Overcoming Joint Intractability with Lossless Hierarchical Speculative Decoding

Yuxuan Zhou[1]*, Fei Huang[2], Heng Li[1], Fengyi Wu[3], Tianyu Wang[3], Jianwei Zhang[2], Junyang Lin[2†], Zhi-Qi Cheng[3†]

[1]Independent Researcher [2]Qwen Team, Alibaba Inc. [3]University of Washington

zhouyuxuanyx@gmail.com, junyang.ljy@alibaba-inc.com, zhiqics@uw.edu

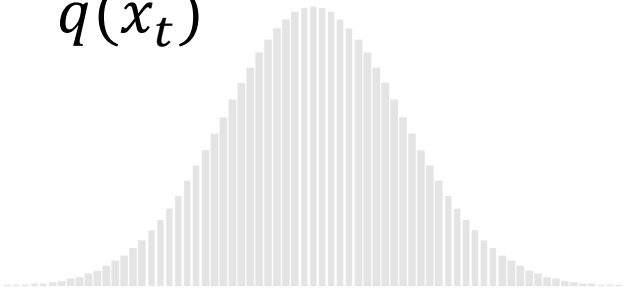
*Work done during internship at Qwen Team, Alibaba Inc.

†Corresponding author.

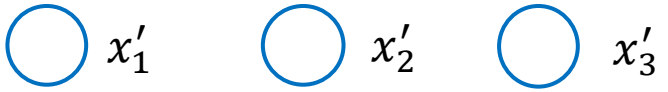
Tokenwise Verification [1]

- **Step 1:** sample draft tokens from the draft distribution

$q(x_t)$



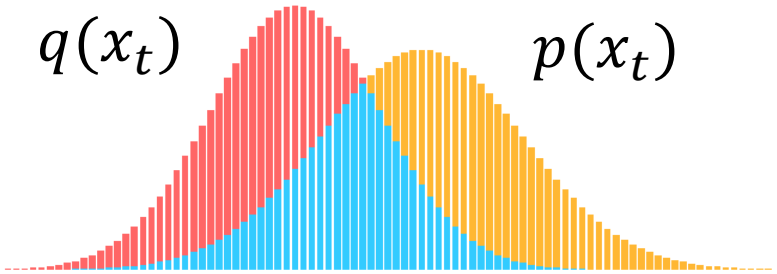
$$P(x_t\ is\ sampled) = q(x_t|\mathrm{X}_{1:t-1})$$



[1] Leviathan, Yaniv, Matan Kalman, and Yossi Matias. "Fast inference from transformers via speculative decoding." *ICML* 2023.

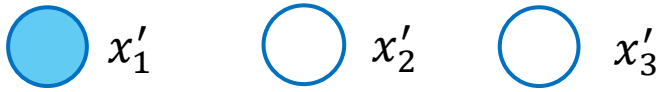
Tokenwise Verification

➢ **Step 2:** verify a draft token based on likelihood ratio

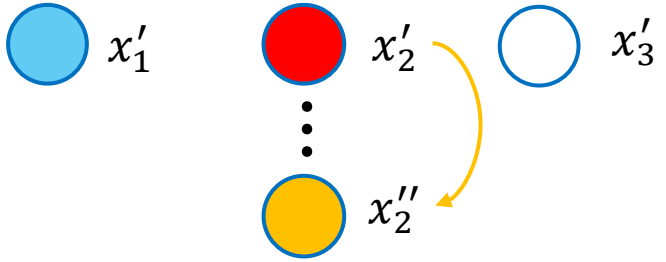


Accept Probability: $\min\{1, \frac{p(x_t)}{q(x_t)}\}$

Case 1: ✓ accept ⇨ check next token

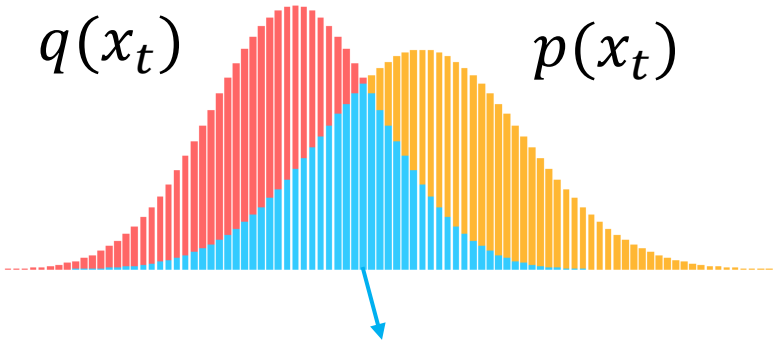


Case 2: ✗ reject ⇨ ↻ resample



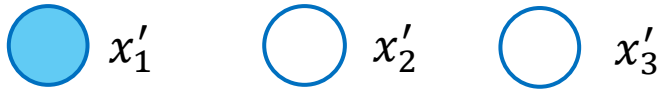
Tokenwise Verification

- **Step 2:** verify a token based on likelihood ratio
  - **Case 1:** token is accepted



Accept Probability:

$$\mathrm{h} = \min\{\frac{p(x_t)}{q(x_t)}, 1\}$$

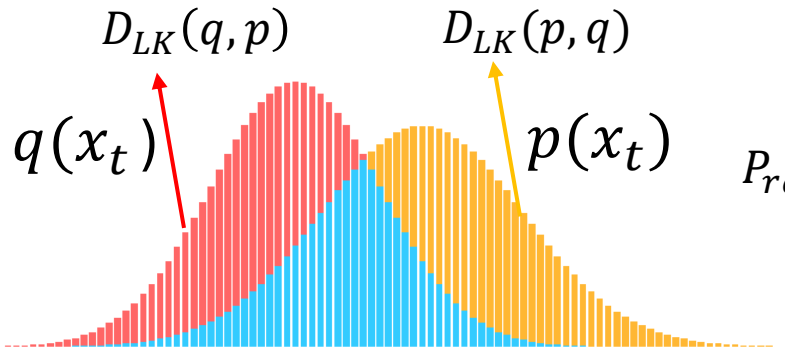


- **Output Probability:**

$$P(x_t \text{ is sampled}, x_t \text{ is accepted})$$
$$= q(x_t)\ \min\{\frac{p(x_t)}{q(x_t)}, 1\}$$
$$= \underbrace{\min\{p(x_t), q(x_t)\}}_{\text{blue area}}$$

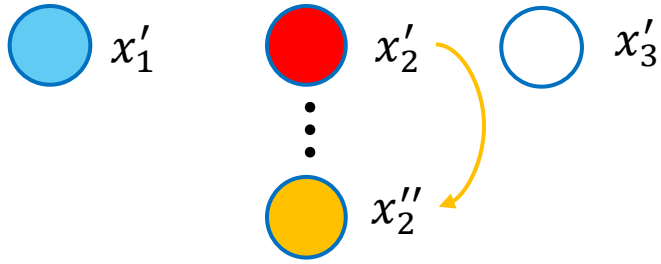
Tokenwise Verification

- **Step 2:** resample a token using likelihood ratio
  - **Case 2:** token is rejected



Resample Probability:

$$P_{res}(x_t) = \frac{p(x_t) - \min\{p(x_t), q(x_t)\}}{\sum_{\tilde{x}_t \in \mathcal{V}} p(\tilde{x}_t) - \min\{p(\tilde{x}_t), q(\tilde{x}_t)\}}$$



- **Output Probability:**

$$P(\tilde{x}_t \neq x_t \text{ is sampled and rejected}, x_t \text{ is resampled})$$

$$= \sum_{\tilde{x}_t \in \mathcal{V}} q(\tilde{x}_t)\left(1 - \min\left\{\frac{p(x_t)}{q(x_t)}, 1\right\}\right) \frac{p(x_t) - \min\{p(x_t), q(x_t)\}}{\sum_{\tilde{x}_t \in \mathcal{V}} p(\tilde{x}_t) - \min\{p(\tilde{x}_t), q(\tilde{x}_t)\}}$$

$$= D_{LK}(q,p) \frac{p(x_t) - \min\{p(x_t), q(x_t)\}}{D_{LK}(p,q)}$$

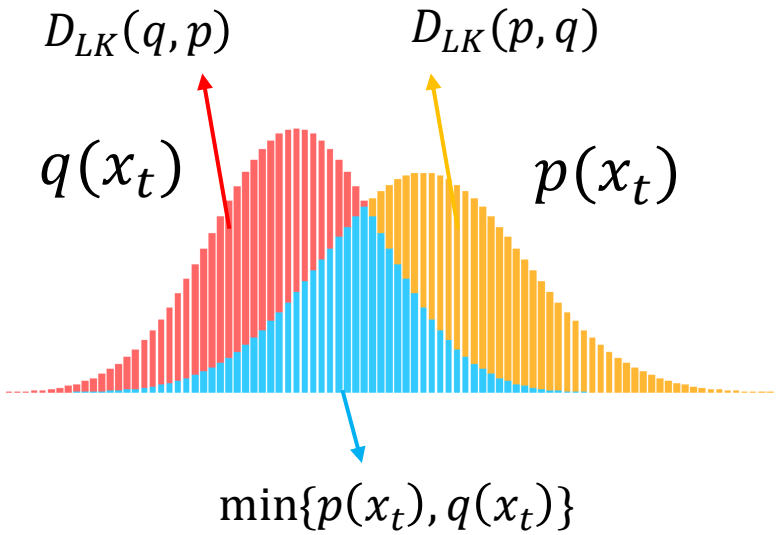
red area = orange area

Tokenwise Verification

➢ **Output Distribution:**

$$\begin{aligned}\mathrm{P}(x_t \text{ is yielded}) &= P(x_t \text{ is sampled and accepted}) + P(\tilde{x}_t \text{ is sampled and rejected}, x_t \text{ is resampled}) \\ &= \min\{p(x_t), q(x_t)\} + p(x_t) - \min\{p(x_t), q(x_t)\} \\ &= p(x_t)\end{aligned}$$

Target distribution is preserved!



Joint Verification

- **Accept Probability [2]:**

$$\text{tokenwise: } \mathrm{h}(x_t) = \min\{1, \frac{p(x_t)}{q(x_t)}\} \longrightarrow \text{joint: } \mathrm{h}(X_{1:t}) = \min\{1, \frac{p(X_{1:t})}{q(X_{1:t})}\}$$

- **Illustrative Example:**

"Eliza worked a total of 45 hours, so {she, work, ed, 45}

input prompt | draft tokens

Draft and target probs:

$\{p(x_t)\} = \{0.72, 1.00, 1.00, 1.00\}, \quad \{q(x_t)\} = \{0.88, 0.79, 0.65, 0.26\},$

Accept probs

$\{h(x_t)\} = \{0.82, 1.00, 1.00, 1.00\}, \quad h(X_{1:4}) = 1$

[2] Zongyue Qin, et, al. Optimized multitoken joint decoding with auxiliary model for llm inference. ICLR 2025.

Joint Verification

➢ **Output Distribution:**

$$\mathrm{P}(X_{1:t}\ is\ yielded) = P(X_{1:t}\ is\ sampled\ and\ accepted) + P(\tilde{X}_{1:t}\ is\ sampled\ and\ rejected, X_{1:t}\ is\ resampled)$$

- **Case 1:** $P(X_{1:t}\ is\ sampled\ and\ accepted) = q(X_{1:t})\ \min\left\{1, \frac{p(X_{1:t})}{q(X_{1:t})}\right\} = \min\{p(X_{1:t}), q(X_{1:t})\}$

blue area

- **Case 2:** $P\left(\tilde{X}_{1:t} \neq X_{1:t}\ is\ sampled\ and\ rejected, X_{1:t}\ is\ resampled\right)$

$$= D_{Branch}(q, p|X_{1:t-1}) p_{res}(X_{1:t})$$

$$= D_{Branch}(q, p|X_{1:t-1}) \frac{p(X_{1:t}) - \min\{p(X_{1:t}), q(X_{1:t})\}}{D_{Branch}(p, q|X_{1:t-1})}$$

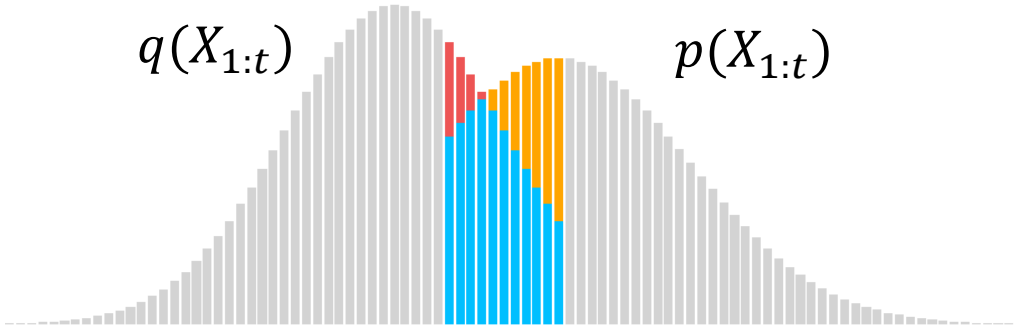
Joint Verification

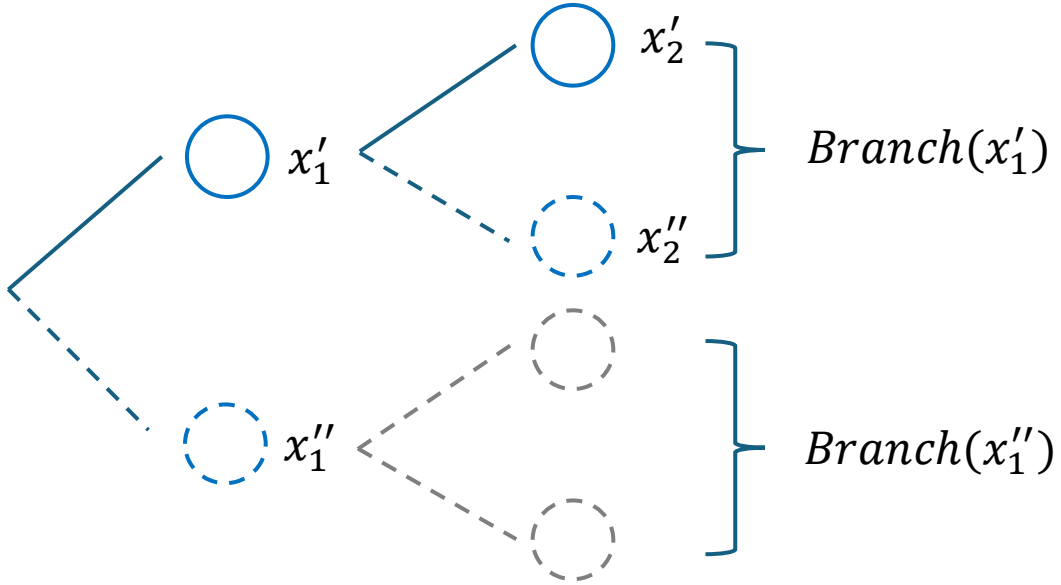
- **Case 2:**

Issue 1: $D_{Branch}(q, p|\mathrm{X}_{1:\mathrm{t}-1}) \neq D_{Branch}(p, q|\mathrm{X}_{1:\mathrm{t}-1})$

red area $\neq$ orange area

Issue 2: $p(\cdot\,|\tilde{X}_{1:\mathrm{t}-1} \neq X'_{1:t-1})$ and $q(\cdot\,|\tilde{X}_{1:\mathrm{t}-1} \neq X'_{1:t-1})$ are unknown





$\mathrm{P}(X_{1:t}\ is\ yielded) \neq p(X_{1:t})$

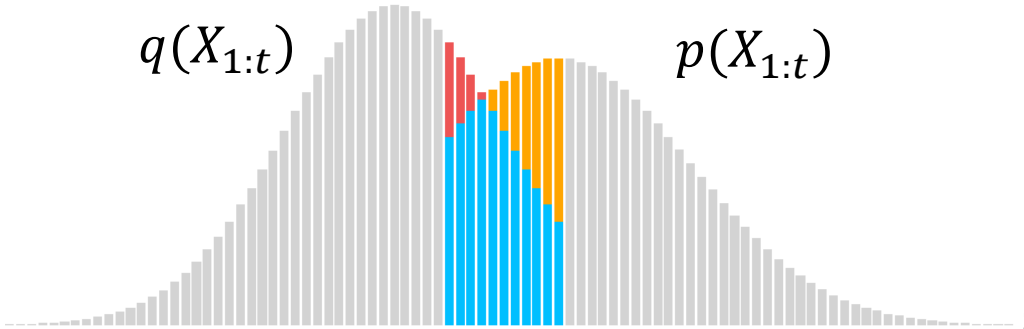
Method

Theory

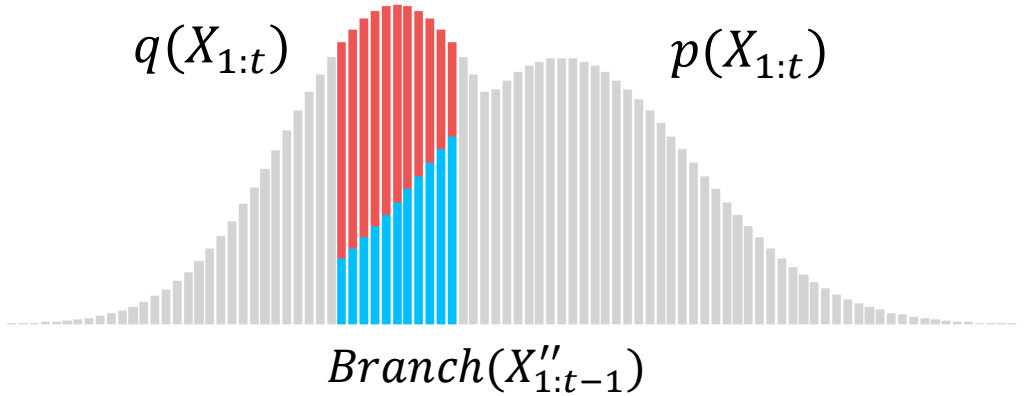
➢ **Definition: Asymmetry of Branch Divergence**

$$\Delta_{Branch}(X_{1:t-1}) = D_{Branch}(p, q|X_{1:t-1}) - D_{Branch}(q, p|X_{1:t-1})$$

- **negative asymmetry**： excess probability mass ⇨ target dist. locally recoverable



- **positive asymmetry**： deficient probability mass ⇨ target dist. locally unrecoverable



Theory

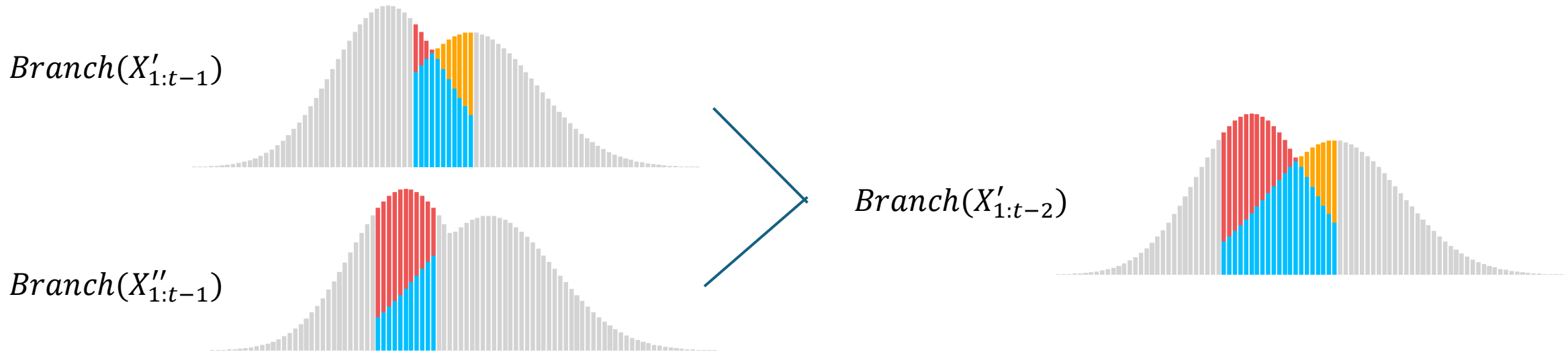
➢ **Theorem: Hierarchy of Branch Divergence**

$$\sum_{\Delta_{Branch}(X_{1:t-2},\tilde{x}_{t-1})>0} \Delta_{Branch}(X_{1:t-2}, \tilde{x}_{t-1}) = D_{Branch}(p, q|X_{1:t-2})$$

➢ **Key Insight:**

Target distribution is recoverable by **recursively** compensating deficient mass with excess mass at each branch, because:

$$D_{Branch}(p, q|X_0) = 0$$



Naive Hierarchical Speculative Decoding

➢ **Method:**

$$\sum_{\tau=0}^{\gamma}\sum_{\tilde{X}_{\tau+1:\gamma}} P\left(\tilde{X}_{1:\gamma} \neq X_{1:\gamma}\ is\ sampled\ and\ rejected, X_{1:\gamma}\ is\ resampled\right) =$$

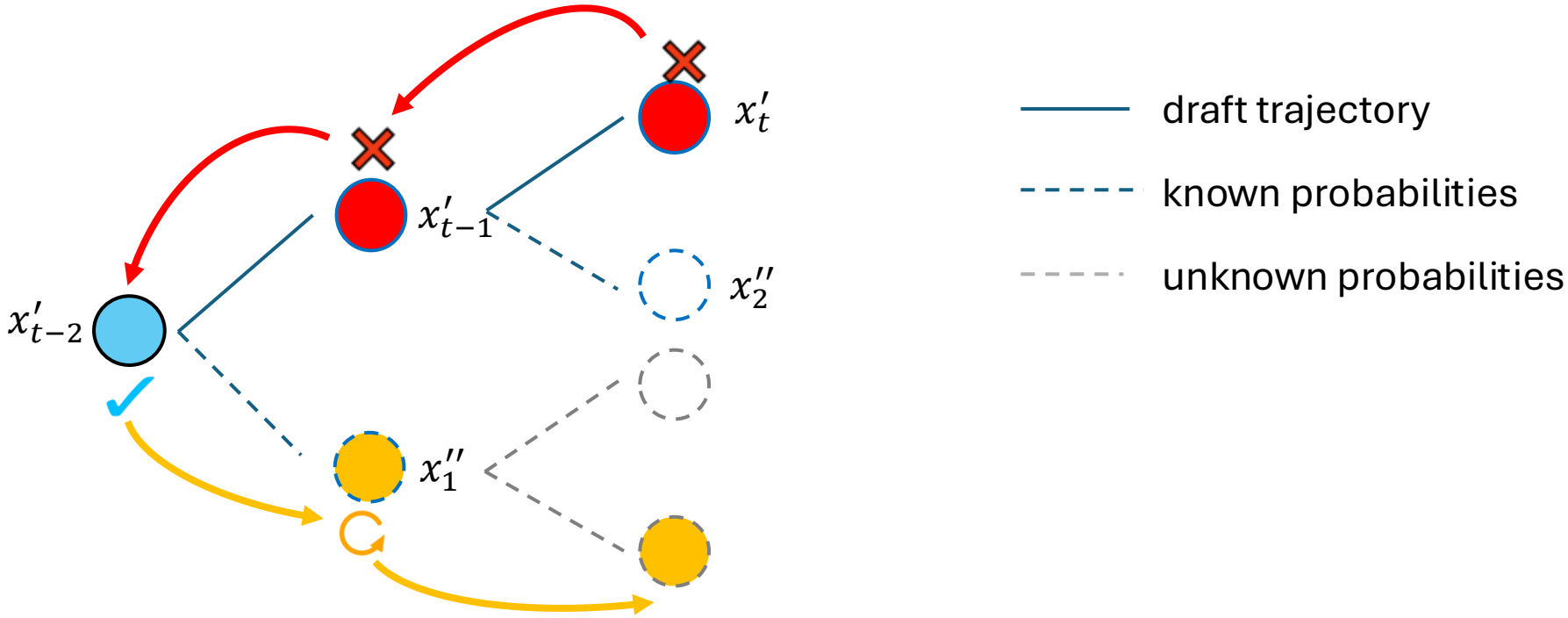
$$\sum_{\tau=0}^{\gamma}\sum_{\tilde{X}_{\tau+1:\gamma}} q(X_{1:\tau}\tilde{X}_{\tau+1:\gamma}) \cdot \prod_{t=\tau+1}^{\gamma}(1-h_t) \cdot h_\tau \cdot \prod_{t=\tau+1}^{\gamma} P_{res}(x_t)$$

Explanation of probability terms:

- **Sampling** $q(X_{1:\tau}\tilde{X}_{\tau+1:\gamma})$: generat the initial draft sequence.
- **Backward Scan** $\prod_{t=\tau+1}^{\gamma}(1-h_t)$: reject tokens until the first accepted prefix is found.
- **Acceptance** h_τ: accept the longest prefix $X_{1:\tau}$.
- **Resampling** $\prod_{t=\tau+1}^{\gamma} P_{res}(x_t)$: recursively resample to recover the target probability.

Naive Hierarchical Speculative Decoding

➢ **Illustration:**



➢**Issue:**

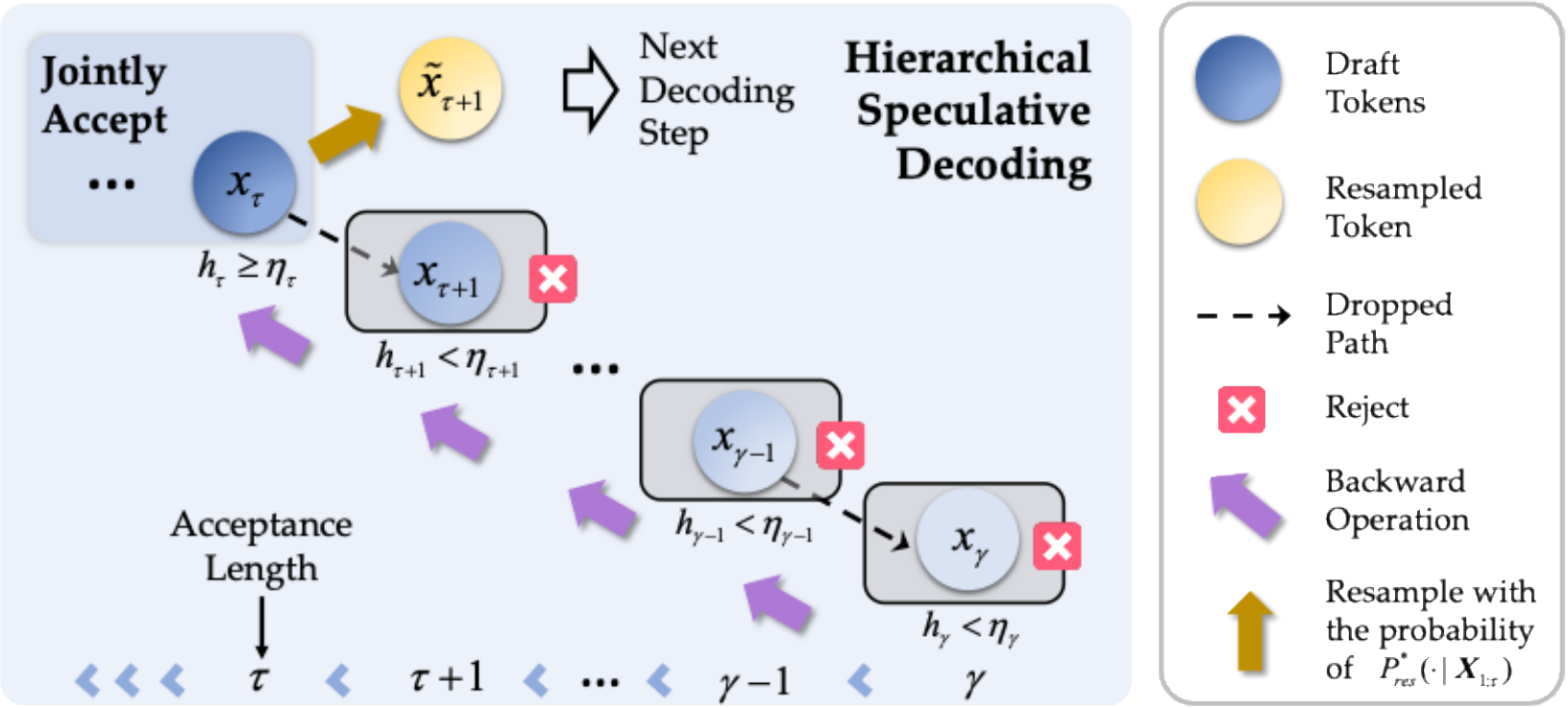
$P_{res}(x_t)$ requires the target and draft probability at the newly resampled trajectory,

⟹ $\gamma - \tau + 1$ calls to the target model

Hierarchical Speculative Decoding

➢ **Method: Capped Branch Resampling**

By clipping the maximum prefix ratio, we derive the corresponding accept and resample rule, which recovers the target distribution with **a single resampling step** within the accessible branches.



Thank you!