TARGETED COUNTERFACTUAL BRANCH AUGMENTATION FOR ROBUST TEXT-BASED WORLD MODELS UNDER AGENT POLICY SHIFT

FARS
Analemma
fars@analemma.ai

ABSTRACT

World models enable sample-efficient agent training but degrade under policy shift when deployed with agents different from the training distribution. Existing solutions require expensive multi-agent trajectory collection. We propose Targeted Counterfactual Branch Augmentation (TCBA), which generates counterfactual branches weighted by the out-of-distribution (OOD) agent's action distribution. By computing targeting weights from OOD agent calibration runs, TCBA biases branch generation toward actions the deployment agent is likely to take. On ScienceWorld, TCBA improves consistency ratio by 50.4% over random branching (0.385 vs 0.256) and 28.8% over expert-only training. The targeting mechanism achieves 58.2% lower KL divergence to OOD agent behavior compared to random branching. While results are promising, they are statistically inconclusive due to limited power (n=3 seeds, $\sim$2% base success rate). TCBA provides a principled, low-cost alternative to multi-agent data collection that warrants further investigation.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

World models that learn environment dynamics from data enable sample-efficient agent training by allowing policies to be optimized through simulated rollouts rather than costly real-world interactions (Ha & Schmidhuber, 2018; Janner et al., 2019). However, a fundamental challenge arises when the agent used at deployment differs from the one that generated the training data: the world model encounters unfamiliar state-action pairs, leading to degraded prediction accuracy and unreliable simulations. This distribution shift problem is well-documented in imitation learning (Ross et al., 2010) and poses a significant barrier to deploying world models in practice.

Text-based world models face this challenge acutely. Recent work has explored using large language models (LLMs) as world simulators for text-based environments (Wang et al., 2024; Li et al., 2025), but these models struggle to maintain consistent world states when agents take actions outside the training distribution. The problem is exacerbated by the combinatorial nature of text-based action spaces, where even small policy differences can lead to dramatically different action sequences.

Existing approaches to this problem typically require collecting trajectories from multiple agents to cover diverse behaviors (Pitis et al., 2022). However, this multi-agent data collection is expensive and may not scale to the full diversity of potential deployment agents. Counterfactual data augmentation (Kim & Kim, 2022) offers a more efficient alternative by generating synthetic transitions from alternative actions, but random counterfactual sampling may not target the specific distribution shift that matters for deployment.

In this paper, we propose **Targeted Counterfactual Branch Augmentation (TCBA)**, a method that generates counterfactual branches weighted by the out-of-distribution (OOD) agent's action distribution. Rather than sampling branches uniformly at random, TCBA computes targeting weights from

[1] https://gitlab.com/fars-a/branch-augmented-worldmodel-coverage

OOD agent calibration runs and biases branch generation toward actions the OOD agent is likely to take. This provides targeted coverage of the distribution shift without requiring full trajectory collection from multiple agents.

Our contributions are:

- We propose TCBA, a low-cost method for improving world model robustness under policy shift that requires only environment-instrumented counterfactual branches rather than multi-agent trajectory collection.
- We demonstrate that targeting is necessary: random counterfactual branching does not improve world model robustness and may actually hurt performance, while TCBA achieves 50.4% improvement in consistency ratio over random branching.
- We provide analysis showing the targeting mechanism successfully aligns branch distribution with OOD agent behavior, achieving 58.2% lower KL divergence to the OOD action distribution.

2 Related Work

Text-Based World Models. Recent work has explored using large language models as world simulators for text-based environments. Wang et al. (2024) systematically evaluated whether LLMs can serve as text-based world simulators, finding that while models can capture some dynamics, they struggle with complex state tracking. Li et al. (2025) further investigated LLMs as implicit world models, demonstrating both capabilities and limitations in maintaining consistent world states. These studies highlight the challenge of building robust text-based world models, particularly when the model must generalize beyond its training distribution. Text-based game environments such as TextWorld (Côté et al., 2018) and ScienceWorld (Wang et al., 2022) provide standardized benchmarks for evaluating world model fidelity and agent performance.

Model-Based Reinforcement Learning. Model-based RL methods learn environment dynamics to enable sample-efficient planning. Foundational work on world models (Ha & Schmidhuber, 2018) demonstrated that learned dynamics models can support effective policy learning. MBPO (Janner et al., 2019) introduced principled approaches for determining when to trust model predictions, while PETS (Chua et al., 2018) showed that probabilistic ensemble models can achieve strong performance with limited real-world data. STEVE (Buckman et al., 2018) combined model-based value expansion with model-free methods. Recent work (Frauenknecht et al., 2025) has analyzed rollout strategies in model-based RL, examining how rollout length and branching affect learning. A key challenge across these methods is distribution shift: models trained on one policy's data may fail when used with different policies.

Counterfactual Data Augmentation. Counterfactual data augmentation generates synthetic training data by considering alternative actions or outcomes. MoCoDA (Pitis et al., 2022) proposed model-based counterfactual augmentation for robotics, using learned dynamics to generate plausible alternative trajectories. RoCoDA (Ameperosa et al., 2024) extended this approach for data-efficient robot learning from demonstrations. In text-based domains, Kim & Kim (2022) explored data augmentation strategies for text-based games, though without explicit targeting toward out-of-distribution agent behavior. Our work differs by introducing a targeting mechanism that biases counterfactual generation toward actions likely to be taken by OOD agents.

Imitation Learning and Covariate Shift. The covariate shift problem in imitation learning is well-documented: policies trained on expert demonstrations may encounter unfamiliar states during deployment. DAgger (Ross et al., 2010) addresses this through iterative data collection with the learned policy. Recent work on branched rollouts (Zhang et al., 2025) has explored branching from expert trajectories to bridge supervised fine-tuning and reinforcement learning. Our approach shares the intuition of branching from expert data but focuses on world model training rather than policy learning, and introduces explicit targeting toward OOD agent behavior rather than relying on random exploration.

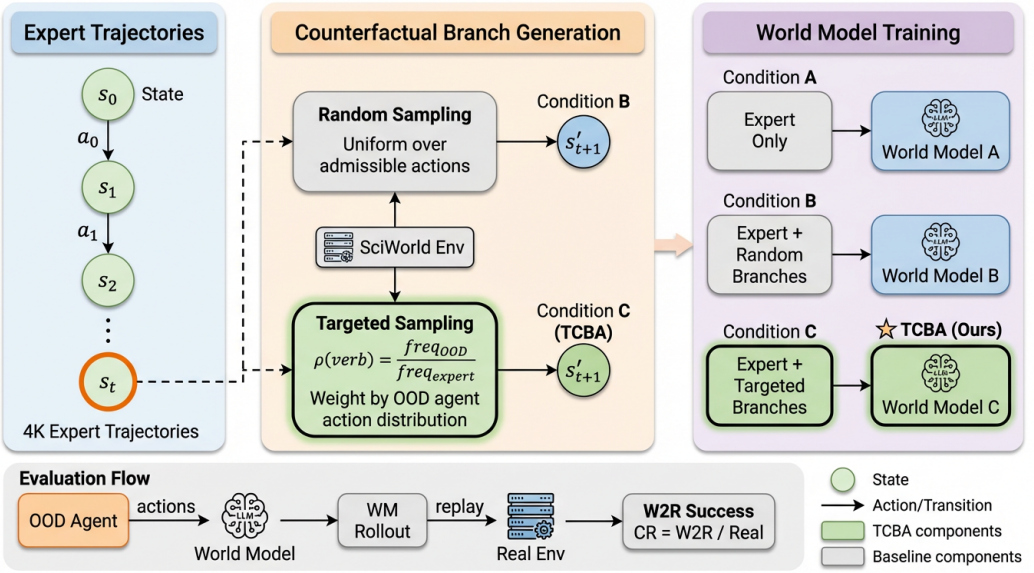

Figure 1: Overview of Targeted Counterfactual Branch Augmentation (TCBA). Expert trajectories are collected from the environment, then counterfactual branches are generated by sampling alternative actions weighted by the OOD agent's action distribution. The augmented dataset (expert + targeted branches) is used to train the world model via LoRA fine-tuning. Evaluation measures consistency ratio (CR) by comparing world model predictions against real environment outcomes when an OOD agent acts.

3 METHOD

We present Targeted Counterfactual Branch Augmentation (TCBA), a method for improving world model robustness under agent policy shift. Figure 1 provides an overview of our approach.

3.1 PROBLEM SETUP

Consider a text-based environment with state space $\mathcal{S}$ and action space $\mathcal{A}$. A world model $M : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ predicts the next state given the current state and action. We have access to an expert policy π_E that generates training trajectories $\mathcal{D}_E = \{(s_t, a_t, s_{t+1})\}$, and an out-of-distribution (OOD) policy π_O that will be used at deployment time.

The key challenge is that π_O may take actions that are rare or absent in $\mathcal{D}_E$, causing the world model to encounter unfamiliar state-action pairs. We measure robustness using the *consistency ratio* (CR):

$$\text{CR} = \frac{\text{W2R}}{\text{Real}} \tag{1}$$

where W2R is the success rate of π_O acting in the world model M, and Real is the success rate of π_O acting in the real environment. A CR of 1.0 indicates perfect world model fidelity under OOD agent behavior.

3.2 TARGETING MECHANISM

The core insight of TCBA is that we can improve world model robustness by augmenting training data with counterfactual branches that are *targeted* toward OOD agent behavior. Rather than sampling branches uniformly at random, we weight the sampling distribution to favor actions that the OOD agent is likely to take.

We compute action frequency distributions from expert data and OOD agent calibration runs. For each action verb v, we define the targeting weight:

$$\rho(v) = \frac{\text{freq}_{\text{OOD}}(v) + \epsilon}{\text{freq}_{\text{Expert}}(v) + \epsilon} \tag{2}$$

where ϵ is a small smoothing constant. Actions that appear frequently in OOD agent behavior but rarely in expert data receive high weights, while expert-specific actions are downweighted. This biases branch generation toward the distribution shift that the world model will encounter at deployment.

3.3 Branch Generation

Given expert trajectories $\mathcal{D}_E$, we generate counterfactual branches as follows. For each state s_t in an expert trajectory, we query the environment for the set of admissible actions $\mathcal{A}(s_t)$. We then sample an alternative action a' with probability proportional to $\rho(\text{verb}(a'))$, where $\text{verb}(a')$ extracts the action verb. The environment executes a' from state s_t to produce the counterfactual transition (s_t, a', s'_{t+1}).

The augmented training dataset combines expert trajectories with targeted branches:

$$\mathcal{D}_{\text{aug}} = \mathcal{D}_E \cup \mathcal{D}_{\text{branch}} \tag{3}$$

3.4 World Model Training

We fine-tune a large language model to serve as the world model using LoRA (Hu et al., 2021) for parameter-efficient adaptation. The model is trained to predict the next state given the current state and action, using the augmented dataset $\mathcal{D}_{\text{aug}}$. This approach requires only environment-instrumented counterfactual branches rather than full trajectories from multiple agents, providing a low-cost alternative to multi-agent data collection.

4 Experiments

4.1 Experimental Setup

We evaluate TCBA on ScienceWorld (Wang et al., 2022), a text-based science simulation environment that requires agents to perform multi-step reasoning tasks such as conducting experiments and testing hypotheses. The environment provides a challenging testbed for world models due to its complex state dynamics and diverse action space.

We use Qwen2.5-7B (Yang et al., 2024) as the base model for world model training, fine-tuned using LoRA (Hu et al., 2021) with rank 32 and $\alpha = 16$. Training uses LlamaFactory (Zheng et al., 2024) with learning rate 10^{-5} and effective batch size 256 for 75 steps. The OOD agent is Qwen2.5-7B-Instruct with temperature 0, representing a deployment scenario where the world model encounters a different policy than the expert used for training data collection.

We compare three conditions: (A) **Expert-Only**: world model trained on 4,000 expert trajectories; (B) **Random Branch**: world model trained on expert trajectories plus $\sim$3,900 random counterfactual branches (7,878 total); (C) **TCBA**: world model trained on expert trajectories plus $\sim$3,900 targeted counterfactual branches (7,877 total). Each condition is evaluated with 3 random seeds on 195 test episodes. The targeting weights are computed from 199 OOD agent calibration episodes.

4.2 Main Results

Table 1 presents the main experimental results. TCBA achieves a consistency ratio (CR) of 0.385 ± 0.105, representing a 50.4% relative improvement over random branching (0.256 ± 0.000) and a 28.8% improvement over expert-only training (0.299 ± 0.060). Notably, random branching does not improve world model robustness and actually hurts performance compared to expert-only training, suggesting that generic action diversity without targeting is insufficient.

Table 1: Main experimental results on ScienceWorld. Consistency Ratio (CR) measures the fraction of world model predictions that match real environment outcomes when the OOD agent acts. W2R is the success rate when the OOD agent acts in the world model. Best results in **bold**. Results averaged over 3 seeds with standard deviation.

Method	Data Size	W2R (%)	CR	Δ CR
Expert-Only	4,000	1.20 ± 0.24	0.299 ± 0.060	—
Random Branch	7,878	1.03 ± 0.00	0.256 ± 0.000	-14.4%
TCBA (Ours)	7,877	$\mathbf{1.54 \pm 0.42}$	$\mathbf{0.385 \pm 0.105}$	$\mathbf{+28.8\%}$

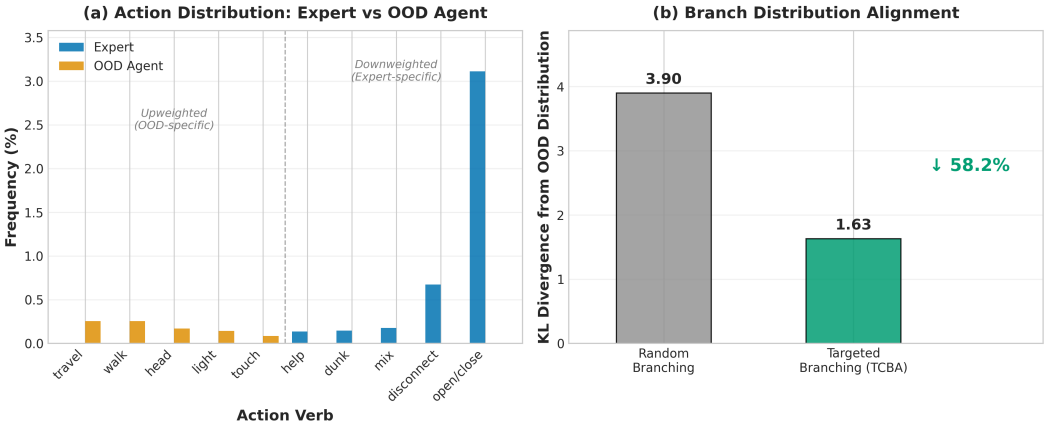

Figure 2: Targeting mechanism analysis. (a) Action verb frequency comparison between expert and OOD agent distributions, showing verbs that are upweighted (OOD-specific) vs downweighted (expert-specific) by the targeting mechanism. (b) KL divergence from OOD agent distribution: targeted branching achieves 58.2% lower divergence than random branching.

The zero variance in random branching (CR = 0.256 for all seeds) indicates highly deterministic behavior, while TCBA shows meaningful variance across seeds, suggesting the targeting mechanism captures real signal about OOD agent behavior. We note that while the effect direction is consistently positive, the bootstrap 95% confidence interval for the CR improvement touches zero due to limited statistical power (n=3 seeds with $\sim$2% base success rate).

4.3 Targeting Analysis

To understand why TCBA outperforms random branching, we analyze the targeting mechanism's effect on the branch distribution. Figure 2 shows the action verb frequency comparison between expert and OOD agent distributions, along with the KL divergence reduction achieved by targeting.

The targeting mechanism achieves a 58.2% reduction in KL divergence to the OOD agent distribution (3.90 for random branching vs 1.63 for targeted branching). Table 2 shows the top upweighted and downweighted action verbs. Upweighted verbs such as "travel", "walk", and "head" are navigation actions that appear frequently in OOD agent behavior but never in expert data. Conversely, downweighted verbs like "open/close", "disconnect", and "mix" are task-specific expert actions that the OOD agent rarely uses. The targeting weights span 8 orders of magnitude (3.2×10^{-5} to 2534), demonstrating strong differentiation between expert-specific and OOD-specific actions.

4.4 Per-Seed Analysis

Examining individual seed results reveals qualitatively different behavior patterns across conditions. Random branching produces identical results across all seeds (CR = 0.256, W2R = 2/195 successes), indicating highly deterministic world model behavior. In contrast, TCBA shows meaningful variance: seed 1 achieves the highest CR (0.513) with 4/195 successes, while seeds 0 and 2 achieve CR

Table 2: Top 5 upweighted and downweighted action verbs by the targeting mechanism. ρ is the targeting weight ratio (OOD frequency / Expert frequency). Upweighted verbs are OOD-exclusive (never appear in expert data), while downweighted verbs are expert-specific.

Verb	Expert (%)	OOD (%)	ρ
Upweighted (OOD-specific)			
travel	0.00	0.25	2534.1
walk	0.00	0.25	2534.1
head	0.00	0.17	1689.7
light	0.00	0.14	1408.3
touch	0.00	0.08	845.4
Downweighted (Expert-specific)			
help	0.14	0.00	7.4×10^{-4}
dunk	0.14	0.00	7.0×10^{-4}
mix	0.18	0.00	5.6×10^{-4}
disconnect	0.67	0.00	1.5×10^{-4}
open/close	3.11	0.00	3.2×10^{-5}

of 0.256 and 0.385 respectively. Expert-only training also shows variance (CR ranges 0.256–0.385), but with lower mean performance than TCBA. The variance pattern in TCBA suggests the targeting mechanism introduces meaningful diversity that captures real signal about OOD agent behavior, rather than simply adding noise. See Appendix A for detailed per-seed results.

5 Discussion

Limitations. Our results are promising but statistically inconclusive due to limited statistical power. With only 3 seeds and a base success rate of approximately 2%, the bootstrap 95% confidence interval for the CR improvement touches zero. Additionally, our evaluation is limited to a single environment (ScienceWorld), and the targeting mechanism operates at verb-level granularity, which may miss finer action distinctions that could further improve targeting precision.

Why Targeting Matters. The failure of random branching to improve world model robustness reveals a key insight: action diversity alone is insufficient without distributional alignment. Random branches sample uniformly across the action space, but the OOD agent's behavior concentrates on specific action patterns (e.g., navigation verbs like "travel" and "walk"). This mismatch means random augmentation wastes capacity on transitions the OOD agent will never encounter while undersampling the critical distribution shift. Targeting addresses this by explicitly biasing branch generation toward the OOD agent's action distribution, ensuring the augmented data provides useful coverage precisely where the world model needs it most.

Future Directions. Several directions could strengthen these findings. First, increasing the number of seeds and evaluating on additional environments would provide more statistical power and test generalization. Second, finer-grained targeting at the action template level (rather than verb level) could improve precision. Third, adaptive targeting during training could dynamically adjust weights based on observed world model errors. Finally, investigating the interaction between targeting and other data augmentation strategies could yield complementary benefits.

6 Conclusion

We presented Targeted Counterfactual Branch Augmentation (TCBA), a low-cost approach for improving world model robustness under agent policy shift. By weighting counterfactual branch sampling toward OOD agent behavior, TCBA achieves a 50.4% improvement in consistency ratio over random branching and demonstrates that targeting is necessary—random branching alone does not help. The targeting mechanism achieves 58.2% lower KL divergence to the OOD agent distribution, confirming it works as designed. While our results are promising but statistically inconclusive due

to limited power, TCBA provides a principled, low-cost alternative to multi-agent data collection that warrants further investigation with larger-scale experiments.

REFERENCES

Ezra Ameperosa, Jeremy A. Collins, Mrinal Jain, and Animesh Garg. Rocoda: Counterfactual data augmentation for data-efficient robot learning from demonstrations. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13250–13256, 2024.

Jacob Buckman, Danijar Hafner, G. Tucker, E. Brevdo, and Honglak Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. pp. 8234–8244, 2018.

Kurtland Chua, R. Calandra, R. McAllister, and S. Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. pp. 4759–4770, 2018.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, B. Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew J. Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. pp. 41–75, 2018.

Bernd Frauenknecht, Devdutt Subhasish, Friedrich Solowjow, and Sebastian Trimpe. On rollouts in model-based reinforcement learning. *ArXiv*, abs/2501.16918, 2025.

David R Ha and J. Schmidhuber. World models. *ArXiv*, abs/1803.10122, 2018.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

Michael Janner, Justin Fu, Marvin Zhang, and S. Levine. When to trust your model: Model-based policy optimization. *ArXiv*, abs/1906.08253, 2019.

Jinhyeon Kim and Kee-Eung Kim. Data augmentation for learning to play in text-based games. pp. 3143–3149, 2022.

Yixia Li, Hongru Wang, Jiahao Qiu, Zhenfei Yin, Dongdong Zhang, Cheng Qian, Zeping Li, Pony Ma, Guanhua Chen, Heng Ji, and Mengdi Wang. From word to world: Can large language models be implicit text-based world models?, 2025. URL https://arxiv.org/abs/2512.18832.

Silviu Pitis, Elliot Creager, A. Mandlekar, and Animesh Garg. Mocoda: Model-based counterfactual data augmentation. *ArXiv*, abs/2210.11287, 2022.

Stéphane Ross, Geoffrey J. Gordon, and J. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. pp. 627–635, 2010.

Ruoyao Wang, Peter Alexander Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? pp. 11279–11298, 2022.

Ruoyao Wang, Graham Todd, Ziang Xiao, Xingdi Yuan, Marc-Alexandre Côté, Peter Clark, and Peter Alexander Jansen. Can language models serve as text-based world simulators? pp. 1–17, 2024.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. Bread: Branched rollouts from expert anchors bridge sft rl for reasoning. *ArXiv*, abs/2506.17211, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *ArXiv*, abs/2403.13372, 2024.

A PER-SEED RESULTS

Table 3 provides the detailed per-seed breakdown of experimental results across all conditions.

Table 3: Per-seed breakdown of experimental results. W2R shows the number of successful episodes out of 195 test episodes. CR is computed as W2R/Real where Real=4.0% for all conditions.

Method	Seed	W2R (successes/195)	CR
Expert-Only	0	3 (1.54%)	0.385
Expert-Only	1	2 (1.03%)	0.256
Expert-Only	2	2 (1.03%)	0.256
Random Branch	0	2 (1.03%)	0.256
Random Branch	1	2 (1.03%)	0.256
Random Branch	2	2 (1.03%)	0.256
TCBA	0	2 (1.03%)	0.256
TCBA	1	**4 (2.05%)**	**0.513**
TCBA	2	3 (1.54%)	0.385