Rooted Absorbed Prefix Trajectory Balance with Submodular Replay for GFlowNet Training

Anonymous Authors[1]

Abstract

Generative Flow Networks (GFlowNets) enable fine-tuning large language models to approximate reward-proportional posteriors, but they remain prone to mode collapse, manifesting as prefix collapse and length bias. We attribute this to two factors: (i) weak credit assignment to early prefixes, and (ii) biased replay that induces a shifted, non-representative training flow distribution. We propose Rooted Absorbed Prefix Trajectory Balance (RapTB), an objective that anchors subtrajectory supervision at the root and propagates sparse terminal task signals to intermediate prefixes via absorbed suffix-based backups, providing dense prefix-level learning signals without extra model forward passes. To mitigate replay-induced distribution shift, we further introduce SubM, a submodular replay refresh strategy that promotes both high reward and diversity. Empirically, on tasks such as molecule generation with LLM using SMILES strings, RapTB combined with SubM consistently improves optimization performance and molecular diversity while preserving high validity.

1. Introduction

Generative Flow Networks (GFlowNets) learn a stochastic policy on a directed acyclic graph (DAG) that constructs objects sequentially, so that completed trajectories are sampled with probability proportional to the rewards (Bengio et al., 2021; 2023; Hu et al., 2023). In contrast to reward-maximizing reinforcement learning, the objective of GFlowNets is distributional: spread probability mass across many high-reward modes in proportion to reward, rather than concentrating on a single optimum (Kaelbling et al., 1996). This method extends naturally to large language models (LLMs) in a terminable prefix-tree formulation (Hu et al., 2024). Every prefix state has an explicit termination edge to a terminal node. The termination edge can be implemented as an EOS action.

In practice, LLM-GFlowNets suffer from mode collapse. We observe two specific failures: (i) **prefix collapse**, entropy drops sharply over early tokens, and many distinct terminals share near-identical prefixes; and (ii) **length bias**, where the model favors sequences that are systematically too short or too long. We trace these issues to two factors: (i) weak substructure credit assignment, since terminal-only task signals provide high-variance and often ambiguous feedback for intermediate choices (Madan et al., 2023), and (ii) replay-induced shifting of training supports, where observing only a tiny fraction of the space leaves many internal flow assignments compatible with the same terminal stop-rewards.

We address these failure modes with two complementary mechanisms: one that strengthens prefix-level credit assignment and the other broadens the support of replay. **RapTB** retains terminal Trajectory Balance (TB) (Malkin et al., 2022) as the primary constraint and adds a lightweight rooted-prefix objective that provides supervision at intermediate prefixes. Concretely, RapTB densifies training signals by propagating sparse terminal task signals to eligible intermediate prefixes via suffix-based backups. In parallel, **SubM** refreshes the replay buffer by selecting a subset of trajectories that maximizes a submodular objective over candidates. The objective jointly encourages high reward, trajectory diversity, and length coverage, expanding the support of the training distribution. Across molecular and arithmetic generation tasks, we find that TB objective (Malkin et al., 2022) can quickly over-concentrate on shared prefixes, while SubTB (Hu et al., 2024) can drift in its termination probabilities. RapTB mitigates both prefix collapse and length bias, and SubM further improves coverage and distribution matching.

Contributions.

- We empirically characterize mode collapse in LLM-GFlowNets as a reproducible combination of prefix collapse and length bias. We provide evidence that it is driven by high-variance terminal credit assignment

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

and replay-induced training distribution shifts.

- We propose RapTB, which augments TB with a rooted prefix objective that propagates sparse terminal task signals to eligible intermediate prefixes via suffix-based backups. It provides dense training signals and reduces variance.
- We introduce Submodular Replay (SubM), a replay refresh rule that balances reward, diversity, and length coverage in one submodular objective. It improves replay coverage and stabilizes training.

Positioning. We focus on TB-family objectives. While TB suffers from high variance under terminal-only task signals (Malkin et al., 2022), existing subtrajectory methods for LLMs (Hu et al., 2024) induce *termination drift* via conflicting overlapping constraints. RapTB resolves this by restricting dense supervision to rooted prefixes using variance-reduced absorbed targets. This formulation eliminates destabilizing boundary conditions and explicitly detaches auxiliary termination gradients to prevent drift, all while preserving the global TB anchor. Orthogonally, SubM stabilizes replay via coverage-aware subset refresh to mitigate the reward-tilted collapse observed in prior work (Shen et al., 2023).

2. Related Work

GFlowNet methodologies and applications. GFlowNets learn generative policies whose terminal distribution is proportional to reward (Bengio et al., 2021; 2023). Trajectory Balance (TB) stabilizes training via global path consistency (Malkin et al., 2022), and subsequent work connects GFlowNets to variational inference (Malkin et al., 2023). To reduce variance, Subtrajectory Balance introduces dense subtrajectory constraints (Madan et al., 2023), but directly adapting it to terminable prefix trees for LLMs can introduce conflicting boundary conditions on the shared termination head, leading to termination drift (Hu et al., 2024). RapTB targets this structural mismatch by restricting dense supervision to rooted prefixes while using suffix-absorbed targets for lower-variance prefix credit.

Experience replay and exploration in GFlowNets. Experience replay improves sample efficiency in GFlowNets, but reward-prioritized replay can induce rich-get-richer collapse and reduce coverage (Shen et al., 2023; Vemgal et al., 2023; Hu et al., 2024). Prior work mitigates this with heavier exploration mechanisms (e.g., local search, evolutionary augmentation, or tree-search-style rollouts). In contrast, we formulate replay refresh as lightweight submodular maximization with a greedy near-optimality guarantee (Kirchhoff & Bilmes, 2014; Kothawade et al., 2022; Killamsetty et al., 2021), selecting a subset that jointly balances reward, diversity, and length coverage.

3. Method

3.1. GFlowNets on Terminable Prefix Trees

We consider a Generative Flow Network (GFlowNet) defined on a directed acyclic graph (DAG) with a unique source s_0 and terminal set $\mathcal{X}$. A trajectory $\xi = (s_0 \to \cdots \to s_\tau \to x)$ ends at a terminal node x, and the target terminal distribution satisfies $p^\star(x) \propto R(x)$ (Bengio et al., 2021; 2023).

Following Hu et al. (2024), we instantiate a GFlowNet on the prefix tree induced by an autoregressive language model augmented with a stop symbol $\top$ (i.e., each prefix has a termination action), such as an End-of-Sentence (EOS) token. Given a prompt, the LLM generates a sequence of tokens. As the prompt is held constant for the generation process, we omit it in the following for simplicity. Therefore, the GFlowNet state is represented by a generated prefix $s_{0:i}$.

From state $s_{0:i}$, the forward policy either emits a token $s_{i+1} \in \mathcal{V}$ (the vocabulary) and transitions to $s_{0:i+1}$, or terminates by emitting $\top$. We parameterize the forward policy by the language model distribution $q_\theta(\cdot \mid s_{0:i})$ over $\mathcal{V} \cup \{\top\}$, and identify

$$P_F^\theta(s_{0:i+1} \mid s_{0:i}) = q_\theta(s_{i+1} \mid s_{0:i}),$$
$$P_F^\theta(s_{0:i}^\top \mid s_{0:i}) = q_\theta(\top \mid s_{0:i}).$$

where $s_{0:i}^\top$ denotes the terminated sequence of $s_{0:i}$ by appending the terminal token $\top$, i.e., $(s_0, s_1, \ldots, s_i, \top)$. Throughout, we identify a stop cut at prefix $s_{0:k}$ with its terminated sequence $s_{0:k}^\top$; when unambiguous, we abbreviate $R(s_{0:k}) \equiv R(s_{0:k}^\top)$ and $P_{\text{ref}}(s_{0:k}) \equiv P_{\text{ref}}(s_{0:k}^\top)$.

Because each non-root prefix has a unique parent, the backward kernel is deterministic, and the state space forms a tree. Therefore, the probability of sampling a terminated sequence $\xi = s_{0:\tau}^\top$ factorizes as

$$q_\theta^\top(\xi) = \Big(\prod_{i=0}^{\tau-1} q_\theta(s_{i+1} \mid s_{0:i})\Big)\, q_\theta(\top \mid s_{0:\tau}).$$

Notation. We use q_θ for the model's autoregressive distribution, and identify the induced forward kernel on the prefix tree as $P_F^\theta(\cdot \mid s) = q_\theta(\cdot \mid s)$. We write a terminal as $x \in \mathcal{X}$ and its corresponding terminated trajectory as $\xi_x = s_{0:\tau}^\top$. When convenient, we write $q_\theta^\top(x)$ as shorthand for $q_\theta^\top(\xi_x)$.

Mixed reward. To stabilize exploration under sparse or high-variance task rewards, for any prefix $s_{0:j}$, we define the stop-reward as a mixture of (i) a frozen reference language model prior and (ii) an external task-specific score, following Hu et al. (2024). Let P_{ref} denote the probability assigned by a fixed, pre-trained reference LM to a terminated sequence, which serves as a prior that regularizes

generation toward fluent and well-formed outputs. For a prefix $s_{0:j}$, we define the mixed stop-reward as

$$\log R(s^{\top}_{0:j}) = \kappa \log P_{\text{ref}}(s^{\top}_{0:j}) + \lambda S(s^{\top}_{0:j}),$$

where $S(s^{\top}_{0:j})$ is the task-only component and $\kappa, \lambda \geq 0$ control the mixture ratio. Depending on the task, S can be dense (defined for any stop cut) or sparse (near-zero for most early stop cuts); the latter regime motivates our absorbed suffix targets. In particular, stopping immediately after the prompt ($k = 0$) is also scored by the same mixed stop-reward. The task-only component $S(\cdot)$ depends on the application; for example, in small-molecule generation it may correspond to a property score (e.g., binding affinity or drug-likeness), while in symbolic expression generation it may reflect functional correctness or coverage. When S is undefined or uninformative for early or invalid stop cuts, we set it to zero and mask those cuts only in the auxiliary prefix losses via an eligibility mask (Appendix C.4); the terminal TB anchor and model logits are unchanged. The exact reward definitions and ablation studies are provided in Appendix C.1.

3.2. Trajectory Balance (TB): Global Path Consistency

Trajectory Balance (TB) (Malkin et al., 2022) enforces global consistency between forward trajectories and stop-rewards by introducing a learnable normalizer $Z_\theta > 0$. For a terminated trajectory $\xi = s^{\top}_{0:\tau}$, the TB log-residual is

$$\begin{aligned}\Delta^{\text{TB}}(\xi) = \log Z_\theta + \sum_{i=0}^{\tau-1} \log P_F^\theta(s_{i+1} \mid s_{0:i}) \\ + \log P_F^\theta(\top \mid s_{0:\tau}) - \log R(s^{\top}_{0:\tau}),\end{aligned} \tag{1}$$

and TB minimizes

$$\mathcal{L}_{\text{TB}} = \mathbb{E}_{\xi \sim q_\theta^{\top}}\left[\Delta^{\text{TB}}(\xi)^2\right].$$

In replay-based training, trajectories are sampled from the buffer-induced off-policy distribution; we therefore minimize the same squared residual under that distribution. Without explicit importance weighting (and full support), we do not claim an exact reward-proportionality guarantee.

Proposition 3.1 (Approximate reward-proportionality from small TB residual)**.** *On a prefix tree, each terminal $x \in \mathcal{X}$ corresponds to a unique terminated trajectory ξ_x. Fix any $Z > 0$ and assume that the TB residual is uniformly bounded: for all $x \in \mathcal{X}$,*

$$\left|\log Z + \log q_\theta^{\top}(\xi_x) - \log R(x)\right| \leq \varepsilon.$$

Then for all $x \in \mathcal{X}$,

$$e^{-\varepsilon}\frac{R(x)}{Z} \leq q_\theta^{\top}(x) \leq e^{\varepsilon}\frac{R(x)}{Z}.$$

TB provides a clean global anchor for reward-proportional sampling, but terminal-only task signals yield high-variance credit assignment on long horizons. In practice, a few high-reward trajectories dominate updates; since many trajectories share early prefixes, these prefixes are repeatedly reinforced while alternatives are under-trained, leading to self-reinforcing prefix collapse and reduced diversity (Shen et al., 2023; Madan et al., 2023).

3.3. Subtrajectory Balance (SubTB): Dense but Over-Constrained Supervision

Subtrajectory Balance (SubTB) (Madan et al., 2023) reduces variance by enforcing TB-style consistency on subtrajectories. Following Hu et al. (2023), for any subtrajectory indexed by $i, j, 0 \leq i < j \leq \tau$, define $\Delta^{\text{SubTB}}_{i\to j}(\xi)$ as

$$\begin{aligned}\Delta^{\text{SubTB}}_{i\to j}(\xi) \triangleq \sum_{k=i}^{j-1} \log P_F^\theta(s_{k+1} \mid s_{0:k}) \\ + \left(\log P_F^\theta(\top \mid s_{0:j}) - \log P_F^\theta(\top \mid s_{0:i})\right) \\ + \left(\log R(s^{\top}_{0:i}) - \log R(s^{\top}_{0:j})\right),\end{aligned} \tag{2}$$

The overall objective is the combination of all subtrajectory objectives, optionally weighted by w_ℓ based on the length of the subtrajectory:

$$\mathcal{L}_{\text{SubTB}}(\xi) \triangleq \frac{\sum_{\ell=1}^{\tau} w_\ell \sum_{i=0}^{\tau-\ell} \left(\Delta^{\text{SubTB}}_{i\to i+\ell}(\xi)\right)^2}{\sum_{\ell=1}^{\tau} w_\ell \sum_{i=0}^{\tau-\ell} 1}, \tag{3}$$

While SubTB provides dense supervision by enforcing consistency on many subtrajectories, in terminable prefix trees it introduces a large number of overlapping constraints that share the same termination head. Each subtrajectory implicitly treats its endpoint as a pseudo-terminal, imposing a distinct boundary condition involving the termination probability $q_\theta(\top \mid s_{0:i})$. These heterogeneous boundary conditions are difficult to satisfy simultaneously, and gradients from many windows accumulate on the shared termination logits. As a result, the model can reduce SubTB residuals by adjusting termination probabilities rather than improving token-level transitions, leading to biased termination behavior such as systematic length drift. This over-constraining effect motivates our more conservative approach to densifying prefix-level supervision.

3.4. RapTB: Rooted Absorbed Prefix Trajectory Balance

RapTB addresses the trade-off between TB's high variance and SubTB's over-constrained objective. It restricts dense supervision to rooted-prefix residuals, reducing destabilizing boundary conditions, and utilizes partial credit as additional training target by "absorbing" credits from suffixes.

Rooted Prefix Residuals. We constrain subtrajectories to be "rooted," originating from s_0. The rooted residual at step k is defined as the difference between the TB residual of the current prefix and the root:

$$\bar{\Delta}_k(\xi) \triangleq \Delta_k^{\mathrm{TB}}(\xi) - \Delta_0^{\mathrm{TB}}(\xi). \tag{4}$$

Here $\Delta_k^{\mathrm{TB}}(\xi)$ is the TB-style log-residual defined at prefix $s_{0:k}$ (including $k=0$, i.e., stopping immediately after the prompt):

$$\Delta_k^{\mathrm{TB}}(\xi) \triangleq \log Z_\theta + \sum_{i=0}^{k-1} \log P_F^\theta(s_{i+1} \mid s_{0:i}) + \log P_F^\theta(\top \mid s_{0:k}) - \log R(s_{0:k}^\top),$$

where the sum is empty when $k=0$. By eliminating the global constant $\log Z_\theta$, this formulation creates a local consistency signal anchored to s_0. Unlike SubTB, which creates conflicting boundary conditions via overlapping windows, our approach provides incremental, step-by-step supervision.

Absorbed Suffix Rewards. To stabilize training against stochastic variance, we introduce the absorbed suffix reward. This approach constructs a lower-variance target by backing up the rewards from the observed suffix, employing an aggregation mechanism (details in Appendix C.3). This distills hindsight information into a smoothed signal, guiding the policy more reliably than terminal feedback (analysis in Appendix C.5). For a trajectory $s_{0:\tau}$, let u_j denote the task-only component at position j, i.e., $\lambda S(s_{0:j}^\top)$. Let K be an auxiliary horizon cap and define $h \triangleq \min(\tau, K)$ (we set $K = L_{\max}$). We define the absorbed target as:

$$u_k^{\max} \triangleq \max_{j \in [k,h]} u_j, \tag{5}$$

$$u_k^{soft} \triangleq \frac{1}{\beta} \log \sum_{j=k}^{h} \exp\Big(\beta u_j - \beta\rho\,(j-k)\Big), \tag{6}$$

$$u_k^{tgt} \triangleq \alpha\, u_k^{\max} + (1-\alpha)\, u_k^{soft}, \qquad \alpha \in [0,1]. \tag{7}$$

The $u_k^{\max}$ enforces a monotone hindsight target along the sampled suffix: it sets the prefix target to be at least the best observed suffix target within the trajectory. The u_k^{soft} term smoothly aggregates multiple suffix rewards in log space, where $\beta > 0,\ \rho \geq 0$. The distance penalty $\rho(j-k)$ downweights distant evidence. Appendix C.5 provides a variance-reduction view: these suffix-backed-up targets act as a lower-variance proxy for the stochastic TB tail term when training early prefixes.

The absorbed target u_k^{tgt} is the task-only partial credit assigned to the prefix $s_{0:k}$. We can then treat it as an "estimated reward" for $s_{0:k}$ and train the model to match it. Operationally, this is equivalent to recomputing the rooted TB residual using a surrogate stop-reward in which the task-only component u_k is replaced by u_k^{tgt} (details in Appendix C). Since $\log R(s_{0:k}^\top) = \kappa \log P_{\mathrm{ref}}(s_{0:k}^\top) + u_k$, this replacement yields the additive correction term $(u_k - u_k^{tgt})$ inside the rooted residual:

$$\mathcal{L}_{\mathrm{aux}}(\xi) \triangleq \frac{\sum_{k=1}^{\tau} w_k \left(\bar{\Delta}_k(\xi) + u_k - u_k^{tgt}\right)^2}{\sum_{k=1}^{\tau} w_k}, \tag{8}$$

where w_k is the length weight. In practice, we use an eligibility mask to downweight very short prefixes and apply STOPGRAD to termination logits inside $\mathcal{L}_{\mathrm{aux}}$ to prevent termination drift; the exact implemented form is in Appendix C.4.

Final Objective. The RapTB objective integrates global TB consistency with this dense guidance.

$$\mathcal{L}_{\mathrm{RapTB}} = \mathbb{E}_{\xi \sim q_\theta^\top}\Big[\underbrace{\Delta^{\mathrm{TB}}(\xi)^2}_{\text{Anchor}} + \underbrace{\eta\, \mathcal{L}_{\mathrm{aux}}(\xi)}_{\text{Partial Credit}}\Big], \tag{9}$$

where η balances global consistency with regularization. The TB term remains the only exact balance condition whose optimum matches the reward-proportional target; the auxiliary term is a variance-reducing regularizer that can bias the solution for finite η, but empirically improves optimization and coverage while the TB anchor stabilizes training.

Compare three losses. TB uses a global objective with sparse supervision, while SubTB increases supervision density but over-constrains termination behavior in prefix trees. RapTB preserves TB as the sole exact balance constraint and adds (i) rooted prefix supervision that avoids heterogeneous window boundaries and (ii) absorbed suffix rewards that reduce variance and improve prefix credit assignment.

3.5. Submodular Replay: Diversity- and Length-balanced Experience Selection

RapTB addresses within-trajectory credit assignment. In parallel, to explicitly enforce diversity in addition to the exploration of GFlowNet, we maintain a fixed-size replay buffer of size B and update it by selecting a representative, diverse, and length-balanced subset from the union of the current buffer and a newly collected batch (details in Appendix D.3.1).

Submodular selection. At each buffer update step, we form the ground set $\mathcal{G}$ as the union of the current buffer and a new generated batch, then select $S \subseteq \mathcal{G}$ with $|S| = B$ by maximizing a monotone submodular function subject

to a cardinality constraint (Kirchhoff & Bilmes, 2014; Kothawade et al., 2022; Killamsetty et al., 2021).

Let $\mathrm{sim}(v, x) \in [0, 1]$ be a task-appropriate similarity. In experiments, we use Morgan fingerprints with Tanimoto similarity for SMILES, and n-gram shingle Jaccard similarity for text generation tasks. We define the facility-location coverage operator, which reflects how well the buffer represents a sample v.

$$\mathrm{msim}(v, S) \triangleq \max_{x \in S} \mathrm{sim}(v, x), \qquad \mathrm{msim}(v, \emptyset) \triangleq 0. \tag{10}$$

The overall submodular objective $f(S)$ is

$$\underbrace{\sum_{x \in S} \mathrm{static}(x)}_{\text{quality / feasibility}} + \underbrace{\lambda_{\mathrm{div}} \sum_{v \in \mathcal{G}} \mathrm{msim}(v, S)}_{\text{facility-location coverage}} + \underbrace{\lambda_{\mathrm{len}} f_{\mathrm{len}}(S)}_{\text{length coverage}}. \tag{11}$$

When weights are nonnegative and $\mathrm{static}(x)$ is shifted to be nonnegative, each term is monotone submodular and so is their sum. Under the fixed-cardinality constraint $|S| = B$, this constant shift does not change the selected subset. $\mathrm{static}(x)$ denotes a fixed per-sample quality/feasibility term (e.g., a property score). We apply validity gating by restricting the items scored in the facility-location term to feasible samples, which preserves monotone submodularity (details in Appendix D.3.1). For length coverage, we discretize samples into bins and use concave-over-counts histogram coverage:

$$f_{\mathrm{len}}(S) \triangleq \sum_{b=1}^{N_{\mathrm{bin}}} \alpha_b \, g\big(c_b(S)\big), \quad g(c) \triangleq \log(1 + c), \tag{12}$$

where N_{bin} is the number of length bins, $c_b(S)$ is the count in bin b, and $\alpha_b \geq 0$ can bias coverage toward desired lengths; implementation details are in Appendix D.3.1.

Greedy update and efficiency. For every gradient step, we update the fixed-size buffer by optimizing $\max_{S \subseteq \mathcal{G}, |S| \leq B} f(S)$ using a greedy algorithm. With cached similarities and histogram counts, one update costs $O(B|\mathcal{G}|)$. In our settings B and $|\mathcal{G}|$ are small, so the overhead is negligible ($\sim 10\mathrm{ms}$ per update).

4. Experiments

4.1. Tasks

Scaffold-conditioned SMILES optimization. We study conditional molecular generation where the conditioning input is a fixed molecular scaffold and the model generates a completion by adding fragments. Each terminal sequence is a SMILES string $x \in \mathcal{X}$ that must (i) be chemically valid and (ii) satisfy the scaffold constraint. We optimize a property objective based on the Estimate of Drug-likeness (Bickerton et al., 2012). Training aims to learn a GFlowNet sampler whose induced terminal distribution assigns higher probability to high-reward scaffold-consistent molecules while maintaining diversity among valid completions.

Expr24 arithmetic expression generation. We also evaluate on a discrete, fully verifiable sparse-reward task: generating an arithmetic expression whose value equals 24. A terminal $x \in \mathcal{X}$ is a variable-length token sequence consisting of digits and operators $\{+, -, \times, \div\}$, evaluated with standard operator precedence. The task score is sparse and exact:

$$R(x) = \mathbb{I}[\mathrm{eval}(x) = 24].$$

This task isolates exploration, credit assignment, termination/length bias, and collapse behavior without domain-specific feasibility issues (e.g., chemical validity).

CommonGen: Concept-to-Sentence Generation. CommonGen (Lin et al., 2020) requires generating a coherent sentence incorporating a given set of concept keywords. We formulate the reward to encourage keyword coverage while maintaining natural language fluency.

4.2. Compared objectives and replay strategies

We compare three objectives adapted to terminable LLM-GFlowNets: TB (Eq. 1), SubTB (Eqs. 2–3), and RapTB. Unless otherwise specified, methods employ the standard **reward-prioritized replay (RP)** from Hu et al. (2024); for ablation purposes, we also include **Reward-prioritized replay training (PRT)** (Shen et al., 2023). We also introduce our proposed **submodular replay** (**SubM**), which acts to explicitly promote diversity among stored high-reward trajectories. All methods share the same model architecture, tokenizer, decoding constraints, and optimizer configuration (details in Appendix D).

Implementation summary. We fine-tune Llama-3.2-1B with LoRA (rank 16) using AdamW (lr 10^{-4}). For RapTB, we apply rooted-prefix auxiliary losses only to eligible prefixes and cap backup depth by $h = \min(\tau, K)$; in the auxiliary branch we stop gradients through the termination head via $\mathrm{stopgrad}(\log q_\theta(\top \mid s_{0:k}))$ to avoid termination drift. For SubM, we periodically refresh a fixed-size replay buffer by greedy maximization of Eq. 11 over a candidate pool, using cached similarities and length histograms. Replay-buffer sizes, refresh cadence, and decoding constraints are task-specific and reported in Appendix D.

4.3. Metrics

We evaluate (i) feasibility and quality (Acc/Score), (ii) diversity (TokEnt for token entropy; FPDiv for SMILES fingerprint diversity), (iii) length/termination calibration,

and (iv) prefix-collapse diagnostics (Surv/Ent/Top1 versus depth). For Expr24, we additionally report coverage (Unique$_\checkmark$, NormCov) and distributional fidelity (KL/JS) against the enumerated oracle. As a termination diagnostic, Expr24 reports the termination log-probability at the sampled stop step $\log p_{\text{term}}(\tau)$, while CommonGen reports a mean termination-logit shift $\Delta \log p_{\text{term}}$ relative to the base model. Unless stated otherwise, we report means over multiple seeds; key tables show 95% confidence intervals, and full metric definitions and protocols are provided in Appendix B.

4.4. Results on Scaffold-conditioned SMILES Optimization

4.4.1. Overall performance

Table 1 shows that RapTB+SubM achieves a strong quality–diversity trade-off while maintaining high validity. In contrast, SubTB suffers from severe validity degradation. TB achieves near-perfect validity but is weaker in reward quality and diversity under the default replay. SubM substantially improves TB by broadening replay coverage (Appendix A.1). This highlights SubM as a generally helpful replay component in our settings for sustaining diverse exploration.

Table 1. **SMILES generation performance.** Unless specified, metrics are computed on valid samples. `FPDiv` represents molecular fingerprint diversity. Len denotes mean terminal length (pre-EOS tokens). Results are mean±95% CI over random seeds; per-length breakdowns are reported in Appendix A.1.

Method	Acc ↑	Score ↑	TokEnt ↑	FPDiv ↑	Len
TB	0.998 ±0.001	0.717 ±0.001	2.503 ±0.026	0.807 ±0.003	3.06 ±0.02
SubTB	0.328 ±0.016	0.755 ±0.004	2.127 ±0.037	0.836 ±0.003	8.35 ±0.06
RapTB	0.996 ±0.001	0.740 ±0.004	2.448 ±0.017	0.860 ±0.001	6.14 ±0.03
RapTB+SubM	0.988 ±0.003	0.844 ±0.001	2.726 ±0.017	0.898 ±0.001	7.44 ±0.05

4.4.2. Per-length analysis

Aggregate metrics can be confounded by length shifts (e.g., diversity can increase mechanically if a method shifts mass to longer sequences). Figure 2 reports the valid-only length histogram and length-conditioned score/diversity. RapTB+SubM remains strong across most lengths, whereas TB concentrates on short lengths and degrades in the long-length regime.

4.4.3. Prefix collapse analysis

Diverse terminals can still share highly concentrated early prefixes and only branch late, a failure mode we refer to as *prefix collapse*. We therefore compute position-wise prefix statistics on all valid samples. Figure 3 reports prefix diagnostics by prefix length k. TB displays rapid survival decay (failure to sustain generation) and sharply increasing top-1 mass (frequency of the most common prefix) at longer prefixes, indicating severe concentration on a few shared partial trajectories. RapTB sustains higher prefix entropy and lower top-1 mass deeper into the trajectory, consistent with earlier and broader branching among valid samples.

4.4.4. Long-horizon stress test

Increasing $L_{\max}$ to 15 exposes severe length collapse in TB, which concentrates mass on short trajectories and fails to reach the long-horizon regime (Table 2). While SubTB improves coverage at the cost of validity, RapTB effectively mitigates this length bias, unlocking access to extended trajectories without compromising accuracy. Crucially, RapTB+SubM achieves the most robust performance: it maximizes long-horizon coverage (Frac(11+)), yields the best quality-diversity trade-off (Score and MacroFP), and exhibits the lowest prefix concentration (Top1), demonstrating superior resistance to mode collapse.

4.5. Variable-length Expr24 under sparse rewards

Enumerable solutions enable controlled diagnostics. `Expr24` disentangles two factors often confounded under sparse terminal-only task signals: *external coverage* (the capacity to discover high-reward modes) and *internal credit assignment* (the fidelity of probability mass allocation and termination calibration under off-policy mixtures). Because the full correct set $\mathcal{Y}^*$ is enumerable, we can (i) control coverage via an **oracle replay** buffer sampling from it, and (ii) directly compare the induced terminal distribution $\pi(x) \triangleq q_\theta^\top(x)$ to a reference p^* using bidirectional KL and token-wise JS, where $\mathrm{KL}(p^* \to \pi)$ highlights mode dropping and $\mathrm{KL}(\pi \to p^*)$ reflects over-concentration.

Expr24 Results. Table 3 shows that RapTB achieves a superior trade-off between correctness and coverage. Under standard **RP**, TB suffers from severe mode collapse (Unique$_\checkmark \approx 5$), whereas RapTB significantly improves diversity without compromising accuracy. This advantage is amplified by **SubM**: RapTB+SubM doubles the normalized coverage of the strongest baseline (0.209 vs. 0.100) while maintaining near-perfect accuracy (>0.99). Finally, the **Oracle** setting verifies the objective's effectiveness: RapTB outperforms TB in both accuracy (0.945 vs. 0.919) and distribution matching (lower KL/JS), indicating that RapTB learns a more precise policy.

Diagnosis of SubTB's abnormal behaviors. In the variable-length Expr24 setting, we observe pronounced termination drift: the log-probability of sampled termination, $\log p_{\text{term}}(\tau)$, degrades to extremely negative values (Table 4). This severely impacts the hit rate when the stopping condi-

Table 2. **Long-horizon stress test on SMILES ($L_{\max} = 15$).** We report length fractions of valid samples (Length Dist.) and diversity metrics. Prefix diagnostics are averaged over k: Surv (fraction of samples reaching length k), Ent (prefix entropy), and Top1 (frequency of most common prefix). MacroFP macro-averages FPDiv across length bins 0–5, 6–10, and 11+. RapTB+SubM achieves the best balance of long-horizon coverage and diversity. Values are point estimates (mean over seeds); 95% CIs are omitted for space.

	Performance		Length Dist.			Prefix Diagnostics			Diversity	
Method	Acc	Score	0–5↓	6–10	11+↑	Surv↑	Ent↑	Top1↓	MacroFP↑	FPDiv↑
TB	**0.999**	0.716	0.858	0.129	0.013	0.207	2.99	0.303	0.653	0.813
SubTB	0.636	0.742	0.286	0.442	0.271	0.561	4.82	0.085	0.716	0.770
RapTB	0.988	0.768	0.113	0.318	0.568	0.681	**5.59**	0.084	0.793	0.810
RapTB+SubM	0.972	**0.849**	**0.094**	0.205	**0.701**	**0.751**	5.32	**0.071**	**0.805**	**0.868**

Table 3. **Expr24 results under different replay schemes.** Mean±95% CI over random seeds.

		Diversity		Quality	Distributional Fidelity		
Replay	**Objective**	Unique$_{\checkmark}$↑	NormCov↑	Acc↑	KL($\pi \to p^*$)↓	KL($p^* \to \pi$)↓	JS_{tok}↓
PRT	TB	103.7±3.2	0.016±0.001	0.999±0.000	1.105±0.002	7.803±0.060	0.292±0.001
	SubTB	292.0±2.9	0.046±0.000	0.311±0.002	0.424±0.010	0.672±0.077	0.107±0.003
	RapTB	129.3±0.4	0.020±0.000	0.992±0.001	0.908±0.003	5.538±0.005	0.230±0.001
RP	TB	5.3±0.4	0.001±0.000	1.000±0.000	1.297±0.001	11.403±0.282	0.339±0.000
	SubTB	324.7±2.7	0.051±0.000	0.229±0.005	0.455±0.005	0.865±0.083	0.109±0.002
	RapTB	246.7±7.1	0.039±0.001	0.991±0.000	0.561±0.001	4.480±0.002	0.147±0.000
SubM	TB	642.0±5.6	0.100±0.001	0.996±0.001	0.182±0.001	0.441±0.005	0.049±0.000
	SubTB	331.3±22.7	0.052±0.004	0.061±0.005	0.149±0.008	0.286±0.070	0.040±0.002
	RapTB	1337.3±7.5	0.209±0.001	0.994±0.001	0.169±0.001	0.623±0.004	0.048±0.000
Oracle	TB	5198.0±5.2	0.812±0.001	0.919±0.001	0.062±0.001	0.066±0.001	0.016±0.000
	SubTB	35.7±2.9	0.006±0.000	0.006±0.000	0.266±0.009	1.491±0.413	0.071±0.003
	RapTB	5220.7±4.3	0.816±0.001	0.945±0.001	0.052±0.001	0.056±0.001	0.013±0.000

tion is a decision variable. We attribute this phenomenon to the enforcement of numerous arbitrary-start windows, which can be partially satisfied by a global shift in $\log q_\theta(\top \mid s_{0:\tau})$ (see analysis in Appendix C.6). To investigate whether termination drift is the dominant failure mode, we employ RootSubTBLogZ, which restricts SubTB windows to be rooted and reintroduces a learnable global normalizer Z_θ. As shown in Table 4, this modification mitigates termination drift and restores accuracy to nearly 1.

4.6. Results on CommonGen

Table 5 reveals that SubTB suffers from policy degeneration, saturating length (20) and distorting stopping logits ($\Delta \log p_{\text{term}} \approx -28.32$) to inflate BLEU. RapTB maintains calibrated stopping behavior. RapTB+SubM achieves the highest BLEU (33.23) with natural lengths; SubM ablations and replay-scheme comparisons are reported on SMILES and Expr24, where we can more cleanly isolate replay effects.

Table 4. **Termination/length calibration diagnostic on Expr24.** More negative values indicate overly suppressed termination. $\log p_{\text{term}}(\tau)$ is the termination log-probability at the sampled stop step, computed from the model's raw $q_\theta(\top \mid s_{0:\tau})$ (no masking or renormalization). Values are point estimates; 95% CIs are omitted for space.

Method	Acc	NormCov	$\log p_{\text{term}}(\tau)$	$\log Z$
RP Replay				
TB	0.999	0.001	-0.000	0.063
SubTB	0.229	0.051	-79.638	–
RapTB	0.991	0.039	-0.065	0.062
RootSubTBLogZ	0.999	0.023	-0.068	0.062
SubM Replay				
RapTB+SubM	0.994	0.209	-0.004	0.062
Oracle Replay				
TB	0.922	0.813	-0.436	0.037
SubTB	0.006	0.006	-86.415	–
RapTB	0.945	0.816	-0.644	0.038
RootSubTBLogZ	0.885	0.727	-1.432	0.036

4.7. Ablations.

Table 6 ablates the key design choices of RapTB and SubM on SMILES. Removing reward absorption degrades both score and diversity, suggesting that suffix evidence provides useful prefix credit. Using only max or only soft backups increases score but reduces diversity, while the mixed backup improves the balance. Detaching termination gradients in the auxiliary branch is also important. Without it, the model collapses to very short sequences (Len 3.40) and the score drops. SubM components are complementary. Reward-only improves score, diversity-only improves FPDiv, and length-

Table 5. **CommonGen performance.** $\Delta \log p_{\text{term}}$ denotes the average difference in termination logits relative to the base model. Values are point estimates; 95% CIs are omitted for space.

Method	TokEnt↑	BLEU-4↑	Len	$\Delta \log p_{\text{term}}$
TB	2.966	5.95	13.86	-1.34
SubTB	3.719	24.39	20.00	-28.32
RapTB	3.933	11.75	15.63	-0.94
RapTB+SubM	**4.102**	**33.23**	11.83	4.89

Table 6. **Ablation study on SMILES generation.**

Variant	Score ↑	FPDiv ↑	TokEnt ↑	Len
RapTB	0.740	0.860	2.448	6.142
RapTB w/o reward absorb	0.716	0.805	2.031	5.296
RapTB absorb max-only	0.821	0.775	1.716	7.431
RapTB absorb soft-only	0.819	0.748	1.516	7.710
RapTB+SubM (Full)	0.844	0.898	**2.726**	7.435
+ length-only SubM	0.773	0.885	1.914	6.569
+ diversity-only SubM	0.741	**0.942**	2.602	5.459
+ reward-only SubM	**0.878**	0.884	1.876	8.122
RapTB w/o detach p_{term}	0.714	0.892	2.072	3.403

only increases long-horizon coverage. Combining them yields the strongest overall trade-off.

5. Discussion and Practical Guidance

Mitigating Replay-Induced Collapse. Standard reward-prioritized replay often induces "rich-get-richer" dynamics (Shen et al., 2023), where the training distribution collapses onto a narrow set of repeated high-reward modes. SubM explicitly counters this by enforcing structural diversity within the buffer. This prevents near-duplicate dominance and ensures the policy learns from a broad, representative landscape rather than degenerate subsets.

Credit Assignment and Consistency. A key limitation of SubTB is that enforcing constraints on arbitrary windows creates conflicting boundary conditions, effectively hardening optimization and destabilizing termination. RapTB resolves this by grounding dense supervision to rooted prefixes, ensuring all partial trajectory updates remain consistent with the global partition function Z. A promising future direction is adaptive subtrajectory selection, where the model learns to identify and prioritize essential substructures online.

6. Conclusion

We studied mode collapse in terminable LLM-GFlowNets and identified two coupled and reproducible failure modes: prefix collapse and length/termination bias under replay-based training. To address unstable credit assignment without inducing termination drift, we proposed RapTB, which augments TB with rooted prefix constraints and suffix-absorbed targets, while stopping gradients through the termination head in the auxiliary branch. To address replay-induced distribution shift and limited external coverage, we introduced SubM, a submodular replay refresh strategy that balances reward with diversity and length support. Across tasks, RapTB+SubM improves long-horizon stability and coverage, yielding better reward–diversity trade-offs and substantially reduced prefix collapse. We hope these results encourage future objectives that explicitly couple coverage-aware replay with selective, adaptively weighted subtrajectory learning for robust autoregressive GFlowNet training.

Impact Statement

This paper advances learning methods for sampling diverse high-quality solutions from various reward distributions. In molecular generation, improved exploration and property optimization may accelerate candidate discovery, but generated molecules are only hypotheses and require downstream validation. We do not foresee direct negative societal impacts from the method itself beyond the general risks of misuse of generative models; appropriate safeguards and domain expert oversight remain necessary.

References

Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. Flow network based generative models for non-iterative diverse candidate generation. *Advances in neural information processing systems*, 34:27381–27394, 2021.

Bengio, Y., Lahlou, S., Deleu, T., Hu, E. J., Tiwari, M., and Bengio, E. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

Hu, E. J., Malkin, N., Jain, M., Everett, K. E., Graikos, A., and Bengio, Y. Gflownet-em for learning compositional latent variable models. In *International Conference on Machine Learning*, pp. 13528–13549. PMLR, 2023.

Hu, E. J., Jain, M., Elmoznino, E., Kaddar, Y., Lajoie, G., Bengio, Y., and Malkin, N. Amortizing intractable inference in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ouj6p4ca60.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., and Iyer, R. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8110–8118, 2021.

Kirchhoff, K. and Bilmes, J. Submodularity for data selection in machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 131–141, 2014.

Kothawade, S., Kaushal, V., Ramakrishnan, G., Bilmes, J., and Iyer, R. Prism: A rich class of parameterized submodular information measures for guided data subset selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10238–10246, 2022.

Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., and Ren, X. Commongen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1823–1840, 2020.

Madan, K., Rector-Brooks, J., Korablyov, M., Bengio, E., Jain, M., Nica, A. C., Bosc, T., Bengio, Y., and Malkin, N. Learning gflownets from partial episodes for improved convergence and stability. In *International Conference on Machine Learning*, pp. 23467–23483. PMLR, 2023.

Malkin, N., Jain, M., Bengio, E., Sun, C., and Bengio, Y. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022.

Malkin, N., Lahlou, S., Deleu, T., Ji, X., Hu, E. J., Everett, K. E., Zhang, D., and Bengio, Y. GFlownets and variational inference. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=uKiE0VIluA-.

Shen, M. W., Bengio, E., Hajiramezanali, E., Loukas, A., Cho, K., and Biancalani, T. Towards understanding and improving GFlowNet training. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30956–30975. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shen23a.html.

Vemgal, N. M., Lau, E., and Precup, D. An empirical study of the effectiveness of using a replay buffer on mode discovery in GFlownets. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. URL https://openreview.net/forum?id=pBk1cRPKBv.

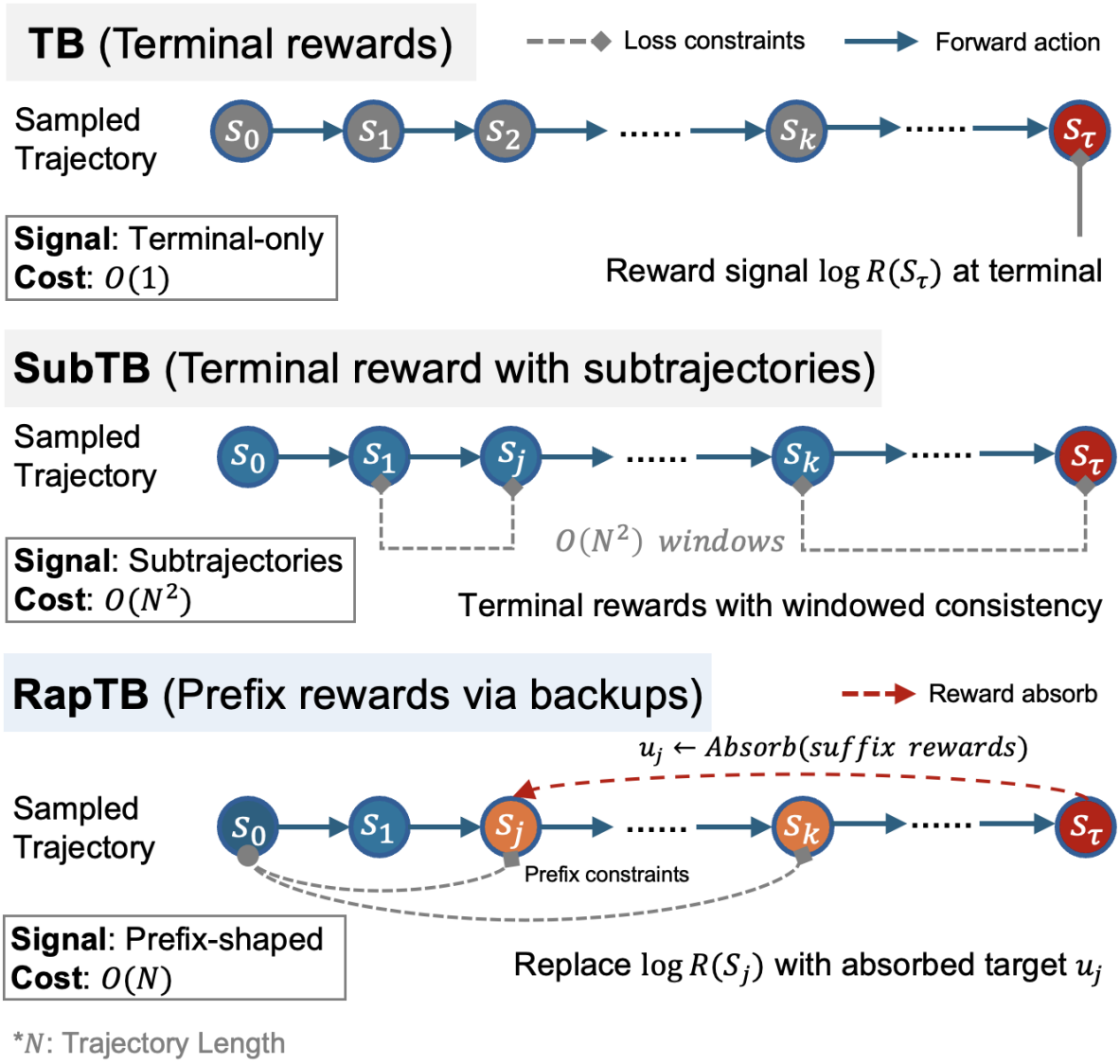
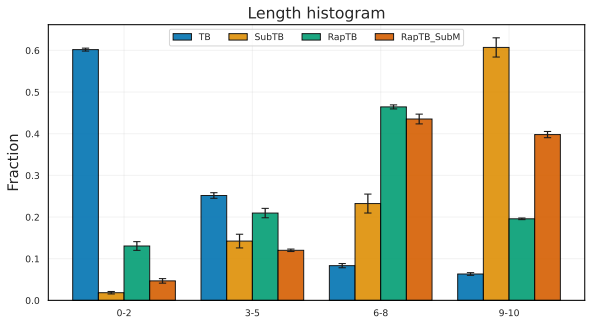
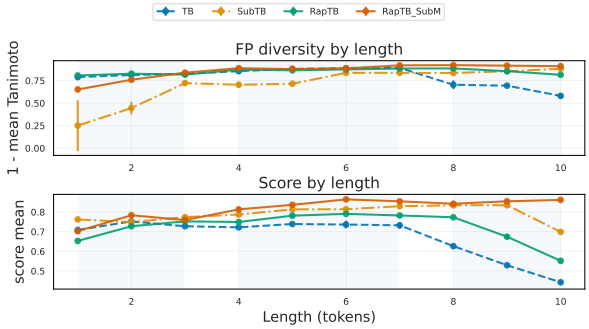


Figure 1. **Training objectives for LLM-GFlowNets.** TB uses only the stop-reward at termination $\log R(s_\tau)$ ($O(1)$). SubTB adds $O(N^2)$ windowed consistency constraints. RapTB replaces prefix stop-rewards with suffix-absorbed targets u_j and applies $O(N)$ rooted prefix constraints. (N: trajectory length.)



(a) Valid-only length histogram.



(b) Valid-only score and FPDiv versus length.

Figure 2. **Length-stratified analysis on SMILES** ($L_{\max} = 10$)**.** (a) Distribution of valid generation lengths. (b) Mean Score and FPDiv conditioned on length.

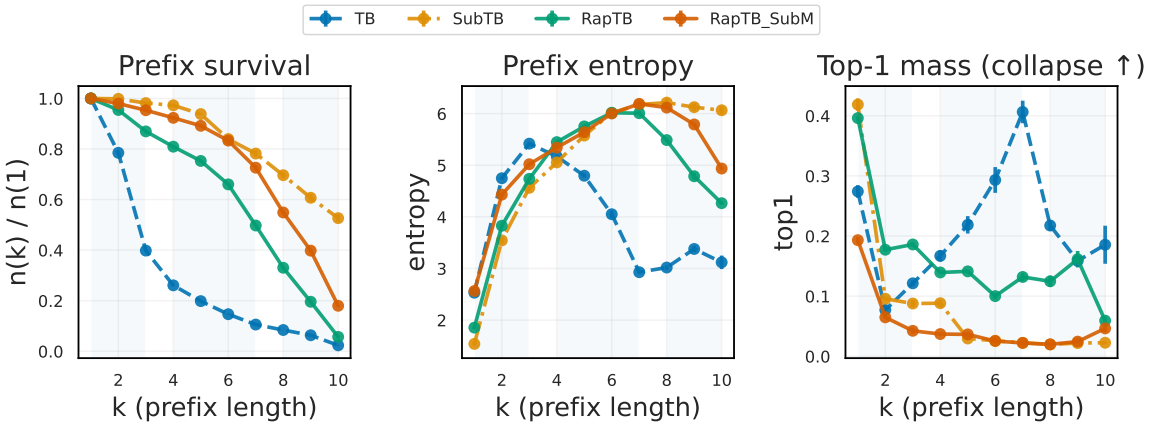


Figure 3. **Prefix-collapse diagnostics on SMILES ($L_{\max} = 10$).** Metrics vs. prefix length k computed on valid samples: prefix survival (fraction of samples reaching length k), entropy, and top-1 mass (frequency of the most common prefix).

A. Additional Results

A.1. SMILES

A.1.1. PER-LENGTH SMILES PERFORMANCE.

Tables 8–9 report valid-only metrics stratified by terminal length L, where the largest differences concentrate in the longest bins ($L \geq 8$) under rewards. Submodular Replay (SubM) substantially improves length-wise coverage for TB, largely eliminating its long-length degradation: for $L = 8$–10, TB's Score increases from $0.626/0.529/0.442$ to $0.861/0.880/0.875$, with diversity recovering simultaneously (TokEnt ≈ 2 and FPDiv ≈ 0.91). Under this stronger coverage regime, the gap between RapTB and TB naturally narrows (a ceiling effect), yet RapTB+SubM remains competitive across lengths and tends to allocate more mass to the longest bin (higher Frac/Count at $L \geq 9$) while maintaining high quality, consistent with RapTB improving the learning signal for suffix credit assignment whereas SubM primarily improves external replay coverage across length/diversity/reward. In contrast, SubTB (with or without SubM) exhibits pronounced length skew, placing disproportionately large mass at $L = 9$–10 alongside a substantial validity drop, consistent with the abnormal termination dynamics discussed in the main text.

All-length averaged summary. Table 7 aggregates metrics across all generated lengths ($L_{\max} = 10$). SubM dramatically changes the effective training distribution of TB: the median length shifts from 2.15 to 7.23 and Frac[0–2] decreases from 0.601 to 0.143, while Frac[9–10] increases from 0.064 to 0.323, yielding a large gain in valid-only average Score ($0.717 \rightarrow 0.842$) and TokEnt ($2.503 \rightarrow 2.775$). With SubM, RapTB+SubM and TB+SubM achieve very similar averaged Score (0.844 vs. 0.842), but RapTB+SubM allocates significantly more probability mass to the longest bin (Frac[9–10] 0.402 vs. 0.323), at a small but noticeable cost in overall validity (Acc 0.988 vs. 0.996). Finally, for SubTB variants, the reported valid-only averages should be interpreted cautiously due to low Acc (≈ 0.30): SubM further concentrates length mass at $L = 9$–10 (Frac > 0.90) without resolving the validity collapse.

Method	Acc	Score	TokEnt	FPDiv	Len_{μ}	Len_{50}	Len_{90}	Frac[0–2]	Frac[3–5]	Frac[6–8]	Frac[9–10]
TB	0.998±0.001	0.717±0.001	2.503±0.026	0.807±0.003	3.06±0.02	2.15±0.03	6.42±0.13	0.601±0.004	0.252±0.007	0.084±0.005	0.064±0.004
TB+SubM	0.996±0.000	0.842±0.001	2.775±0.002	0.889±0.002	6.56±0.09	7.23±0.07	9.84±0.06	0.143±0.006	0.186±0.015	0.348±0.013	0.323±0.008
RapTB	0.996±0.001	0.740±0.004	2.448±0.017	0.860±0.001	6.14±0.03	6.58±0.03	9.01±0.04	0.134±0.007	0.204±0.009	0.466±0.017	0.197±0.015
RapTB+SubM	0.988±0.003	0.844±0.001	2.726±0.017	0.898±0.001	7.44±0.05	7.91±0.03	9.84±0.02	0.046±0.005	0.119±0.003	0.433±0.012	0.402±0.007
SubTB	0.328±0.016	0.755±0.004	2.127±0.037	0.836±0.003	8.35±0.06	9.22±0.05	9.97±0.03	0.006±0.001	0.047±0.005	0.076±0.005	0.871±0.001
SubTB+SubM	0.298±0.006	0.736±0.005	2.165±0.006	0.851±0.002	8.73±0.06	9.59±0.08	9.99±0.01	0.005±0.001	0.029±0.003	0.057±0.001	0.909±0.003

Table 7. All-length averaged SMILES performance and induced length distribution ($L_{\max} = 10$; mean±95% CI over 6 runs). Acc is computed over all samples; Score/TokEnt/FPDiv and Frac[·] are computed on valid samples only.

L	TB				SubTB				RapTB				TB+SubM				SubTB+SubM				RapTB+SubM			
	Acc	Score	Frac	Count	Acc	Score	Frac	Count	Acc	Score	Frac	Count	Acc	Score	Frac	Count	Acc	Score	Frac	Count	Acc	Score	Frac	Count
1	1±0	0.708±0.004	0.215±0.007	687±22.7	1±0	0.762±0.007	0.002±0.001	1.7±0.7	1±0	0.651±0.004	0.042±0.002	133±4.9	1±0	0.712±0.007	0.055±0.007	175±22.6	1±0	0.766±0.001	0.004±0.003	3.7±2.6	1±0	0.702±0.006	0.02±0.002	64.3±7.5
2	0.999±0.001	0.752±0.002	0.387±0.008	1235.7±28.1	1±0	0.748±0.003	0.017±0.003	17.7±3.5	0.999±0.002	0.725±0.004	0.092±0.006	294.7±17.4	1±0	0.79±0	0.088±0.003	282±7.9	1±0	0.75±0.014	0.012±0.004	11.7±4	1±0	0.783±0.005	0.026±0.003	83.3±10.8
3	0.998±0.002	0.727±0.002	0.138±0.003	439.3±10.5	1±0	0.774±0.01	0.009±0.004	9.3±4	1±0	0.754±0.001	0.056±0.002	177±6.8	1±0	0.787±0.003	0.058±0.008	184.3±26.7	1±0	0.759±0.008	0.008±0.003	8±3	0.996±0.008	0.759±0.006	0.03±0.006	95±19.3
4	0.998±0.003	0.722±0.001	0.062±0.002	199.3±6.8	1±0	0.787±0.003	0.035±0.003	36.3±2.4	0.995±0.006	0.748±0.003	0.058±0.004	185.7±13.6	0.997±0.005	0.793±0.003	0.046±0.005	145.3±16.8	1±0	0.789±0.006	0.024±0.003	23.3±2.4	1±0	0.813±0.012	0.031±0.002	99.3±5.3
5	0.996±0.008	0.739±0.003	0.052±0.002	165.3±5.8	1±0	0.813±0.001	0.099±0.011	103.7±9.9	0.998±0.004	0.78±0.009	0.09±0.005	288.3±14.6	0.997±0.002	0.851±0.004	0.083±0.004	263.7±13.6	1±0	0.815±0.001	0.066±0.011	63±10.4	0.995±0.006	0.836±0.003	0.059±0.004	185.3±12.1
6	0.995±0.005	0.736±0.019	0.041±0.005	131±17.4	1±0	0.814±0.001	0.057±0.015	60±13.8	0.997±0.001	0.79±0.004	0.155±0.007	494.7±23.1	0.998±0.002	0.844±0.005	0.094±0.004	301±13.1	1±0	0.819±0.001	0.051±0.007	48.7±6.2	0.995±0.005	0.864±0.003	0.107±0.008	338.3±24.9
7	0.995±0.009	0.732±0.017	0.021±0.002	68.7±7.4	1±0	0.829±0.003	0.086±0.009	90±9.8	0.997±0.002	0.785±0.004	0.17±0.003	540.3±8.2	0.997±0.002	0.86±0.004	0.122±0.006	387.7±18.9	1±0	0.831±0.005	0.064±0.002	60.7±3.5	0.993±0.003	0.854±0.002	0.177±0.009	560±26.8
8	0.995±0.011	0.626±0.012	0.021±0.002	66.3±5.7	1±0	0.834±0.002	0.089±0.007	93.7±4.7	0.994±0.005	0.775±0.006	0.141±0.009	449±29.3	0.998±0.002	0.861±0.001	0.133±0.007	422.7±23.6	1±0	0.837±0.005	0.077±0.006	73±5.2	0.992±0.006	0.842±0.001	0.151±0.015	477.3±46.2
9	1±0	0.529±0.008	0.04±0.002	127.3±5.1	1±0	0.834±0.002	0.081±0.002	84.7±4	0.996±0.001	0.666±0.005	0.141±0.012	448.3±37.9	0.998±0.003	0.88±0.003	0.139±0.004	443.3±13.6	1±0	0.838±0.005	0.051±0.01	48.3±10.3	0.997±0.003	0.854±0.002	0.218±0.007	689.3±24.3
10	0.979±0.028	0.442±0.013	0.023±0.003	74.7±8.5	0.205±0.019	0.699±0.004	0.527±0.024	554±52	0.98±0.019	0.561±0.01	0.055±0.006	176±19.8	0.988±0.002	0.875±0.002	0.183±0.012	583.3±38.4	0.215±0.007	0.69±0.009	0.643±0.016	614.3±22.5	0.956±0.007	0.861±0.002	0.18±0.003	568.3±7.9

Table 8. Per-length valid-only core metrics of SMILES generation (mean±95% CI, $L_{max} = 10$).

L	TB		SubTB		RapTB		TB+SubM		SubTB+SubM		RapTB+SubM	
	TokEnt	FPDiv	TokEnt	FPDiv	TokEnt	FPDiv	TokEnt	FPDiv	TokEnt	FPDiv	TokEnt	FPDiv
1	2.39±0.07	0.789±0.012	0.23±0.45	0.25±0.283	2.28±0.11	0.824±0.005	1.35±0.09	0.628±0.014	0.17±0.33	0.2±0.226	1.32±0.08	0.651±0.027
2	2.54±0.01	0.812±0.005	1.05±0.33	0.445±0.069	2.33±0.03	0.827±0.003	2.12±0.07	0.84±0.004	0.99±0.09	0.435±0.056	1.82±0.13	0.759±0.022
3	2.44±0.02	0.825±0.001	1.1±0.17	0.722±0.022	1.99±0.03	0.818±0.003	2.21±0.07	0.877±0.002	0.87±0.43	0.604±0.091	1.76±0.04	0.838±0.002
4	2.46±0.05	0.854±0.002	1.35±0.14	0.703±0.03	2.02±0.06	0.867±0.005	2.39±0.01	0.909±0.005	1.31±0.36	0.739±0.05	2.19±0.07	0.888±0.007
5	2.41±0.02	0.88±0.005	1.19±0.11	0.714±0.01	2.21±0.01	0.859±0.003	2.48±0.04	0.906±0.004	1.15±0.05	0.721±0.004	2.21±0.12	0.877±0.015
6	2.42±0.08	0.886±0.002	1.7±0.06	0.836±0.008	2.15±0.04	0.88±0.005	2.44±0.05	0.92±0.001	1.69±0.04	0.845±0.001	2.37±0.03	0.886±0.007
7	2.35±0.02	0.891±0.003	1.5±0.05	0.836±0.007	2.18±0.02	0.884±0.001	2.49±0.01	0.926±0.001	1.68±0.02	0.864±0.008	2.47±0.03	0.919±0.003
8	1.61±0.23	0.703±0.049	1.62±0.03	0.834±0.012	2.17±0.02	0.882±0.001	2.54±0.01	0.918±0.001	1.55±0.06	0.829±0.01	2.59±0.03	0.921±0.001
9	1.35±0.05	0.693±0.016	1.71±0.07	0.853±0.002	1.98±0.04	0.848±0.006	2.36±0.02	0.909±0.001	1.67±0.05	0.857±0.008	2.28±0.05	0.916±0.001
10	1.06±0.11	0.58±0.023	2.16±0.04	0.881±0	1.75±0.03	0.815±0.004	2.08±0.02	0.906±0.001	2.18±0.01	0.884±0.003	2.16±0.02	0.91±0

Table 9. Per-length valid-only diversity metrics of SMILES generation(mean±95% CI, $L_{max} = 10$).

L	TB					SubTB					RapTB					TB+SubM					SubTB+SubM					RapTB+SubM				
	UniqStr	UniqMol	UniqRateStr	UniqRateMol	UniqStr	UniqMol	UniqRateStr	UniqRateMol	UniqStr	UniqMol	UniqRateStr	UniqRateMol	UniqStr	UniqMol	UniqRateStr	UniqRateMol	UniqStr	UniqMol	UniqRateStr	UniqRateMol	UniqStr	UniqMol	UniqRateStr	UniqRateMol	UniqStr	UniqMol	UniqRateStr	UniqRateMol
1	29.3±1.1	25.3±1.5	0±0	0±0	1.3±0.4	1.3±0.4	0.8±0.2	0.8±0.2	17.3±0.8	15.7±0.4	0.1±0	0.1±0	13.7±0.8	13.7±0.8	0.1±0	0.1±0	1.3±0.4	1.3±0.4	0.5±0.3	0.5±0.3	8.3±0.8	8.3±0.8	0.1±0	0.1±0				
2	184.7±2.1	115.7±3	0.1±0	0.2±0	6.3±0.8	6±0.7	0.4±0	0.3±0	77±1.9	71±2.5	0.3±0	0.2±0	60.3±3	59.7±2.9	0.2±0	0.2±0	5±0.7	5±0.7	0.4±0	0.4±0	27.3±1.5	27.3±1.5	0.3±0	0.3±0				
3	224±7.5	215.7±7.3	0.5±0	0.5±0	6.3±1.7	6.3±1.7	0.7±0.1	0.7±0.1	78.3±2.2	76.3±2.2	0.4±0	0.4±0	52.3±5.8	52.3±5.8	0.3±0	0.3±0	4.7±1.8	4.7±1.8	0.6±0.1	0.6±0.1	27.7±3	27.7±3	0.3±0	0.3±0				
4	166±4	164±3.6	0.8±0	0.8±0	17±1.2	16.7±0.8	0.5±0.1	0.5±0	111±3.1	107.3±3	0.6±0	0.6±0	58±0.7	57.7±1.1	0.4±0	0.4±0	14.7±2.2	13.7±1.8	0.6±0.1	0.6±0.1	41±6.1	41±6.1	0.4±0.1	0.4±0.1				
5	156±5.9	153.3±5.8	0.9±0	0.9±0	41.3±3.7	38.3±3.9	0.4±0	0.4±0	187±7.6	181.7±7.9	0.6±0	0.6±0	101.3±3.5	100.7±3.3	0.4±0	0.4±0	31.3±4.2	28.3±3.5	0.5±0	0.4±0	74±8.1	73.7±7.9	0.4±0	0.4±0				
6	129.7±10.2	128.7±9.5	1±0	1±0	43.3±6.6	40±6.6	0.7±0	0.7±0	332.3±4.2	322±5	0.7±0	0.7±0	130.3±12.8	130±12.4	0.4±0	0.4±0	39±5.4	37±5.2	0.8±0	0.8±0	123.3±6	123.3±6	0.4±0	0.4±0				
7	68±4.7	68±4.7	1±0	1±0	60±3.1	55±2.5	0.7±0	0.6±0	448.7±7.2	411.7±3.3	0.8±0	0.8±0	157±8.7	156.3±8.4	0.4±0	0.4±0	52.3±3	50.3±1.8	0.9±0	0.8±0	235±6.1	220.7±0.4	0.4±0	0.4±0				
8	35±5	35±5	0.5±0.1	0.5±0.1	73.3±8.1	69.3±7	0.8±0.1	0.7±0.1	409.7±15.1	400±12.5	0.9±0	0.9±0	163±6.8	162.3±6.5	0.4±0	0.4±0	58.7±3.2	56.3±3.2	0.8±0	0.8±0	299±13.6	298.3±13.7	0.6±0	0.6±0				
9	38.7±1.1	38.7±1.1	0.3±0	0.3±0	76±6.2	72±5.2	0.9±0.1	0.8±0	246.3±6	244.3±5.4	0.6±0	0.5±0	159.7±4.3	159.3±4.7	0.4±0	0.4±0	45.7±5.4	43.3±5.5	0.9±0	0.9±0	353.7±5	291.7±5.3	0.5±0	0.6±0				
10	33±4.4	33±4.4	0.4±0	0.4±0	482.3±31.8	435.3±14.1	0.9±0	0.9±0	96.3±3.3	95.7±2.9	0.5±0	0.5±0	156.3±4.6	145.3±3.7	0.3±0	0.3±0	558±8.7	467.3±4.2	0.9±0	0.9±0	242±5.9	225.3±2.9	0.4±0	0.4±0				

Table 10. Per-length valid-only uniqueness metrics of SMILES generation (mean±95% CI, $L_{\max} = 10$).

A.1.2. Per-length prefix collapse analysis of SMILES generation.

Table 11 reports prefix distributions at depth k (mean±95% CI), where Survival should be read jointly with concentration metrics (Entropy/Eff/Top1/UniqueRate). Without SubM, TB exhibits rapid attrition (e.g., Survival drops to 0.105 at k=7 and 0.023 at k=10) and increased concentration among the remaining deep prefixes (Top1 reaches 0.406 at k=7), consistent with prefix collapse. RapTB substantially improves Survival at larger k (e.g., 0.497 at k=7) while reducing deep-prefix concentration (Top1 0.132 at k=7), indicating more sustained branching beyond early decisions. Enabling SubM (Table 12) further alleviates collapse for both TB and RapTB by increasing deep-prefix diversity (higher Entropy/Eff, lower Top1) and improving Survival at large k (e.g., TB: $0.023 \rightarrow 0.183$ and RapTB: $0.057 \rightarrow 0.180$ at k=10).

k	TB					SubTB					RapTB				
	Survival	Ent	Eff	Top1	UniqueRate	Survival	Ent	Eff	Top1	UniqueRate	Survival	Ent	Eff	Top1	UniqueRate
1	1±0	2.531±0.022	12.57±0.28	0.274±0.01	0.013±0.001	1±0	1.539±0.036	4.66±0.17	0.419±0.01	0.017±0.001	1±0	1.841±0.025	6.3±0.15	0.4±0.004	0.01±0
2	0.785±0.005	4.748±0.019	115.34±2.25	0.077±0.001	0.135±0.002	0.998±0	3.544±0.063	34.69±2.15	0.096±0.004	0.096±0.002	0.958±0.001	3.838±0.023	46.46±1.05	0.174±0.005	0.076±0.002
3	0.398±0.002	5.418±0.027	225.52±6.1	0.121±0.005	0.45±0.005	0.982±0.002	4.563±0.044	95.94±4.19	0.088±0.006	0.24±0.003	0.866±0.005	4.717±0.031	111.95±3.52	0.184±0.005	0.215±0.005
4	0.261±0.002	5.17±0.038	176.02±6.81	0.167±0.009	0.566±0.012	0.973±0.004	5.054±0.035	156.78±5.48	0.088±0.006	0.355±0.003	0.81±0.006	5.45±0.047	233.1±11.25	0.135±0.007	0.357±0.01
5	0.198±0.002	4.796±0.041	121.2±4.96	0.219±0.014	0.599±0.009	0.938±0.005	5.575±0.048	264.06±12.39	0.03±0.003	0.471±0.004	0.752±0.004	5.747±0.051	313.88±16.34	0.138±0.007	0.474±0.01
6	0.146±0.003	4.051±0.075	57.65±4.37	0.293±0.021	0.533±0.007	0.839±0.011	6.012±0.046	408.87±18.66	0.026±0.001	0.638±0.006	0.662±0.006	6.027±0.044	414.97±17.93	0.102±0.004	0.575±0.012
7	0.105±0.002	2.928±0.02	18.7±0.37	0.406±0.019	0.363±0.008	0.782±0.012	6.175±0.07	482.25±33.04	0.022±0.002	0.738±0.01	0.506±0.004	6.011±0.05	408.6±20.46	0.132±0.004	0.637±0.01
8	0.084±0.001	3.018±0.042	20.48±0.84	0.217±0.008	0.246±0.012	0.696±0.011	6.213±0.081	501.27±39.04	0.02±0	0.803±0.016	0.337±0.003	5.492±0.072	243.57±17.66	0.126±0.011	0.6±0.013
9	0.063±0.002	3.378±0.059	29.37±1.76	0.159±0.008	0.27±0.009	0.607±0.015	6.124±0.076	458.44±33.86	0.022±0.001	0.83±0.015	0.196±0.009	4.774±0.031	118.44±3.61	0.16±0.01	0.474±0.012
10	0.023±0.002	3.119±0.131	22.88±3.02	0.185±0.032	0.442±0.045	0.527±0.015	6.065±0.076	431.97±31.82	0.022±0.003	0.871±0.01	0.055±0.004	4.276±0.022	71.97±1.57	0.049±0.007	0.549±0.023

Table 11. Prefix statistics by depth. Mean±95% CI.

k	TB+SubM					SubTB+SubM					RapTB+SubM				
	Survival	Ent	Eff	Top1	UniqueRate	Survival	Ent	Eff	Top1	UniqueRate	Survival	Ent	Eff	Top1	UniqueRate
1	1±0	2.715±0.017	15.11±0.25	0.165±0.006	0.011±0	1±0	1.559±0.037	4.76±0.18	0.394±0.01	0.017±0.001	1±0	2.563±0.004	12.98±0.05	0.193±0.005	0.009±0
2	0.945±0.004	4.778±0.004	118.92±0.46	0.052±0.003	0.094±0.001	0.996±0.002	3.669±0.041	39.26±1.58	0.094±0.005	0.108±0.004	0.98±0.001	4.43±0.015	83.96±1.27	0.065±0.001	0.076±0.001
3	0.857±0.004	5.224±0.01	185.6±1.81	0.033±0.002	0.164±0.005	0.984±0.002	4.723±0.024	112.55±2.69	0.089±0.004	0.278±0.008	0.953±0.003	5.019±0.022	151.34±3.33	0.042±0.006	0.141±0.001
4	0.799±0.005	5.434±0.014	229.16±3.26	0.034±0.002	0.211±0.006	0.975±0.004	5.275±0.021	195.48±4.12	0.09±0.004	0.437±0	0.923±0.007	5.342±0.036	209.02±7.37	0.037±0.002	0.196±0.004
5	0.753±0.008	5.667±0.021	289.19±6	0.024±0.002	0.254±0.007	0.951±0.004	5.765±0.014	318.94±4.35	0.029±0.001	0.563±0	0.892±0.006	5.641±0.047	282.07±13.11	0.036±0.004	0.254±0.006
6	0.671±0.01	5.75±0.019	314.29±5.93	0.019±0.001	0.29±0.006	0.885±0.007	6.136±0.02	462.51±9.06	0.025±0.002	0.704±0.014	0.833±0.004	6.002±0.019	404.41±7.56	0.025±0.001	0.326±0.002
7	0.576±0.01	5.617±0.009	275.21±2.49	0.022±0.001	0.299±0.003	0.834±0.006	6.29±0.021	539.52±11.44	0.015±0.002	0.793±0.013	0.726±0.009	6.189±0.023	487.47±11.36	0.022±0.003	0.404±0.001
8	0.455±0.007	5.363±0.015	213.33±3.24	0.028±0.001	0.293±0.002	0.771±0.005	6.292±0.016	540.19±8.66	0.015±0.002	0.832±0.012	0.549±0.005	6.116±0.022	453.29±9.84	0.02±0.002	0.456±0.003
9	0.322±0.005	4.997±0.014	148.03±2.02	0.039±0.001	0.284±0.006	0.694±0.008	6.242±0.014	513.7±6.98	0.016±0.002	0.865±0.011	0.398±0.005	5.79±0.017	327.02±5.68	0.024±0.003	0.445±0.005
10	0.183±0.008	4.447±0.025	85.38±2.14	0.053±0.007	0.268±0.012	0.643±0.01	6.262±0.013	524.33±6.63	0.01±0.001	0.909±0.01	0.18±0.002	4.936±0.021	139.26±2.9	0.046±0.006	0.426±0.007

Table 12. Prefix statistics by depth (Continue). Mean±95% CI.

A.1.3. Per-length SMILES performance on long horizon.

Table 13–15 show that increasing the horizon amplifies length-wise failure modes. TB rapidly loses support on long valid trajectories: its Frac/Count becomes negligible beyond $L \geq 12$ (e.g., Frac 0.003 at $L = 12$ and effectively zero thereafter), and the corresponding Score degrades sharply at long lengths (e.g., $0.544/0.471/0.433/0.375$ for $L = 9$–12), indicating severe long-horizon under-coverage. RapTB maintains substantially higher terminal quality on long trajectories (Score ≈ 0.75–0.81 for $L = 10$–14 with near-perfect Acc), but still under-allocates mass to the extreme tail ($L = 15$) as reflected by a small Frac/Count. Combining RapTB with SubM shifts probability mass back to long lengths without sacrificing quality, yielding strong tail performance (at $L = 15$, Acc 0.84 and Score 0.85 with a much larger Frac/Count), and also improves long-length diversity/uniqueness compared to RapTB alone (Tables 14, 15). In contrast, SubTB places substantial mass on very long lengths (e.g., large Frac at $L = 15$) but exhibits low Acc/Score there, consistent with the termination/length instability discussed in the main text.

A.1.4. Per-depth prefix collapse on long horizon ($L_{\max} = 15$).

Tables 16–17 report prefix statistics at depth k, where concentration metrics (Entropy/Eff/Top1/UniqueRate) must be interpreted jointly with Survival. On the long horizon, TB exhibits a textbook collapse pattern: Survival drops from 0.758

L	TB				SubTB				RapTB				RapTB+SubM			
	Acc	Score	Frac	Count	Acc	Score	Frac	Count	Acc	Score	Frac	Count	Acc	Score	Frac	Count
1	1±0	0.707±0.003	0.242±0.012	774.7±37.9	1±0	0.765±0	0.002±0.001	1.7±0.7	1±0	0.7±0.005	0.033±0.004	105.7±11.6	1±0	0.694±0.012	0.018±0.001	56.5±2.9
2	1±0	0.748±0	0.34±0.01	1086.7±33.2	1±0	0.738±0.009	0.05±0.002	51.3±2.6	1±0	0.723±0.003	0.033±0.005	103±16.7	1±0	0.798±0.004	0.018±0.008	57.5±26.5
3	0.999±0.002	0.724±0.007	0.128±0.005	409±16.3	1±0	0.723±0.016	0.03±0.006	31±6	1±0	0.747±0.012	0.011±0.001	34.3±2.4	1±0	0.824±0.003	0.016±0.004	50.5±12.7
4	0.999±0.002	0.739±0	0.088±0.006	281.3±20.3	1±0	0.796±0.008	0.069±0.011	70.7±11.3	0.995±0.01	0.758±0.005	0.017±0.002	55.3±7.5	1±0	0.853±0.017	0.016±0.006	51±17.6
5	0.997±0.003	0.749±0.001	0.06±0.003	192.3±10.3	1±0	0.795±0.003	0.135±0.004	137±5.2	1±0	0.779±0.01	0.019±0.008	60±25.5	1±0	0.888±0.01	0.025±0.001	77±2
6	1±0	0.736±0.011	0.043±0.002	138.3±7.3	1±0	0.794±0.009	0.066±0.012	66.7±11.4	0.993±0.013	0.798±0.014	0.017±0.002	55.3±6.8	1±0	0.879±0.004	0.031±0.001	95.5±4.9
7	0.995±0.01	0.741±0.021	0.022±0.002	69.3±7.5	0.997±0.005	0.811±0.004	0.138±0.018	140.7±17.4	0.99±0.011	0.799±0.012	0.03±0.002	95.7±5.7	0.996±0.008	0.893±0.006	0.034±0.008	107±23.5
8	0.982±0.018	0.705±0.021	0.012±0.001	37±4.5	1±0	0.821±0.004	0.107±0.009	108.7±9.2	0.987±0.006	0.776±0.009	0.031±0.003	99±7.9	0.995±0.011	0.863±0.013	0.029±0.001	91±2
9	0.997±0.005	0.544±0.024	0.032±0.006	101±19.7	1±0	0.808±0.003	0.07±0.003	71±3	0.997±0.005	0.706±0.003	0.084±0.009	266.3±28	1±0	0.871±0.02	0.045±0.009	141±27.4
10	1±0	0.471±0.002	0.021±0.001	65.7±2.4	1±0	0.818±0.007	0.062±0.004	63.3±4	0.997±0.001	0.751±0.008	0.155±0.008	490.3±22.9	1±0	0.862±0.001	0.066±0.004	204±13.7
11	1±0	0.433±0.034	0.009±0.002	28±6.3	1±0	0.826±0.011	0.041±0.006	42±5.9	0.995±0.001	0.757±0.002	0.154±0.003	487.7±10.5	0.988±0.017	0.847±0.008	0.094±0.004	292±11.8
12	1±0	0.375±0.014	0.003±0.001	10±3.4	1±0	0.803±0.013	0.023±0.003	23.3±3.5	0.995±0.004	0.787±0.002	0.158±0.007	498.7±24.2	0.996±0.003	0.853±0.002	0.126±0.01	391.5±32.3
13	1±0	0.711±0	0±0	1±0	1±0	0.797±0.013	0.011±0.003	11.7±3.6	0.995±0.001	0.807±0.002	0.155±0.011	489±33.3	0.993±0.002	0.832±0.005	0.16±0.003	499±9.8
14	1±0	0.332±0.112	0.001±0	3.7±1.3	1±0	0.832±0.014	0.009±0.004	8.7±3.6	0.997±0.003	0.795±0.007	0.08±0.009	254±27.5	0.992±0.006	0.853±0.005	0.192±0.003	596±7.8
15	–	–	–	–	0.247±0.01	0.494±0.02	0.187±0.007	190.3±7.7	0.748±0.041	0.763±0.013	0.022±0.001	68.7±4	0.84±0.023	0.85±0.006	0.129±0.005	400±17.6

Table 13. Per-length valid-only core metrics of SMILES generation (mean±95% CI, $L_{\max} = 15$).

L	TB		SubTB		RapTB		RapTB+SubM	
	TokEnt	FPDiv	TokEnt	FPDiv	TokEnt	FPDiv	TokEnt	FPDiv
1	2.45±0.04	0.82±0.011	0±0	0±0	1.8±0.06	0.625±0.009	0.92±0.03	0.455±0.024
2	2.57±0	0.813±0.004	0.83±0.21	0.344±0.071	2.07±0.12	0.802±0.004	1.04±0.2	0.661±0.025
3	2.49±0	0.834±0	1.49±0.05	0.793±0.005	1.6±0.14	0.706±0.02	1.21±0.24	0.682±0.027
4	2.18±0.03	0.791±0.009	1.21±0.08	0.673±0.012	1.27±0.23	0.748±0.04	1.17±0.06	0.653±0.026
5	2.42±0.04	0.868±0.003	1.25±0.02	0.748±0.007	1.72±0.05	0.825±0.008	1.32±0.19	0.793±0.017
6	2.5±0.06	0.879±0.004	1.62±0.02	0.841±0.005	1.53±0.19	0.79±0.028	1.73±0.04	0.879±0.003
7	2.32±0	0.879±0.003	0.93±0.05	0.705±0.015	1.55±0.11	0.826±0.006	1.69±0.1	0.857±0.013
8	2.23±0.03	0.881±0.005	1.19±0.09	0.796±0.007	1.82±0.02	0.835±0.003	2.05±0.1	0.873±0.007
9	1.51±0.17	0.718±0.019	1.43±0.03	0.846±0.001	1.75±0.04	0.808±0.005	1.91±0.06	0.872±0.009
10	1.18±0.04	0.648±0.006	1.2±0.13	0.782±0.019	1.72±0.02	0.833±0.003	2.17±0.03	0.891±0.007
11	0.82±0.23	0.524±0.048	1.33±0.07	0.835±0.006	1.74±0.01	0.835±0.001	2.08±0	0.883±0.006
12	0.46±0.03	0.422±0.029	1.38±0.14	0.846±0.003	1.57±0.02	0.819±0.001	2.05±0.01	0.9±0.001
13	0±0	–	1.24±0.3	0.83±0.017	1.38±0.02	0.797±0.003	2.02±0.06	0.894±0.003
14	0.22±0.35	0.39±0.209	1.09±0.12	0.813±0.015	1.46±0.05	0.82±0.002	2±0.03	0.897±0
15	–	–	2.29±0.05	0.884±0.001	1.57±0.07	0.825±0.007	1.85±0.02	0.889±0

Table 14. Per-length valid-only diversity metrics of SMILES generation (mean±95% CI, $L_{\max} = 15$).

L	TB				SubTB				RapTB				RapTB+SubM			
	UniqStr	UniqMol	UniqRateStr	UniqRateMol	UniqStr	UniqMol	UniqRateStr	UniqRateMol	UniqStr	UniqMol	UniqRateStr	UniqRateMol	UniqStr	UniqMol	UniqRateStr	UniqRateMol
1	31±1.2	24.7±3	0±0	0±0	1±0	1±0	0.7±0.2	0.7±0.2	14±1.4	11.3±1.5	0.1±0	0.1±0	5.5±0.6	5.5±0.6	0.1±0	0.1±0
2	189±3.6	128.7±4.3	0.2±0	0.3±0	7.3±0.4	6.3±0.4	0.1±0	0.1±0	35.7±4.2	33±4.5	0.3±0	0.3±0	8.5±1.7	8.5±1.7	0.1±0	0.1±0
3	215.7±2.7	203.7±1.1	0.5±0	0.5±0	14.3±0.8	14.3±0.8	0.5±0.1	0.5±0.1	20.7±1.5	19.7±1.8	0.6±0	0.6±0	8.5±0.6	8.5±0.6	0.2±0	0.2±0
4	170±9.3	164.7±8.6	0.6±0	0.6±0	22±1.4	20.7±1.8	0.3±0	0.3±0	32±4.5	31.3±4.8	0.6±0.1	0.6±0.1	15.5±1.7	15.5±1.7	0.3±0	0.3±0
5	178.7±6.6	176±6.8	0.9±0	0.9±0	47±1.2	42.3±0.4	0.3±0	0.3±0	51.3±11	48.7±7.7	0.9±0	0.8±0.1	23±0	23±0	0.3±0	0.3±0
6	133±5.4	131.7±5.7	1±0	1±0	43.7±3	43±2.9	0.7±0	0.7±0.1	41.3±5	40±5.7	0.7±0.1	0.7±0.1	36.5±0.6	36.5±0.6	0.4±0	0.4±0
7	69.3±4.8	69.3±4.8	1±0	1±0	55.3±3	48±3.8	0.4±0	0.3±0	76.7±2.3	72.3±3.3	0.8±0	0.8±0	45±7.9	44.5±7.4	0.4±0	0.4±0
8	37±2.9	36.7±2.5	1±0	1±0	68±2.6	63.3±1.7	0.6±0	0.6±0	94.7±3.5	93±2.6	1±0	0.9±0	39.5±4	39.5±4	0.4±0	0.4±0
9	40±3.3	40±3.3	0.4±0	0.4±0	55.3±1.5	54.7±1.7	0.8±0	0.8±0	167.3±4.7	157.7±2.2	0.6±0	0.6±0	63±4.5	63±4.5	0.4±0	0.4±0
10	26.3±1.5	26.3±1.5	0.4±0	0.4±0	41.7±1.5	41.3±1.7	0.7±0	0.7±0	358.7±7.2	324.7±6.5	0.7±0	0.7±0	101±0	101±0	0.5±0	0.5±0
11	14.7±1.1	14.7±1.1	0.5±0.1	0.5±0.1	35.7±3.6	35.7±3.6	0.8±0	0.8±0	415.7±9.9	397.3±8.8	0.9±0	0.8±0	139.5±6.2	139.5±6.2	0.5±0	0.5±0
12	5.7±0.8	5.7±0.8	0.6±0.1	0.6±0.1	22.3±2.3	22.3±2.3	1±0.1	1±0.1	391±10.3	360±10	0.8±0	0.7±0	226.5±4	226.5±4	0.6±0	0.6±0
13	1±0	1±0	1±0	1±0	11.3±2.2	11.3±2.2	1±0	1±0	357.7±12.8	328.7±11.5	0.7±0	0.7±0	334.5±18.7	334±19.2	0.7±0	0.7±0
14	2.3±0.4	2.3±0.4	0.7±0.2	0.7±0.2	8.7±2.3	8.7±2.3	1±0	1±0	219.7±12.6	211.3±12.5	0.9±0	0.8±0	303±9.1	261.5±11.9	0.5±0	0.5±0
15	–	–	–	–	186±5.7	186±5.7	1±0	1±0	67±2.1	66.7±1.8	1±0	1±0	217.5±11.9	217.5±11.9	0.5±0	0.5±0

Table 15. Per-length valid-only uniqueness metrics of SMILES generation (mean±95% CI, $L_{\max} = 15$).

at k=2 to 0.034 at k=10 and is essentially zero for $k \geq 12$, while Top1 peaks at $0.424/0.457$ for k=7/8, indicating that only a tiny fraction of trajectories reach deep prefixes and those prefixes are highly shared. RapTB substantially improves deep-prefix survival (e.g., 0.723 at k=10) and keeps Top1 low at depth ($\approx$ 0.06–0.07 for k=7–10), consistent with sustained branching and reduced prefix concentration; RapTB+SubM further lowers deep-prefix Top1 while maintaining high Survival, suggesting that improved replay coverage helps prevent the buffer from over-focusing on a few dominant prefixes under long-horizon generation. At $L_{\max} = 15$, we only report RapTB+SubM as the combined setting due to the additional training cost of re-running all baseline+SubM variants at this horizon. Finally, although SubTB shows strong prefix-level dispersion (low Top1 with high Entropy/Eff and high Survival), this alone does not imply better terminal quality or validity; thus these prefix statistics should be interpreted together with the per-length terminal metrics in Tables 13–15, where SubTB degrades on long-length performance.

k	TB					SubTB				
	Survival	Ent	Eff	Top1	UniqueRate	Survival	Ent	Eff	Top1	UniqueRate
1	1±0	2.637±0.021	13.97±0.3	0.236±0.009	0.013±0	1±0	1.235±0.016	3.44±0.05	0.508±0.007	0.012±0.001
2	0.758±0.007	4.775±0.027	118.61±3.24	0.082±0.005	0.152±0.008	0.998±0	3.085±0.009	21.87±0.2	0.164±0.009	0.07±0.003
3	0.418±0.005	5.495±0.048	243.8±11.8	0.108±0.008	0.447±0.01	0.948±0.002	3.987±0.031	53.93±1.7	0.098±0	0.179±0.004
4	0.29±0.006	5.452±0.052	233.74±12.43	0.145±0.008	0.591±0.016	0.917±0.002	4.466±0.035	87.04±3	0.082±0.004	0.262±0.003
5	0.202±0.004	5.178±0.092	178.4±16.91	0.208±0.014	0.682±0.02	0.848±0.007	5.041±0.093	155.54±14.12	0.058±0.008	0.375±0.016
6	0.142±0.006	4.373±0.141	80.31±11.33	0.294±0.02	0.607±0.025	0.714±0.008	5.393±0.093	221.07±20.27	0.066±0.011	0.522±0.016
7	0.099±0.005	3.278±0.175	27.03±4.53	0.424±0.027	0.457±0.031	0.648±0.008	5.46±0.069	235.93±16.3	0.071±0.012	0.593±0.011
8	0.077±0.004	2.639±0.112	14.11±1.54	0.457±0.002	0.316±0.031	0.51±0.005	5.702±0.033	299.69±10.07	0.03±0.006	0.727±0.008
9	0.065±0.003	2.835±0.08	17.1±1.37	0.309±0.03	0.28±0.024	0.403±0.004	5.654±0.036	285.55±10.24	0.038±0.007	0.815±0.012
10	0.034±0.001	2.977±0.088	19.73±1.72	0.154±0.018	0.333±0.031	0.333±0.002	5.547±0.039	256.65±9.97	0.045±0.007	0.864±0.013
11	0.013±0.002	2.415±0.052	11.21±0.59	0.326±0.015	0.441±0.031	0.271±0.004	5.51±0.03	247.38±7.36	0.019±0.001	0.933±0.008
12	0.004±0.001	1.737±0.216	5.86±1.35	0.365±0.045	0.517±0.046	0.23±0.004	5.398±0.029	221.18±6.45	0.021±0.003	0.967±0.005
13	0.001±0	0.347±0.43	1.61±0.76	0.833±0.207	0.428±0.205	0.207±0.006	5.304±0.039	201.39±7.93	0.024±0.004	0.976±0.004
14	0.001±0	0.745±0.224	2.18±0.52	0.6±0.172	0.689±0.215	0.195±0.005	5.252±0.037	191.04±7.2	0.024±0.003	0.978±0.005
15	0±0	0±0	0±0	0±0	–	0.187±0.004	5.205±0.035	182.38±6.38	0.025±0.003	0.977±0.005

Table 16. Prefix statistics by depth on SMILES generation (mean±95% CI, $L_{\max} = 15$): TB vs. SubTB.

k	RapTB					RapTB+SubM				
	Survival	Ent	Eff	Top1	UniqueRate	Survival	Ent	Eff	Top1	UniqueRate
1	1±0	1.693±0.014	5.44±0.08	0.375±0.004	0.008±0	1±0	1.938±0.001	6.94±0.01	0.342±0	0.007±0
2	0.967±0.002	3.438±0.033	31.14±1.03	0.204±0.006	0.051±0.002	0.982±0.001	3.659±0.011	38.83±0.43	0.125±0.009	0.039±0
3	0.934±0.001	4.542±0.04	93.95±3.74	0.106±0.006	0.138±0.006	0.963±0.005	4.227±0.033	68.53±2.26	0.126±0.01	0.09±0.003
4	0.923±0.001	5.249±0.026	190.42±4.86	0.071±0.002	0.238±0.004	0.947±0.003	4.829±0.009	125.11±1.11	0.094±0.003	0.148±0.002
5	0.906±0.001	5.822±0.019	337.73±6.41	0.07±0.005	0.345±0.003	0.931±0	5.184±0.001	178.44±0.24	0.095±0.003	0.204±0.002
6	0.887±0.005	6.239±0.013	512.44±6.71	0.071±0.005	0.445±0.004	0.906±0	5.701±0.007	299.31±2.04	0.044±0.002	0.263±0.002
7	0.869±0.005	6.559±0.021	705.68±14.92	0.073±0.005	0.542±0.002	0.875±0.001	6.07±0.015	432.83±6.29	0.02±0.001	0.322±0.002
8	0.839±0.004	6.78±0.016	879.89±14.09	0.058±0.004	0.614±0.001	0.841±0.004	6.151±0.022	469.45±10.24	0.018±0.002	0.357±0.003
9	0.808±0.003	6.89±0.013	982.52±12.91	0.053±0.004	0.658±0.002	0.812±0.003	6.203±0.001	494.29±0.41	0.019±0.002	0.382±0.004
10	0.723±0.008	7.041±0.005	1142.49±5.93	0.024±0.001	0.718±0.006	0.766±0.008	6.313±0.017	551.78±9.32	0.02±0.002	0.432±0.001
11	0.568±0.011	6.93±0.016	1022.69±16.25	0.024±0	0.769±0.007	0.701±0.006	6.386±0.007	593.32±4.26	0.022±0.002	0.481±0.002
12	0.414±0.01	6.638±0.006	763.56±4.76	0.03±0.004	0.775±0.007	0.607±0.008	6.371±0.018	584.42±10.32	0.016±0.003	0.52±0.004
13	0.257±0.012	6.201±0.01	493.35±5.11	0.046±0.007	0.793±0.009	0.481±0.003	6.205±0.02	495.45±9.87	0.02±0.004	0.548±0
14	0.102±0.005	5.587±0.029	267.14±7.85	0.024±0.004	0.889±0.021	0.32±0.001	5.677±0.083	292.79±24.31	0.031±0.005	0.511±0.019
15	0.022±0.001	4.195±0.03	66.4±2.02	0.029±0.001	0.976±0.005	0.129±0.003	4.899±0.098	134.67±13.23	0.077±0.015	0.543±0.016

Table 17. Prefix statistics by depth on SMILES generation (mean±95% CI, $L_{\max} = 15$): RapTB vs. RapTB+SubM.

A.2. Expr24

We provide results with 95% CI and per-length termination probability analysis of different objectives under various replay strategy.

B. Metrics: Formal Definitions and Protocol

Sampling and aggregation protocol. For each run, we draw N=6400 i.i.d. terminal samples $\{x_i\}_{i=1}^{N}$ from the learned sampler. Let $\mathbb{I}_{\text{valid}}(x) \in \{0, 1\}$ indicate whether x satisfies task constraints. Let $\mathcal{D}$ denote the multiset of all samples and let

$$\mathcal{D}_{\text{valid}} \triangleq \{x_i \in \mathcal{D} : \mathbb{I}_{\text{valid}}(x_i) = 1\}, \qquad n_{\text{valid}} \triangleq |\mathcal{D}_{\text{valid}}|.$$

Unless explicitly stated otherwise, all metrics *except* `Acc` are computed on $\mathcal{D}_{\text{valid}}$. We report a metric as 0 if its denominator is 0 (e.g., $n_{\text{valid}} = 0$).

Table 18. Expr24 results under four replay schemes. All the experiments are run under 3 different random seeds with 95% CI. Per-run sample size is 6400.

Replay	Objective	Unique$_\checkmark$	NormCov	Acc	KL($\pi \to p^*$)	KL($p^* \to \pi$)	JS$_{\text{tok}}$
PRT	TB	103.7±3.2	0.016±0.001	0.999±0.000	1.105±0.002	7.803±0.060	0.292±0.001
	SubTB	292.0±2.9	0.046±0.000	0.311±0.002	0.424±0.010	0.672±0.077	0.107±0.003
	RapTB	129.3±0.4	0.020±0.000	0.992±0.001	0.908±0.003	5.538±0.005	0.230±0.001
RP	TB	5.3±0.4	0.001±0.000	1.000±0.000	1.297±0.001	11.403±0.282	0.339±0.000
	SubTB	324.7±2.7	0.051±0.000	0.229±0.005	0.455±0.005	0.865±0.083	0.109±0.002
	RapTB	246.7±7.1	0.039±0.001	0.991±0.000	0.561±0.001	4.480±0.002	0.147±0.000
SubM	TB	642.0±5.6	0.100±0.001	0.996±0.001	0.182±0.001	0.441±0.005	0.049±0.000
	SubTB	331.3±22.7	0.052±0.004	0.061±0.005	0.149±0.008	0.286±0.070	0.040±0.002
	RapTB	1337.3±7.5	0.209±0.001	0.994±0.001	0.169±0.001	0.623±0.004	0.048±0.000
Oracle	TB	5198.0±5.2	0.812±0.001	0.919±0.001	0.062±0.001	0.066±0.001	0.016±0.000
	SubTB	35.7±2.9	0.006±0.000	0.006±0.000	0.266±0.009	1.491±0.413	0.071±0.003
	RapTB	5220.7±4.3	0.816±0.001	0.945±0.001	0.052±0.001	0.056±0.001	0.013±0.000

Replay	Objective	$\ell = 3$	$\ell = 5$	$\ell = 7$	$\ell = 9$
PRT	TB	–	–	-0.000	-0.001
	SubTB	-5.220	-1.292	-1.064	-79.638
	RapTB	–	-2.451	-2.319	-0.065
	RootSubTBLogZ	-0.709	-0.606	-0.411	-0.068
Oracle	TB	–	-4.391	-1.779	-0.441
	SubTB	-1.017	-2.803	-4.530	-86.415
	RapTB	–	-8.312	-3.341	-0.644
	RootSubTBLogZ	-0.442	-0.417	-0.354	-1.432

Table 19. Per-length $\log p_{\text{term}}$ on Expr24 (PRT and Oracle replay).

Across random seeds, we report the mean and a two-sided 95% confidence interval. With S seeds and per-seed values $m_1, \ldots, m_S$, we report

$$\bar{m} \pm t_{S-1,0.975} \frac{\mathrm{sd}(m_1, \ldots, m_S)}{\sqrt{S}}.$$

B.1. Terminal-level metrics

Accuracy / validity rate (`Acc`). `Acc` measures the fraction of valid samples among all N draws:

$$\texttt{Acc} \triangleq \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\text{valid}}(x_i). \tag{13}$$

For SMILES, $\mathbb{I}_{\text{valid}}(x) = 1$ requires chemical validity and scaffold consistency. For Expr24, $\mathbb{I}_{\text{valid}}(x) = 1$ requires syntactic validity and $\mathrm{eval}(x) = 24$.

Task score on valid samples (`Score`). Let $s(x)$ be the task score.

$$\texttt{Score} \triangleq \frac{1}{n_{\text{valid}}} \sum_{x \in \mathcal{D}_{\text{valid}}} s(x). \tag{14}$$

For Expr24 with binary reward, $s(x) = \mathbb{I}_{\text{valid}}(x)$, so `Score` numerically equals `Acc`.

Pre-EOS length and length statistics. Each terminal x is a variable-length token sequence with pre-EOS length $\ell(x)$ (number of tokens before EOS/$\top$). We report

$$\texttt{Len} \triangleq \frac{1}{n_{\text{valid}}} \sum_{x \in \mathcal{D}_{\text{valid}}} \ell(x), \qquad \texttt{Len}_{50},\ \texttt{Len}_{90} \text{ as percentiles of } \{\ell(x) : x \in \mathcal{D}_{\text{valid}}\}.$$

Length-bin fractions and counts. Given an integer bin $[a, b]$ (inclusive), define

$$\texttt{Frac}[a\text{–}b] \triangleq \frac{1}{n_{\text{valid}}} \sum_{x \in \mathcal{D}_{\text{valid}}} \mathbb{I}[a \le \ell(x) \le b], \qquad \texttt{Count}[a\text{–}b] \triangleq \sum_{x \in \mathcal{D}_{\text{valid}}} \mathbb{I}[a \le \ell(x) \le b].$$

For an open-ended bin $[a, +\infty)$, replace the indicator with $\mathbb{I}[\ell(x) \ge a]$.

Termination calibration ($\log p_{\text{term}}(\tau)$). For each sampled trajectory, let τ_i be the sampled stop step (the position where EOS/$\top$ is taken). Define the per-sample termination log-probability

$$\log p_{\text{term}}(\tau_i) \triangleq \log q_\theta(\top \mid s_{0:\tau_i}),$$

evaluated from the model's raw termination head (no masking/renormalization). We report the mean over all samples:

$$\log p_{\text{term}}(\tau) \triangleq \frac{1}{N} \sum_{i=1}^{N} \log p_{\text{term}}(\tau_i).$$

More negative values indicate overly suppressed termination.

Uniqueness metrics for SMILES. Let $\text{canon}(x)$ denote canonicalization used in evaluation (e.g., canonical SMILES / canonical molecule identity). We define

$$\texttt{UniqStr} \triangleq |\{x : x \in \mathcal{D}_{\text{valid}}\}|, \qquad \texttt{UniqRateStr} \triangleq \texttt{UniqStr}/n_{\text{valid}}.$$

For molecule-level uniqueness, define $\texttt{UniqMol} \triangleq |\{\text{canon}(x) : x \in \mathcal{D}_{\text{valid}}\}|$ and $\texttt{UniqRateMol} \triangleq \texttt{UniqMol}/n_{\text{valid}}$.

B.2. Token entropy

Ragged token entropy (`TokEnt`). Let $\mathcal{D}_{\text{valid}} = \{x_i\}_{i=1}^{n_{\text{valid}}}$ and let $\ell_i \triangleq \ell(x_i)$. For each position $t \ge 1$, consider the survivor index set $\mathcal{I}_t \triangleq \{i : \ell_i \ge t\}$ with $n_t \triangleq |\mathcal{I}_t|$. If $n_t \le 1$, skip this position. Otherwise define the empirical marginal

$$\hat{p}_t(v) \triangleq \frac{1}{n_t} \sum_{i \in \mathcal{I}_t} \mathbb{I}[x_{i,t} = v]$$

and its entropy (natural log)

$$H_t \triangleq -\sum_v \hat{p}_t(v) \log(\hat{p}_t(v) + \epsilon), \qquad \epsilon = 10^{-10}.$$

Let $\mathcal{T} \triangleq \{t : n_t > 1\}$ and report

$$\texttt{TokEnt} \triangleq \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} H_t.$$

Length-bucketed token entropy (`TokEnt`(ℓ)). Group valid samples by their pre-EOS length ℓ and compute `TokEnt` on each fixed-length bucket after truncation to ℓ. We set $\texttt{TokEnt}(\ell) = 0$ if the bucket has ≤ 1 sample or $\ell \le 0$.

Fingerprint diversity (`FPDiv`) for SMILES. Let $f(x)$ be a fingerprint and $\text{sim}(x, x')$ a similarity (Tanimoto in our SMILES experiments). We report

$$\texttt{FPDiv} \triangleq 1 - \frac{2}{n_{\text{valid}}(n_{\text{valid}} - 1)} \sum_{\substack{x, x' \in \mathcal{D}_{\text{valid}} \\ x < x'}} \text{sim}(x, x'). \tag{15}$$

Macro-averaged fingerprint diversity (`MacroFP`). Given length bins $\mathcal{B}$ (e.g., 0–5, 6–10, 11+), let $\mathcal{D}_{\text{valid}}^{(b)}$ be the valid subset in bin b. Define

$$\texttt{MacroFP} \triangleq \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \texttt{FPDiv}(\mathcal{D}_{\text{valid}}^{(b)}),$$

where $\texttt{FPDiv}(\cdot) = 0$ if a bin contains fewer than 2 samples.

B.3. Prefix-collapse metrics

Prefixes and survivors. For a terminal $x = (x_1, \ldots, x_{\ell(x)})$, define its length-$k$ prefix $s_{0:k}(x) \triangleq (x_1, \ldots, x_k)$ for $k \le \ell(x)$. At depth k, the valid survivor multiset is

$$\mathcal{D}_{\text{valid},k} \triangleq \{x \in \mathcal{D}_{\text{valid}} : \ell(x) \ge k\}, \qquad n_k \triangleq |\mathcal{D}_{\text{valid},k}|.$$

Prefix survival ($\texttt{Surv}(k)$).

$$\texttt{Surv}(k) \triangleq \frac{n_k}{n_{\text{valid}}}.$$

Prefix entropy ($\texttt{Ent}(k)$) and effective prefix count ($\texttt{Eff}(k)$). Let $\hat{p}_k(p)$ be the empirical frequency of prefix p among survivors:

$$\hat{p}_k(p) \triangleq \frac{1}{n_k} \sum_{x \in \mathcal{D}_{\text{valid},k}} \mathbb{I}[s_{0:k}(x) = p].$$

Define

$$\texttt{Ent}(k) \triangleq -\sum_p \hat{p}_k(p) \log \hat{p}_k(p), \qquad \texttt{Eff}(k) \triangleq \exp(\texttt{Ent}(k)).$$

Top-1 prefix mass ($\texttt{Top1}(k)$).

$$\texttt{Top1}(k) \triangleq \max_p \hat{p}_k(p).$$

Unique prefix rate ($\texttt{UniqueRate}(k)$).

$$\texttt{UniqueRate}(k) \triangleq \frac{|\{s_{0:k}(x) : x \in \mathcal{D}_{\text{valid},k}\}|}{n_k}.$$

B.4. Distribution metrics for Expr24 (oracle reference)

For Expr24, valid means correct. Let $\mathcal{U}_{\text{valid}} \triangleq \{\text{tuple}(x) : x \in \mathcal{D}_{\text{valid}}\}$ be the set of unique valid sequences. Let the enumerated oracle set be $\mathcal{Y}^\star$. We report

$$\texttt{Unique}_\checkmark \triangleq |\mathcal{U}_{\text{valid}}|, \qquad \texttt{CovCount} \triangleq |\mathcal{U}_{\text{valid}} \cap \mathcal{Y}^\star|, \qquad \texttt{Cov} \triangleq \texttt{CovCount}/|\mathcal{Y}^\star|,$$

and the sampling-cap-normalized variant

$$\texttt{NormCov} \triangleq \texttt{CovCount}/\min(N, |\mathcal{Y}^\star|).$$

Position-wise token marginals. Given a multiset of sequences $\mathcal{S}$ (oracle or sampled), define survivors at 0-indexed position t: $\mathcal{S}_t \triangleq \{x \in \mathcal{S} : |x| > t\}$. Let $C_t(v)$ be the count of token v at position t among $\mathcal{S}_t$ and $Z_t \triangleq \sum_v C_t(v)$. If $Z_t > 0$, define the empirical marginal $q_t(v) \triangleq C_t(v)/Z_t$.

Stabilized KL/JS. Let $\epsilon = 10^{-9}$ and let $\mathcal{A}_t$ be the union support of oracle and sampled marginals at position t. Define

$$\text{KL}_\epsilon(p\|q) \triangleq \sum_{v \in \mathcal{A}_t} (p(v) + \epsilon) \log \frac{p(v) + \epsilon}{q(v) + \epsilon},$$

and $\text{JS}_\epsilon(p, q) \triangleq \frac{1}{2}\text{KL}_\epsilon(p\|m) + \frac{1}{2}\text{KL}_\epsilon(q\|m)$ where $m = \frac{1}{2}(p + q)$.

Scalar divergences. Let $p_t^\star$ be the oracle marginal at position t computed from $\mathcal{Y}^\star$, and π_t the sampled marginal at t computed from $\mathcal{D}_{\text{valid}}$ (duplicates kept). Let $T_{\max}$ be the maximum length across oracle and sampled sequences. We report

$$\text{KL}(\pi \to p^\star) \triangleq \frac{1}{T_{\max}} \sum_{t=0}^{T_{\max}-1} \text{KL}_\epsilon(\pi_t\|p_t^\star), \quad \text{KL}(p^\star \to \pi) \triangleq \frac{1}{T_{\max}} \sum_{t=0}^{T_{\max}-1} \text{KL}_\epsilon(p_t^\star\|\pi_t),$$

$$\text{JS}_{\text{tok}} \triangleq \frac{1}{T_{\max}} \sum_{t=0}^{T_{\max}-1} \text{JS}_\epsilon(\pi_t, p_t^\star).$$

C. Derivations and Objective Details

C.1. Reference-prior reward shaping

We stabilize exploration by mixing (i) a frozen reference-LM prior as a log-regularizer and (ii) an external task score. Following Method 3.1, we represent the LLM-GFlowNet state by a generated prefix. Let $s_{0:k}$ denote the length-k prefix state (with $s_{0:0} \equiv s_0$ as the root), and let termination occur by emitting $\top$ at $s_{0:\tau}$. Let $q_{\mathrm{ref}}(\cdot \mid s_{0:k})$ be a frozen reference LM over $\mathcal{V} \cup \{\top\}$.

Reference log-score at a stop cut. We define the reference log-probability of stopping at $s_{0:k}$ as

$$\log P_{\mathrm{ref}}(s_{0:k}) \triangleq \sum_{t=0}^{k-1} \log q_{\mathrm{ref}}(s_{t+1} \mid s_{0:t}) + \log q_{\mathrm{ref}}(\top \mid s_{0:k}) \tag{16}$$

Mixed stop-reward and task-only component. Given an external task score $S(s_{0:k})$ (e.g., validity/QED/Expr-hit), we define the mixed log stop-reward

$$\log R(s_{0:k}) \triangleq \kappa \log P_{\mathrm{ref}}(s_{0:k}) + \lambda S(s_{0:k}), \tag{17}$$

where κ is a fixed scaling (typically κ=1) and λ sets the task term scale (empirically, $\lambda = 50$). Equivalently, define the *task-only* log component

$$u(s_{0:k}) \triangleq \log R(s_{0:k}) - \kappa \log P_{\mathrm{ref}}(s_{0:k}), \tag{18}$$

which is the part we "absorb" in RapTB (the reference-derived baseline remains intact).

Ablation: removing the reference prior. Table 20 shows that dropping the reference-prior term can severely destabilize training (here shown for TB on SMILES), causing sharp validity collapse and degenerate length behavior.

Table 20. SMILES generation ablations (TB reference). Unless specified, all metrics are computed on valid samples. `Len` denotes the mean token length of valid samples ($L_{\max} = 10$).

Method	Acc ↑	Score ↑	TokEnt ↑	FPDiv ↑	Len
TB	**0.998**	**0.717**	**2.503**	**0.807**	3.065
TB w/o ref	0.381	0.601	0.418	0.425	10.000

C.2. Terminable prefix-tree specialization of TB

Following Method 3.1, $q_\theta(\cdot \mid s_{0:k})$ parametrizes the forward policy over $\mathcal{V} \cup \{\top\}$. For readability, write $q_\theta(s_{t+1} \mid s_{0:t})$ for token actions and $q_\theta(\top \mid s_{0:k})$ for termination at prefix $s_{0:k}$. On the prefix tree, each non-root prefix has a unique parent, hence backward factors are deterministic and vanish in log-space.

Terminal TB residual. For a realized termination index τ, the TB log-residual is

$$\Delta^{\mathrm{TB}}(\xi) = \log Z_\theta + \sum_{t=0}^{\tau-1} \log q_\theta(s_{t+1} \mid s_{0:t}) + \log q_\theta(\top \mid s_{0:\tau}) - \log R(s_{0:\tau}) \tag{19}$$

We minimize $\mathcal{L}_{\mathrm{TB}} = \mathbb{E}_\xi[\Delta^{\mathrm{TB}}(\xi)^2]$.

Residual at an intermediate prefix. In our implementation, the model outputs per-prefix stop logits $\log p_{\mathrm{term}}[k] \equiv \log q_\theta(\top \mid s_{0:k})$ and per-prefix stop-reward logs $\log r[k] \equiv \log R(s_{0:k})$ for *all* $k \in \{0, \ldots, L{-}1\}$, including $k = 0$ (stopping immediately after the prompt). Token-transition log-probabilities are stored as $\log p_F[t] \equiv \log q_\theta(s_{t+1} \mid s_{0:t})$ for steps $t \in \{0, \ldots, L{-}2\}$.

Define the TB-style residual at prefix $s_{0:k}$ by:

$$\Delta_k^{\mathrm{TB}}(\xi) \triangleq \log Z_\theta + \sum_{t=0}^{k-1} \log p_F[t] + \log p_{\mathrm{term}}[k] - \log r[k], \qquad k \in \{0, \ldots, L{-}1\}, \tag{20}$$

where $\Delta_0^{\mathrm{TB}}(\xi) = \log Z_\theta + \log p_{\mathrm{term}}[0] - \log r[0]$.

RapTB uses the *rooted* version (cancels $\log Z_\theta$):

$$\bar{\Delta}_k(\xi) \triangleq \Delta_k^{\mathrm{TB}}(\xi) - \Delta_0^{\mathrm{TB}}(\xi), \qquad k \geq 1. \tag{21}$$

Incremental rooted form. For a terminable trajectory $\xi = (s_0 \to s_1 \to \cdots)$, we introduce the per-step increment for $t \geq 0$:

$$\delta_t(\xi) \triangleq \big(\log r[t] - \log r[t{+}1]\big) + \log p_F[t] + \big(\log p_{\mathrm{term}}[t{+}1] - \log p_{\mathrm{term}}[t]\big), \tag{22}$$

and its cumulative sum

$$C_0(\xi) \triangleq 0, \qquad C_k(\xi) \triangleq \sum_{t=0}^{k-1} \delta_t(\xi), \quad k \geq 1. \tag{23}$$

By construction, the reward-difference and termination-difference terms telescope, yielding the closed form

$$C_k(\xi) = \sum_{t=0}^{k-1} \log p_F[t] + \log p_{\mathrm{term}}[k] - \log r[k] - \Big(\log p_{\mathrm{term}}[0] - \log r[0]\Big). \tag{24}$$

Comparing with the residual definition in Eq. (20), we have $C_k(\xi) = \Delta_k^{\mathrm{TB}}(\xi) - \Delta_0^{\mathrm{TB}}(\xi)$, i.e., for $k \geq 1$,

$$C_k(\xi) = \bar{\Delta}_k(\xi). \tag{25}$$

C.3. Absorbed suffix backups in RapTB

RapTB "absorbs" the *task-only* component $u(s_{0:k})$ in Eq. (18), while leaving the reference-derived baseline intact. Along a sampled suffix, denote $u_k \triangleq u(s_{0:k})$.

Finite horizon. Let $K \in \{1, \ldots, L_{\max}\}$ be the auxiliary horizon cap (maximum prefix depth) used for auxiliary supervision, and define

$$h \triangleq \min(\tau, K). \tag{26}$$

All backup targets below are defined over indices $j \in [k, h]$.

Max and soft backups. Define

$$u_k^{\max} \triangleq \max_{j \in [k,h]} u_j, \tag{27}$$

$$u_k^{soft} \triangleq \frac{1}{\beta} \log \sum_{j=k}^{h} \exp\Big(\beta u_j - \beta\rho(j-k)\Big), \qquad \beta > 0,\ \rho \geq 0, \tag{28}$$

$$u_k^{tgt} \triangleq \alpha\, u_k^{\max} + (1-\alpha)\, u_k^{soft}, \qquad \alpha \in [0,1]. \tag{29}$$

where $\beta > 0$ controls softness and $\rho \geq 0$ penalizes distant evidence. We compute u_k^{soft} using the LogSumExp trick to ensure numerical stability and prevent overflow. In our implementation, absorption is applied to the external task-score component only.

C.4. Practical details: prefix eligibility, scheduling, and stop gradients

Gating and weights. Let $k_{\min} \geq 1$ be a minimum prefix depth for auxiliary supervision and define

$$w_k \triangleq \mathbf{1}[k \geq k_{\min}]. \tag{30}$$

Let $L_{\max}$ be the decoding maximum length; in our experiments we set the auxiliary horizon cap $K = L_{\max}$. Define the stop index τ as the first step where $\top$ is sampled; if $\top$ is never sampled before $L_{\max}$, decoding forces termination and we set $\tau = L_{\max}$. Let K be the auxiliary horizon cap and define $h \triangleq \min(\tau, K)$. We only consider rooted prefixes within the realized trajectory and within the horizon:

$$M_k(\xi) \triangleq \mathbf{1}[1 \leq k \leq h] \cdot \mathbf{1}[k \geq k_{\min}]. \tag{31}$$

(Equivalently, $M_k(\xi) = 1$ iff prefix $s_{0:k}$ is eligible to contribute to $\mathcal{L}_{aux}$.)

Selective absorption (a refinement of the same gating). Recall the task-only component $u_k \triangleq u(s_{0:k}) = \log R(s_{0:k}) - \kappa \log P_{\mathrm{ref}}(s_{0:k})$. Absorption is only applied to *eligible* prefixes that (i) have no task signal at the current prefix and (ii) admit sufficient suffix evidence under the cap K. We introduce an absorption gate

$$A_k(\xi) \triangleq M_k(\xi) \cdot \mathbf{1}[\tau \geq K] \cdot \mathbf{1}[k < K] \cdot \mathbf{1}\big[|u_k| \leq \varepsilon_{\mathrm{ab}}\big], \tag{32}$$

where $\varepsilon_{\mathrm{ab}} > 0$ is a small numerical threshold (capturing the "no task reward" case). We also define the distance discount weight, restricted to the same eligible set,

$$d_k(\xi) \triangleq M_k(\xi)\, \gamma^{(h-k)}, \qquad \gamma \in (0, 1). \tag{33}$$

Absorbed correction on rooted residuals. Let u_k^{tgt} be the suffix target (max/soft/mix) computed over $j \in [k, h]$ (Appendix C.3). We add a discounted absorbed correction to the rooted residual:

$$\mathrm{corr}_k(\xi) \triangleq A_k(\xi)\, d_k(\xi)\, \big(u_k - u_k^{tgt}\big), \qquad C_k^{ab}(\xi) \triangleq C_k(\xi) + \mathrm{corr}_k(\xi). \tag{34}$$

This modification affects only the task-only component inside the auxiliary rooted constraints: when $A_k(\xi) = 1$, the "missing" u_k at the current prefix is replaced by a suffix-derived proxy u_k^{tgt}, while the reference-derived baseline remains unchanged. The combined gate $A_k(\xi)$ enforces that absorption is used only when it is both *eligible* and *evidence-supported* (requiring $\tau \geq K$ and $k < K$), and the factor $d_k(\xi)$ discounts distant evidence by γ^{h-k}.

Stop-gradient on the termination head in the auxiliary branch. To prevent auxiliary constraints from being satisfied by systematically drifting length calibration, we stop gradients through the termination log-probabilities *only in the auxiliary branch*. Concretely, in the construction of the rooted cumulative residuals $C_k(\xi)$ (and thus $C_k^{ab}(\xi)$), we replace every occurrence of $\log q_\theta(\top \mid s_{0:k})$ by $\mathrm{stopgrad}(\log q_\theta(\top \mid s_{0:k}))$:

$$\log q_\theta(\top \mid s_{0:k}) \mapsto \mathrm{stopgrad}(\log q_\theta(\top \mid s_{0:k})) \qquad \text{inside } C_k(\xi) \text{ and } C_k^{ab}(\xi). \tag{35}$$

The terminal TB anchor $\Delta^{\mathrm{TB}}(\xi)$ in Eq. (19) is optimized with full gradients.

Normalized auxiliary loss and "TB-only" fallback. Using the eligibility mask $M_k(\xi)$ in Eq. (31) and the length weights w_k in Eq. (30), we define the per-trajectory auxiliary objective as

$$\mathcal{L}_{aux}(\xi) \triangleq \frac{\sum_{k=1}^{h} M_k(\xi)\, w_k \left(C_k^{ab}(\xi)\right)^2}{\sum_{k=1}^{h} M_k(\xi)\, w_k + \epsilon}, \qquad \epsilon > 0, \tag{36}$$

with the convention $\mathcal{L}_{aux}(\xi) = 0$ if $\sum_{k=1}^{h} M_k(\xi)\, w_k = 0$.

Final objective (additive). Let $\eta \geq 0$ be the global auxiliary weight. We use an additive objective

$$\mathcal{L}_{RapTB} \triangleq \mathbb{E}_{\xi \sim q_\theta}\left[\Delta^{\mathrm{TB}}(\xi)^2 + \eta\, \mathcal{L}_{aux}(\xi)\right], \tag{37}$$

so trajectories with no eligible auxiliary prefixes reduce to pure terminal TB.

C.5. Variance reduction view of RapTB

TB tail as a stochastic regression target. Fix a prefix cut at index m along a sampled terminable trajectory $\xi = (s_{0:0} \to s_{0:1} \to \cdots \to s_{0:\tau})$ on the prefix tree. Starting from the terminal TB residual in Eq. (1), we decompose it as

$$\Delta^{\mathrm{TB}}(\xi) = X(s_{0:m}(\xi)) - Y_m(\xi), \tag{38}$$

where the *prefix term*

$$X(s_{0:m}(\xi)) \triangleq \log Z_\theta + \sum_{t=0}^{m-1} \log q_\theta(s_{t+1} \mid s_{0:t}), \tag{39}$$

Algorithm 1 RapTB auxiliary targets (single trajectory)

Roll out a terminable trajectory $\xi = (s_{0:0} \to \cdots \to s_{0:\tau})$ with token log-probs $\log p_F[t]$, termination log-probs $\log p_{\text{term}}[k]$, and stop-rewards $\log R(s_{0:k})$.
Decompose $\log R(s_{0:k}) = \kappa \log P_{\text{ref}}(s_{0:k}) + \lambda S(s_{0:k})$ (task-only $\lambda S(s_{0:k})$).
For eligible prefixes $k \le h = \min(\tau, K)$, compute suffix targets u_k^{tgt}
    via Appendix C.3 (Eq. (27)).
Form rooted residuals C_k using Appendix C.4.
    In this branch, apply STOPGRAD to $\log p_{\text{term}}[\cdot]$ as in Eq. (35).
Replace u_k by u_k^{tgt} when the absorption gate is on (Eq. (32)).
    This yields C_k^{ab} via Eq. (34).
Compute $\mathcal{L}_{\text{aux}}$ as the normalized weighted sum of $(C_k^{\text{ab}})^2$, and train with $\mathcal{L}_{\text{TB}} + \eta \mathcal{L}_{\text{aux}}$.

is deterministic given the prefix $s_{0:m}$, and the *tail term*

$$Y_m(\xi) \triangleq \log R(s_{0:\tau}) - \log q_\theta(\top \mid s_{0:\tau}) + \sum_{t=m}^{\tau-1} \log q_\theta(s_{t+1} \mid s_{0:t}), \tag{40}$$

depends only on the sampled suffix $(s_{0:m} \to \cdots \to s_{0:\tau})$.

Conditioning on $s_{0:m}$, TB minimizes a conditional least-squares error:

$$\mathbb{E}\Big[\big(X(s_{0:m}) - Y_m(\xi)\big)^2 \,\Big|\, s_{0:m}\Big] = \big(X(s_{0:m}) - \mu(s_{0:m})\big)^2 + \mathrm{Var}(Y_m(\xi) \,|\, s_{0:m})\,, \tag{41}$$

where $\mu(s_{0:m}) \triangleq \mathbb{E}[Y_m(\xi) \mid s_{0:m}]$ is the minimum-MSE target for that prefix.

Connecting to RapTB. Eq. (41) highlights a core difficulty under terminal-only task signals: early prefixes are trained through a single stochastic tail target $Y_m(\xi)$ whose conditional variance can be large, so credit assignment to early decisions is noisy and can favor rich-get-richer flow allocations (prefix collapse). A natural response is to add subtrajectory consistency, but enforcing arbitrary-start windows introduces state-dependent boundary terms; in terminable prefix trees, the commonly used flow-free/windowed SubTB ties these boundaries to $-\log q_\theta(\top \mid s)$, so heterogeneous starts combined with discontinuous rewards can be absorbed by the shared termination head, inducing termination/length drift. RapTB takes a conservative middle ground: it keeps terminal TB unchanged as the global anchor, and adds only *root-start* residuals $C_k = \Delta_k^{\text{TB}} - \Delta_0^{\text{TB}}$ (Appendix C.2), providing $O(\tau)$ prefix-local supervision without introducing a separate flow head or arbitrary-start boundary heterogeneity. To further reduce variance in auxiliary targets, RapTB densifies only the *external* component of the stop-reward by absorbing high-reward evidence along the sampled suffix (Appendix C.3), while leaving any reference-derived baseline intact; this yields a lower-variance proxy for the reward term inside rooted constraints without changing the terminal TB semantics. Finally, we stop gradients through termination logits in the auxiliary branch so these additional prefix constraints cannot be satisfied by globally shifting $q_\theta(\top \mid s)$, preventing auxiliary-driven length bias.

C.6. Why naive LLM-SubTB can fail under terminal-only task signals

Key point. General SubTB constraints starting at arbitrary intermediate prefixes $s_{0:i}$ require a *state-dependent* normalizer/flow (e.g., $F_\theta(s_{0:i})$ or a local partition $Z(s_{0:i})$). In terminal-only-task-signal LLM settings, most prefixes have negligible task-only signal $u(s) \approx 0$, so a flow-free objective can inadvertently use the termination head $q_\theta(\top \mid s)$ as a "baseline sink", inducing systematic termination drift (length bias) and hurting accuracy.

Sketch in our notation. Define the (rooted) stop residual without $\log Z_\theta$:

$$\mathrm{res}(k) \triangleq \sum_{t=0}^{k-1} \log q_\theta(s_{t+1} \mid s_{0:t}) + \log q_\theta(\top \mid s_{0:k}) - \log R(s_{0:k}).$$

Then $\delta_t = \mathrm{res}(t{+}1) - \mathrm{res}(t)$ and $C_k = \mathrm{res}(k) - \mathrm{res}(0)$. Arbitrary-start subtrajectory consistency effectively compares $\mathrm{res}(j)$ and $\mathrm{res}(i)$ for $i > 0$ under an *implicit constant baseline*, whereas correct balance from $s_{0:i}$ needs a *local* baseline

reflecting downstream reward mass. When $\log R(s)$ carries little task signal for most prefixes, optimizing these constraints can primarily shift $\log q_\theta(\top \mid s)$ to absorb missing baselines, resulting in miscalibrated stopping probabilities and length bias. RapTB avoids this failure mode by (i) restricting auxiliary constraints to root-start prefixes, and (ii) stopping gradients through $\log q_\theta(\top \mid s)$ in the auxiliary branch.

D. Implementation Details and Reproducibility

Config provenance. The hyperparameters in Tables 21–23 are taken from the task configs used to produce our main results.

D.1. Model, decoding, and context-free grammar (CFG)

Backbone and parameterization. We fine-tune an autoregressive LLM with a terminable action at every prefix state. The forward kernel is parameterized as in Sec. 3 by (i) a token head $p_F(\cdot \mid s)$ and (ii) a termination head $p_{\text{term}}(s)$. We use parameter-efficient fine-tuning (LoRA) unless otherwise specified.

LoRA fine-tuning details. We use parameter-efficient fine-tuning with LoRA on the LLM backbone. Concretely, we apply LoRA to the attention and MLP projection modules {`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `down_proj`, `up_proj`} with rank r=16, α=16, dropout 0.1, and no bias parameters. The backbone is `meta-llama/Llama-3.2-1B`. We optimize with AdamW (lr 10^{-4}). Any auxiliary schedulers used in training (e.g., temperature/replay schedules) are reported explicitly in Table 23.

Molecular fingerprints for similarity/diversity. For SMILES similarity used by SubM, we compute RDKit Morgan fingerprints with radius 2 and 2048 bits, and use Tanimoto similarity for the facility-location diversity term. These settings are fixed across all SMILES experiments.

Context-free grammar (CFG) constrained decoding. We enforce syntactic constraints during generation using a grammar processor with an incremental EBNF parser. At each step, the parser consumes the current prefix and returns the set of next tokens that keep the prefix valid under the grammar. The processor restricts *which* tokens are feasible, **without masking or modifying their logits**. This is a decoding-time feasibility filter shared by all objectives (TB/SubTB/RapTB), so it does not change the training objective while substantially reducing invalid roll-outs. Figures 4–5 list the grammars used for SMILES and Expr24, respectively.

Table 21. Core training and decoding hyperparameters. TB and SubTB share all hyperparameters with RapTB except the loss definition (and RapTB-specific coefficients).

	SMILES	**Expr24**
Trainer steps (max)	5000	5000
Precision	Bf16	Bf16
Grad clip	0.5	0.5
Grad accumulation	4	4
Optimizer	AdamW (lr $= 10^{-4}$)	AdamW (lr $= 10^{-4}$)
Sampling (train)		
# trajectories per update (n_{samples})	32	32
p_F temperature mix	$T_{\text{hi}} : 1.5 \to 1.0$; $T_{\text{lo}} : 0.8 \to 1.0$	$T_{\text{hi}} : 1.5 \to 1.0$; $T_{\text{lo}} : 0.8 \to 1.0$
low-temp probability	0.666	0.666
replay mixture ratio	0.7	0.7
CFG decoding (grammar)		
Min/max length	1 / 10 (15)	3 / 9

D.2. SMILES constraints and validity evaluation

Constrained decoding and legal token list. We apply the grammar processor at *every* decoding step. In addition, we use a fixed allowlist of tokenizer tokens deemed legal for SMILES generation. The combination of EBNF parsing and token

```
root ::= smiles

smiles ::= atom ( chain | branch )*

chain ::= (dot atom | bond? ( atom | ring_closure ) )+

branch ::= "(" ( ( dot | bond )? smiles )+ ")"

atom ::= organic_symbol | aromatic_symbol | atom_spec | wildcard

bond ::= "" | "=" | "#" | "$" | ":" | "@" | "@@"

dot ::= "."

wildcard ::=  "*"

atom_spec ::= "[" ( "se" | "as" | aromatic_symbol | element_symbol | wildcard ) chiral_class? h_count? ( charge | class? ) "]"

organic_symbol ::= "B" | "C" | "N" | "O" | "P" | "S" | "F" | "I" | "Br" | "Cl" | "At" | "Ts"

aromatic_symbol ::= "b" | "c" | "n" | "o" | "p" | "s"

element_symbol  ::= "A" ( "c" | "g" | "l" | "m" | "r" | "s" | "t" | "u" ) |
                    "B" ( "a" | "e" | "h" | "i" | "k" | "r" )? |
                    "C" ( "a" | "d" | "e" | "f" | "l" | "m" | "n" | "o" | "r" | "s" | "u" )? |
                    "D" ( "b" | "s" | "y" ) |
                    "E" ( "r" | "s" | "u" ) |
                    "F" ( "e" | "l" | "m" | "r" )? |
                    "G" ( "a" | "d" | "e" ) |
                    "H" ( "e" | "f" | "g" | "o" | "s" )? |
                    "I" ( "n" | "r" )? |
                    "K" "r"? |
                    "L" ( "a" | "i" | "r" | "u" | "v" ) |
                    "M" ( "c" | "g" | "n" | "o" | "t" ) |
                    "N" ( "a" | "b" | "d" | "e" | "h" | "i" | "o" | "p" )? |
                    "O" ( "g" | "s" )? |
                    "P" ( "a" | "b" | "d" | "m" | "o" | "r" | "t" | "u" )? |
                    "R" ( "a" | "b" | "e" | "f" | "g" | "h" | "n" | "u" ) |
                    "S" ( "b" | "c" | "e" | "g" | "i" | "m" | "n" | "r" )? |
                    "T" ( "a" | "b" | "c" | "e" | "h" | "i" | "l" | "m" | "s" ) |
                    "U" | "V" | "W" | "Xe" | "Y" "b"? |
                    "Z" ( "n" | "r" )

ring_closure ::= "%" [1-9] [0-9] | [0-9]

chiral_class ::= ( "@" ( "@" | "TH" [1-2] | "AL" [1-2] | "SP" [1-3] | "TB" ( "1" [0-9]? | "2" "0"? | [3-9] ) | "OH" ( "1"

[0-9]? | "2" [0-9]? | "3" "0"? | [4-9] ) )? )?

charge   ::= "-" ( "-" | "0" | "1" [0-5]? | [2-9] )? | "+" ( "+" | "0" | "1" [0-5]? | [2-9] )?

h_count   ::= "H" [0-9]?

class    ::= ":" [0-9]+
```

Figure 4. EBNF grammar used for constrained SMILES decoding.

```
root  ::= expr4
expr4 ::= num | num op num | num op num op num | num op num op num op num
| num op num op num op num op num | num op num op num op num op num op num
op    ::= "+" | "-" | "*" | "/"
num   ::= [0-9]
```

Figure 5. EBNF grammar used for constrained Expr24 decoding.

Table 22. RapTB-specific hyperparameters. $k_{\min}$ is linearly scheduled by training step.

	SMILES	Expr24
Aux weight η	0.25	0.25
Distance discount γ	0.99	0.99
Detach p_{term} in aux	True	True
Absorb gate threshold ε_{ab}	10^{-6}	10^{-6}
Target mode	mix	mix
Mix weight α (max vs soft)	0.5	0.8
Soft backup β	5.0	3.0
Distance penalty ρ	0.1	0.5
$k_{\min}$ schedule	$5 \to 2$ (5000 steps)	$7 \to 3$ (5000 steps)
Aux horizon cap K	$L_{\max}$	$L_{\max}$

allowed lists substantially reduces invalid generations, without changing the learning objective.

Validity and scoring. We use RDKit-based validation to identify valid SMILES and to compute the task reward (QED in our setup). Invalid generations receive an invalidity shaping schedule as configured in the reward module.

D.3. Replay buffers

Reward-prioritized replay buffer (RP). We maintain a replay buffer keyed by the decoded prompt string. For each prompt, the buffer stores up to B trajectories using a min-heap over the reward proxy, thus keeping the current top-B items. Exact duplicates (identical decoded strings) are discarded. To prevent near-duplicate high-reward items from dominating the buffer, we additionally apply a near-duplicate filter using edit distance between the tokenized answers (excluding the termination token): a new item is rejected if it is too similar to an existing buffer item, unless it has a higher reward (or is explicitly forced to be added). When enabled, we use a small buffer-augmentation trick for forced/validated items to ensure they are preferred. In our configs, the per-prompt buffer capacity is $B = 200$, and the near-duplicate tolerance is 0.25.

Reward-prioritized replay training (PRT). The buffer is sufficiently populated, and replay sampling follows a two-tier scheme inspired by Shen et al. (2023): An α fraction of each replay minibatch is sampled uniformly from the top-β_{tier} reward tier (i.e., the highest-reward $\lceil \beta|B| \rceil$ items), and the rest is sampled uniformly from the remaining items; sampling falls back to uniform replay when prioritization is disabled or infeasible.

D.3.1. Submodular Replay Details

Submodular replay (SubM): objective and selection. SubM replaces the "keep top-B by reward" rule with a *submodular subset selection* step that refreshes the buffer to maximize a weighted objective combining quality, diversity, and length coverage. for each candidate item x, we define a *static score* $s(x) = w_{\text{rew}}\, r(x) + w_{\text{val}} \mathbf{1}[\text{valid}(x)]$ and maximize a facility-location style diversity term together with a concave length-bin coverage term. Concretely, during greedy selection we use the per-candidate marginal gain

$$\Delta(x \mid S) = s(x) + w_{\text{div}} \sum_{u \in \mathcal{G}} \max\left\{0,\ \text{sim}(u,x) - \text{msim}(u,S)\right\} + w_{\text{len}} \alpha_{b(x)} \Big(\log(1 + c_{b(x)}(S) + 1) - \log(1 + c_{b(x)}(S)) \Big), \tag{42}$$

where $\text{sim}(\cdot,\cdot) \in [0,1]$ is a task-dependent similarity, $b(x)$ is the length-bin index of x, $c_b(S)$ is the current count in bin b, and $\alpha_b \geq 0$ controls how strongly we bias coverage toward specific length regimes. In our configs we use uniform weights, so $\alpha_b = 1$ for all bins. Importantly, the implementation uses token length rather than raw string length to define bins, avoiding the common mismatch between tokenizer length and character length.

Similarity backends. For SMILES, we canonicalize valid molecules with RDKit and compute Morgan fingerprints (radius 2, 2048 bits); similarity is Tanimoto computed by RDKit bulk routines. For Expr24 (string-domain tasks), we use k-gram shingles (default k=2) and Jaccard similarity between shingle sets. These similarities are only used by the diversity term in Eq. (42).

Efficiency: cached coverage updates and greedy variants. To compute the facility-location marginal efficiently, we maintain $\mathrm{msim}(u, S)$ for each ground element $u \in \mathcal{G}$. When evaluating a candidate x, we compute $\{\mathrm{sim}(u, x)\}_{u\in\mathcal{G}}$ via a bulk similarity call and accumulate $\sum_u \max\{0,\ \mathrm{sim}(u, x) - \mathrm{msim}(u, S)\}$, then update $\mathrm{msim}(u, S) \leftarrow \max\{\mathrm{msim}(u, S), \mathrm{sim}(u, x)\}$ after selecting x. We use standard greedy algorithm in our submodular replay method.

Validity gating for diversity. SubM optionally restricts the diversity optimization to a valid-heavy candidate pool: it keeps all valid items and only adds enough invalid items (ranked by static score) to ensure the valid proportion is at least the specified ratio. This prevents invalid strings from consuming the diversity budget while still allowing the buffer to remain populated when valid samples are scarce. In SMILES we set the validity-gating ratio to 0.0, while in Expr24 we set the ratio to 1.0, so the facility-location term is computed only among valid items.

Table 23. Factor schedulers used in our training runs. Each factor uses linear interpolation from `start` to `end` over a fixed `horizon` (in steps).

Task	Factor	start	end	horizon
SMILES	replay_buffer	0.50	0.25	5000
SMILES	k_min	5	2	5000
Expr24	replay_buffer	0.50	0.25	5000
Expr24	oracle_buffer	0.75	0.25	5000
Expr24	k_min	7	3	5000

Table 24. Submodular replay hyperparameters (SMILES). We select a buffer of size B using greedy maximization of a facility-location + length objective.

Hyperparameter	Value
Buffer capacity B	200
Similarity backend	RDKit bulk Tanimoto (Morgan r=2, 2048-bit)
Weights $(w_{\mathrm{rew}}, w_{\mathrm{val}}, w_{\mathrm{div}}, w_{\mathrm{len}})$	$(1.0, 1.0, 1.0, 1.0)$
Length bin size	1 token
Length alpha mode / power	uniform / 1.0
Validity gating ratio	0.0
Selection strategy	standard greedy

Table 25. Submodular replay hyperparameters (Expr24). We select a buffer of size B using greedy maximization of a facility-location + length objective.

Hyperparameter	Value
Buffer capacity B	200
Refresh period K_{buf}	1
Similarity backend	k-gram shingles + Jaccard
Weights $(w_{\mathrm{rew}}, w_{\mathrm{val}}, w_{\mathrm{div}}, w_{\mathrm{len}})$	$(1.0, 1.0, 1.0, 0.0)$
Length bin size	-
Length alpha mode / power	-
Validity gating ratio	1.0
Selection strategy	standard greedy