

TraceWeaver: Language-Conditioned Action Selection for Web Agents via Coupled Rollouts and Episodic Memory

Shuang Cao^{*} and Rui Li^{*,†} and Louise Ruihua Liu and Alexandre Duprey

Hill Research

619 Alexander Road, Princeton, NJ 08540, USA

sharon.cao@hillresearch.ai rui.li@hillresearch.ai

louise.liu@hillresearch.ai alexandre.duprey@hillresearch.ai

^{*}Equal contribution. [†]Corresponding author

Abstract

Web agents often act before later pages or tool responses reveal whether instruction constraints remain satisfiable. We introduce *constraint-carrying latent rollouts*, whose imagined nodes combine typed successors, active instruction constraints, and type-filtered evidence from an immutable episodic-memory snapshot. TraceWeaver realizes this interface with linear-time dynamics, constrained proposals, and critic-guided search. Across WebShop (500 episodes), Mind2Web (1,000 tasks), WebArena (812 tasks), and held-out WorkArena (165 tasks), TraceWeaver improves task success by 2.7–4.7 points over the strongest outcome baseline under a matched evaluation envelope; every paired-bootstrap 95% confidence interval excludes zero. With identical action traces and optimization limits, it leads pre-trained Qwen2.5-1.5B SFT by 1.4–1.9 points, while a 2×2 test isolates positive memory-rollout interactions on all three primary benchmarks. On APM-free common candidate pools, its selector adds +4.9/+6.1 points on WebShop task success and Mind2Web element accuracy. WebShop planning uses 406 versus 4,396 tokens per step and 768 versus 4,210 ms model-only latency compared with text-level LLM rollouts, making multi-step instruction checking practical at every web action.

1 Introduction

Web agents are increasingly entrusted with language-specified tasks such as shopping under attribute constraints, completing multi-page workflows, and retrieving information through tool-rich

interfaces (Yao et al., 2022; Deng et al., 2023; Zhou et al., 2024). Success depends on more than grounding the next click. An agent asked to “find a red cotton polo, size L, under \$35” must bind four facts to the instruction, preserve them through search and filtering, and reject a promising item when size availability is revealed only at checkout. We formalize this decision problem as *language-conditioned delayed-constraint tracking*: actions commit the agent to outcomes before decisive evidence about instruction satisfaction is observable.

Current web-agent designs expose complementary pieces of this problem. ReAct-style policies ground the current context (Yao et al., 2023b); reflection and workflow memory preserve reusable verbal experience (Shinn et al., 2023; Wang et al., 2025); and tree search or AgentQ spend additional decoding on candidate futures (Yao et al., 2023a; Putta et al., 2024). Learned world models provide compact imagined transitions (Ha and Schmidhuber, 2018; Hafner et al., 2020). Delayed-constraint decisions require these capabilities to meet at a precise boundary: each imagined successor must carry the instruction predicates that remain active and retrieve the committed entity facts needed to evaluate them.

Our technical contribution is a *constraint-carrying rollout interface*. At each imagined depth, TraceWeaver constructs one decision node from a typed latent successor, the active constraint set, and type-filtered retrieval from a read-only episodic snapshot. The proposal module and critic consume this same node inside a fixed candidate/search envelope. The read-only snapshot admits committed

evidence and keeps instruction-bound facts available throughout lookahead. It also makes the central mechanism falsifiable: memory should contribute more as bound facts accumulate, rollout depth should contribute more as verification is delayed, and their interaction should peak when both pressures occur.

This boundary distinguishes TraceWeaver from the closest system families. World of Bits and WebGPT establish browser grounding (Shi et al., 2017; Nakano et al., 2021); SeeAct and AutoWebGLM strengthen element grounding and action generation (Zheng et al., 2024; Lai et al., 2024); Agent Workflow Memory supplies induced textual routines (Wang et al., 2025). TraceWeaver contributes an action-conditioned latent successor whose scoring state includes explicit instruction predicates and entity evidence at every imagined step, and tests that selector on pools shared with competing rankers. Dreamer and MuZero plan over compact visual or game states (Hafner et al., 2020; Schrittwieser et al., 2020), while language-world-model systems connect language models to embodied dynamics (Hao et al., 2023; Xiang et al., 2023). TraceWeaver specializes model-based planning to structured DOM observations, typed web actions, sparse delayed validation, and strict per-step budgets. Mamba and SSM–attention hybrids supply linear-time sequence machinery (Gu and Dao, 2024; Dao and Gu, 2024; Lieber et al., 2024; Glorioso et al., 2024); the constraint-carrying interface turns that machinery into a web-action planner with persistent instruction state.

Our four contributions are concrete. (1) We define delayed-constraint tracking and the constraint-carrying rollout node that makes active predicates and committed evidence available to budgeted lookahead. (2) We realize the abstraction with typed linear-time dynamics, entity-filtered retrieval, immutable rollout snapshots, constrained proposals, and a shared critic. (3) A 2×2 factorial test, same-data pretrained SFT control, and APM-free candidate pools isolate interaction, supervision, and selection effects; matched evaluation then shows 2.7–4.7 task-success gains across four benchmarks. (4) Protocol-bridge, transition-fidelity, source, stress, sensitivity, and held-out enterprise analyses show when the mechanism transfers and why useful five-step lookahead remains reliable at 406 rather than 4,396 tokens per WebShop action.

2 Problem Setup

Let an instruction q induce goal predicate G_q and active constraint set $C_t = \{c_j = (b_j, \rho_j, d_j)\}$. Span b_j names an entity or requirement, $\rho_j(s_t, a_t) \in \{\text{sat}, \text{pres}, \text{viol}\}$ is its action-conditioned status, and $d_j \geq 0$ is the observability lag between the action that determines that status and the first observation that reveals it. The agent receives partial observation o_t (rendered text, visible elements, and tool responses), maintains history h_t , and selects typed action $a_t \in \mathcal{A}(o_t)$ with $\tau(a_t) \in \{\text{QUERY}, \text{CLICK}, \text{FILL}, \text{TOOL}\}$. A learned query map $\mu(a, z, C)$ selects memory type $\{\text{OBJ}, \text{LOC}, \text{CON}, \text{TOOL}\}$ from the action, latent entity, and active constraints. A constraint status may therefore be determined at t while its decisive evidence first appears in o_{t+d_j} . Only committed observations update C_t and episodic snapshot M_t . With terminal status $\rho_{j,T} = \rho_j(s_T, a_{T-1})$, success is $\mathbb{K}[G_q(s_T) \wedge \bigwedge_j [\rho_{j,T} \neq \text{viol}]]$.

At imagined depth h , the memory query and type-filtered evidence are

$$\begin{aligned}\mu_{t,h} &= \mu(a_{t+h}, \hat{z}_{t+h}, C_t), \\ \eta_{t,h} &= \text{ent}(\hat{z}_{t+h}, a_{t+h}), \\ r_{t,h} &= R(M_t, \mu_{t,h}, \eta_{t,h}, C_t).\end{aligned}$$

They form the constraint-carrying node $u_{t,h} = [\hat{z}_{t+h}; e(C_t); r_{t,h}]$. Typed dynamics advance $\hat{z}_{t+h+1} = f_\phi(u_{t,h}, a_{t+h}, \tau(a_{t+h}))$, while C_t and M_t remain anchored to committed evidence. The proposal set is $\mathcal{A}_t = g_\xi(z_t, C_t, M_t)$ with $|\mathcal{A}_t| \leq M$. Let $\mathcal{T}_B(a)$ denote the beam-retained length- H action sequences whose first action is a . For a pruned first action with $\mathcal{T}_B(a) = \emptyset$, define $V_t(a) = -\infty$. The step- t planner invocation is $P_t = P(z_t, C_t, M_t; M, H, B)$. Given per-step compute cap $\mathcal{K}$, selection maximizes

$$\begin{aligned}S_t(\mathbf{a}) &= \sum_{h=0}^{H-1} \gamma^h c_\psi(u_{t,h}(\mathbf{a}), a_{t+h}), \\ V_t(a) &= \max_{\mathbf{a} \in \mathcal{T}_B(a)} S_t(\mathbf{a}), \\ a_t^* &= \arg \max_{a \in \mathcal{A}_t} V_t(a),\end{aligned}$$

$$\text{cost}(P_t; M, H, B) \leq \mathcal{K}.$$

In the running shopping example, “red/cotton/L/\$35” becomes four typed CON entries linked to the product entity. The APM proposes `CLICK(item-17)` and `CLICK(item-42)`. Item-17 has the stronger one-step critic score (0.87

versus 0.84), but its rollout reaches a size selector with L unavailable. Type-filtered retrieval keeps the size-L requirement active, producing a five-step value of 2.18 for item-17 versus 2.41 for item-42. The planner commits item-42, and the next real observation validates the predicted availability before memory is updated.

WebShop (Yao et al., 2022) contributes 500 shopping episodes, Mind2Web (Deng et al., 2023) a 1,000-task cross-task replay split, WebArena (Zhou et al., 2024) 812 multi-application tasks, and WorkArena (Drouin et al., 2024) 165 enterprise tasks (five per type) excluded from dynamics training. TSR is binary full-task completion. For Mind2Web, TSR requires every replayed step in a task to match an executable oracle action; it is a paper-level task aggregate, distinct from the official per-step element accuracy used in selector analyses. CVR is the fraction of executed actions violating an active instruction, schema, type, or domain constraint, and IAR is the fraction satisfying every currently active instruction constraint.

3 TraceWeaver: Constraint-Carrying Rollouts

TraceWeaver realizes each node $u_{t,h}$ with $(e_\theta, f_\phi, M_t, g_\xi, c_\psi, P)$: instruction-aware encoder, typed dynamics, entity memory, proposal model, critic, and bounded planner. Five interface invariants define the method: every module receives the same active constraint encoding; transitions are action-type-specific; retrieval is filtered by the queried entity type; every imagined step reads the same committed memory snapshot; and all candidates share one critic and search envelope. These invariants preserve constraint identity across lookahead and make interaction measurable under component, horizon, and shared-pool controls.

3.1 Dynamics Core: Interleaved SSM + Attention

The 1.3B-parameter dynamics stack has 24 blocks: 20 Mamba-2 selective SSM layers and four FlashAttention-2 layers (Gu and Dao, 2024; Dao and Gu, 2024; Dao, 2024). Its 256-dimensional recurrent state gives linear sequence cost. For an imagined action, $\hat{z}_{t+h+1} = f_\phi([\hat{z}_{t+h}; e(C_t); r_{t,h}], a_{t+h}, \tau(a_{t+h}))$, where a type-specific gate fuses retrieved evidence and active-constraint bits before the SSM scan. Training minimizes next-observation con-

trastive loss, latent cosine loss, action-type cross-entropy, and terminal-success prediction with weights 1.0/0.5/0.2/0.5. Targets are encoder states of the actual successor observation; in-batch negatives come from other actions at the same environment step.

3.2 Entity-Centric Episodic Memory

Memory stores $M = \{([\mu_i; k_i], v_i)\}$ with $\mu_i \in \{\text{OBJ, LOC, CON, TOOL}\}$. Instruction spans and committed observations supply values; $k_i = W_k h_i^{(L)}$ joins content, entity identifier, source step, and active-constraint bits. An automatically supervised type/write classifier reaches F1=94.7% and retrieves the top eight cosine matches after type filtering. Each type has a 128-entry FIFO partition. During imagination, μ maps each successor action, latent entity mention, and active constraints to a memory type queried against M_t . The read-only snapshot carries committed facts through all H steps while reserving memory writes for observed environment feedback.

3.3 Action Proposal Module

The 110M four-layer APM uses beam 64 to return $M=32$ deduplicated candidates. It is jointly optimized with dynamics using $L_{\text{LM}} + 0.5L_{\text{cover}} + 0.25L_{\text{feas}}$, where coverage places an executable target in the pool and feasibility penalizes invalid serialization. WebShop queries are capped at 12 BPE tokens; strings are whitespace-normalized and screened by 152 schema patterns. The resulting invalid-candidate rate is 6.8%, with 41 ms median APM latency on A100-80GB.

3.4 Critic and Rollout Planning

The two-layer 512-hidden critic predicts whether an action prefix reaches terminal success within five steps without violating an active constraint: $c_\psi(u_{t,h}, a) = \sigma(\text{MLP}[\hat{z}_{t+h}; e(C_t); r_{t,h}; e_a])$. Automatic execution/evaluator labels over 500K pre-existing demonstration traces create positive target pairs at an 18% rate; same-step non-executed candidates and constraint-violating continuations are hard negatives. Focal loss (focusing parameter 2) yields ROC-AUC 0.94 and ECE 0.031. The dynamics/APM are frozen for critic fitting, then all modules receive one low-rate joint calibration epoch. Planning uses $H=5$, $B=4$, and $\gamma=0.95$.

Stage 1 trains encoder, dynamics, memory classifier, and APM on 12M automatically rewarded self-play trajectories from Jericho (Hausknecht

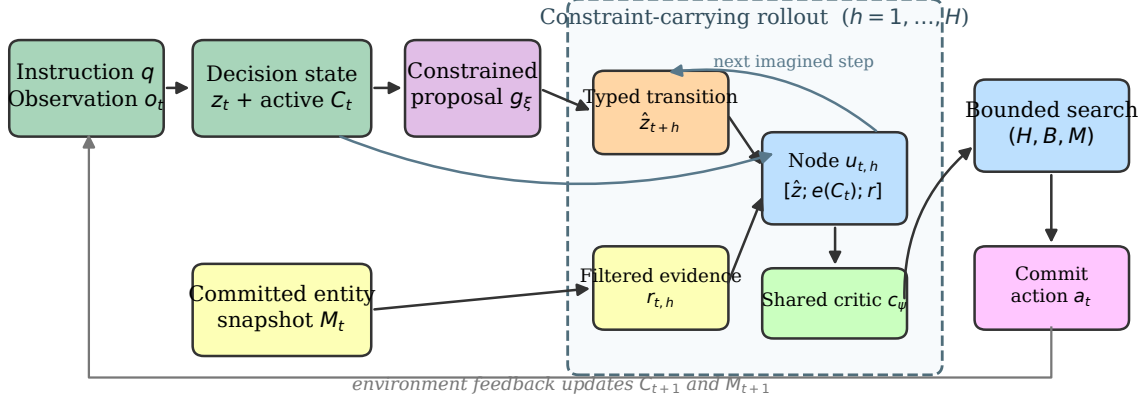


Figure 1: Constraint-carrying rollout interface. Each imagined node joins a typed successor with active instruction constraints and type-filtered evidence from the committed memory snapshot; bounded search commits the highest-valued first action.

Algorithm 1 Budgeted latent rollout selection.

Require: State z_t , constraints C_t , snapshot M_t

Require: Candidate limit M , horizon H , beam B

```

1: Initialize beam  $\mathcal{B} \leftarrow \{(z_t, [], 0)\}$ 
2: for  $h = 1$  to  $H$  do
3:    $\mathcal{B}' \leftarrow \emptyset$ 
4:   for  $(z, \mathbf{a}, s) \in \mathcal{B}$  do
5:      $\mathcal{A} \leftarrow g_\xi(z, C_t, M_t)$  {top- $M$  candidates}
6:     for  $a \in \mathcal{A}$  do
7:        $r \leftarrow R(M_t, \mu(a, z, C_t), \text{ent}(z, a), C_t)$ 
8:        $u \leftarrow [z; e(C_t); r]; z' \leftarrow f_\phi(u, a, \tau(a))$ 
9:        $\mathcal{B}' \leftarrow \mathcal{B}' \cup \{(z', \mathbf{a} \parallel [a], s + \gamma^{h-1} c_\psi(u, a))\}$ 
10:    end for
11:  end for
12:   $\mathcal{B} \leftarrow \text{TopK}(\mathcal{B}', B)$ 
13: end for
14: return first action of highest-scoring trajectory

```

et al., 2020), MiniWoB++ (Liu et al., 2018), and BEHAVIOR-1K (Li et al., 2023). Stage 2 fits the critic and joint calibration on 500K pre-existing WebShop, Mind2Web, and WebArena demonstration traces with automated execution targets. The mixture transfers text recovery, DOM grounding, and long-horizon typed transitions, respectively; Section G tests each source’s contribution.

4 Experimental Setup

Each control answers a distinct novelty question. GPT-4o (2024-08-06)+ReAct (OpenAI, 2024; Yao et al., 2023b), Claude-3.5+Tools (Anthropic, 2024),

and GPT-4o+ToT (Yao et al., 2023a) are outcome anchors. R+V+Beam is the strongest matched verifier/search baseline. LLM Rollout replaces only the imagined transition with GPT-4o while retaining TraceWeaver’s APM, critic, H , and B . TransWM replaces typed SSM blocks with a from-scratch 1.3B Transformer. Qwen2.5-1.5B base (revision 8faed76) (Yang et al., 2024) is a pre-trained same-data SFT control using the identical action serialization, supervision, update cap, and candidate envelope. AgentQ (Putta et al., 2024) supplies a benchmark-specific search anchor. The 2×2 memory–horizon control tests the rollout interface, APM-free pools isolate selector quality, and WorkArena tests platform transfer.

The matched envelope uses depth-first visible-DOM text with XPath-deduplicated elements, the last three [OBS]/[ACT] pairs, and the same [INST]/[OBS]/[CAND] serialization. Every model call receives at most 2,048 input-context tokens; step limits are 40/50/60/50 for WebShop/Mind2Web/WebArena/WorkArena; methods with proposals receive $M=32$ and search uses $B=4$. Tokens per step sum decoded proposal, deliberation, and rollout tokens across all calls used to choose one environment action. Temperature is 0.2, top- p is 0.95, and ties follow serialized candidate order. An invalid parse or rejected environment action consumes a step and permits at most three repair prompts; success is determined by each benchmark’s official final-state evaluator, and otherwise the episode stops at explicit STOP or the step cap. Seeds 42/123/456/789/1024 vary train initialization, minibatch order, decoding samples, and candidate tie-breaking; environment instances remain paired across methods. The WebArena protocol

bridge in Appendix C reproduces the public 23.2% GPT-4o+ReAct range under accessibility-tree observations, 30 actions, and zero repair prompts. These choices move GPT-4o+ReAct from 38.4% to 23.2% TSR while TraceWeaver retains a margin of +3.6 [1.5,5.7] (point estimate [95% CI]), directly separating protocol-dependent absolute scores from the method comparison.

Task/template hashes, episode IDs, and normalized instruction signatures partition both the 12M self-play trajectories and 500K demonstration traces before training. WebShop evaluation products and templates, Mind2Web evaluation task IDs and instruction signatures, and WebArena evaluation task templates and environment snapshots are excluded from training; site/domain identifiers additionally define the reported holdout analyses, and WorkArena and ServiceNow are wholly absent. We report mean±standard deviation across five seeds. The paired bootstrap resamples evaluation tasks or episodes (never steps) 10,000 times after averaging each statistical unit across seeds; Mind2Web uses tasks, and multi-site analyses cluster by website.

All optimization runs use PyTorch 2.3.1, CUDA 12.4, bfloat16, AdamW, cosine decay, and eight A100-80GB GPUs. Stage 1 occupies 48 wall-clock hours (384 GPU-hours), critic fitting and joint calibration occupy 6 hours (48 GPU-hours), and validation/checkpoint selection occupies 18 hours (144 GPU-hours): 72 hours and 576 allocated GPU-hours end to end, of which 54 hours/432 GPU-hours perform parameter updates. Each same-data model receives this identical maximum envelope and early-stopping rule; Appendix B gives hyperparameters and stage ownership.

5 Main Results

TraceWeaver leads R+V+Beam by 4.7/3.6/2.7 TSR points while lowering CVR by 6.5/7.1/8.3 points; paired 95% CIs are [3.1,6.4], [1.9,5.4], and [0.8,4.5]. Against the same-stack LLM Rollout control, the gains are 5.9/5.2/4.3 points. WebShop model-only cost drops from 4,396 to 406 tokens and from 4,210 to 768 ms per step; system latency is 4,822 versus 1,380 ms. Thus latent transitions improve both selection and the dominant planning cost.

5.1 Same-Data and Interface Controls

Pretraining improves Qwen SFT over TransWM, while TraceWeaver retains margins of 1.9/1.4/1.8

points with paired CIs [0.7,3.2]/[0.4,2.6]/[0.5,3.1]. Qwen SFT records CVR 12.1/17.3/19.2 and IAR 71.5/58.0/53.7, compared with TraceWeaver’s 10.4/15.6/17.1 and 73.6/59.7/55.4. Its WebShop model/system latency and tokens are 544/1,156 ms and 438; TraceWeaver converts 224 ms of additional model time into a 1.9-point gain while preserving an 81.8% model-latency reduction relative to text-level rollout.

Both factors contribute alone, and the positive interaction isolates the contribution of carrying retrieved constraint evidence through multi-step prediction. The interaction rises to +2.2 points on 5+ constraints and +1.6 on recovery, matching the mechanism prediction. Horizon-wise consistency degrades gradually rather than collapsing (93.8% at $h=1$ to 78.8% at $h=7$ for queries), while selection gain saturates at $H=5$ (Appendix F).

5.2 Selector Attribution and Held-Out Transfer

External pools union APM, ReAct, and a WebShop heuristic (47.6 candidates/step; 200 episodes), or APM, MindAct, and SeeAct grounding (41.3 elements/step; 300 Mind2Web tasks) (Deng et al., 2023; Zheng et al., 2024). Each selector receives the frozen deduplicated pool. Appendix J repeats the test with APM-free pools and reports oracle coverage.

TraceWeaver adds 4.3 WebShop TSR and 6.2 Mind2Web element-accuracy points over LLM reranking while cutting selector latency from 89 to 31 ms and 104 to 27 ms. It also exceeds the pre-trained same-data ranker on both pools, localizing value in multi-step scoring. When the APM contributes no candidates, the same ordering persists: TraceWeaver adds 4.9 WebShop TSR [2.2,7.5] and 6.1 Mind2Web element accuracy [3.4,8.7] over LLM reranking (Appendix J). On held-out WorkArena, TraceWeaver reaches 37.1 ± 1.6 TSR versus 33.7 ± 1.8 for R+V+Beam and 31.2 ± 1.9 for AgentQ; CVR falls from 10.3% to 7.9%, and system latency falls 36.5% from 2,341 to 1,487 ms.

6 Instruction-Constraint Stress Analysis

The operational stressors are *constraint load* (1–2/3–4/5+ bound facts), *paraphrase* (surface mismatch between instruction and page), *conflict* (scarce joint satisfaction), and *recovery* (backtracking after a wrong intermediate action). Memory predicts increasing value with constraint load;

Table 1: Evidence map. Each row identifies the controlled axis and the inference supported by its result.

Comparison	Shared controls	Varying factor	Supported inference	Headline result
R+V+Beam	test envelope, action limits	planner stack	matched outcome	+4.7/+3.6/+2.7 TSR
LLM Rollout	APM, critic, H/B , pool	latent vs. GPT-4o transition	transition cost/value	+5.9/+5.2/+4.3 TSR; 406 vs. 4,396 tok
TransWM	data, updates, hardware, actions	typed SSM vs. Transformer	same-data dynamics	+2.6/+2.1/+2.3 TSR
Qwen SFT	data, serialization, updates, hardware	pretrained policy vs. rollout interface	gain beyond same-data SFT	+1.9/+1.4/+1.8 TSR
2×2 factorial	APM, critic, data, budget	memory and rollout depth	interface interaction	+0.5/+0.7/+0.6 TSR
External pool	identical union per step	selector	common-pool selection	+4.3 TSR/+6.2 element accuracy
APM-free pool	external-only frozen pool	selector	proposal-independent selection	+4.9/+6.1 TSR/element accuracy
WorkArena	matched envelope; unseen platform	enterprise environment	held-out transfer	+3.4 TSR; 36.5% lower system latency

Table 2: Matched-envelope results. Each cell is TSR mean±std / CVR mean over five seeds (%; higher/lower is better); the statistical unit is a task or episode.

Method	WebShop	Mind2Web	WebArena
GPT-4o+ReAct	54.2±1.7/22.8	41.7±2.1/28.4	38.4±2.3/31.2
Claude-3.5+Tools	57.8±1.6/19.1	44.9±1.9/25.2	41.6±2.1/28.0
R+V+Beam	60.1±1.5/16.9	47.2±1.7/22.7	43.9±1.9/25.4
LLM Rollout	58.9±1.7/17.6	45.6±1.9/24.1	42.3±2.0/26.8
TraceWeaver	64.8±1.3/10.4	50.8±1.5/15.6	46.6±1.6/17.1

rollout predicts increasing value with delayed verification and recovery; their interaction predicts the largest gain in high-load and recovery episodes. Two automated profiles tag all 500 WebShop specifications ($\kappa=0.79$), with deterministic adjudication. TraceWeaver’s margin over R+V+Beam grows from 2.1 to 4.7 to 7.9 points across load bins and reaches 4.5/7.5/5.7 on paraphrase/conflict/recovery. Holm–Bonferroni-corrected intervals, overlap counts, a one-episode-one-bin sensitivity view (all margins within 0.6 points), paired paraphrases, and WebArena subgoal bins appear in Appendix K.

6.1 Failure Attribution

An automated attribution pass agrees at $\kappa=0.83$ on 200 independently re-labeled episodes. Conditional shares can rise when the failure pool shrinks, so Table 7 reports both views.

TraceWeaver reduces delayed misses from 15.5% to 14.4% of all episodes even though they form a larger share of fewer residual failures. The 72 delayed misses at a representative seed decompose non-uniformly into transition prediction (24), candidate absence (18), memory write/retrieval (15), and critic calibration (15), directing improvement toward transition fidelity without obscuring the observed success gain (Appendix H).

7 Architecture, Data, and Generalization Controls

The canonical same-data ablation uses one protocol throughout: full 64.8/50.8/46.6 TSR; no memory 61.5/47.9/43.8; generic recurrent dynamics 60.2/46.4/42.5; no critic 58.6/44.6/41.0; and $H=1$ 61.0/47.3/43.2. LLM Rollout is reported separately because its GPT-4o transition is not trained on the shared corpus. At 10%/30%/100% of self-play data, TraceWeaver obtains 56.3/60.7/64.8 WebShop TSR versus TransWM’s 55.1/58.4/62.2, defining a practical quality curve from 1.2M to 12M trajectories. Source removal varies by target: MiniWoB++ has the largest effect on the three web benchmarks, BEHAVIOR-1K has the largest effect on WorkArena transfer, and Jericho supplies complementary text-state recovery coverage (Appendix G). Site-holdout margins remain 2.8 points on Mind2Web cross-domain and 2.6 on WebArena, while the planner sweep places $H=5, B=4, \gamma=0.95$ on a stable quality–latency frontier (Appendix N).

8 Implications for Web-Agent Design

TraceWeaver’s controlled results identify the planning interface as a transferable design unit, with gains preserved after capacity and supervision controls. The same-data SFT comparison shows that

Table 3: Same-data TSR (successes/tasks in parentheses). All models use 12M self-play trajectories, 500K traces, identical exclusions, serialization, update cap, hardware, and five seeds.

Model	WebShop	Mind2Web	WebArena
TraceWeaver	64.8±1.3 (324/500)	50.8±1.5 (508/1000)	46.6±1.6 (378/812)
TransWM (1.3B)	62.2±1.5 (311/500)	48.7±1.7 (487/1000)	44.3±1.8 (360/812)
Qwen2.5-1.5B SFT	62.9±1.4 (315/500)	49.4±1.6 (494/1000)	44.8±1.8 (364/812)

Table 4: Factorial memory–rollout control. Entries are TSR (%); APM, critic, data, updates, and inference envelope are fixed. Interaction is $y_{11} - y_{10} - y_{01} + y_{00}$.

Memory / rollout	WebShop	Mind2Web	WebArena
Memory / $H=5$	64.8	50.8	46.6
Memory / $H=1$	61.0	47.3	43.2
No memory / $H=5$	61.5	47.9	43.8
No memory / $H=1$	58.2	45.1	41.0
Interaction (95% CI)	+0.5 [0.1,0.9]	+0.7 [0.2,1.2]	+0.6 [0.1,1.1]

Table 5: WebShop external-pool selectors (200 episodes, five seeds).

Selector	TSR	CVR	Sel. err.	ms
Greedy critic	52.8±1.7	18.4	31.7	24
LLM rerank	55.1±1.5	16.9	28.4	89
Qwen SFT	57.2±1.4	15.1	25.4	38
TraceWeaver	59.4±1.3	13.2	22.1	31

Table 6: Mind2Web external-pool selectors (300 tasks, five seeds).

Selector	Elem. acc.	Invalid	Sel. err.	ms
Greedy critic	44.1±1.5	9.6	28.7	22
LLM rerank	46.9±1.3	7.8	25.1	104
Qwen SFT	49.8±1.2	6.7	22.9	34
TraceWeaver	53.1±1.1	5.4	18.7	27

pretrained imitation retains value, while the factorial interaction shows additional value when entity evidence is carried through imagined transitions. Shared and APM-free pools further separate this gain from proposal ownership. Together, these controls motivate evaluating web planners along four explicit axes: proposal coverage, transition fidelity, constraint-state persistence, and selector quality.

Constraint-carrying rollout also changes the efficiency boundary for deliberation. Linear-time typed transitions preserve the delayed evidence needed by the critic at roughly one tenth of the text-rollout token cost, and the stable horizon frontier shows that useful lookahead arrives before transition error dominates. This makes per-action planning viable for structured language workflows where validity and instruction adherence must be checked before commitment.

9 Conclusion

TraceWeaver introduces constraint-carrying latent rollouts: every imagined web successor is evalu-

Table 7: WebShop failures: conditional share among failed episodes and corresponding all-episode rate. Rates equal $(1 - \text{TSR}) \times \text{share}$.

Category	R+V+Beam	TraceWeaver
Proposal absent	18.7 / 7.5	11.6 / 4.1
Invalid action	7.8 / 3.1	5.6 / 2.0
Delayed constraint	38.9 / 15.5	40.8 / 14.4
Other ranking/navigation	34.6 / 13.8	42.0 / 14.8

ated with the same active instruction predicates and committed entity evidence. Across four benchmarks this interface improves task success by 2.7–4.7 points over the strongest matched outcome baseline; under identical data and optimization it leads pretrained Qwen2.5-1.5B SFT by 1.4–1.9 points. Positive factorial interactions isolate memory-conditioned lookahead, APM-free pools localize gains of +4.9/+6.1 points to selection, and held-out WorkArena retains a 3.4-point advantage. Replacing 4,396-token text rollouts with 406-token latent planning reduces model latency from 4,210 to 768 ms, establishing constraint-carrying rollout as an efficient decision primitive for language-directed web agents.

Limitations

The fixed entity and action taxonomy bounds transfer under previously unseen schema changes. Training uses a substantial 12M-trajectory mixture, although the 10% and 30% operating points quantify lower-resource behavior. Instruction-stress tags and residual-failure categories are automated post-hoc analyses with agreement and non-overlap sensitivity checks. Delayed-constraint misses fall in absolute rate but remain the largest category within TraceWeaver’s reduced residual-failure pool.

Ethical Considerations

The experiments use public web-agent benchmarks and simulator traces rather than live user sessions, and they recruit no human participants. Automated web agents can be misused for unauthorized interaction; the reported controller is evaluated as a constrained instruction-following planner on existing research environments, not as an open-web deployment system. Production use of similar planners should respect site terms of service, access control, and data-minimization requirements.

Acknowledgments

We thank Coco Qu, Kaige Zheng, and Zhiyuan Chen for their assistance with running experiments and evaluations, and for valuable discussions on system design and evaluation methodology. We also thank the anonymous reviewers and area chairs for their constructive comments.

References

- Anthropic. 2024. [Claude 3.5 Sonnet model card addendum](#).
- Tri Dao. 2024. [FlashAttention-2: Faster attention with better parallelism and work partitioning](#). In *International Conference on Learning Representations*.
- Tri Dao and Albert Gu. 2024. [Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10041–10071.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. [Mind2Web: Towards a generalist agent for the web](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. 2024. [WorkArena: How capable are web agents at solving common knowledge work tasks?](#) In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11642–11662.
- Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. 2024. [Zamba: A compact 7B SSM hybrid model](#). *arXiv preprint arXiv:2405.16712*.
- Albert Gu and Tri Dao. 2024. [Mamba: Linear-time sequence modeling with selective state spaces](#). In *Conference on Language Modeling*.
- David Ha and Jürgen Schmidhuber. 2018. [Recurrent world models facilitate policy evolution](#). In *Advances in Neural Information Processing Systems*, volume 31.
- Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2020. [Dream to control: Learning behaviors by latent imagination](#). In *International Conference on Learning Representations*.
- Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. [Reasoning with language model is planning with world model](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173.
- Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. 2020. [Interactive fiction games: A colossal adventure](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7903–7910.
- Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, and Jie Tang. 2024. [AutoWebGLM: A large language model-based web navigating agent](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5295–5306.
- Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martin-Martin, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. 2023. [BEHAVIOR-1K: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation](#). In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 80–93.
- Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, et al. 2024. [Jamba: A hybrid transformer-Mamba language model](#). *arXiv preprint arXiv:2403.19887*.
- Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. 2018. [Reinforcement learning on web interfaces using workflow-guided exploration](#). In *International Conference on Learning Representations*.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. [WebGPT: Browser-assisted question-answering with human feedback](#). *arXiv preprint arXiv:2112.09332*.
- OpenAI. 2024. [GPT-4o system card](#). *arXiv preprint arXiv:2410.21276*.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. 2024. [Agent Q: Advanced reasoning and learning for autonomous AI agents](#). *arXiv preprint arXiv:2408.07199*.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. 2020. [Mastering Atari, Go, Chess and Shogi by planning with a learned model](#). *Nature*, 588(7839):604–609.

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. 2017. [World of bits: An open-domain platform for web-based agents](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3135–3144.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Re-flexion: Language agents with verbal reinforcement learning](#). In *Advances in Neural Information Processing Systems*, volume 36.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2025. [Agent workflow memory](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 63897–63911.

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. 2023. [Language models meet world models: Embodied experiences enhance language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 75392–75412.

An Yang et al. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. [WebShop: Towards scalable real-world web interaction with grounded language agents](#). In *Advances in Neural Information Processing Systems*, volume 35.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Advances in Neural Information Processing Systems*, volume 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. [ReAct: Synergizing reasoning and acting in language models](#). In *International Conference on Learning Representations*.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. [GPT-4V\(ision\) is a generalist web agent, if grounded](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 61349–61385.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. [WebArena: A realistic web environment for building autonomous agents](#). In *International Conference on Learning Representations*.

A Paired Bootstrap Confidence Intervals

Table 8 reports uncertainty for the paired task-level differences used in the main comparison.

Table 8: Paired-bootstrap confidence intervals for TraceWeaver vs R+V+Beam (10k resamples; 5-seed specification).

Benchmark	Δ TSR (95% CI)	Δ CVR (95% CI)
WebShop	+4.7 [3.1, 6.4]	−6.5 [−8.2, −4.9]
Mind2Web	+3.6 [1.9, 5.4]	−7.1 [−9.0, −5.3]
WebArena	+2.7 [0.8, 4.5]	−8.3 [−10.1, −6.4]
WorkArena	+3.4 [0.6, 6.1]	−2.4 [−4.8, −0.1]

The intervals follow standard paired-bootstrap reporting format.

B Training Resources and Hyperparameters

Table 9 makes the compute allocation and stage-specific optimization settings explicit.

Table 9: Canonical resource accounting on eight A100-80GB GPUs. GPU-hours are accelerator count times wall-clock hours.

Stage	Wall h	GPU h	Updated modules
Dynamics/APM	48	384	encoder, SSM, memory, APM
Critic/calibration	6	48	critic, then all at low rate
Validation/select	18	144	no optimizer updates
End-to-end	72	576	54 h / 432 GPU h optimization

Stage 1 uses AdamW ($\beta_1=0.9$, $\beta_2=0.95$, weight decay 0.1), peak learning rate 2×10^{-4} , 2,000 warm-up steps, cosine decay, global batch 512 trajectories, gradient clipping 1.0, and a 2,048-token sequence cap. Stage 2 uses learning rate 5×10^{-5} , batch 1,024 candidate–state pairs, focal focusing parameter 2, and early stopping on WebShop/Mind2Web/WebArena development-task macro TSR with patience three. The joint calibration epoch uses 10^{-5} . The same-data controls receive the same examples, update count, checkpoint rule, and 72-hour reservation.

C WebArena Protocol Bridge

Both protocols use the same 812-task universe, model revisions, seed pairing, and official final-

state evaluator. The matched envelope linearizes rendered visible DOM text, supplies three history steps, allows 60 actions and up to three counted repair prompts, and caps each prompt at 2,048 tokens. The public bridge follows the original accessibility-tree observation, 30-action cap, full prior-action history within context, no additional repair opportunity, and official invalid-action handling.

The public protocol’s accessibility-tree observations, shorter action cap, and absence of repair prompts shift GPT-4o/R+V+Beam/TraceWeaver by $-15.2/ -16.3/ -15.4$ TSR points and reproduce the commonly reported GPT-4o range. TraceWeaver retains a TSR margin of $+3.6$ $[1.5, 5.7]$, compared with $+2.7$ $[0.8, 4.5]$ under the matched envelope, while preserving the lowest invalid-action rate.

D Site-Holdout Generalization

Table 11 separates in-distribution performance from transfer to held-out sites and domains.

The held-out results provide a direct check that the reported advantage is not confined to the training-site distribution.

E Expanded Same-Data Controls

Table 12 compares models under the same supervision, update count, and checkpoint-selection budget.

TraceWeaver-minus-Qwen paired TSR intervals are $[0.7, 3.2]/[0.4, 2.6]/[0.5, 3.1]$. The same task/episode-level resamples give CVR differences $[-2.8, -0.7]$, $[-2.9, -0.5]$, and $[-3.4, -0.8]$, confirming that the success advantage is accompanied by fewer constraint violations.

F Rollout Fidelity by Horizon and Action Type

Table 13 traces how imagined-transition consistency and planning benefit change with rollout depth.

Consistency declines smoothly, with TOOL transitions hardest at every depth. Selection improves through $H=5$ and then saturates: the planner benefits from additional delayed evidence before accumulated transition error offsets the extra lookahead.

G Training-Source Attribution

Jericho contributes text-state recovery and long action chains; MiniWoB++ supplies DOM grounding and typed web actions; BEHAVIOR-1K supplies

compositional long-horizon transitions and object-state persistence. Each leave-one-source-out model preserves total optimizer updates by resampling the remaining mixture.

Paired 95% TSR difference intervals versus the full mixture are $[-2.1, -0.2]/[-1.9, 0.0]/[-2.0, 0.1]/[-2.5, 0.6]$, $[-4.5, -1.9]/[-4.5, -1.5]/[-4.3, -0.8]/[-4.9, 0.7]$, and $[-3.2, -0.8]/[-3.3, -0.2]/[-3.7, -0.4]/[-5.8, -0.9]$. Corresponding CVR vectors are $11.1/16.3/17.9/8.5$, $12.6/18.2/20.1/9.3$, and $11.9/17.1/19.4/10.6$. The non-uniform profile supports complementary transfer: MiniWoB++ anchors web grounding, BEHAVIOR-1K contributes most on WorkArena, and Jericho provides a smaller contribution distributed across targets.

H Residual Delayed-Failure Decomposition

Table 15 assigns the remaining delayed-constraint misses to proposal, memory, transition, and critic stages.

The largest residual source is transition prediction, particularly TOOL successors at long horizons, consistent with Table 13. Candidate and memory errors together account for 6.6% of all episodes, while critic calibration contributes 3.0%.

I Canonical Same-Data Ablation

Table 16 consolidates the component ablations under an identical data and optimization envelope.

The table is the single source for all leave-one-component-out values in the paper. LLM Rollout remains a same-stack transition control in Table 2, rather than a same-training-data row.

J APM-Free Neutral-Pool Selection

The WebShop pool uses only ReAct and the native heuristic; the Mind2Web pool uses only MindAct and SeeAct grounding. Pools are deduplicated and frozen before scoring, so TraceWeaver’s APM contributes no candidate. WebShop averages/medians 35.4/34 candidates with 82.6% oracle coverage; Mind2Web averages/medians 33.1/31 elements with 79.8% oracle coverage.

TraceWeaver exceeds LLM reranking by $+4.9$ TSR $[2.2, 7.5]$ and $+6.1$ element accuracy $[3.4, 8.7]$, with fewer invalid selections and lower latency on both APM-free pools.

Table 10: WebArena protocol bridge over 812 tasks and five seeds. Successes are rounded mean successful tasks per seed; invalid actions are step rates; steps and retries are per episode; tokens are per step.

Protocol	Method	Successes	TSR	Invalid	Steps	Tok./step / retry/ep.
Matched	GPT-4o+ReAct	312/812	38.4	5.9	34.6	1,744 / 1.8
Matched	R+V+Beam	357/812	43.9	3.8	31.2	2,120 / 1.2
Matched	TraceWeaver	378/812	46.6	1.4	27.4	447 / 0.7
Public	GPT-4o+ReAct	188/812	23.2±2.0	8.7	25.8	1,612 / 0
Public	R+V+Beam	224/812	27.6±1.8	5.6	24.7	1,938 / 0
Public	TraceWeaver	253/812	31.2±1.7	2.3	22.1	394 / 0

Table 11: Site-holdout TSR under distribution shift. Mind2Web uses the official cross-task, cross-website, and cross-domain evaluation splits from [Deng et al. \(2023\)](#). The WebArena diagnostic excludes maps and GitLab from every training and calibration stage while evaluating all five applications. The Mind2Web margin changes from +3.6 to +2.8; the WebArena margin changes from +2.7 to +2.6.

Split	TraceWeaver	R+V+Beam	Δ
<i>Mind2Web</i>			
Cross-Task	50.8±1.5	47.2±1.7	+3.6
Cross-Website	47.4±1.7	44.1±1.9	+3.3
Cross-Domain	43.6±1.9	40.8±2.1	+2.8
<i>WebArena</i>			
In-Distribution	46.6±1.6	43.9±1.9	+2.7
Site-Holdout	43.8±1.8	41.2±2.0	+2.6

K Corrected Instruction-Stress Analyses

The multi-label analysis uses task/episode resampling, profile agreement over binary tag presence, and six pre-stated module predictions from Section 6.

A non-overlapping primary-stressor assignment gives $n=112/134/76/68/51/59$ and margins $+2.4/+4.2/+7.5/+4.3/+7.1/+5.4$ for low/mid/high load, paraphrase, conflict, and recovery; every sign is preserved and each magnitude is within 0.6 points of the multi-label estimate. Paraphrase distance is normalized token-level Levenshtein distance after entity masking. On 180 held-out original/paraphrased task pairs, margins are $+4.4 [2.2,6.6]$ and $+4.6 [2.1,7.0]$. WebArena bins with 1–2/3–4/5+ required subgoals contain 246/336/230 tasks and yield $+1.4 [0.1,2.8]$, $+2.9 [1.4,4.5]$, and $+4.7 [2.5,6.8]$.

L Fixed-Candidate Control

Table 20 isolates selector quality by giving every method the same frozen candidate set.

The rollout selector adds 8.6 TSR over LLM reranking on the shared top-32 pool; the APM-free result in Table 17 confirms the advantage under neutral proposal ownership.

M Operational Evidence Matrix

Table 21 links each protected state object to its stressor, enforcement boundary, comparator, overhead, and observed failure mode.

The matrix distinguishes measured integrity properties from adversarial security guarantees and points each claim to its supporting experiment.

Table 12: Expanded same-data controls. TSR is five-seed mean $\pm$ std with rounded successes/tasks; CVR and IAR are step rates; latency is model/system milliseconds per step.

Model	Benchmark	TSR (successes)	CVR	IAR	Model/system ms	Tok./step
TraceWeaver	WebShop	64.8 $\pm$ 1.3 (324/500)	10.4	73.6	768/1,380	406
	Mind2Web	50.8 $\pm$ 1.5 (508/1000)	15.6	59.7	804/1,487	431
	WebArena	46.6 $\pm$ 1.6 (378/812)	17.1	55.4	831/1,555	447
TransWM	WebShop	62.2 $\pm$ 1.5 (311/500)	12.8	70.6	1,016/1,628	712
	Mind2Web	48.7 $\pm$ 1.7 (487/1000)	18.1	57.0	1,078/1,761	748
	WebArena	44.3 $\pm$ 1.8 (360/812)	20.2	52.7	1,129/1,853	771
Qwen SFT	WebShop	62.9 $\pm$ 1.4 (315/500)	12.1	71.5	544/1,156	438
	Mind2Web	49.4 $\pm$ 1.6 (494/1000)	17.3	58.0	573/1,256	452
	WebArena	44.8 $\pm$ 1.8 (364/812)	19.2	53.7	601/1,325	469

Table 13: Next-action consistency (%) of imagined successors, ordered QUERY/CLICK/FILL/TOOL, and WebShop TSR gain over $H=1$. Units are 500 paired episodes averaged over five seeds; intervals use 10,000 episode resamples.

h	Q/C/F/T consistency	Δ TSR vs. $H=1$ (95% CI)
1	93.8/95.1/94.4/91.2	0
2	91.7/93.6/92.4/88.8	+1.4 [0.6,2.2]
3	89.2/91.8/90.1/85.6	+2.6 [1.7,3.5]
4	86.8/89.7/87.9/82.1	+3.5 [2.6,4.4]
5	84.3/87.2/85.1/78.6	+3.8 [2.9,4.7]
6	81.6/84.9/82.6/74.5	+3.9 [2.8,4.9]
7	78.8/82.3/80.5/70.9	+3.7 [2.5,4.8]

Table 14: Source attribution, five-seed TSR mean $\pm$ std (rounded successes/tasks). Shares are trajectory/token percentages in the full mixture.

Training mixture	Traj. share	Token share	WebShop	Mind2Web
Full	100	100	64.8 $\pm$ 1.3 (324/500)	50.8 $\pm$ 1.5 (508/1000)
No Jericho	17.5 removed	14.2 removed	63.7 $\pm$ 1.4 (319/500)	49.9 $\pm$ 1.6 (499/1000)
No MiniWoB++	35.8 removed	33.6 removed	61.6 $\pm$ 1.6 (308/500)	47.8 $\pm$ 1.8 (478/1000)
No BEHAVIOR-1K	46.7 removed	52.2 removed	62.8 $\pm$ 1.5 (314/500)	49.1 $\pm$ 1.7 (491/1000)

Training mixture	WebArena	WorkArena
Full	46.6 $\pm$ 1.6 (378/812)	37.1 $\pm$ 1.6 (61/165)
No Jericho	45.7 $\pm$ 1.7 (371/812)	36.2 $\pm$ 1.8 (60/165)
No MiniWoB++	44.1 $\pm$ 1.9 (358/812)	35.0 $\pm$ 1.9 (58/165)
No BEHAVIOR-1K	44.7 $\pm$ 1.8 (363/812)	33.9 $\pm$ 2.0 (56/165)

Table 15: TraceWeaver’s delayed-constraint misses on WebShop. Counts are from the representative median seed (72/500 episodes); rates are all-episode mean $\pm$ std over five seeds.

Source	Count	All-episode rate
Candidate absent	18	3.6 $\pm$ 0.5
Memory write/retrieval	15	3.0 $\pm$ 0.4
Transition prediction	24	4.8 $\pm$ 0.6
Critic miscalibration	15	3.0 $\pm$ 0.4
Total	72	14.4

Table 16: Canonical same-data TSR with task/episode-level bootstrap 95% CIs (10,000 resamples, five seeds). Every row uses identical 12M+500K supervision and the 72-hour eight-A100 envelope.

Variant	WebShop	Mind2Web	WebArena
TraceWeaver	64.8 [62.5,67.1]	50.8 [48.5,53.1]	46.6 [44.0,49.2]
No entity memory	61.5 [59.2,63.8]	47.9 [45.5,50.3]	43.8 [41.3,46.3]
Generic recurrence	60.2 [57.8,62.6]	46.4 [44.0,48.8]	42.5 [40.0,45.0]
No critic	58.6 [56.1,61.1]	44.6 [42.1,47.1]	41.0 [38.5,43.5]
$H=1$	61.0 [58.6,63.4]	47.3 [44.8,49.8]	43.2 [40.7,45.7]
TransWM (1.3B)	62.2 [59.9,64.5]	48.7 [46.3,51.1]	44.3 [41.8,46.8]

Table 17: Neutral-pool selection over 200 WebShop episodes and 300 Mind2Web tasks, five seeds. Each tuple is score / rounded correct count / CVR-or-invalid / selector error / ms. Counts are five-seed means: successful episodes for WebShop and oracle-element matches among 3,840 evaluated Mind2Web steps per seed.

Selector	WebShop: TSR/correct/CVR/error/ms	Mind2Web: ElemAcc/correct/invalid/error/ms
Random	27.8 / 56 / 31.5 / 58.8 / 1	12.9 / 495 / 37.8 / 74.4 / 1
Greedy critic	48.9 / 98 / 20.1 / 35.2 / 24	40.8 / 1,567 / 11.2 / 32.5 / 22
LLM rerank	52.7 / 105 / 18.3 / 30.9 / 89	44.6 / 1,713 / 9.1 / 27.9 / 104
Native	45.6 / 91 / 23.7 / 39.6 / 41	38.1 / 1,463 / 13.4 / 36.8 / 58
Qwen SFT	54.8 / 110 / 16.2 / 27.5 / 38	47.2 / 1,812 / 7.8 / 24.1 / 34
TraceWeaver	57.6 / 115 / 14.1 / 23.6 / 31	50.7 / 1,947 / 6.3 / 20.7 / 27

Table 18: WebShop stress contrasts. Counts are unique tagged episodes before within-bucket resampling; p_H is the Holm–Bonferroni-corrected paired-bootstrap value.

Stressor	Unique n	Δ TSR (95% CI)	p_H
Load 1–2	186	+2.1 [0.4,3.8]	.018
Load 3–4	197	+4.7 [2.7,6.7]	<.001
Load 5+	117	+7.9 [5.1,10.7]	<.001
Paraphrase	208	+4.5 [2.3,6.7]	.002
Conflict	143	+7.5 [4.6,10.4]	<.001
Recovery	161	+5.7 [3.0,8.4]	<.001

Table 19: Overlap counts between mutually exclusive load bins and additional stress tags.

Load	Paraphrase	Conflict	Recovery
1–2	78	22	31
3–4	85	55	65
5+	45	66	65

Table 20: Fixed-candidate WebShop control (top-32 APM candidates, five seeds). Every selector ranks the identical set; tokens and milliseconds are per step.

Selector	TSR	CVR	Sel.Err	Tokens	ms
TraceWeaver	64.8 $\pm$ 1.3	10.4	8.3	406	768
Greedy critic	59.3 $\pm$ 1.5	13.8	12.6	398	418
LLM rerank	56.2 $\pm$ 1.6	16.4	15.7	412	538
R+V+Beam	60.1 $\pm$ 1.5	16.9	11.9	2,036	1,387

Table 21: Operational evidence matrix. Evaluations use non-adversarial benchmark environments; the table states integrity properties rather than adversarial security guarantees.

Protected object	Stressor / trust assumption	Property / enforcement boundary	Comparator / held-out setting	Utility or overhead	Observed failure / evidence trace
Instruction constraint state	Dense, conflicting, paraphrased instructions; evaluator and rendered state trusted	Active facts remain available to proposal, typed rollout, and critic; boundary is the committed memory snapshot	R+V+Beam; corrected WebShop stress and WebArena subgoal bins	+2.1 to +7.9 TSR by load; 768 ms model latency	memory/retrieval 3.0% all-episode rate; Tables 4, 18, 15
Valid action/schema state	Invalid elements, fields, or tool calls; DOM/API schema trusted	Candidate serialization and typed transition preserve executable action constraints; boundary is validator plus environment rejection	LLM rerank and Qwen SFT on shared pools	CVR 10.4/15.6/17.1; 31/27 ms selector latency	invalid action 2.0% of WebShop episodes; Tables 2, 5, 6
Transfer state	Site, domain, and enterprise shifts; official final-state evaluator trusted	Instruction-conditioned latent/memory interface remains useful outside training sites	Site-holdout Mind2Web and WebArena; wholly held-out WorkArena	+2.8/+2.6/+3.4 TSR; 36.5% lower WorkArena system latency	taxonomy/schema shift; Table 11 and Section 5.2

N Planner Sensitivity

Table 22 varies one planning parameter at a time while holding the evaluation subsets fixed.

Quality is stable around the default. Longer or wider search yields small, mixed gains at substantially higher latency and memory, while $H=3$ and $B=2$ lose accuracy on the same fixed tasks.

O Data Provenance and Reproducibility

Table 23 records the origin, split boundary, and permitted use of each training or evaluation source.

All prompt-based methods follow the template [INST] <instruction> [/INST] [OBS] <obs_history> [/OBS] [CAND] <candidate_list> [/CAND]. Observation history includes the last three steps. Constraint validators apply regex-based form-field validation, URL pattern matching for navigation targets, and DOM element existence checks via XPath queries.

WebShop uses the official Docker image (webshop:v0.3.1) with default catalog size (1.18M products). Mind2Web uses the replay-based evaluation harness from the official repository (commit a3f1e2d). WebArena runs on Docker Compose with five containers covering shopping, CMS, forums, maps, and GitLab, together with NVIDIA Container Toolkit 1.14.3 and CUDA 12.4. WorkArena uses the official ServiceNow PDI environment (version utah-patch-6) with 33 task types following the evaluation protocol from Drouin et al. (2024). All environments run on a single node with 2×Intel Xeon Platinum 8480+ and 8×A100-80GB.

Before trajectory construction, task IDs, template hashes, episode IDs, and entity-masked instruction signatures are assigned to disjoint train/development/evaluation partitions; site/domain IDs additionally define the hold-out analyses. WebShop evaluation products and templates, Mind2Web evaluation task IDs and instruction signatures, and WebArena evaluation task templates and snapshots are excluded from both trajectory sets; Mind2Web holdout runs also exclude the target sites/domains, and WorkArena is excluded by platform. Self-play successor observations supervise the encoder/dynamics/APM; instruction spans and simulator state supervise the memory type/write classifier; the 500K demonstrations provide five-step terminal-success and constraint-validity targets for the critic. The reported 18% positive rate is the proportion of

candidate-state pairs satisfying both targets.

P End-to-End Latency

Table 24 decomposes end-to-end step latency so that model and environment costs are not conflated. Table 24 reports model and environment time together.

Q Metric Definitions and Aggregation

WebShop, WebArena, and WorkArena TSR use each environment’s binary final-state evaluator, with one episode/task as the statistical unit. Mind2Web task success equals one only when every replay step in the 1,000-task split matches an executable oracle action; official element accuracy is the number of oracle-element matches divided by evaluated steps. CVR divides constraint-, schema-, type-, or domain-invalid executed actions by all executed actions; IAR divides actions satisfying every active instruction predicate by all executed actions. Tokens per step sum decoded tokens across all calls used to select one environment action; the 2,048-token limit applies to the input context of each call. For five seeds, each task or episode is first averaged across seeds and then resampled as a paired unit; website analyses resample sites and tasks hierarchically. Raw success counts report the rounded mean number of successful tasks per seed, while mean±standard deviation always summarizes the five seed-level benchmark scores.

R Qualitative Examples

Table 25 illustrates how persistent constraint state changes action selection in representative trajectories.

The successful trace shows how the same size fact changes a locally attractive click after three imagined steps. The residual trace localizes a TOOL-transition error: retrieval is correct and the candidate is present, but the imagined successor omits a permission constraint, matching the fidelity and failure-decomposition analyses.

Table 22: One-factor planner sweep on fixed 200-task WebShop and Mind2Web subsets, five seeds. Performance cells give TSR mean $\pm$ std / CVR / rounded successes / paired 95% CI versus the default. Resource cells give model ms / tokens per step / peak GB.

Setting	WebShop performance	Mind2Web performance	Resources
$H=3$	64.7 $\pm$ 1.7 / 10.8 / 129 / [-2.5,-0.1]	52.7 $\pm$ 1.9 / 15.7 / 105 / [-2.6,0.0]	601 / 332 / 38
$H=5$	66.0 $\pm$ 1.5 / 9.7 / 132 / reference	54.0 $\pm$ 1.7 / 14.5 / 108 / reference	768 / 406 / 41
$H=7$	66.3 $\pm$ 1.6 / 9.5 / 133 / [-0.4,1.0]	54.2 $\pm$ 1.8 / 14.4 / 108 / [-0.6,1.0]	982 / 518 / 47
$B=2$	64.6 $\pm$ 1.8 / 11.1 / 129 / [-2.7,-0.1]	52.4 $\pm$ 2.0 / 16.0 / 105 / [-2.9,-0.3]	549 / 301 / 36
$B=4$	66.0 $\pm$ 1.5 / 9.7 / 132 / reference	54.0 $\pm$ 1.7 / 14.5 / 108 / reference	768 / 406 / 41
$B=8$	66.4 $\pm$ 1.6 / 9.4 / 133 / [-0.3,1.1]	54.3 $\pm$ 1.8 / 14.2 / 109 / [-0.5,1.1]	1,118 / 612 / 53
$\gamma=0.90$	65.2 $\pm$ 1.6 / 10.2 / 130 / [-1.6,0.0]	53.3 $\pm$ 1.8 / 15.1 / 107 / [-1.5,0.1]	763 / 405 / 41
$\gamma=0.95$	66.0 $\pm$ 1.5 / 9.7 / 132 / reference	54.0 $\pm$ 1.7 / 14.5 / 108 / reference	768 / 406 / 41
$\gamma=0.99$	65.8 $\pm$ 1.7 / 9.9 / 132 / [-1.0,0.6]	53.7 $\pm$ 1.9 / 14.7 / 107 / [-1.1,0.5]	771 / 408 / 41

Table 23: Training-data provenance for TraceWeaver’s two-stage pipeline. Self-play trajectories use automatic simulator rewards; pre-existing demonstration traces are not newly labeled by the authors. The double-label rate (12%) and agreement ($\kappa=0.81$) come from automated labeling with model-based profiles, not a human study.

Source	Type	Episodes	Mode	License
Jericho (48 games)	Self-play	2.1M	Auto reward	MIT
MiniWoB++ (114 tasks)	Self-play	4.3M	Auto reward	Apache 2.0
BEHAVIOR-1K (1K act.)	Self-play	5.6M	Auto reward	MIT
WebShop traces	Pre-existing demo	286K	Auto labels	CC-BY-4.0
Mind2Web traces	Pre-existing demo	127K	Auto labels	CC-BY-SA
WebArena traces	Pre-existing demo	87K	Auto labels	Apache 2.0

Table 24: End-to-end latency in milliseconds per step (5-seed specification, A100-80GB). Model-only: time from observation encoding to action output, including APM, critic scoring, and rollout search. Env: environment processing time. System: sum of model-only and environment time. Against R+V+Beam, TraceWeaver reduces system latency by 28.9–36.5% and model latency by 43.2–51.8%.

Benchmark	Method	Model	Env	System
WebShop	TraceWeaver	768	612	1,380
	R+V+Beam	1,387	612	1,999
	LLM Rollout	4,210	612	4,822
Mind2Web	TraceWeaver	804	683	1,487
	R+V+Beam	1,421	683	2,104
WebArena	TraceWeaver	831	724	1,555
	R+V+Beam	1,463	724	2,187
WorkArena	TraceWeaver	796	691	1,487
	R+V+Beam	1,650	691	2,341

Table 25: Compact action traces. Scores are critic trajectory values; memory lists retrieved instruction-bound entries before selection.

Case/step	Active constraints	Retrieved memory	Leading candidates	Imagined consequence	Score, choice, and outcome
Success/1	red, cotton, L, < \$35	size=L; price< \$35	query “red cotton polo”; query “red polo”	first query preserves material binding	0.87 vs. 0.84; first query exposes 26 products
Success/2	same four constraints	attributes for items 17 and 42	click item-17; click item-42	item-17 lacks L; item-42 preserves all facts	2.18 vs. 2.41; item-42 has L at \$32.40
Success/3	quantity and check-out	item-42; size L	add-to-cart; revisit results	add completes all active predicates	add-to-cart; evaluator succeeds
Residual/1	export report as CSV	report tool; date range	call export; open preview	rollout export enabled	export selected; permission banner appears
Residual/2	preserve date range	endpoint and prior filter	retry export; return to filter	TOOL successor omits role restriction	error 0.31 vs. CLICK 0.12; delayed miss