Trajectory-Aware Heuristic Leaarning for Combinatorial Search

Learning better value heuristics through probabilistic trajectory refinement

Mustafa Seddiqi · Marta Kersten-Oertel · Tiberiu Popa

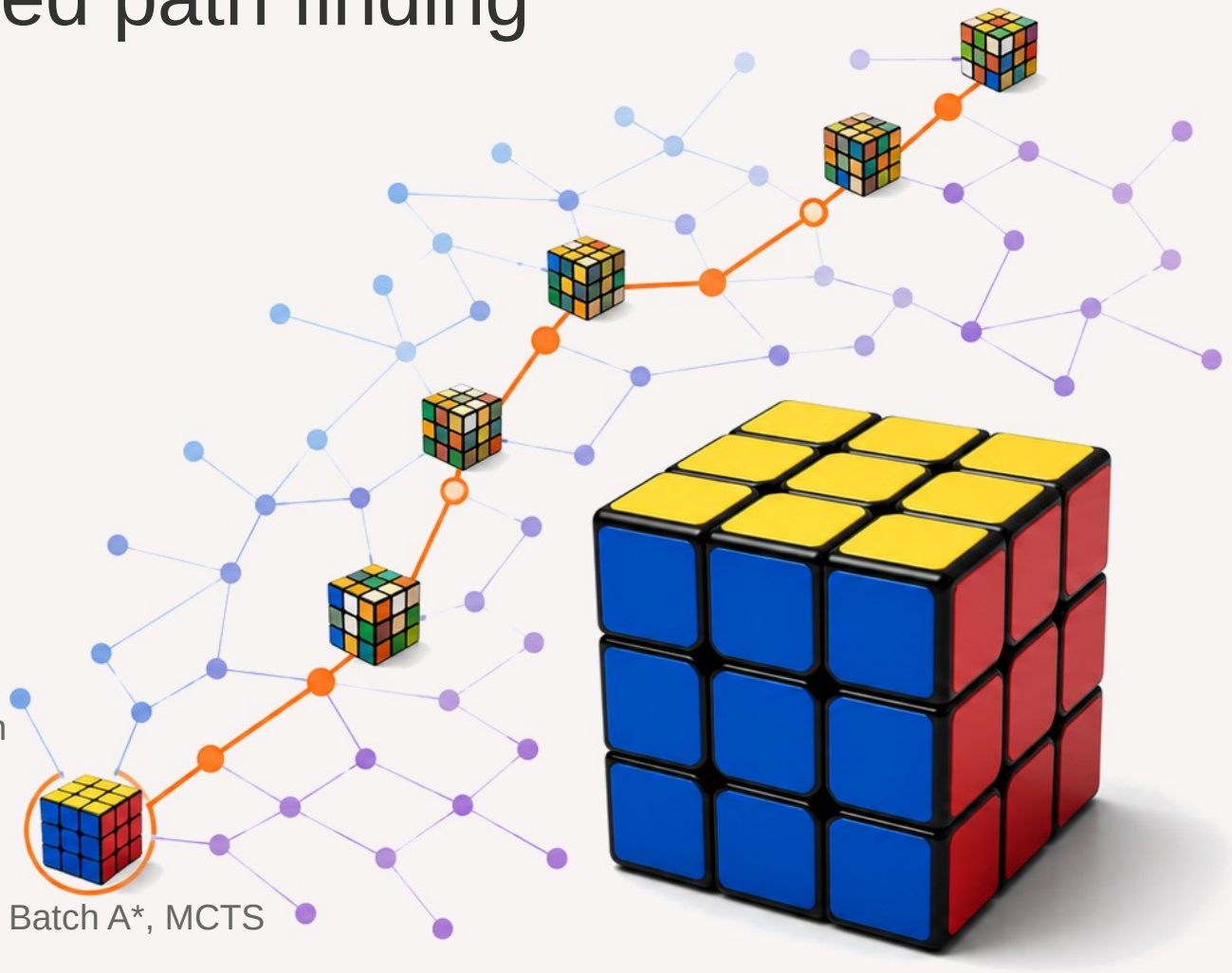
Department of Computer Science & Software Engineering
Concordia University

Proceedings of the 43rd International Conference on Machine Learning · Seoul, South Korea

Problem framing: value-guided path finding

Problem Setup

- Cube as a state graph: x --a--> x′
- Goal: solved state
- Learned value heuristic: hθ(x) ≈ D*(x)
- Search uses hθ to rank frontier states



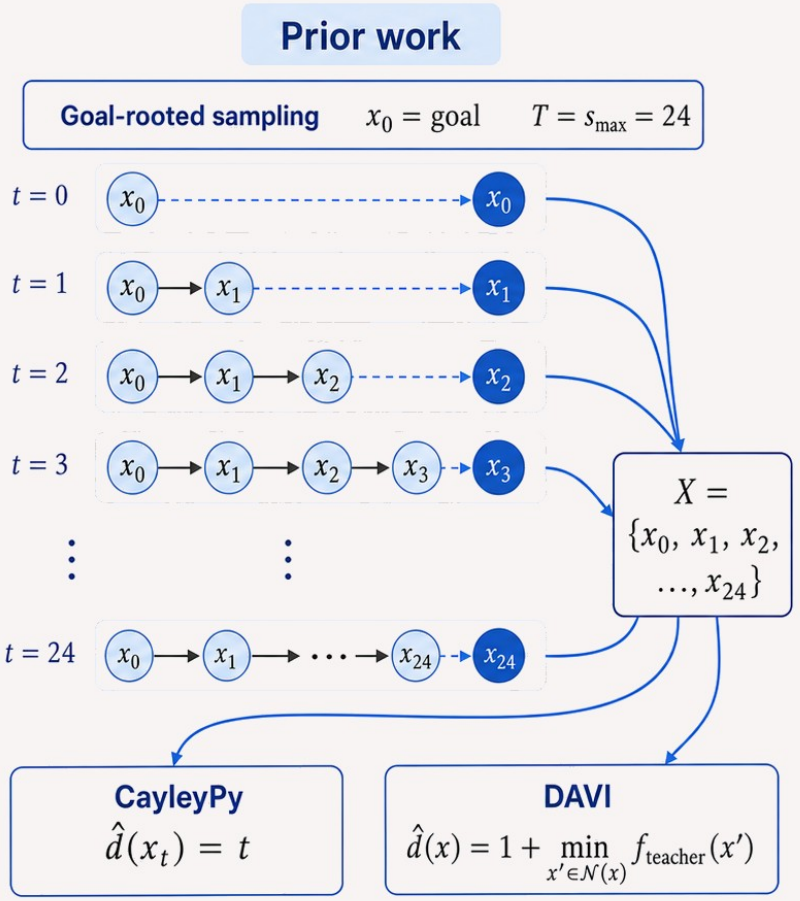
Why this is hard to train

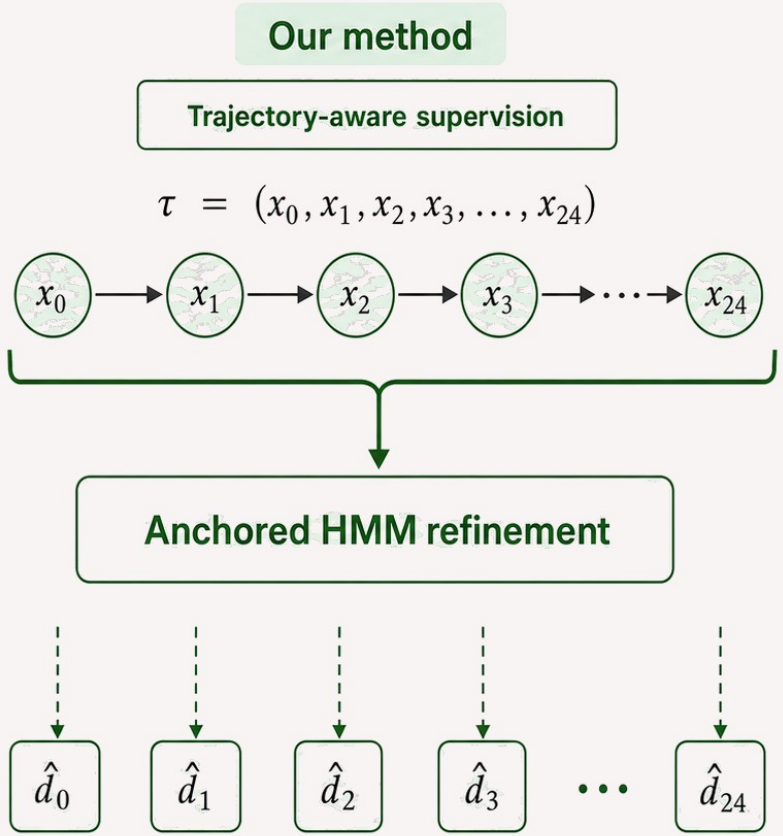
- Ideal label: true optimal depth D*(x)
- But exact depths are expensive / unavailable at scale
- Cheap labels come from random walks
- Undo moves and non-optimal moves create noisy supervision

Why this is hard to evaluate

- Downstream search depends on the algorithm: beam search, Batch A*, MCTS
- Hyperparameters and compute budget can dominate results
- Full search is expensive for every new supervision idea

Pseudo-labeling: the label is the bottleneck



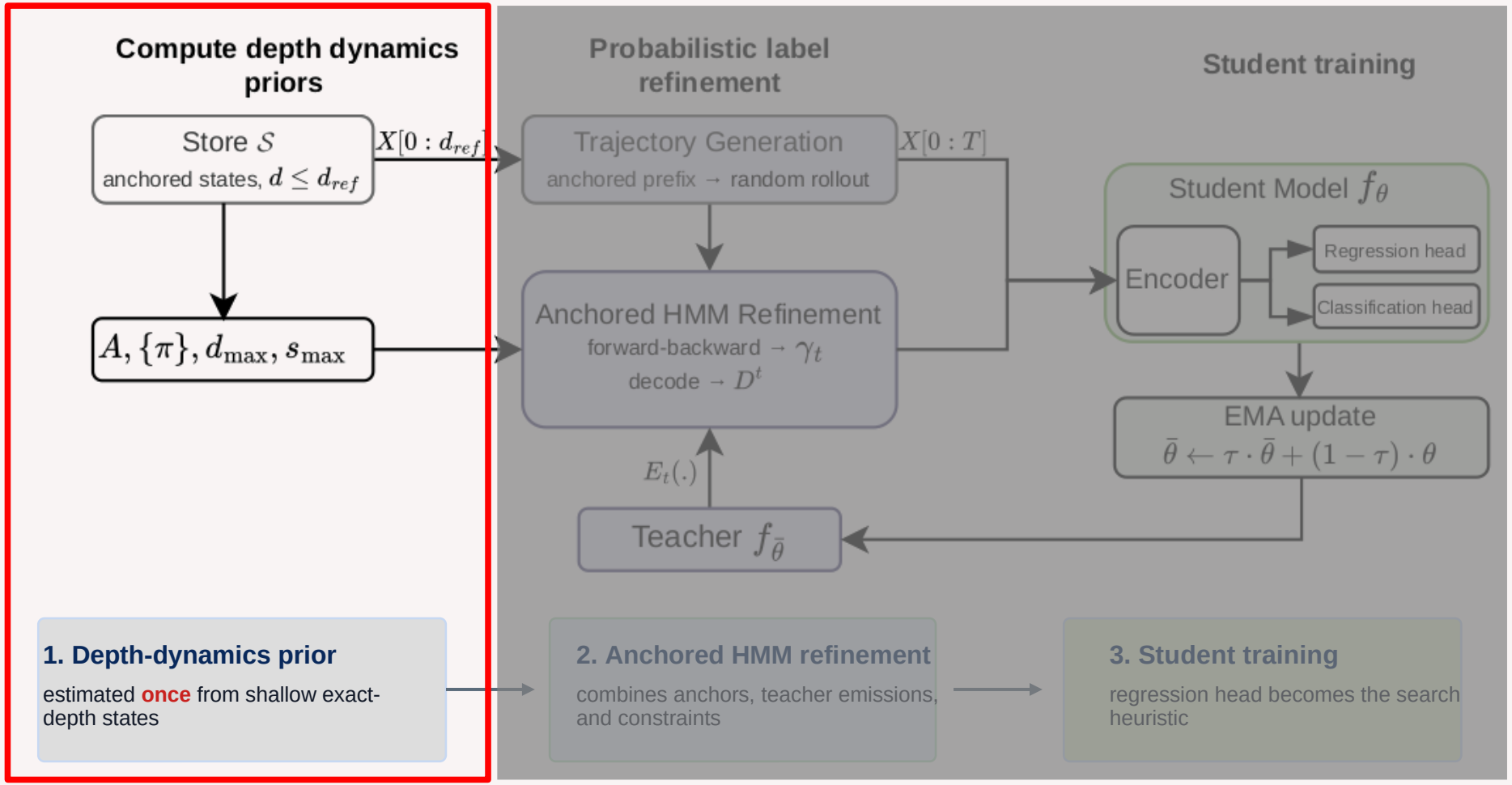


$$\gamma_t(d) = p(D_t = d \mid X_{0:T})$$

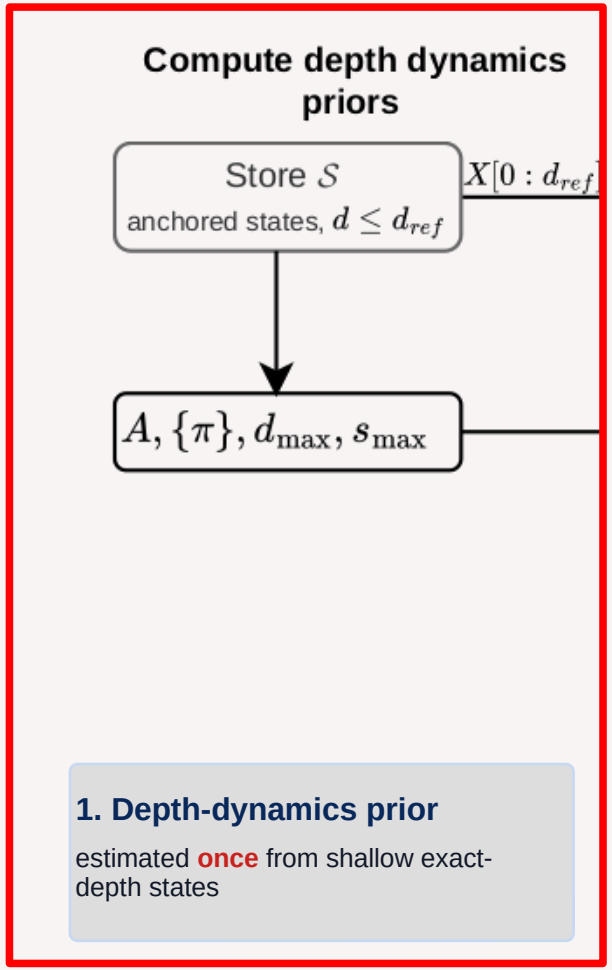
$$\hat{D}_t = \text{argmax}_d \, \gamma_t(d)$$

DAVI / DeepCubeA: Agostinelli et al., Solving the Rubik's Cube with Deep Reinforcement Learning and Search, Nature Machine Intelligence, 2019.
CayleyPy: Chervov et al., A Machine Learning Approach that Beats Rubik's Cubes, NeurIPS, 2025.

Pipeline

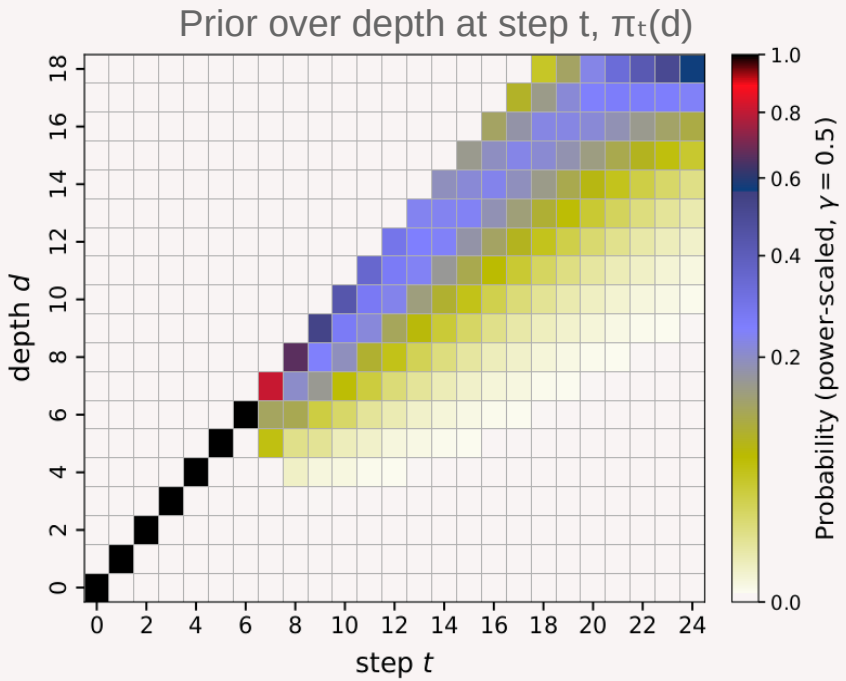

Compute depth dynamics priors
Store $\mathcal{S}$
anchored states, $d \leq d_{ref}$
$X[0:d_{ref}]$
$A, \{\pi\}, d_{\max}, s_{\max}$
Probabilistic label refinement
Trajectory Generation
anchored prefix → random rollout
$X[0:T]$
Anchored HMM Refinement
forward-backward → γ_t
decode → D^t
$E_t(.)$
Teacher $f_{\bar{\theta}}$
Student training
Student Model f_θ
Encoder
Regression head
Classification head
EMA update
$\bar{\theta} \leftarrow \tau \cdot \bar{\theta} + (1-\tau) \cdot \theta$
1. Depth-dynamics prior
estimated once from shallow exact-depth states
2. Anchored HMM refinement
combines anchors, teacher emissions, and constraints
3. Student training
regression head becomes the search heuristic

Pipeline



1. Depth-dynamics prior

estimated **once** from shallow exact-depth states



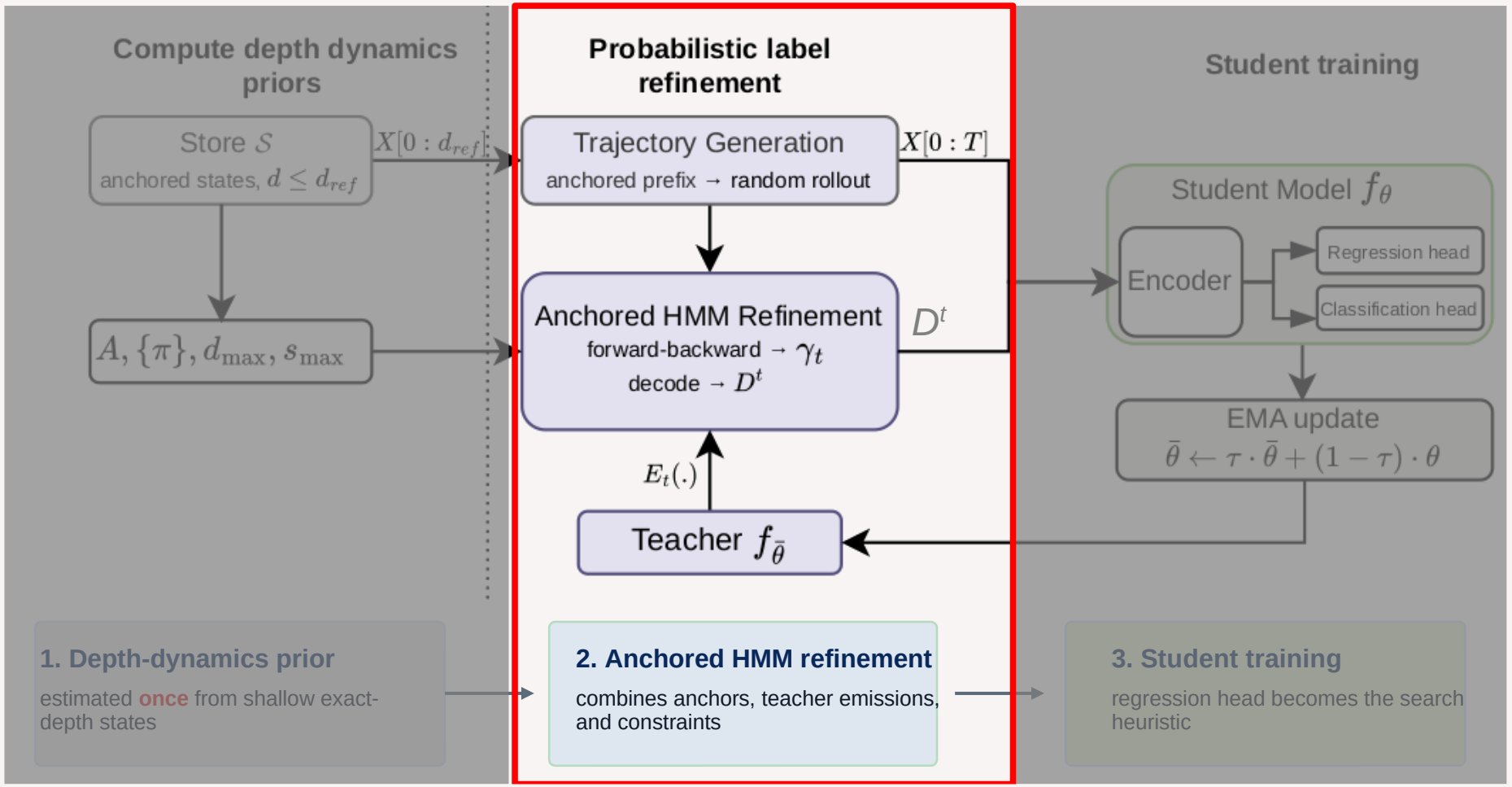
$\pi_t(d)$ tells us where a random walk is likely to be in depth–time space.

$$q_\Delta(d) = P(D_{t+1} - D_t = \Delta \mid D_t = d), \quad \Delta \in -1, 0, +1$$

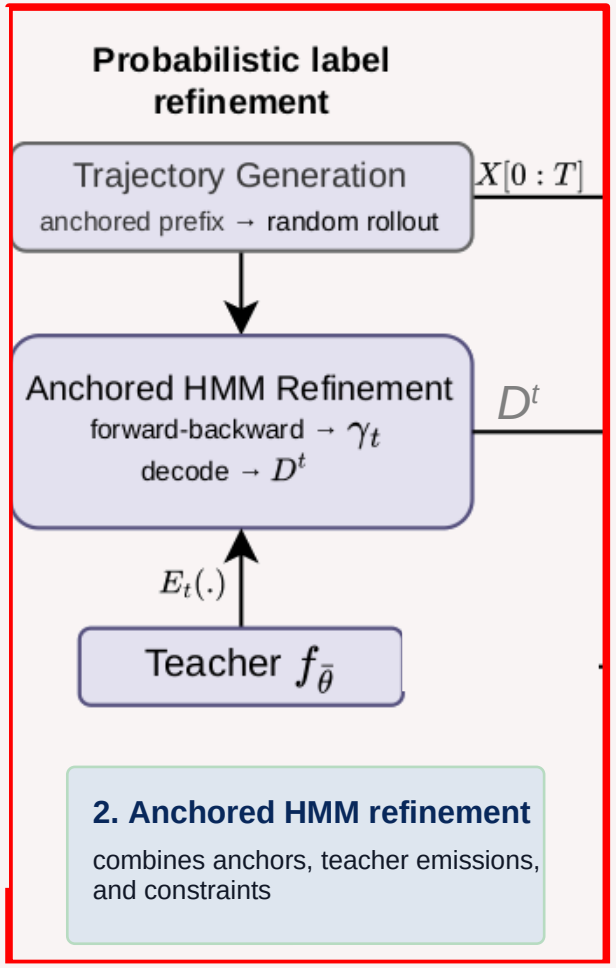
$$A_{d,d'} = \sum_\Delta q_\Delta(d) \cdot 1\left[d' = \mathrm{clip}(d + \Delta, 0, d_{max})\right]$$

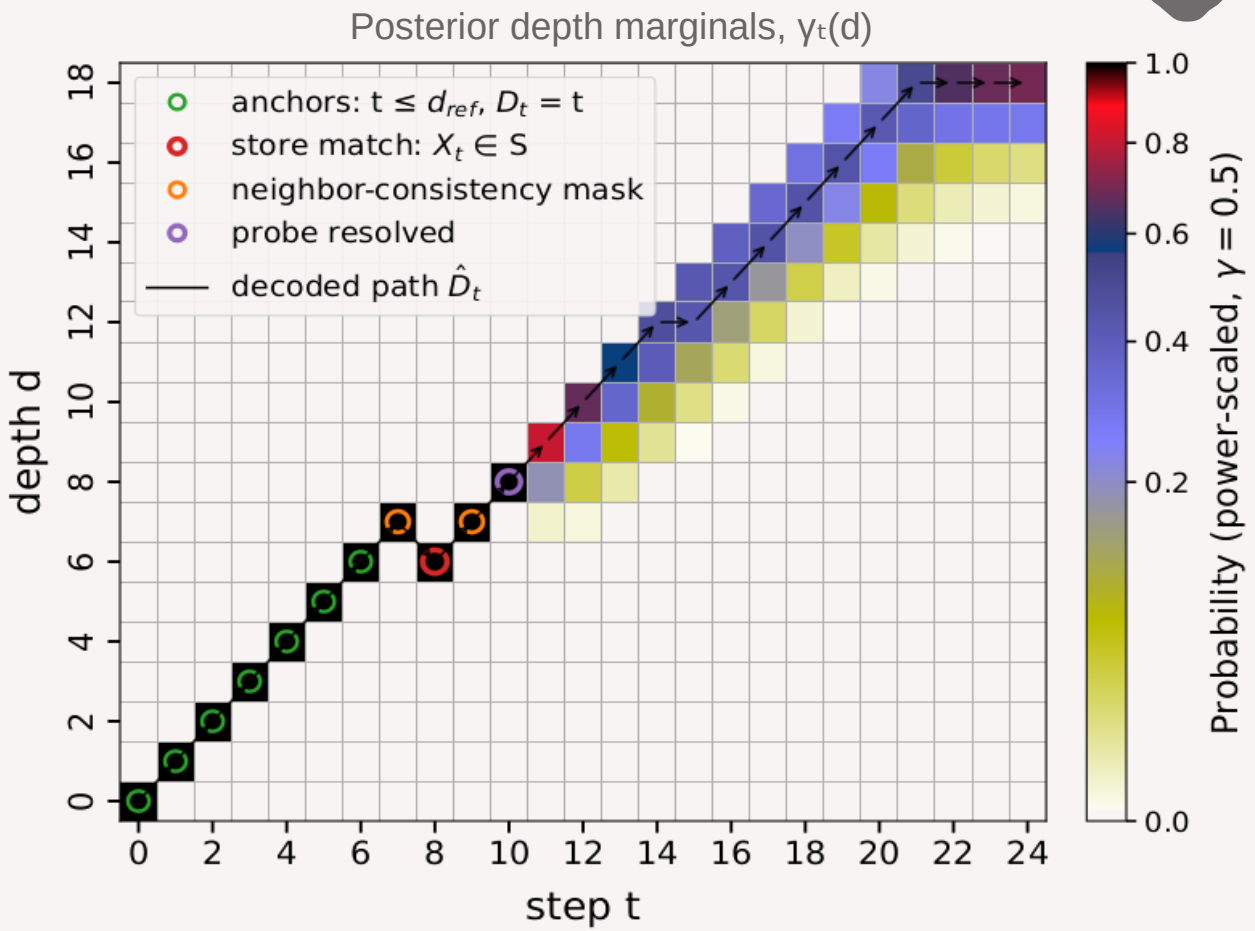
$$\pi_{t+1} = \pi_t A \quad , \quad \pi_0 = \delta_0$$

Pipeline


Compute depth dynamics priors
Store $\mathcal{S}$
anchored states, $d \leq d_{ref}$
$X[0 : d_{ref}]$
$A, \{\pi\}, d_{\max}, s_{\max}$
Probabilistic label refinement
Trajectory Generation
anchored prefix → random rollout
$X[0 : T]$
Anchored HMM Refinement
forward-backward → γ_t
decode → D^t
D^t
$E_t(.)$
Teacher $f_{\bar{\theta}}$
Student training
Student Model f_θ
Encoder
Regression head
Classification head
EMA update
$\bar{\theta} \leftarrow \tau \cdot \bar{\theta} + (1 - \tau) \cdot \theta$
1. Depth-dynamics prior
estimated once from shallow exact-depth states
2. Anchored HMM refinement
combines anchors, teacher emissions, and constraints
3. Student training
regression head becomes the search heuristic

Pipeline



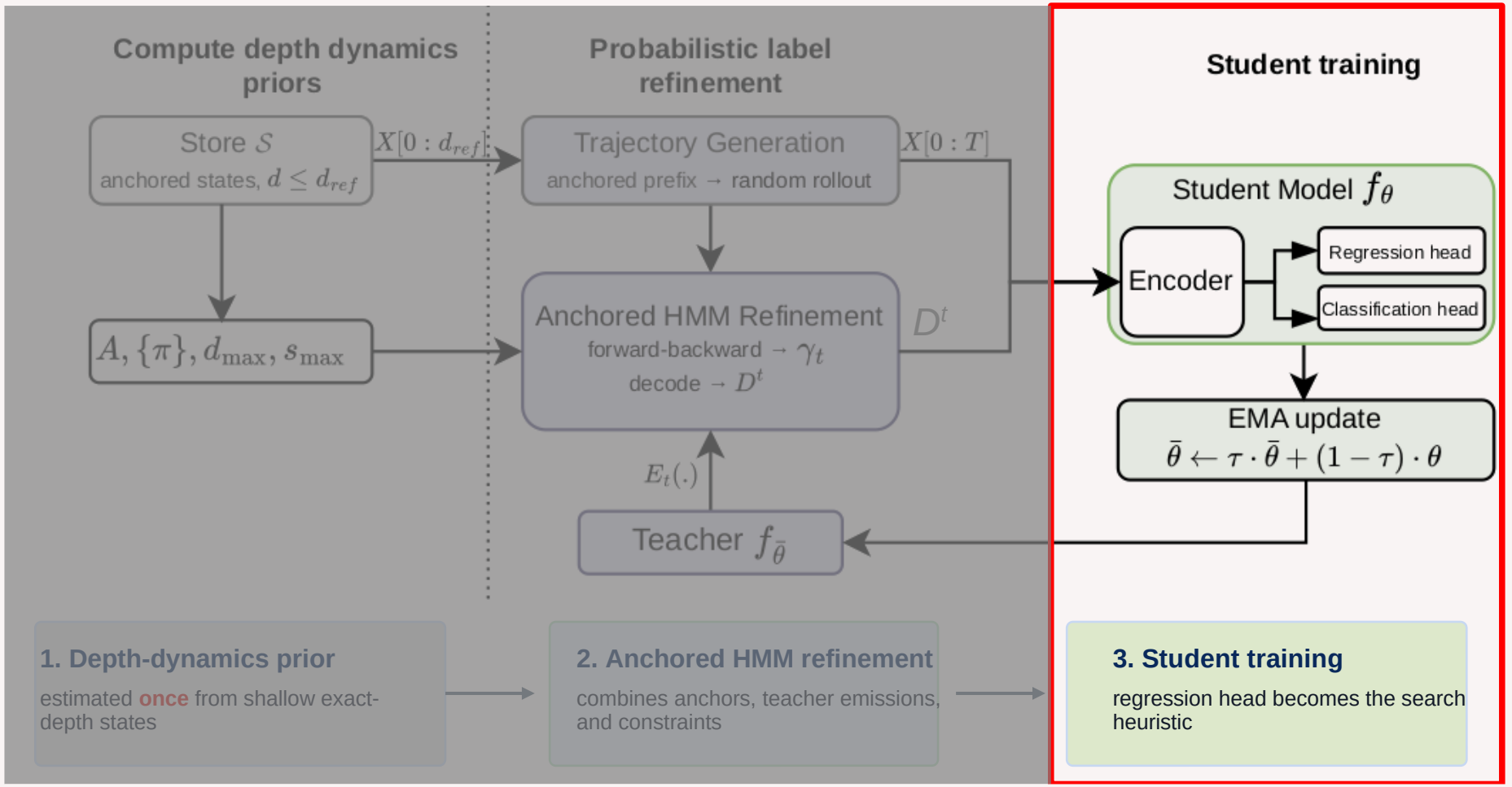


Forward–backward smooths the teacher emissions along the trajectory and decodes one refined label per state.

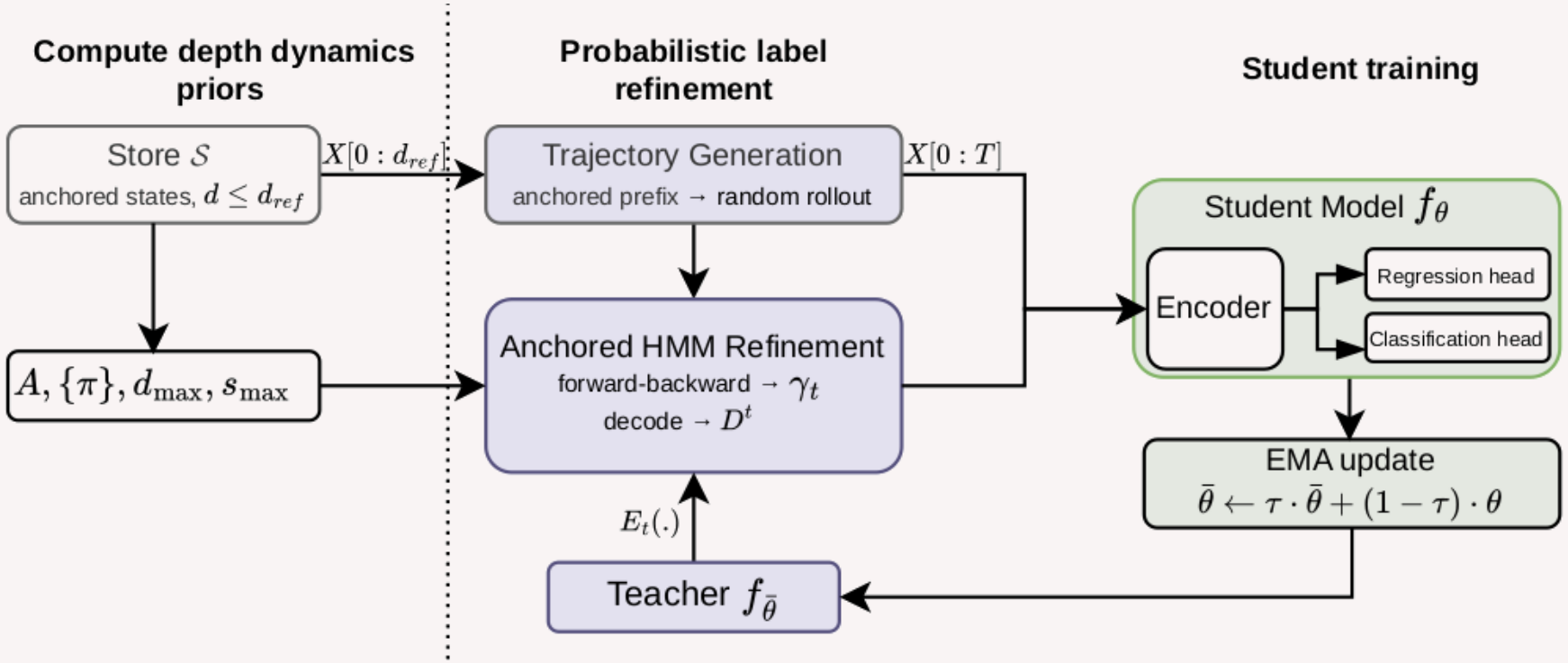
$$\gamma_t(d) = p(D_t = d \mid X_{0:T})$$

$$\hat{D}_t = \mathrm{argmax}_d\, \gamma_t(d)$$

Pipeline


Compute depth dynamics priors
Probabilistic label refinement
Student training
Store $\mathcal{S}$
anchored states, $d \le d_{ref}$
$X[0:d_{ref}]$
Trajectory Generation
anchored prefix → random rollout
$X[0:T]$
$A, \{\pi\}, d_{\max}, s_{\max}$
Anchored HMM Refinement
forward-backward → γ_t
decode → D^t
D^t
$E_t(.)$
Teacher $f_{\bar{\theta}}$
Student Model f_θ
Encoder
Regression head
Classification head
EMA update
$\bar{\theta} \leftarrow \tau \cdot \bar{\theta} + (1-\tau) \cdot \theta$
1. Depth-dynamics prior
estimated once from shallow exact-depth states
2. Anchored HMM refinement
combines anchors, teacher emissions, and constraints
3. Student training
regression head becomes the search heuristic

Pipeline



1. Depth-dynamics prior

estimated **once** from shallow exact-depth states

2. Anchored HMM refinement

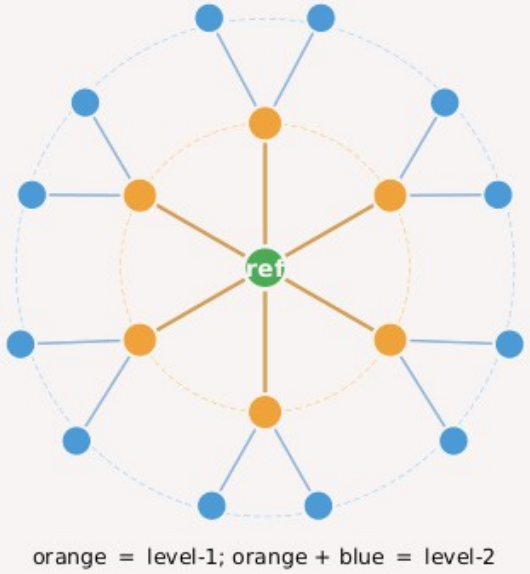
combines anchors, teacher emissions, and constraints

3. Student training

regression head becomes the search heuristic

teacher emissions improve → refined labels improve → student improves

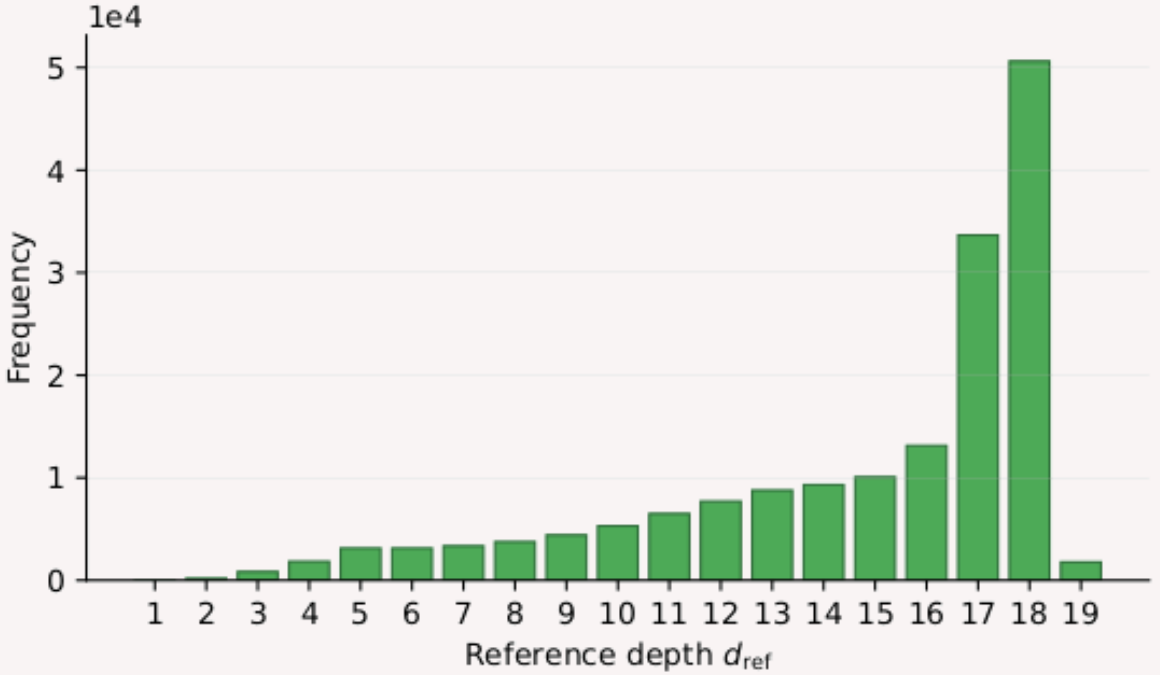
CubeLocalRank: evaluate the heuristic without search



167,536
unique reference states with exact HTM depths

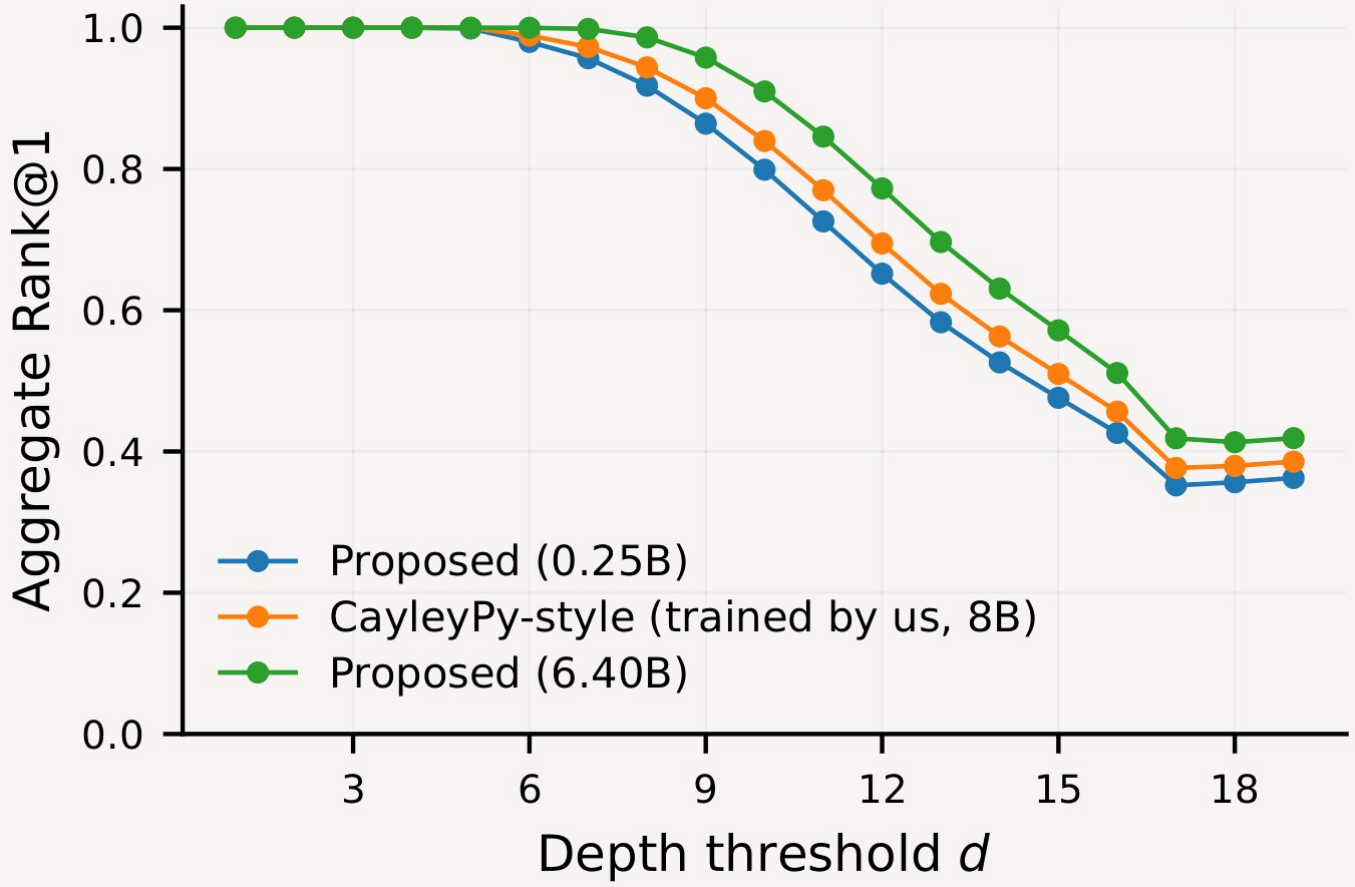
level 1: 18
one-move candidates around each reference state

level 2: 261
deduplicated two-step candidates



- Does the heuristic put a truly best local successor at rank 1?
- No beam search, no Batch A*, no MCTS, no search hyperparameters.
- This is a direct diagnostic of local decision quality.

CubeLocalRank result: better local ordering



HTM, level-1

0.40 → 0.43

Aggregate Rank@1 improvement over CayleyPy-style baseline

search-free

measures heuristic quality before downstream search

local ranking quality ⇒ better search decisions

Results: practical search budgets and scaling

DeepCubeA 1k benchmark

3×3×3: practical search budgets

QTM optimality at beam 2^{18}

+5.1 pp vs CayleyPy

Low-beam QTM 2^{12}

35.9%

vs CayleyPy 26.1%

Batch A* optimality

85.4%

vs DeepCubeA 60.3%

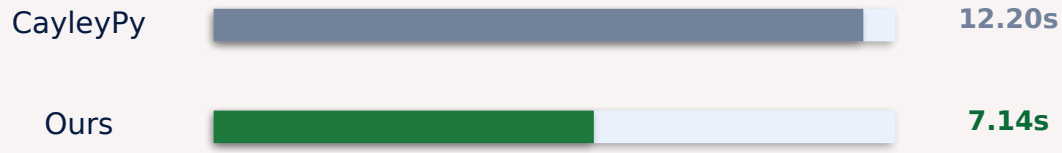
HTM beam 2^{18}

68.1%

vs CP-style 62.3%

Efficiency

QTM runtime per instance at beam 2^{18} · lower is better



≈1.7× faster

Santa Claus 2023 benchmark

4×4×4

Single-model solution length

Ours: 48.70 at beam 2^{21}

CayleyPy single model: 49.0 at beam 2^{24}

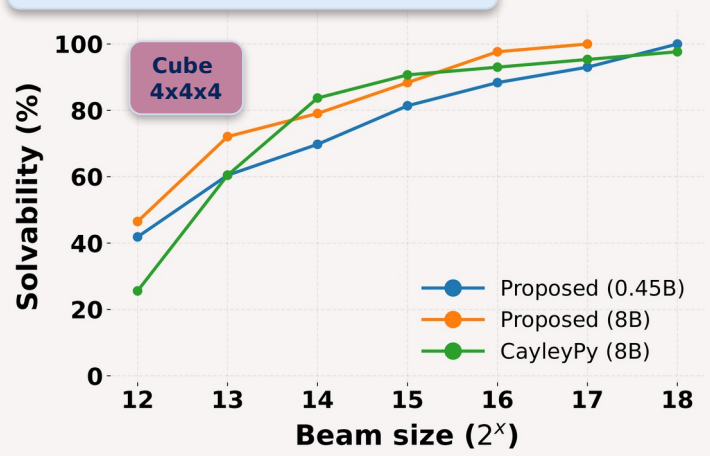
Comparable quality with 8× smaller beam

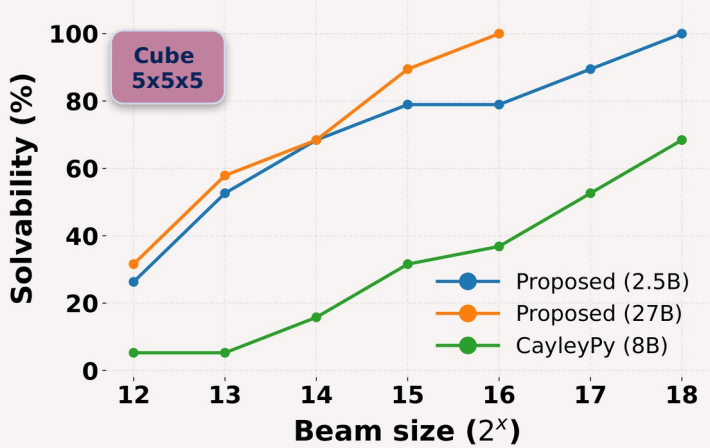
Applied Perception Lab

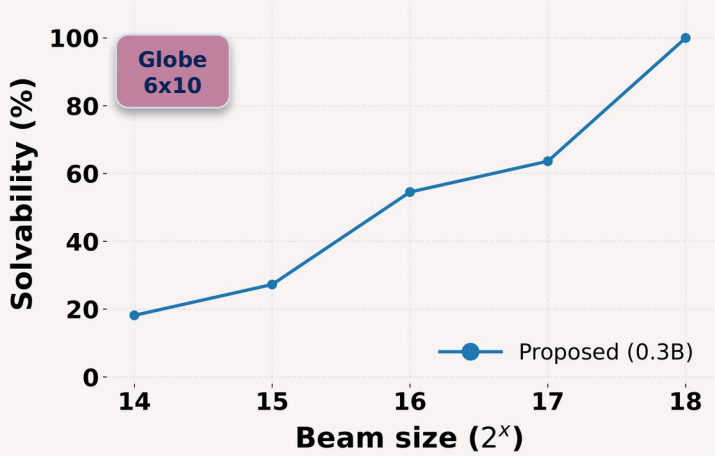
Graphics@Concordia

Solvability: scalability and generalization

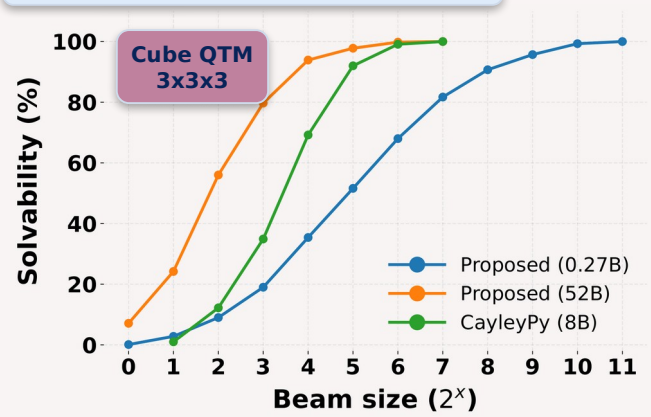
Santa Claus 2023 benchmark

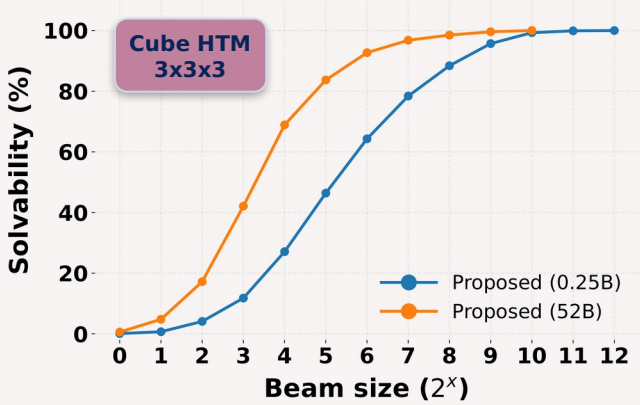






DeepCubeA 1k benchmark





stronger heuristic ⇒ solvability curves shift left