration diversity with the validity of teacher supervision.

Table 6: Performance on more student models.

	Method	AIME24		AIME25		MATH
		P@8	P@1	P@8	P@1	
Qwen2.5-7B	SFT	23.33	**10.42**	16.67	4.58	62.60
	ImitKD	20.00	5.83	**26.67**	5.83	58.60
	SKD	23.33	8.06	24.44	7.36	64.27
	MOTAB	**30.00**	**10.42**	**26.67**	**7.92**	**70.40**
Llama3.2-3B	SFT	16.67	3.33	6.67	0.83	44.00
	ImitKD	16.67	2.92	6.67	1.67	45.80
	SKD	17.78	2.92	8.89	2.22	46.47
	MOTAB	**23.33**	**5.00**	**16.67**	**5.83**	**52.60**

Table 7: Results on datasets from other domains.

	Method	IFEval		GPQA		
		P.S.	Avg.	P@8	P@4	P@1
Qwen3-32B	SFT	29.92	37.23	85.35	68.35	31.06
	ImitKD	32.53	41.02	84.85	68.34	39.93
	SKD	35.86	43.38	85.86	73.97	41.98
	MOTAB	**37.89**	**46.80**	**89.39**	**77.18**	**46.97**
gpt-oss-120b	SFT	32.16	40.37	84.34	72.60	44.44
	ImitKD	32.35	40.15	76.26	65.12	39.39
	SKD	34.57	42.71	88.89	79.14	44.38
	MOTAB	**41.59**	**50.01**	**91.41**	**79.42**	**50.51**

Sensitivity analysis. We conduct a sensitivity analysis on the threshold γ_0 to evaluate its impact on reasoning performance and intervention frequency. As illustrated in Fig. 5, increasing γ_0 inherently enforces a stricter safety boundary, leading to a monotonic rise in the backtracking rate. However, the reasoning performance on AIME25 Pass@1 exhibits an inverted-U trend, peaking at a moderate value of $\gamma_0 = 0.3$. This suggests that an overly low γ_0 is too lenient, failing to catch severe logical drifts and thus permitting error cascades. Conversely, an excessively high γ_0 becomes overly rigid, constantly interrupting the student's natural exploration and discarding acceptable sub-optimal paths, which undermines the mitigation of exposure bias.

5 Related Works

In this section, we briefly summarize the literature on LLM reasoning distillation. Generally, prior reasoning distillation methods follow an *off-policy paradigm* (Ho et al., 2023; Hsieh et al., 2023; Magister et al., 2023; Kou et al., 2026). This process involves sampling complex problems (Ye et al., 2025; Muennighoff et al., 2025), utilizing a teacher model to generate multiple responses via temperature sampling (Chen et al., 2023; Yan et al., 2026), and conducting rigorous filtering based on criteria, *e.g.,* response length (Wu et al., 2025; Chen et al., 2025c), answer correctness (Chen et al., 2025b; Li et al., 2025), and the student's adaptability (Yuan et al., 2024; Wang et al., 2026).

However, as analyzed in Sec. 2, early research also identified that such off-policy methods suffer from exposure bias, *i.e.,* train-test mismatch (Ranzato et al., 2016; Lin et al., 2020). Theoretically motivated by imitation learning (Ross et al., 2011; Zhang & Cho, 2017), recent methods propose *on-policy distillation* (Agarwal et al., 2024; Song & Zheng, 2026). Unlike previous approaches, this paradigm allows the student model to generate its own reasoning trajectories while receiving dense, real-time feedback from the teacher model at each token level, *e.g.,* f-divergence (Agarwal et al., 2024), log probabilities (Gu et al., 2024), and the combination of forward and backward KL divergences (Jung et al., 2025). In parallel, recent reinforcement learning methods have further demonstrated the efficacy of on-policy data through rollouts by LLM itself (Shao et al., 2024; Chen et al., 2025a). In this work, we address the dilemma between off-policy and on-policy distillation, *i.e., dual exposure biases*. The most intuitive approach involves identifying a mixture distribution of the two, *e.g.,* through the scheduled sampling (Bengio et al., 2015). Recent speculative distillation strategies (Xu et al., 2025; Kim et al., 2025) have similarly adopted this rationale; however, our analysis in Sec. 2.4 demonstrates that such schemes still fail to satisfy the ideal unbiased condition, which ultimately motivates our proposed MOTAB.

6 Conclusion

In this paper, we identify a dilemma between the off-policy and on-policy paradigms in LLM reasoning distillation, *i.e., dual exposure biases*. We demonstrate that while off-policy distillation suffers from exposure bias due to training-inference mismatches, on-policy distillation inevitably introduces a reversed exposure bias, wherein the teacher model fails to provide reliable supervision when conditioned on straying, student-generated trajectories. To resolve this dilemma, we introduced MOTAB, a novel trajectory monitoring and backtracking framework. By employing an adaptive safety boundary, MOTAB allows the student model to explore and learn from its own errors, thereby mitigating exposure bias, while explicitly backtracking and applying teacher intervention before the context degrades irreparably, circumventing reversed exposure bias. Extensive experiments demonstrate that MOTAB enhances reasoning capabilities and mitigates dual exposure biases.

AI USE STATEMENT

LLMs were utilized solely for grammatical refinement and linguistic polishing of the manuscript; the core ideas, methodology, and results were developed independently by the authors.

REFERENCES

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

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Annual Conference on Neural Information Processing Systems*, pp. 1171–1179, 2015.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. MCC-KD: multi-cot consistent knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 6805–6820, 2023.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *CoRR*, abs/2510.18874, 2025a.

Xiaoshu Chen, Sihang Zhou, Ke Liang, Xiaoyu Sun, and Xinwang Liu. Skip-thinking: Chunk-wise chain-of-thought distillation enable smaller language models to reason better and faster. In *Conference on Empirical Methods in Natural Language Processing*, pp. 12142–12157, 2025b.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. In *Annual Conference on Neural Information Processing Systems*, 2025c.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *AAAI Conference on Artificial Intelligence*, pp. 30377–30385, 2026.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *CoRR*, abs/2603.25562, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. In *First Workshop on Foundations of Reasoning in Language Models*, 2025.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Annual Meeting of the Association for Computational Linguistics*, pp. 14852–14882, 2023.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL*, pp. 8003–8017, 2023.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. Todi: Token-wise distillation via fine-grained divergence control. In *Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8091, 2025.

Jaehoon Kim, Kwangwook Seo, and Dongha Lee. In their own words: Reasoning traces tailored for small models make them better reasoners. *CoRR*, abs/2509.22230, 2025.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *International Conference on Machine Learning*, pp. 24872–24895, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. In *International Conference on Machine Learning*, 2025.

Zhiqiang Kou, Junyang Chen, Xin-Qiang Cai, Xiaobo Xia, Ming-Kun Xie, Dong-Dong Wu, Biao Liu, Yuheng Jia, Xin Geng, Masashi Sugiyama, and Tat-Seng Chua. Positive-unlabeled reinforcement learning distillation for on-premise small models. *CoRR*, abs/2601.20687, 2026.

Yang Li, Youssef Emad, Karthik Padthe, Jack Lanchantin, Weizhe Yuan, Thao Nguyen, Jason E Weston, Shang-Wen Li, Dong Wang, Ilia Kulikov, et al. Naturalthoughts: Selecting and distilling reasoning traces for general reasoning tasks. In *NeurIPS Workshop on Efficient Reasoning*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Conference on Empirical Methods in Natural Language Processing*, pp. 6121–6133, 2020.

Kaiyuan Liu, Shaotian Yan, Rui Miao, Bing Wang, Chen Shen, Jun Zhang, and Jieping Ye. Where did this sentence come from? tracing provenance in LLM reasoning distillation. In *International Conference on Learning Representations*, 2026.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Annual Conference on Neural Information Processing Systems*, pp. 6379–6390, 2017.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adámek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Annual Meeting of the Association for Computational Linguistics*, pp. 1773–1781, 2023.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925, 2025.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, and Laura Toni. A survey of temporal credit assignment in deep reinforcement learning. *Transactions on Machine Learning Research*, 2024, 2024.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

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

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *CoRR*, abs/2604.00626, 2026.

Bing Wang, Rui Miao, Chen Shen, Shaotian Yan, Kaiyuan Liu, Ximing Li, Xiaosong Yuan, Sinan Fan, Jun Zhang, and Jieping Ye. On the step length confounding in LLM reasoning data selection. *CoRR*, abs/2604.06834, 2026.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. In *Annual Conference on Neural Information Processing Systems*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Annual Conference on Neural Information Processing Systems*, 2022.

Xiaojun Wu, Xiaoguang Jiang, Huiyang Li, Jucai Zhai, Dengfeng Liu, Qiaobo Hao, Huang Liu, Zhiguo Yang, Ji Xie, Ninglun Gu, Jin Yang, Kailai Zhang, Yelun Bao, and Jun Wang. Beyond scaling law: A data-efficient distillation framework for reasoning. *CoRR*, abs/2508.09883, 2025.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Zhihao Zhang, Wenxiang Chen, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Agentprm: Process reward models for LLM agents via step-wise promise and progress. In *ACM Web Conference*, pp. 4184–4195, 2026.

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

Shaotian Yan, Kaiyuan Liu, Chen Shen, Bing Wang, Sinan Fan, Jun Zhang, Yue Wu, Zheng Wang, and Jieping Ye. Distribution-aligned sequence distillation for superior long-cot reasoning. *CoRR*, abs/2601.09088, 2026.

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

Yongjin Yang, Yinghui He, Jiarui Liu, and Zhijing Jin. Making complex reasoning student-friendly: A hybrid LLM-to-SLM distillation framework. In *The 1st Workshop on Scaling Post-training for LLMs*, 2026.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. In *Conference on Language Modeling*, 2025.

Xiaosong Yuan, Chen Shen, Shaotian Yan, Xiaofeng Zhang, Liang Xie, Wenxiao Wang, Renchu Guan, Ying Wang, and Jieping Ye. Instance-adaptive zero-shot chain-of-thought prompting. In *Annual Conference on Neural Information Processing Systems*, volume 37, pp. 125469–125486, 2024.

Xiaosong Yuan, Chen Shen, Shaotian Yan, Kaiyuan Liu, Xiaofeng Zhang, Sinan Fan, Liang Xie, Wenxiao Wang, Renchu Guan, Ying Wang, and Jieping Ye. Differential fine-tuning large language models towards better diverse reasoning abilities. In *International Conference on Learning Representations*, 2026.

Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In *AAAI Conference on Artificial Intelligence*, pp. 2891–2897, 2017.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911, 2023.

A THEORETICAL ANALYSIS

In this section, we provide a formal and mathematical derivation of the dual exposure biases in LLM reasoning distillation, and prove how our proposed MOTAB framework theoretically resolves them. Specifically, we answer the following three basic questions:

RQ1. Why does off-policy distillation fail during inference?

RQ2. Why does on-policy distillation corrupt the teacher's supervision?

RQ3. Why does a mixture of off-policy and on-policy distributions fail to resolve this dilemma?

RQ4. How does MOTAB formally guarantee distribution coverage and supervision validity?

LLM reasoning is a strict full-history-dependent auto-regressive process. Formally, let $\mathcal{V}$ denote the vocabulary space. For an initial question $\mathbf{q}$, a reasoning trajectory of length L is a sequence of reasoning steps $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L)$, and each step is a sequence of tokens $\mathbf{y}_l = (y_{l,1}, y_{l,1}, \ldots, y_{l,|\mathbf{y}_l|})$. At any step l, the context is the entire historical prefix $\mathbf{s}_{t-1} = [\mathbf{q}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{l-1}]$. A policy LLM $\boldsymbol{\pi}$ defines a conditional probability distribution over the next token $\boldsymbol{\pi}(y_{l,i} \mid \mathbf{s}_{t-1}, \mathbf{y}_{l,<i})$.

The probability of a reasoning trajectory under policy $\boldsymbol{\pi}$ is given by the chain rule of probability:

$$P_{\boldsymbol{\pi}}(\mathcal{Y} \mid \mathbf{q}) = \prod_{l=1}^{L} \prod_{i=1}^{|\mathbf{y}_l|} \boldsymbol{\pi}(y_{l,i} \mid \mathbf{s}_{l-1}, \mathbf{y}_{l,<i}).$$

Let $d^{\boldsymbol{\pi}}$ denote the marginal distribution of all possible historical prefixes $\mathbf{s}$ generated by iteratively unrolling policy $\boldsymbol{\pi}$.

A.1 RQ1: EXPOSURE BIAS IN OFF-POLICY DISTILLATION

Off-policy distillation minimizes the single-step Kullback-Leibler divergence exclusively over the teacher's historical prefixes. We assume the student model $\boldsymbol{\pi}_S$ is optimized such that its expected single-step error on the teacher's manifold p_T is bounded by ϵ:

$$\mathbb{E}_{\mathbf{s}_{l-1} \sim p_T} \left[D_{\mathrm{TV}}\big(\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l-1}) \parallel \boldsymbol{\pi}_S(\cdot \mid \mathbf{s}_{l-1})\big) \right] \leq \epsilon, \quad \forall l \in [1, L]. \tag{12}$$

Theorem 1. *Under the training condition in Eq. 12, during free-form inference where prefixes are generated by the student p_S, the sequence-level divergence $D_{\mathrm{TV}}\big(p_T(\mathbf{s}_T) \parallel p_S(\mathbf{s}_T)\big)$ scales at an $\mathcal{O}(L^2)$ rate in the worst case, demonstrating the exposure bias.*

Proof. Using the chain rule of probability, the prefix distributions can be recursively decomposed: $p(\mathbf{s}_l) = p(\mathbf{s}_{l-1})\boldsymbol{\pi}(\mathbf{y}_l \mid \mathbf{s}_{l-1})$, where $\boldsymbol{\pi}(\mathbf{y}_l \mid \mathbf{s}_{l-1}) = \prod_{i=1}^{|\mathbf{y}_l|} \boldsymbol{\pi}(y_{l,i} \mid \mathbf{s}_{l-1}, \mathbf{y}_{l,<i})$. By the telescoping property of the total variation distance for sequence distributions, the trajectory-level divergence at the final step L can be expanded as the sum of step-level conditional divergences evaluated over the *student's* prefix distribution:

$$D_{\mathrm{TV}}\big(p_T(\mathbf{s}_L) \parallel p_S(\mathbf{s}_L)\big) \leq \sum_{l=1}^{L} \mathbb{E}_{\mathbf{s}_{l-1} \sim p_S} \left[D_{\mathrm{TV}}\big(\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l-1}) \parallel \boldsymbol{\pi}_S(\cdot \mid \mathbf{s}_{l-1})\big) \right] = \sum_{l=1}^{L} \mathcal{L}_{\mathrm{FR}}^{(l)}. \tag{13}$$

Here, $\mathcal{L}_{\mathrm{FR}}^{(l)}$ denotes the free-running expected divergence. Notice that $\mathcal{L}_{\mathrm{FR}}^{(l)}$ is evaluated over prefixes $\mathbf{s}_{l-1} \sim p_S$, while the training objective in Eq. 12 assumes $\mathbf{s}_{l-1} \sim p_T$. To formally relate them, we apply a change-of-measure expansion to $\mathcal{L}_{\mathrm{FR}}^{(l)}$:

$$\begin{aligned} \mathcal{L}_{\mathrm{FR}}^{(l)} &= \sum\nolimits_{\mathbf{s}_{l-1}} p_S(\mathbf{s}_{l-1}) D_{\mathrm{TV}}\big(\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l-1}) \parallel \boldsymbol{\pi}_S(\cdot \mid \mathbf{s}_{l-1})\big) \\ &= \mathbb{E}_{\mathbf{s}_{l-1} \sim p_T} \left[\frac{p_S(\mathbf{s}_{l-1})}{p_T(\mathbf{s}_{l-1})} D_{\mathrm{TV}}\big(\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l-1}) \parallel \boldsymbol{\pi}_S(\cdot \mid \mathbf{s}_{l-1})\big) \right]. \end{aligned} \tag{14}$$

Because generation is strictly history-dependent, if the student makes an early mistake at step $k < l$, the subsequent prefix $\mathbf{s}_{l-1}$ drifts out of the teacher's data manifold. Mathematically, this implies $p_T(\mathbf{s}_{l-1}) \to 0$, causing the density ratio $\frac{p_S(\mathbf{s}_{l-1})}{p_T(\mathbf{s}_{l-1})} \to \infty$. Consequently, for step l, the expected divergence $\mathcal{L}_{\mathrm{FR}}^{(l)}$ is no longer bounded by ϵ. Instead, in the worst-case out-of-distribution scenario, the divergence reaches its theoretical maximum $D_{\mathrm{TV}} \to 1$. Substituting this cascading error into Eq. 13 yields an expected compounding sequence error that grows quadratically $\mathcal{O}(L^2\epsilon)$, effectively proving the *exposure bias*. □

A.2 RQ2: Reversed Exposure Bias in On-Policy Distillation

To prevent the density ratio explosion seen in Eq. 14, on-policy distillation enforces the student to learn directly from its own rollout distribution: $\mathbb{E}_{\mathbf{s}_{l-1}\sim p_S}\left[D_{\mathrm{KL}}(\boldsymbol{\pi}_S \parallel \boldsymbol{\pi}_T)\right]$.

Theorem 2. *Let $\mathbf{s}_{l-1} \sim p_S$ be a sub-optimal reasoning history such that it deviates fundamentally from logical coherence. Conditioning the teacher on this state causes its predictive distribution over the next step to collapse to a uniform prior, maximizing entropy. Consequently, the teacher merely provides a suppressive signal rather than constructive guidance, destroying valid supervision.*

Proof. By Bayes' theorem, the teacher's conditional probability for the next reasoning step $\mathbf{y}_l$ given the history $\mathbf{s}_{l-1}$ is defined by the joint probability of the sequence:

$$\boldsymbol{\pi}_T(\mathbf{y}_l \mid \mathbf{s}_{l-1}) = \frac{p_T(\mathbf{s}_{l-1}, \mathbf{y}_l)}{p_T(\mathbf{s}_{l-1})}. \tag{15}$$

In practice, LLM distributions are smoothed by a prior to prevent absolute zero probabilities. Over a reasoning step of length $K = |\mathbf{y}_l|$, the uniform prior is $U(\mathbf{y}_l) = \frac{1}{|\mathcal{V}|^K}$. We approximate the implicit generative distribution as a mixture with a smoothing factor $\lambda \in (0, 1)$:

$$\tilde{p}_T(\mathbf{s}_{l-1}, \mathbf{y}_l) = (1-\lambda)p_T(\mathbf{s}_{l-1}, \mathbf{y}_l) + \lambda\frac{p_T(\mathbf{s}_{l-1})}{|\mathcal{V}|^K}.$$

When $\mathbf{s}_{l-1}$ is a sub-optimal prefix generated by the student, the true logical continuation density $p_T(\mathbf{s}_{l-1}, \mathbf{y}_l) \to 0$. Substituting this into the conditional distribution yields:

$$\boldsymbol{\pi}_T(\mathbf{y}_l \mid \mathbf{s}_{l-1}) \approx \frac{(1-\lambda)\cdot 0 + \lambda\frac{p_T(\mathbf{s}_{l-1})}{|\mathcal{V}|^K}}{p_T(\mathbf{s}_{l-1})} = \frac{\lambda}{|\mathcal{V}|^K} \propto \frac{1}{|\mathcal{V}|^K}.$$

This indicates that in the sub-optimal regime, the teacher lacks valid priors, and its conditional distribution degenerates toward a uniform distribution over the sequence space. The conditional entropy of the teacher for step l then diverges to:

$$\mathcal{H}(\boldsymbol{\pi}_T \mid \mathbf{s}_{l-1}) = -\sum\nolimits_{\mathbf{y}_l} \boldsymbol{\pi}_T(\mathbf{y}_l \mid \mathbf{s}_{l-1}) \log \boldsymbol{\pi}_T(\mathbf{y}_l \mid \mathbf{s}_{l-1}) \longrightarrow K \log |\mathcal{V}|.$$

Minimizing $D_{\mathrm{KL}}(\boldsymbol{\pi}_S \parallel \boldsymbol{\pi}_T)$ against this maximum-entropy distribution implies that the teacher assigns uniformly low confidence to all possible next steps. Rather than providing active, constructive guidance (i.e., a sharp distribution peaking at the correct logical deduction), the teacher acts purely as a penalty mechanism, suppressing the student's generation across the board. This forces the student model to unlearn deterministic logical reasoning, formally defining the *reversed exposure bias*. □

Beyond the theoretical reversed exposure bias, relying on strict on-policy supervision introduces severe practical and computational bottlenecks. First, calculating the token-level KL divergence fundamentally requires the teacher and student models to share the exact same tokenizer vocabulary $\mathcal{V}$ (Yan et al., 2026). This constraint severely restricts the flexibility of cross-architecture distillation. Second, on-policy distillation forces the massive teacher model to synchronously evaluate or generate target distributions conditioned on the student's real-time rollouts. This step-by-step dependency introduces prohibitive computational latency and memory overhead, rendering large-scale on-policy distillation highly inefficient in practice.

A.3 RQ3: The Failure of Naive Distribution Mixing

Given the complementary flaws of off-policy and on-policy distillation, an intuitive baseline would be to optimize the student over a mixture distribution, *e.g.,* SKD (Xu et al., 2025), or combine the forward and reversed KL divergences (Jung et al., 2025). Let $p_{\text{mix}} = \alpha p_S + (1-\alpha)p_T$ be a linearly mixed historical context distribution with a mixture coefficient $\alpha \in (0, 1)$. The expected distillation objective becomes:

$$\mathcal{J}_{\text{mix}}(\boldsymbol{\pi}_S) = \mathbb{E}_{\mathbf{s}_{l-1}\sim p_{\text{mix}}}\left[D_{\mathrm{KL}}\big(\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l-1}) \parallel \boldsymbol{\pi}_S(\cdot \mid \mathbf{s}_{l-1})\big)\right].$$

Theorem 3 (zero-sum trade-off). *Optimizing over a naive mixture distribution p_{mix} cannot concurrently resolve the dual exposure biases. The expected inference error is strictly lower-bounded by*

a convex combination of the forward error cascade (Theorem 1) and the reversed entropy collapse (Theorem 2). Consequently, any choice of α merely trades one form of bias for another.

Proof. By expanding the definition of the expectation over the discrete state space, we substitute the definition of the mixed density $p_{\text{mix}}(\mathbf{s}_{l-1}) = \alpha p_S(\mathbf{s}_{l-1}) + (1-\alpha)p_T(\mathbf{s}_{l-1})$ into the objective:

$$\begin{aligned}\mathcal{J}_{\text{mix}} &= \sum\nolimits_{\mathbf{s}_{l-1}} p_{\text{mix}}(\mathbf{s}_{l-1}) D_{\text{KL}}\big(\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l-1}) \parallel \boldsymbol{\pi}_S(\cdot \mid \mathbf{s}_{l-1})\big) \\ &= \sum\nolimits_{\mathbf{s}_{l-1}} \big[\alpha p_S(\mathbf{s}_{l-1}) + (1-\alpha)p_T(\mathbf{s}_{l-1})\big] D_{\text{KL}}\big(\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l-1}) \parallel \boldsymbol{\pi}_S(\cdot \mid \mathbf{s}_{l-1})\big).\end{aligned}$$

By distributing the summation, the mixture objective $\mathcal{J}_{\text{mix}}$ is explicitly decoupled into two independent expectations:

$$\mathcal{J}_{\text{mix}} = \alpha \underbrace{\mathbb{E}_{\mathbf{s}_{l-1}\sim p_S}\left[D_{\text{KL}}\big(\boldsymbol{\pi}_T \parallel \boldsymbol{\pi}_S\big)\right]}_{\text{On-policy constraint}} + (1-\alpha) \underbrace{\mathbb{E}_{\mathbf{s}_{l-1}\sim p_T}\left[D_{\text{KL}}\big(\boldsymbol{\pi}_T \parallel \boldsymbol{\pi}_S\big)\right]}_{\text{Off-policy constraint}}.$$

This exact decoupling reveals a structural contradiction between the two terms.

First, consider the α component. For historical prefixes sampled directly from the student's rollout ($\mathbf{s}_{l-1} \sim p_S$), the trajectories inevitably contain out-of-distribution trajectories where $p_T(\mathbf{s}_{l-1}) \to 0$. As rigorously established in Theorem 2, the teacher's conditional entropy diverges. Thus, minimizing the KL divergence forces the student to fit a maximum-entropy distribution, inflicting a supervision penalty proportional to $\mathcal{H}(\boldsymbol{\pi}_T \mid \mathbf{s}_{l-1}) \to K \log |\mathcal{V}|$.

Second, consider the $(1-\alpha)$ component. To avoid the aforementioned entropy collapse, one must heavily weight the off-policy data by setting $\alpha \to 0$. However, as proven in Theorem 1, optimizing exclusively over p_T leaves the student vulnerable to compounding sequence-level errors during test-time free-running generation. This inference error cascade is mathematically bounded by the trajectory-level divergence $D_{\text{TV}}\big(p_T(\mathbf{s}_L) \parallel p_S(\mathbf{s}_L)\big) \geq \Omega(L^2\epsilon)$.

Formally, we define the total sub-optimality risk functional $\mathcal{R}_{\text{total}}$ as the weighted sum of the expected training supervision corruption and the expected inference sequence deviation. The risk is strictly bounded from below:

$$\mathcal{R}_{\text{total}} \geq \alpha \cdot \mathbb{E}_{\mathbf{s}_{l-1}\sim p_S}\big[\mathcal{H}(\boldsymbol{\pi}_T \mid \mathbf{s}_{l-1})\big] + (1-\alpha) \cdot D_{\text{TV}}\big(p_T(\mathbf{s}_L) \parallel p_S(\mathbf{s}_L)\big).$$

Because the two penalty terms are mutually antagonistic, lowering α to secure valid supervision inherently amplifies the sequential divergence, and raising α to cover the student's distribution actively corrupts the reasoning targets. Thus, the structural deadlock remains unbroken. The theoretical oracle requires valid supervision *conditioned on* sub-optimal contexts, a property that naive distribution mixing mathematically fails to provide. □

A.4 RQ4: Theoretical Guarantees of MOTAB

We formally prove how MOTAB satisfies both the *coverage* and *validity* conditions. Let $p^{\gamma}_{\boldsymbol{\pi}_S}$ denote the truncated distribution of prefixes generated under the MOTAB adaptive boundary mechanism.

Theorem 4 (coverage). *The truncated distribution $p^{\gamma}_{\boldsymbol{\pi}_S}$ strictly covers the authentic student distribution p_S up to a tightly bounded divergence, eliminating the exposure bias without exposing the teacher to fatal logical collapses.*

Proof. The synthesized distribution $p^{\gamma}_{\boldsymbol{\pi}_S}$ matches p_S until a truncation occurs, triggered when $V_T(\mathbf{s}_{l-1}, \mathbf{y}^S_l) < \gamma(\mathbf{s}_{l-1})$. The distance between the true student distribution and the truncated distribution is bounded by the expected cumulative probability of the truncation over the L steps:

$$D_{\text{TV}}(p_S \parallel p^{\gamma}_{\boldsymbol{\pi}_S}) \leq \sum\nolimits_{l=1}^{L} \mathbb{P}_{\mathbf{s}_{l-1}\sim p_S, \mathbf{y}_l \sim \boldsymbol{\pi}_S}\left(V_T(\mathbf{s}_{l-1}, \mathbf{y}_l) < \gamma(\mathbf{s}_{l-1})\right). \tag{16}$$

Recall our adaptive boundary formulation:

$$\gamma(\mathbf{s}_{l-1}) = \gamma_0 \exp\big(-\alpha \mathcal{H}(\boldsymbol{\pi}_T \mid \mathbf{s}_{l-1})\big).$$

When the teacher identifies multiple valid reasoning paths, *i.e.,* high entropy $\mathcal{H}$, $\gamma(\mathbf{s}_{l-1})$ exponentially decays. This ensures the condition $V_T < \gamma$ is *strictly conservative* and triggers only when $V_T \to 0$ (an irrecoverable logical error), rather than for stylistic variances. Since $\gamma(\mathbf{s}_{l-1}) \leq \gamma_0$, the total

variation in Eq. 16 is rigorously constrained by the base scalar γ_0. Therefore, $p^{\gamma}_{\boldsymbol{\pi}_S} \approx p_S$ across recoverable reasoning manifolds, verifying the *coverage condition*. □

Theorem 5 (validity). MOTAB *unconditionally prevents the entropy collapse described in Theorem 2, providing valid oracle supervision and fully circumventing the reversed exposure bias.*

Proof. In MOTAB, upon detecting a boundary breach at step l, the framework strictly avoids querying $\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_l)$. Instead, credit assignment backtracks to a previous step l^*, and the off-policy target $\mathcal{Y}^T$ is sampled recursively as:

$$\mathcal{Y}^T_{l^*:L} \sim \prod_{\tau=l^*}^{L} \boldsymbol{\pi}_T(\mathbf{y}_\tau \mid \mathbf{s}^T_{\tau-1}), \quad \text{where } \mathbf{s}^T_{l^*-1} = \mathbf{s}_{l^*-1}.$$

Because the state $\mathbf{s}_{l^*-1}$ was verified as safe prior to backtracking, it satisfies the boundary constraint $V_T(\mathbf{s}_{l^*-2}, \mathbf{y}_{l^*-1}) \geq \gamma(\mathbf{s}_{l^*-2})$. This guarantees a high transition likelihood, ensuring the root context remains strictly within the teacher's dense data manifold:

$$p_T(\mathbf{s}_{l^*-1}) \geq \delta \gg 0.$$

Following the derivation in Eq. 15, because $p_T(\mathbf{s}_{l^*-1})$ is substantially strictly positive, the conditional distribution $\boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l^*-1})$ remains sharp and highly confident. The step-level entropy $\mathcal{H}(\boldsymbol{\pi}_T \mid \mathbf{s}_{l^*-1})$ is thus bounded securely away from the uniform maximum $K \log |\mathcal{V}|$.

Finally, the asynchronous stitching objective constructs the loss over the `[REV]` semantic transition:

$$\mathcal{L}_{\text{stitch}} = \mathbb{E}_{\mathcal{Y}^T}\left[-\log \boldsymbol{\pi}_S\big(\texttt{[REV]} \oplus \mathcal{Y}^T \mid \mathbf{s}_l^{\text{flawed}}\big)\right].$$

This isolates the teacher's generation process from the flawed context $\mathbf{s}_l^{\text{flawed}}$. The teacher operates purely on the pristine root $\mathbf{s}_{l^*-1}$, while the student learns a conditional mapping from its own flawed exploration to a pristine corrective sequence via the `[REV]` bottleneck. Consequently, the teacher's supervision signal remains unconditionally valid, neutralizing the reversed exposure bias. □

B ALGORITHMIC PIPELINE OF MOTAB

In this section, we present the detailed algorithmic procedure of our proposed data synthesis framework, MOTAB. The process is iteratively executed over a set of unlabelled reasoning questions to construct a refined supervised fine-tuning dataset.

As outlined in Alg. 1, the pipeline requires a teacher policy $\boldsymbol{\pi}_T$, a student policy $\boldsymbol{\pi}_S$, and an initial set of questions $\mathcal{Q}$. For each question, the student auto-regressively generates reasoning steps while the teacher synchronously evaluates the transition likelihood V_T and context entropy $\mathcal{H}$. If the generation remains within the adaptive safety boundary γ, the step is accepted, preserving the student's authentic on-policy distribution. Once a critical logical drift is detected ($V_T < \gamma$), the algorithm triggers the credit assignment mechanism to trace back the temporal difference errors δ_k. It identifies the optimal safe bifurcation state $\mathbf{s}_{l^*-1}$, prompts the teacher to generate a pristine correction $\mathcal{Y}^T_{l^*:T}$, and stitches the flawed exploration with the oracle correction using a specific `[REV]` token. The resulting trajectories explicitly empower the student model with both robustness to its own stylistic variances and the capability for logical self-correction.

C COMPUTATIONAL COMPLEXITY ANALYSIS

In addition to resolving the theoretical dual exposure biases, MOTAB introduces significant computational advantages during the data synthesis pipeline. In this section, we formally analyze the time complexity of MOTAB and compare it against the SKD baseline introduced in Sec. 2.4.

To formulate the computational overhead, let Λ denote the total token length of a reasoning trajectory. For an LLM parameterized by $\boldsymbol{\theta}$, we distinguish between two primary operations: (1) *Auto-regressive generation* ($\mathcal{O}_{\text{gen}}(\boldsymbol{\theta}, \Lambda)$), which requires sequential decoding and saving KV cache, which is memory-bandwidth bound; (2) *Parallel verification and forward pass* ($\mathcal{O}_{\text{vfy}}(\boldsymbol{\theta}, \Lambda)$), which computes logits over a given sequence in parallel. Due to parallelized KV-caching and matrix multiplications, verification is significantly faster than generation, *i.e.,*, $\mathcal{O}_{\text{vfy}} \ll \mathcal{O}_{\text{gen}}$.

Complexity of SKD. In standard SKD pipelines, the teacher model $\boldsymbol{\pi}_T$ continuously monitors the student $\boldsymbol{\pi}_S$ at the token or granular step level. Whenever the student's logic slightly deviates, SKD

Algorithm 1 Complete pipeline of the data synthesis via MOTAB

Input: Question set $\mathcal{Q}$, Teacher policy $\boldsymbol{\pi}_T$, Student policy $\boldsymbol{\pi}_S$, Base threshold γ_0, Scaling factor α.
Output: Synthesized supervised fine-tuning dataset $\mathcal{D}_{\text{SFT}}$.

Initialize dataset $\mathcal{D}_{\text{SFT}} \leftarrow \varnothing$;
for each question $\mathbf{q} \in \mathcal{Q}$ **do**
  Initialize context $\mathbf{s}_0 \leftarrow \mathbf{q}$, step index $l \leftarrow 1$;
  Flag $is_safe \leftarrow$ True;
  while is_safe **and** not reached terminal state **do**
    Sample student reasoning step: $\mathbf{y}_l^S \sim \boldsymbol{\pi}_S(\cdot \mid \mathbf{s}_{l-1})$;
    Compute step value $V_T(\mathbf{s}_{l-1}, \mathbf{y}_l^S)$ via Eq. 5;
    Compute teacher entropy $\mathcal{H}(\boldsymbol{\pi}_T \mid \mathbf{s}_{l-1})$;
    Calculate adaptive boundary $\gamma(\mathbf{s}_{l-1}) \leftarrow \gamma_0 \cdot \exp(-\alpha \cdot \mathcal{H})$;
    if $V_T(\mathbf{s}_{l-1}, \mathbf{y}_l^S) \geq \gamma(\mathbf{s}_{l-1})$ **then**
      ▷ *Safe step, continue on-policy unrolling*
      Update context $\mathbf{s}_l \leftarrow [\mathbf{s}_{l-1}, \mathbf{y}_t^S]$;
      $l \leftarrow l + 1$;
    else
      ▷ *Boundary breached, trigger backtracking*
      $is_safe \leftarrow$ False;
      for $k = 1$ **to** t **do**
        Compute temporal difference error $\delta_k = V_T(\mathbf{s}_{k-1}, \mathbf{y}_k^S) - V_T(\mathbf{s}_{k-2}, \mathbf{y}_{k-1}^S)$;
      end for
      Identify optimal safe root $l^* \leftarrow \arg\min_{k \leq l}(\delta_k)$ s.t. prior state is safe;
      ▷ *Off-policy correction and stitching*
      Sample corrective trajectory $\mathcal{Y}_{l^*:T}^T \sim \boldsymbol{\pi}_T(\cdot \mid \mathbf{s}_{l^*-1})$;
      Construct stitched trajectory $\boldsymbol{\tau} \leftarrow [\mathbf{q}, \mathcal{Y}_{1:l}^S] \oplus \texttt{[REV]} \oplus \mathcal{Y}_{l^*:T}^T$;
      $\mathcal{D}_{\text{SFT}} \leftarrow \mathcal{D}_{\text{SFT}} \cup \{\boldsymbol{\tau}\}$;
    end if
  end while
  if is_safe **then**
    ▷ *Successfully finished without correction*
    $\mathcal{D}_{\text{SFT}} \leftarrow \mathcal{D}_{\text{SFT}} \cup \{\mathbf{s}_l\}$;
  end if
end for
return $\mathcal{D}_{\text{SFT}}$

forces the teacher to intervene and generate a replacement. Assuming an intervention rate $\alpha \in (0, 1)$, the total time complexity per trajectory is:

$$\mathcal{O}_{\text{SKD}} \approx \mathcal{O}_{\text{gen}}(\boldsymbol{\theta}_S, \Lambda) + \mathcal{O}_{\text{vfy}}(\boldsymbol{\theta}_T, \Lambda) + \alpha \cdot \mathcal{O}_{\text{gen}}(\boldsymbol{\theta}_T, \Lambda).$$

Because SKD enforces strict alignment, the intervention rate α is inherently high. The frequent context switching and repeated auto-regressive generation by the massively parameterized teacher ($\boldsymbol{\theta}_T > \boldsymbol{\theta}_S$) result in a severe computational bottleneck.

Complexity of MOTAB. Conversely, MOTAB isolates operations to minimize the teacher's generation burden. Assuming a boundary breach occurs at step l (corresponding to token length Λ_{fail}), and the safe point is identified at step l^*, the complexity of MOTAB is decomposed as follows:

- **Phase 1 (monitoring):** The student freely generates the prefix, and the teacher evaluates the steps using Eq. 5 and 6. Because the student's text is already generated, the teacher performs highly efficient parallel forward passes. The cost is $\mathcal{O}_{\text{gen}}(\boldsymbol{\theta}_S, \Lambda_{\text{fail}}) + \mathcal{O}_{\text{vfy}}(\boldsymbol{\theta}_T, \Lambda_{\text{fail}})$.
- **Phase 2 (backtracking):** Identifying the root cause l^* relies on computing the temporal difference error δ_k. Since the step-wise values V_T are already cached during Phase 1, this operation requires zero neural network passes. It is purely algebraic, incurring a negligible complexity of $\mathcal{O}(1)$.
- **Phase 3 (stitching):** The teacher generates the corrective suffix of length Λ_{correct} from the root cause. The cost is $\mathcal{O}_{\text{gen}}(\boldsymbol{\theta}_T, \Lambda_{\text{correct}})$.

Summing these phases, the total time complexity for MOTAB is:

$$\mathcal{O}_{\text{MOTAB}} \approx \mathcal{O}_{\text{gen}}(\boldsymbol{\theta}_S, \Lambda_{\text{fail}}) + \mathcal{O}_{\text{vfy}}(\boldsymbol{\theta}_T, \Lambda_{\text{fail}}) + \mathcal{O}_{\text{gen}}(\boldsymbol{\theta}_T, \Lambda_{\text{correct}}). \tag{17}$$

The fundamental efficiency advantage of MOTAB lies in the adaptive boundary. By tolerating sub-optimal explorations, MOTAB drastically reduces the necessity for teacher interventions, shifting the teacher's computational workload from expensive auto-regressive generation ($\mathcal{O}_{\text{gen}}$) to highly efficient parallel verification ($\mathcal{O}_{\text{vfy}}$). Furthermore, unlike external verification methods, *e.g.,* process reward models, that require training an additional critic network, our backtracking by temporal difference error reuses the teacher's innate confidence, effectively achieving precise credit assignment at zero additional inference cost.

Empirical computational budgets. To validate the practical efficiency of MOTAB, we measure the real-world data synthesis speed on a hardware setup consisting of 8×A800 GPUs. We compare the average generation time of MOTAB against the SKD baseline across three distinct hyperparameter configurations (which control the intervention rates by the teacher). Empirically, both methods demonstrate comparable synthesis efficiency, each requiring approximately 3 to 4 hours to generate 1k reasoning trajectories on average. This result confirms that while MOTAB introduces sophisticated mechanisms, *e.g.,* backtracking with temporal difference error and trajectory stitching, to theoretically resolve the dual exposure biases, it does not incur noticeable empirical overhead in practice.

D DETAILED EXPERIMENTAL SETTINGS

In this section, we provide a comprehensive overview of the experimental setups, including the specific datasets, model architectures, baseline methodologies, and implementation details.

D.1 TRAINING DATASETS

Our training phase leverages high-quality reasoning problem sets to facilitate effective knowledge distillation. Specifically, the training corpus is constructed from two primary sources:

- ***LIMO-v2*** (Ye et al., 2025): We utilize problem sets from the *LIMO-v2* repository[1], which is specifically curated to elicit deep mathematical reasoning capabilities. This dataset has undergone rigorous quality filtering and comprises 800 mathematical problems, and we generate five distinct responses for each problem.
- ***AceReason*** (Chen et al., 2025c): To introduce further diversity and scale to our training distribution, we randomly sample a subset of 10,000 problem instances from the *AceReason* repository[2]. For each problem, we generate one response.

D.2 EVALUATION BENCHMARKS

To rigorously assess the generalization and reasoning capability of our trained student models, we employ a suite of five challenging benchmarks, categorized into domain-specific mathematical reasoning and cross-domain capabilities:

- **Mathematical reasoning:** We evaluate our student models on *AIME2024* and *AIME2025*, which consist of highly complex, competition-level problems from the American Invitational Mathematics Examination. Each test set comprises 30 challenging mathematical problems. During evaluation, we generate 8 responses per problem and report the pass@k accuracy, where at least one correct response is identified among k samples, for $k \in \{8, 4, 1\}$. Additionally, we use *MATH500* (Lightman et al., 2024), a comprehensive benchmark encompassing various difficulty levels and subfields of mathematics. We evaluate the student models across varying difficulty levels, ranging from Level 1 to Level 5. The respective test sets consist of 43, 90, 105, 128, and 134 problems, for which we report accuracy scores individually.

[1] https://huggingface.co/datasets/GAIR/LIMO-v2.
[2] https://huggingface.co/datasets/nvidia/AceReason-1.1-SFT.

- **Cross-domain capabilities:** Our models are trained on two mathematical reasoning datasets. To ensure that our approach generalizes to other domains, *e.g.,* general instruction-following and diverse knowledge reasoning, we evaluate our models on the *IFEval* benchmark (Zhou et al., 2023), which assesses rigorous, verifiable instruction adherence. This benchmark comprises 541 prompts. We report both the widely used prompt-level-strict accuracy and the mean across all four metrics, including prompt-level-strict, inst-level-strict, prompt-level-loose, and inst-level-loose. Furthermore, we evaluate our models on *GPQA* (Rein et al., 2023), a challenging graduate-level benchmark covering physics, biology, and chemistry. This benchmark consists of 198 challenging scientific questions. Consistent with our evaluation protocol on the AIME datasets, we perform eight samplings per problem and report the pass@k metrics.

D.3 Model Architectures

We evaluate our method across LLMs of varying scales and architectural families.

Teacher LLMs. We employ *Qwen3-32B*[3] and *gpt-oss-120b*[4] as our teacher models. These large-scale models possess extensive parameter counts and demonstrate state-of-the-art proficiency in generating high-quality, accurate, and complex reasoning trajectories.

Student LLMs. To evaluate the efficacy and scalability of our method, we select *Qwen3-4B-Instruct*[5], *Qwen2.5-7B-Instruct*[6], and *Llama3.2-3B-Instruct*[7] as our student models. These models represent lightweight, efficient architectures (ranging from 3B to 7B parameters) that are highly suitable for real-world deployments requiring low-latency inference.

D.4 Baseline Methods

We compare our proposed approach against four representative baselines spanning standard fine-tuning and knowledge distillation paradigms.

- **Supervised fine-tuning (SFT):** The student model is directly fine-tuned on reasoning trajectories generated by the teacher, serving as a fundamental reference baseline.
- **Rejection sampling fine-tuning (RFT):** Multiple candidate trajectories are sampled from the student model; those leading to incorrect final answers are discarded, and the student is fine-tuned exclusively on the surviving correct rollouts.
- **Imitation-based knowledge distillation (ImitKD)** (Lin et al., 2020): An autoregressive imitation learning method in which, at each decoding step, the generation is stochastically sourced from either the teacher or the student.
- **Speculative knowledge distillation (SKD)** (Xu et al., 2025): A step-level, speculative distillation scheme. At each reasoning step l, the student first generates a candidate step $\mathbf{a}_l^S$, which is then evaluated by both the teacher and the student via their respective geometric mean token probabilities:

$$\bar{p}_M\left(\mathbf{a}_l^S \mid \mathbf{s}_{l-1}\right) = \exp\left(\frac{1}{|\mathbf{a}_l^S|}\sum_{i=1}^{|\mathbf{a}_l^S|}\log p_M(y_i \mid \mathbf{s}_{l-1}, \mathbf{y}_{<i})\right), \quad M \in \{S, T\}.$$

  The student's step is accepted if the teacher's likelihood sufficiently dominates the student's:

$$\bar{p}_T\left(\mathbf{a}_l^S \mid \mathbf{s}_{l-1}\right) \;>\; \beta \cdot \bar{p}_S\left(\mathbf{a}_l^S \mid \mathbf{s}_{l-1}\right),$$

  where $\beta \in (0, 1]$ is a margin factor controlling acceptance strictness. When this condition fails, the student's step is discarded, and the teacher regenerates a replacement, steering the trajectory back toward high-quality reasoning.

[3] https://huggingface.co/Qwen/Qwen3-32B.
[4] https://huggingface.co/openai/gpt-oss-120b.
[5] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.
[6] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.
[7] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

Meanwhile, we also compare several recent on-policy distillation methods, which are described as follows:

- **Top-k on-policy distillation (top-k OPD)** (Fu et al., 2026): This method mitigates the train-inference distribution mismatch by training the student model on its own self-generated sequences rather than a fixed dataset. It leverages token-level teacher probabilities as expert feedback on these student outputs, accommodating flexible divergence measures and seamless integration with reinforcement learning. Unlike GKD (Agarwal et al., 2024), which calculates the divergence over the entire vocabulary-level distribution, this method utilizes the top-k tokens to enhance computational efficiency.
- **Distribution-aligned sequence distillation (DASD)** (Yan et al., 2026): Designed to enhance sequence-level distillation for long chain-of-thought reasoning, this method integrates temperature-scheduled learning and divergence-aware sampling to better cover and align with the teacher's distribution. To explicitly combat exposure bias, it employs a mixed-policy strategy that corrects student-generated prefixes with targeted teacher continuations.

D.5 Implementation Details

Data generation. All models are served via SGLang. During the data generation pipeline of MOTAB, the student generates reasoning steps delimited by the stop sequence `".\n\n"`, with each step capped at 8,192 tokens and the full trajectory bounded at 32,768 tokens. The teacher's predictive entropy $\mathcal{H}(\boldsymbol{\pi}_T \mid \mathbf{s}_{t-1})$ is estimated from the renormalized top-20 next-token logprobs. Upon an unsafe point, the backtracking based on the temporal difference error identifies the point l^* (subject to the constraint that step $k-1$ was still above its adaptive threshold), after which the teacher generates a full corrective suffix from the pristine context s_{l^*-1}, and the final SFT trajectory is assembled as: student exploration + `"However,"` + teacher correction. We set $\gamma_0 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and temperature $\tau = 0.6$ for Qwen3-32B or 1.0 for gpt-oss-120b, and generate $N = 5$ trajectories per question in *LIMO-v2* dataset and $N = 1$ for *AceReason*.

Student model training. The student models are trained with full-parameter SFT using LLaMA-Factory on 8 GPUs via `torchrun`, with *DeepSpeed ZeRO-Stage 3*, *Flash Attention 2*, and the *Liger Kernel* enabled. To accelerate training, we employ sequence packing, where shorter training samples are concatenated into sequences of up to 32,768 tokens to maximize context utilization. We train for 6 epochs with a per-device batch size of 4 (effective global batch size 32), using AdamW with β_1=0.9, β_2=0.95, ϵ=10^{-8}, weight decay 0.1, and gradient clipping at norm 1.0. The learning rate is 5×10^{-5} with a cosine decay schedule (minimum 10^{-6}) and a linear warmup over 5% of total steps.

Implementation of on-policy distillation. We distill from a Qwen3-8B teacher into a Qwen3-1.7B student using the framework from Fu et al. (2026). The three components of our on-policy pipeline: *rollout generation*, *teacher forward pass*, and *student optimization*, execute in a time-sliced fashion on the same set of GPUs, with each component activated only during its designated phase to accommodate all three models within the available memory budget. At each training iteration, the rollout engine generates responses for a batch of 32 prompts, sampling 5 responses per prompt with temperature 0.6 and top-p 0.95. The maximum sequence length is set to 32,768 tokens. The teacher model then performs a forward pass over these rollout sequences, and the resulting logits are used to supervise the student in the subsequent training phase. The student is trained with a reverse-KL distillation objective, with a global batch size of 16 and a per-device micro-batch size of 1. We use the AdamW optimizer with a peak learning rate of 2×10^{-6} and a linear warmup over the first 5% of training steps. Training runs for 10 epochs with gradient checkpointing enabled throughout. A key memory challenge in on-policy distillation arises from materialising student and teacher logit tensors simultaneously: for a 32k-token response and a vocabulary of approximately 152k tokens, storing both logit matrices together with their gradients requires a number of GPU memory. To address this, top-k OPD (Fu et al., 2026) calculates the KL divergence using only the top-32 tokens in the vocabulary. Experimental results demonstrate that this approach significantly reduces the computational burden without compromising the overall performance.

E More Experimental Results

In this section, we provide more experimental results and analysis for MOTAB.

Table 8: Complete experimental results on more student models using Qwen3-32B as the teacher model. The bold results indicate the best scores.

Method	AIME24			AIME25			MATH500					IFEval	GPQA
	P@8	P@4	P@1	P@8	P@4	P@1	L1	L2	L3	L4	L5	Avg.	P@8
Student LLM: Qwen3-1.7B													
SFT	36.67	25.90	9.58	26.67	22.29	11.25	**90.70**	84.44	73.33	67.97	43.28	43.38	68.18
RFT	13.33	11.38	5.83	33.33	24.67	12.08	81.40	75.56	74.29	66.41	41.04	41.02	60.10
ImitKD	36.67	30.19	16.25	36.67	28.33	12.50	81.40	80.00	80.00	59.38	48.51	43.31	63.13
SKD ($\beta = 0.6$)	23.33	17.57	7.92	26.67	22.33	9.59	**90.70**	83.33	77.14	66.41	45.52	44.36	61.62
SKD ($\beta = 0.8$)	43.33	34.71	14.58	36.67	28.24	13.33	88.37	85.56	76.19	68.75	44.03	44.36	**73.74**
SKD ($\beta = 1.0$)	23.33	18.57	10.83	33.33	23.90	9.17	86.05	76.67	69.52	60.16	41.04	44.19	62.63
MOTAB (ours)	**50.00**	**41.38**	**22.08**	**40.00**	**32.05**	**14.58**	**90.70**	**93.33**	**82.86**	**78.91**	**58.21**	**46.80**	**73.74**
Student LLM: Llama3.2-3B-Instruct													
SFT	16.67	11.19	3.33	6.67	3.33	0.83	79.07	53.33	51.43	42.19	22.39	42.30	58.59
RFT	10.00	7.38	2.50	6.67	3.33	0.83	74.42	58.89	59.05	42.19	20.90	41.16	53.54
ImitKD	16.67	9.76	2.92	6.67	4.76	1.67	76.74	66.67	57.14	40.62	17.91	43.35	66.67
SKD ($\beta = 0.6$)	16.67	8.33	2.08	6.67	4.95	2.08	76.64	**68.89**	58.10	42.97	15.67	41.57	64.65
SKD ($\beta = 0.8$)	20.00	11.67	2.92	13.33	9.05	2.92	**88.37**	63.33	59.05	**50.00**	23.88	41.65	67.68
SKD ($\beta = 1.0$)	16.67	12.38	3.75	6.67	5.24	1.67	83.72	58.89	52.38	34.38	17.91	39.79	71.21
MOTAB (ours)	**23.33**	**13.76**	**5.00**	**16.67**	**11.67**	**5.83**	86.05	**68.89**	**60.95**	48.44	**28.36**	**47.67**	**73.74**
Student LLM: Qwen2.5-7B-Instruct													
SFT	23.33	19.76	**10.42**	16.67	10.00	4.58	83.72	74.44	69.52	60.16	44.78	34.78	69.70
RFT	16.67	12.86	5.42	20.00	14.00	5.00	88.37	82.22	80.95	58.59	35.07	34.27	56.06
ImitKD	20.00	16.19	5.83	**26.67**	16.57	5.83	74.42	70.00	72.38	57.81	35.82	35.32	70.71
SKD ($\beta = 0.6$)	23.33	19.24	10.00	23.33	17.86	**7.92**	83.72	83.33	75.24	65.62	33.58	36.18	70.71
SKD ($\beta = 0.8$)	26.67	22.05	10.00	**26.67**	18.81	7.50	88.37	81.11	73.33	**69.53**	35.07	33.85	76.77
SKD ($\beta = 1.0$)	20.00	13.33	4.17	23.33	17.14	6.67	86.05	77.78	72.38	64.06	41.79	34.63	77.27
MOTAB (ours)	**30.00**	**23.33**	**10.42**	**26.67**	**19.52**	**7.92**	**90.70**	**84.44**	**81.90**	67.19	**48.51**	**37.61**	**79.29**

E.1 EXPERIMENTS ON MORE STUDENT MODELS

To further substantiate the robustness and generalizability of our proposed framework, Table 8 provides comprehensive experimental results across a broader spectrum of student models, utilizing Qwen3-32B as the teacher. Specifically, we evaluate MOTAB on diverse model scales and architectures, including Qwen3-1.7B, Llama3.2-3B-Instruct, and Qwen2.5-7B-Instruct. Across all three student models, MOTAB demonstrates consistent and remarkable superiority over both standard off-policy (*e.g.,* SFT, RFT) and on-policy (*e.g.,* ImitKD, SKD) baselines. For instance, on the extremely lightweight Qwen3-1.7B model, MOTAB dramatically improves the AIME24 Pass@1 from 9.58% (SFT) to 22.08%, and elevates the most challenging MATH500 Level 5 accuracy to 58.21%, significantly eclipsing all baseline methods. Furthermore, the results on Llama3.2-3B highlight the efficacy of MOTAB in cross-architecture distillation scenarios. Despite the inherent stylistic and distribution discrepancies between the Qwen teacher and the Llama student, our method successfully bridges this gap, consistently pushing the performance upper bound. Notably, we observe that heuristic-based on-policy methods like SKD exhibit severe sensitivity to their manual thresholds β, with their optimal configurations fluctuating heavily depending on the specific student architecture and evaluation metric. In contrast, the entropy-driven adaptive boundary in MOTAB seamlessly adjusts to the student's intrinsic distribution, systematically yielding optimal performance without requiring exhaustive hyperparameter tuning. Beyond mathematical reasoning, MOTAB also delivers robust enhancements on general instruction following (IFEval) and advanced scientific knowledge (GPQA) tasks. This comprehensively verifies that the self-correction capabilities and error-recovery mechanisms instilled by our trajectory stitching and backtracking are domain-agnostic, ultimately fostering a more versatile and reliable reasoning policy across varying base models.

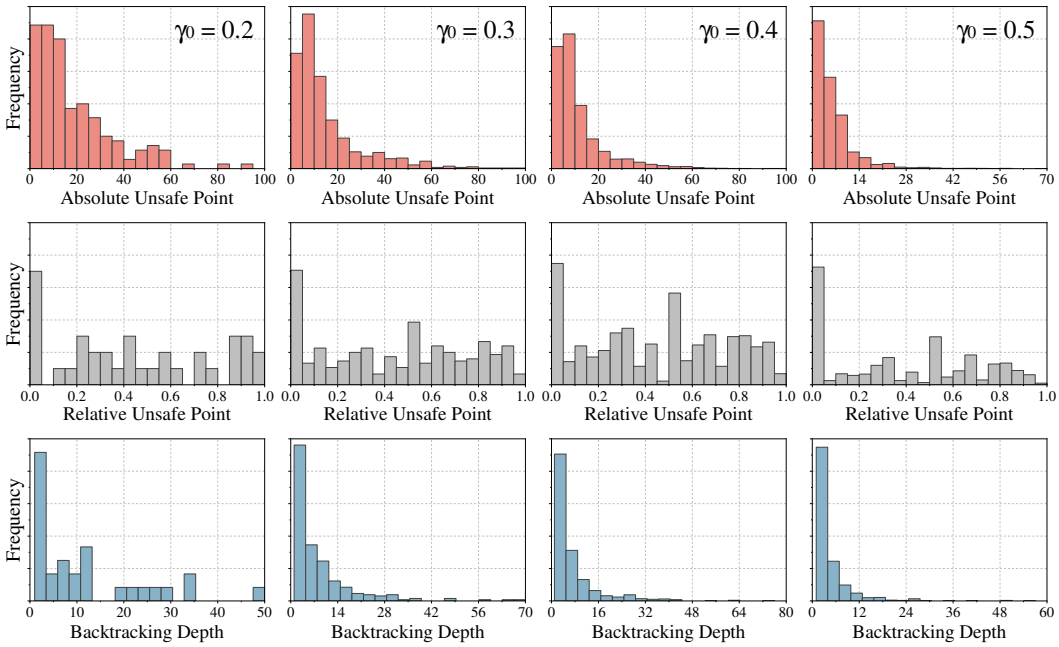

Figure 5: Distributions of absolute unsafe points (top row), relative unsafe points (middle row), and backtracking depths (bottom row) across varying base safety thresholds $\gamma_0 \in \{0.2, 0.3, 0.4, 0.5\}$.

E.2 Analysis of Backtracking Dynamics

To better understand the intervention behavior of MOTAB, we visualize the distributions of detected unsafe points and the corresponding backtracking depths across varying base thresholds ($\gamma_0 \in \{0.2, 0.3, 0.4, 0.5\}$). As illustrated in Fig. 5, while the absolute unsafe points (top row) tend to cluster in the earlier steps of generation, the relative unsafe points (middle row) exhibit a remarkably uniform distribution across the entire lifespan of the trajectories. This uniformity underscores that logical deviations are not strictly confined to the initial planning phases but can emerge dynamically at any stage of complex reasoning, thereby necessitating the continuous, step-by-step monitoring implemented in our framework. More importantly, the distribution of backtracking depths (bottom row) provides compelling empirical validation for our credit assignment mechanism. Across all threshold settings, a substantial proportion of interventions require rewinding by multiple steps (depth > 1) rather than merely correcting the immediately preceding step. This phenomenon corroborates our core hypothesis regarding the delayed manifestation of reasoning errors: a critical boundary breach detected at step t is frequently rooted in a flawed premise established several steps prior. By successfully tracing the temporal difference error back to the genuine safe bifurcation point, the credit assignment strategy ensures that the teacher model intervenes exclusively within a pristine, uncorrupted context, thereby fundamentally circumventing the reversed exposure bias.

E.3 Dynamics of Teacher Confidence and Safety Margins

To further investigate how the safety boundary influences the reasoning process, we analyze the dynamics of the teacher's step-level value V_T across different base thresholds ($\gamma_0 \in \{0.1, 0.3, 0.5\}$). As depicted in the left column of Fig. 6, V_T generally exhibits an upward trend as the step index increases, indicating that logical certainty solidifies as the trajectory approaches the final answer. Notably, as γ_0 increases, V_T converges to 1.0 much more rapidly. This demonstrates that a stricter boundary compellingly restricts the student to the teacher's high-confidence manifold, accelerating logical convergence but concurrently compressing the exploratory space. This observation is corroborated by the V_T distributions (center column), which become increasingly heavily skewed toward 1.0 at higher γ_0 values. Furthermore, we evaluate the safety margin, defined as $V_T - \gamma$, to compare trajectories that triggered backtracking against those that did not (right column). As γ_0 increases, the discriminative gap in the safety margin between the "Back" and "No Back" data progressively widens.

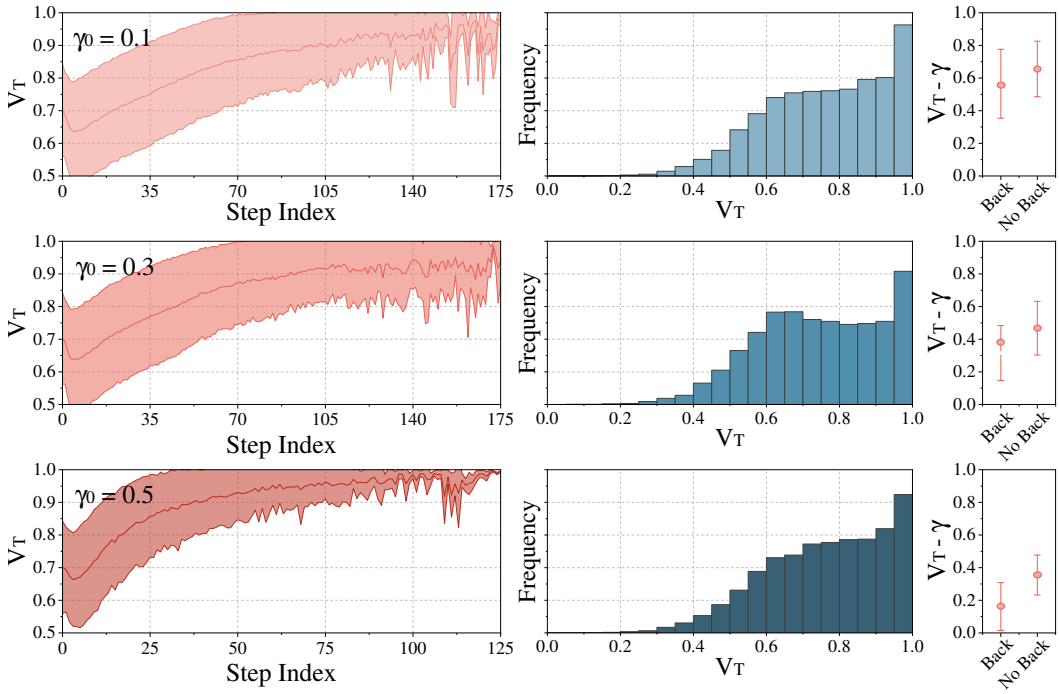

Figure 6: Evolution of the step-level value V_T across reasoning steps (left), the overall distribution of V_T (center), and the safety margins ($V_T - \gamma$) for backtracked versus non-backtracked trajectories (right), evaluated under different base thresholds $\gamma_0 \in \{0.1, 0.3, 0.5\}$.

This indicates that a tighter boundary enhances the framework's sensitivity, more decisively distinguishing fundamentally flawed deviations from valid logical progressions. However, synthesizing this with our main results reveals a crucial trade-off: while a larger γ_0 provides sharper error detection and faster confidence convergence, it prematurely truncates acceptable sub-optimal explorations, thereby hindering the mitigation of exposure bias. Consequently, a moderate boundary (*e.g.*, $\gamma_0 = 0.3$) is vital to balance discriminative error detection with adequate distributional coverage.

E.4 Qualitative Case Studies

To intuitively demonstrate how the MOTAB framework operates in practice, we present two qualitative case studies. In the first case, illustrated in Fig. 7, the student is tasked with an arithmetic word problem. The student correctly calculates an intermediate result of 33.75 women; however, because a human count must logically be an integer, the student subsequently spirals into confusion, generating an uncertain and rambling interpretation that violates the safety boundary ($0.42 < 0.49$). Instead of discarding the entire attempt or allowing the error to cascade, our framework backtracks to the precise point before the logical collapse. It then stitches the flawed exploration with a teacher-generated correction that rigorously formalizes the impossibility of the scenario. In the second case, presented in Fig. 8, the student attempts to find the minimum of a quadratic function within a parameterized interval $[-2, a]$. After successfully calculating the vertex, the student makes a premature and incomplete logical leap regarding the boundary condition, causing the value score to drop below the threshold ($0.41 < 0.48$). MOTAB traces the temporal difference error back to the vertex calculation and injects a comprehensive case-by-case analysis. By seamlessly stitching this pristine, teacher-guided correction to the student's flawed attempt using the semantic token "**However**", the resulting trajectory explicitly teaches the student how to recognize its own premature conclusions and pivot toward a more systematic analytical approach. Together, these cases illustrate how MOTAB effectively preserves the student's natural reasoning style while providing timely, high-quality interventions precisely when the intrinsic logic breaks down.

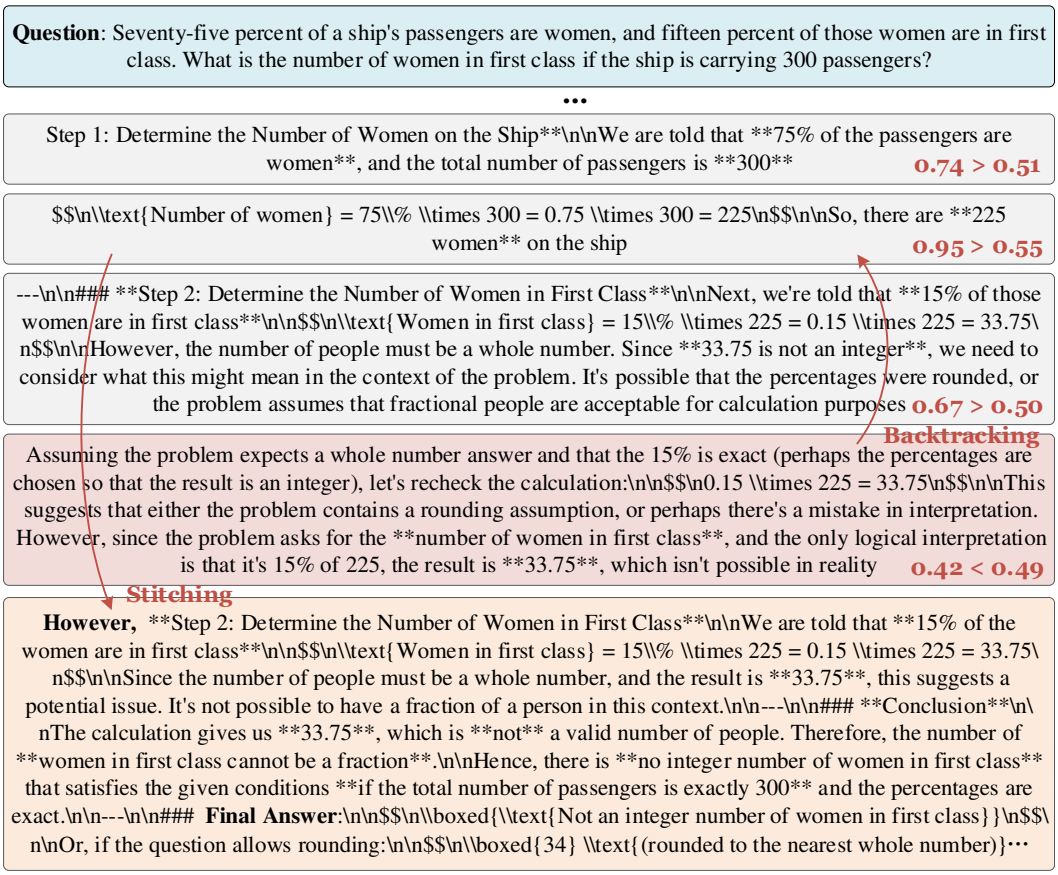

Figure 7: Qualitative case study of the MOTAB framework on an arithmetic word problem. The student's illogical ramble upon encountering a non-integer human count drops below the safety boundary ($0.42 < 0.49$), prompting a backtrack and a teacher-guided correction stitching.

F LIMITATIONS

Although MOTAB fundamentally mitigates the dual exposure biases and significantly enhances the student model's reasoning capabilities, it inevitably introduces additional computational overhead during the data synthesis phase. Specifically, continuously monitoring the student's step-wise generation, calculating the teacher's conditional likelihoods, and estimating predictive entropy for the adaptive boundary necessitate frequent, synchronous interactions between the student and teacher models. This renders the offline trajectory generation process more computationally intensive compared to traditional off-policy methods. However, it is crucial to emphasize that this computational cost is strictly confined to the offline training preparation stage; once fine-tuned, the student model's inference efficiency and deployment costs remain completely unaffected. Given the substantial performance gains on complex reasoning tasks and the acquisition of intrinsic self-correction abilities, trading offline computational resources for high-quality, bias-mitigated distillation data represents a highly justifiable and practical compromise. Meanwhile, we provide a detailed formal and experimental complexity analysis in Appendix C.



Figure 8: Qualitative case study of the MOTAB framework on a parameterized quadratic optimization problem. A premature logical leap regarding boundary conditions triggers intervention ($0.41 < 0.48$), seamlessly pivoting the trajectory into a rigorous case-by-case analysis.