CausalCache: Conditional High-Fidelity Restoration for Long-Horizon GUI Agents

Jiaxuan Luo[1], **Zhanfeng Liao**[2], **Jiayao Teng**[1], **Yuan Wang**[2]
[1]Johns Hopkins University
[2]Tsinghua University
jluo50@jhu.edu, zliao08@mails.tsinghua.edu.cn,
jteng68@jhu.edu, ywang80@mails.tsinghua.edu.cn

Abstract

Long-horizon GUI agents can retain complete action histories as compact text, but only a few historical screenshots fit in active context. We formulate this as *budgeted fidelity restoration*: every event remains summarized, while a fixed budget B determines which events regain their archived screenshots. Recent-B assigns all visual slots to the latest events. CausalCache instead scores the complete history and swaps in an older event only when its predicted utility exceeds that of a recent event. A history-gated key/value adapter modifies only restored history-image tokens and is exactly bypassed when no history image is active, preserving current-screen processing. The adapter and selector are trained with matched-budget interventions on desktop trajectories and evaluated zero-shot on mobile. On OSWorld-Verified, activating historical screenshots improves success by about 13 percentage points over summary-only memory. Under the official 15-step limit, CausalCache and Recent-4 are statistically indistinguishable; in a 30-step diagnostic, CausalCache achieves 46.7% success versus 42.4% (+4.3 points). Zero-shot on 117 MobileWorld tasks, CausalCache improves over Recent-4 from 30.2% to 36.8%. The gain is concentrated on a predefined cross-app memory-candidate split (30.6% vs. 19.4%, +11.2 points), while single-app controls show no detectable difference (43.6% vs. 42.4%). These results show that selecting *which* past events regain pixels is more effective than spending a fixed visual budget entirely on recency.

Introduction

Autonomous agents that follow natural-language instructions and operate graphical user interfaces promise broad practical value: they can automate repetitive digital chores, carry out multi-application workflows on behalf of their users, and extend what a single person can accomplish on desktop and mobile devices (Xie et al. 2024; Rawles et al. 2025; Xu et al. 2026). Delivering that value on long tasks turns memory into a first-order design problem: GUI agents accumulate screenshots, actions, text arguments, coordinates, and interface changes over many decisions. A complete action trace persists cheaply as compressed text, but keeping every screenshot active is costly and distracting. Each past event therefore has two policy-visible fidelities: a summary-only record and, when needed, that same record with its archived screenshot reattached. The memory-control question is not only which events remain recorded, but which recorded events should be exposed again in high-fidelity pixels (Lu et al. 2025; Zeng et al. 2026; Shi et al. 2026).

Our central claim is that the object of control is *event fidelity*, not event inclusion. Existing controllers retain recent events, retrieve visually similar screenshots, or learn salience from task supervision (Zeng et al. 2026; Shi et al. 2026; Liu et al. 2026). Yet the value of promoting a summarized event back to pixels depends on which images are already active: without recent images, a distant screenshot may lack local grounding; with a saturated recent window, another image may be redundant. Moreover, the action policy must consume the reattached pixels. Our experiments show that a strong frozen policy exhibits weak and unreliable *average* selectivity under matched replacement probes, although it can still benefit from restored evidence on a subset of states; HGKV sharpens this content sensitivity and provides a more reliable utility signal for training and deployment.

CausalCache accordingly treats GUI memory as *conditional high-fidelity restoration* ("causal" denotes policy-specific effects of matched prompt interventions, rather than structural causal discovery). Every event remains in the complete low-fidelity action trace, and an archived screenshot is reattached only when that event is promoted into the active visual context; no pixels are reconstructed from text. Under a budget of B high-fidelity history images, Recent-B promotes the latest B events. CausalCache reallocates the same B promotions over the complete trace, with recent events as the default: a distant event enters only by displacing a recent image whose conditional marginal utility it exceeds. Because every compared allocation carries the same number of images, the contrast isolates fidelity *allocation* from visual capacity, and how many non-recent events are actually promoted is measured as a result, never preset. Figure 1 illustrates this reallocation on a real cross-application episode.

Learning to *use* reattached pixels must not corrupt the base agent. Full-layer adaptation simultaneously changes current-screen understanding, action syntax, and grounding. CausalCache instead trains a history-gated KV interface (HGKV) that touches only restored history-image tokens and is structurally bypassed at zero budget, with a per-arm-anchored difference-of-differences (DiD) objective in which every arm is anchored to its own frozen score: uniform amplification of any history cancels exactly, and only *selective* use of target-relevant content reduces the loss.

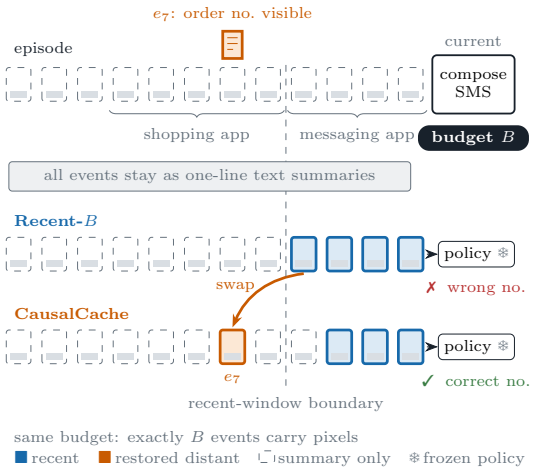

Figure 1: CausalCache on a real cross-app episode (schematic; drawn with B=4). The agent must compose an SMS containing the order number it saw earlier in a shopping app. Every event persists as a one-line text summary; a budget of B events may additionally carry their archived pixels. After the app switch the evidence frame e_7 lies outside the recent window: Recent-B spends every slot on the latest events and types the wrong number, while CausalCache swaps one recent slot for e_7 at the same budget and succeeds.

Our contributions are:

- a conditional fidelity-restoration formulation of GUI memory: all events persist as compressed action records, while Recent-B and CausalCache allocate the same B summary-plus-image promotions over that complete trace;
- a policy-preserving HGKV interface trained on matched-budget replacement groups with per-arm-anchored DiD supervision, exact no-history parity, and pre-specified drift caps;
- a budget-aware restoration selector over the complete summarized trace, trained on desktop data only and evaluated zero-shot on mobile and desktop benchmarks.

Related Work

Visual GUI agents and benchmarks. WebShop, WebArena, Mind2Web, WorkArena, WorkArena++, VisualWebArena, and REAL span grounded shopping, real websites, enterprise workflows, and visual web interaction (Yao et al. 2022; Zhou et al. 2024; Deng et al. 2023; Drouin et al. 2024; Boisvert et al. 2024; Koh et al. 2024; Garg et al. 2025). AITW, AndroidControl, AndroidWorld, OSWorld, and GUI-Odyssey extend executable evaluation to mobile and desktop interfaces (Rawles et al. 2023; Li et al. 2024a; Rawles et al. 2025; Xie et al. 2024; Lu et al. 2025). Screenshot-native agents learn grounding and actions directly from pixels (Hong et al. 2024; Cheng et al. 2024; Lin et al. 2025; Xu et al. 2026). These works measure long-horizon interaction; our question is which completed event should re-enter a bounded prompt in pixels. MobileWorld further stresses long-horizon, cross-application mobile workflows under reproducible functional evaluation (Kong et al. 2026).

Agent memory and reusable experience. RAG retrieves external memory, ReAct interleaves reasoning with action, and later agents preserve reflections, experience, or evolving long-term stores (Lewis et al. 2020; Yao et al. 2023; Shinn et al. 2023; Zhao et al. 2024; Zhong et al. 2024; Gutiérrez et al. 2024). ICAL retrieves multimodal programs of thought (Sarch et al. 2024). These systems write or retrieve artifacts across interactions or tasks; CausalCache instead holds a within-episode summary trace fixed and controls whether an existing event is exposed as summary only or summary plus image.

Long-horizon GUI memory. GUI-Odyssey resamples prior screenshots, MementoGUI learns memory operators, and AndroTMem retrieves causally linked anchor states (Lu et al. 2025; Zeng et al. 2026; Shi et al. 2026). Their primary variable is which item is retained or retrieved. CausalCache keeps the compressed trace fixed, controls event fidelity conditioned on the already-active visual context, and trains the interface that consumes restored pixels.

Context and visual-token compression. Prompt compressors remove textual redundancy (Jiang et al. 2023, 2024; Pan et al. 2024); KV methods retain, evict, select, or quantize cache states (Xiao et al. 2024; Zhang et al. 2023; Liu et al. 2023; Li et al. 2024b; Liu et al. 2024; Corallo and Papotti 2024); and visual methods merge or prune tokens within an image (Bolya et al. 2023; Chen et al. 2024; Yang et al. 2025). Their unit is a token or KV position. CausalCache instead preserves the summary trace and allocates whole archived screenshots across events; HGKV controls only how promoted history-image tokens are consumed.

Interventional memory value. Recent work studies memory through controlled interventions or active state influence (Srivastava 2026; Liu et al. 2026). Our intervention is narrower and executable: a selected event regains one archived post-action image while the remaining prompt stays fixed, yielding a policy-specific behavioral surrogate rather than an environment-level causal effect. Set models parameterize coalition-conditioned scores (Zaheer et al. 2017; Lee et al. 2019), whose marginals connect to cooperative-game attribution (Shapley 1953); we use controlled restorations to train a deployable selector for one fixed GUI policy.

Method

Problem Formulation

At decision t, the agent receives instruction g, current observation x_t, and history $H_t = \{e_1, \ldots, e_{t-1}\}$ with $e_j = (x_j, a_j, x_{j+1})$, where each event has a compact low-fidelity representation $e_j^{\text{low}} = \tilde{e}_j$, its one-line action summary, and an available high-fidelity representation $e_j^{\text{high}} = (r_j, v_j)$: the official retained-turn form pairing the step's verbatim policy response r_j with its reattached archived post-action screenshot v_j. The official protocol pairs every retained observation

with the response that produced it, so promotion and demotion move an event between these two forms as a unit — a fidelity allocation reassigns retained turns, not images alone.

Fixed active-fidelity budget. Every prompt contains the goal g, the *complete* textual action summaries $\{\tilde{e}_1, \ldots, \tilde{e}_{t-1}\}$, and the current screenshot x_t. An active visual-context budget B limits how many summarized events may be promoted by reattaching their archived pixels; it is not a limit on cold archival storage. A restoration allocation is a set $S \subseteq \{1, \ldots, t-1\}$ with $|S| = \min(B, t{-}1)$: the budget is always filled, and reallocation is over *which* events occupy it. Let $Q(S)$ denote the policy's mean target-token log-likelihood of the successful next action when events in S are exposed in summary-plus-image form. The standard baseline is $S_{\text{recent}} = \text{Recent-}B$, the B most recent distinct events promoted to high fidelity. CausalCache targets

$$S^* = \arg \max_{\substack{S \subseteq \{1,\ldots,t-1\} \\ |S| = \min(B,t-1)}} Q(S), \qquad Q(S^*) \geq Q(S_{\text{recent}}), \tag{1}$$

so recency is one candidate fidelity allocation, not a separate modality. We use *causal* in a deliberately narrow, operational sense: for a fixed policy, current observation, summarized trace, and active-image budget, we intervene on the fidelity allocation and measure the resulting change in policy behavior. The conditional marginal utility so defined is policy- and prompt-specific; it is not an estimate of an environment-level structural causal effect. The realized replacement intensity $k = |S \setminus \text{Recent-}B|$ is a measured statistic, not a preset parameter. All comparisons at a given B hold the image count, resolutions, underlying event trace, and prompt structure fixed: every allocation renders exactly $|S|$ retained turns and keeps every other event at summary fidelity; only which events occupy the high-fidelity form differs. Cross-B comparisons are analysis, not the causal contrast. Given a partial allocation S, conditional marginals $\Delta_t(j \mid S) = Q(S \cup \{j\}) - Q(S)$ drive selection, and a replacement is justified only when the distant event's restoration marginal exceeds that of the recent visual realization it displaces. Every formally reported set utility reruns the complete restored set; we never report a sum of singleton gains as the utility of a coalition.

Policy-Preserving History-Gated KV Adapter

Let h_l be the input to layer l's attention projections. In the last eight language-model layers, CausalCache adds rank-8, α=16 low-rank residuals (Hu et al. 2022) only to key and value projections:

$$\begin{aligned} K_l &= W_l^K h_l + M_{\text{hist}}\, \Delta W_l^K h_l, \\ V_l &= W_l^V h_l + M_{\text{hist}}\, \Delta W_l^V h_l. \end{aligned} \tag{2}$$

M_{hist} is a token mask that is true only for image tokens reattached to promoted historical events. It excludes instruction tokens, event-summary text, the current observation image, and action tokens. The mask is constructed fail-closed from the multimodal token-type sequence and per-image patch geometry; any mismatch between declared images and encoded image blocks aborts the forward pass. When no historical event is promoted, the adapter context is absent and the projection hook returns the original module output, so $\pi_{\text{HG}}(\cdot \mid S{=}\varnothing) = \pi_0(\cdot \mid S{=}\varnothing)$ up to bitwise equality in our implementation.

Per-B Fidelity-Restoration Supervision

Training states are decision points of successful desktop trajectories in which the target action a_t^* recurs earlier in the same trajectory under full-action equivalence (type, arguments, and coordinates within tolerance); the screenshot preceding the earlier occurrence is the target-specific restoration candidate. For each state and each budget B we construct three matched-budget fidelity allocations with identical summaries, current screen, and target—every prompt carries exactly B history images:

- **Recent**: Recent-B, the B most recent distinct events promoted to summary-plus-image form;
- **Relevant restore**: Recent-$(B{-}1)$ + $\{v_{j^+}\}$, the oldest recent visual slot replaced by the target-specific archived screenshot (age $\geq B{+}2$, strictly older than the whole window);
- **Wrong restore**: the same slot replaced by an age-matched same-trajectory screenshot whose following action is *not* equivalent to a_t^*.

Positives do not require the base action to be wrong: both information addition and evidence amplification are admissible mechanisms.

Why single-slot replacement (k=1). Training teaches only the k=1 replacement interface: offline mining finds simultaneous multi-frame demand rare, and a matched-budget probe shows the first replacement is worth its slot while the paired increments of a second and third are indistinguishable from zero or negative. Larger-k allocations remain expressible at deployment because the selector composes exact-B sets with recent fallback, so k stays a result statistic; mining details and the full probe are in the supplement.

Let ℓ denote mean target log-likelihood, superscripted by adapter state (active or frozen bypass). Each arm is anchored to *its own* frozen score:

$$\begin{aligned} &A_s = \ell^{\text{on}}_{\text{relevant}} - \ell^{\text{off}}_{\text{relevant}}, \qquad A_r = \ell^{\text{on}}_{\text{recent}} - \ell^{\text{off}}_{\text{recent}}, \\ &A_n = \ell^{\text{on}}_{\text{wrong}} - \ell^{\text{off}}_{\text{wrong}}, \end{aligned} \tag{3}$$

and the DiD objective is

$$\begin{aligned} \mathcal{L} = {} & [m - (A_s - A_r)]_+ + [m - A_s]_+ \\ & + [m - (A_s - A_n)]_+ \\ & + \lambda_{\text{cap}}\big([|A_r| - \epsilon]_+ + [|A_n| - \epsilon]_+\big) + \lambda \|\Delta W\|_2^2, \end{aligned} \tag{4}$$

with m=0.01, ϵ=0.02, λ_{cap}=2, λ=10^{-4}, and no action cross-entropy. On the identity adapter every increment is zero, so $A_s - A_r$ is exactly zero: the frozen policy's own preference for target-relevant restorations cannot masquerade as adapter skill, and uniformly amplifying all restored pixels buys nothing. The dead-zone caps bound recent-window and wrong-history drift with gradients on the same scale as the selection hinges. Losses and evaluation are stratified by B; a pooled mean could hide a method that helps at B=4 but fails at B=1.

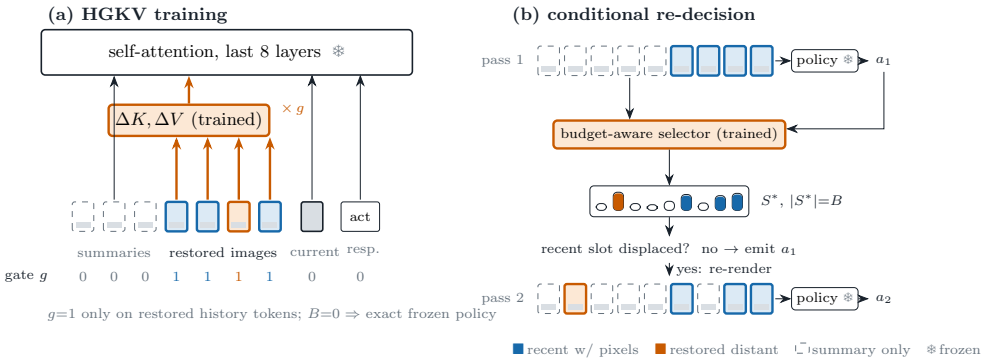

Figure 2: The two learned components of CausalCache. (a) The history-gated K/V interface (HGKV): a low-rank $\Delta K, \Delta V$ branch on the last eight frozen self-attention layers is multiplied by a binary token gate g that is 1 only on restored history-image tokens, so current-screen, text, and response tokens pass through unchanged, and at B=0 the gate vanishes and the policy is bitwise-identical to the frozen base. (b) Proposal-conditioned re-decision: pass 1 renders the default Recent-B allocation and yields a proposal a_1; a two-tower scorer conditioned on per-event features with a_1 as witness and on budget context scores the *complete* summarized trace and returns the exact-B set S^*. If S^* keeps the recent tail, a_1 is emitted with no extra compute; only when a recent event is displaced is the prompt re-rendered under S^* and the *same* frozen weights generate the final action a_2. Event cards follow Figure 1: dashed gray = summary-only, blue = recent with pixels, orange = promoted distant event; orange-outlined boxes are the trained modules; snowflakes denote frozen components.

Budget-Aware Restoration Selector

After the history interface is frozen, selector labels are true policy utilities: singleton utilities for the *entire* summarized event pool, matched $Q(\text{Recent-}B)$ anchors, and set utilities along teacher paths, each fully re-scored as a complete prompt through the frozen HGKV policy. The selector is a set-conditional marginal scorer $\widehat{\Delta}_t(j \mid S)$ trained on the implied marginals with a joint regression and within-state ranking loss; it scores each candidate event conditioned on the current action proposal, event recency, and the already-selected set (the full feature list is in the supplement). Deployment composes an exact-B allocation with a small beam search initialized at Recent-B: a distant event enters only by beating the predicted marginal of the recent realization it evicts, realizing Eq. 1 as a relative comparison—declining every replacement recovers Recent-B exactly. Selector labels cover $B \in \{1, 2, 4\}$; the realized replacement intensity k is reported per budget, never preset.

Reference actions. Part of the selector's input asks whether the action taken right after an archived event matches a reference action. Training uses the gold teacher-forced target, an offline oracle unavailable at deployment; CausalCache therefore re-grounds the reference in the policy's own tentative action: the policy first acts on the Recent-B prompt, and re-decides on the re-allocated prompt only when a recent visual realization is displaced. This proposal reference is statistically indistinguishable from the gold oracle on held-out desktop groups, whereas weaker references—no action signal, or instruction-only similarity—degrade selection and over-replace; a single-pass variant reusing the previously executed action removes the second call but transfers worse across platforms (reference ladder and efficiency analysis in the supplement).

Experiments

Experimental Setup

Training Data and Model Selection. Policy adaptation and selector training use only desktop data: successful AgentNet/OpenCUA Ubuntu trajectories (5,000 screened, 2,293 successful, 6,003 decision points) yield fixed-budget replacement groups at B=1/2/4 of 969/764/470 (2,203 in total; 1,774 training units, 211 development groups, and 218 held-out test groups, trajectory-disjoint), complemented by a high-precision OSWorld witness seed mined from official successful trajectories. A further 160 groups at B=8 are constructed but deliberately excluded from training, so B=8 evaluates out-of-training budget extrapolation. Checkpoints and thresholds are selected only on the desktop development split under pre-specified gates: positive DiD selection with a positive bootstrap lower bound, and drift caps $|A_r|, |A_n| < \epsilon$. No mobile benchmark data touches training or selection. All adapter training, offline scoring, and closed-loop serving run on NVIDIA H200 Tensor Core GPUs.

The Fixed-Budget Fidelity-Restoration Design. Figure 2 shows the two learned components and the same-budget re-decision that defines the core comparison. Every compared allocation carries exactly B history images with identical resolutions, summaries, and prompt structure, so the contrast isolates *which summarized events are promoted to high fidelity*. B=4 is the primary deployment-like setting; the realized number of non-recent promotions k is measured rather than preset.

Policies and Fidelity-Allocation Baselines. The frozen policy is GUI-Owl-1.5-8B-Instruct (Xu et al. 2026). Under identical data, steps, and cadence we compare the frozen policy, an ungated KV control (the same last-eight layers,

Benchmark / stratum	Recent-B	CausalCache	Δ
MobileWorld full (117)	30.2	36.8	$+6.6^{*}$
memory-critical (62)	19.4	30.6	$+11.2^{*}$
single-app control (55)	42.4	43.6	$+1.2$
split×method interact.	—	—	$+10.1$
OSWorld-Verified (361)	33.0	33.3	$+0.3$
30 max steps	42.4	46.7	$+4.3^{*}$

Table 1: Primary closed-loop contrasts at the deployment budget $B{=}4$ (task-paired bootstrap; MobileWorld three-run means; the OSWorld row uses the official 15-step cap). * the 95% confidence interval strictly excludes zero (full CI tables in the supplement; the interaction interval accompanies them). The gain concentrates exactly on the construction-defined memory-critical stratum, with no detectable effect on controls; on desktop the allocations tie at the official 15-step cap and separate when it is doubled to 30.

K/V targets, rank, and alpha as HGKV but active on all tokens), and HGKV itself (token-gated, structurally bypassed at $B{=}0$). Fidelity-allocation baselines are Recent-B, Random, OCR/RGB similarity, and the learned budget-aware restoration selector (exact-B, recent fallback); a beam oracle over the candidate pool bounds headroom. Every method receives the same summaries and the same active image budget.

Zero-Shot Benchmarks and Metrics. The policy interface and selector, trained on desktop data only, are evaluated (i) on held-out desktop decision groups (offline, per-B), (ii) zero-shot on an uncontaminated OSWorld task roster (Xie et al. 2024) with official evaluator scores (tasks whose trajectories seeded the witness corpus are excluded from zero-shot claims and reported separately as in-domain diagnostics), and (iii) zero-shot on a cross-app mobile memory benchmark of 117 runnable tasks (62 cross-app memory candidates vs. 55 single-app controls, split mechanically by the benchmark's task metadata), which contributes no data to any training or selection decision (Kong et al. 2026). Offline metrics are teacher-forced log-likelihood margins and full-action equivalence under an argument and coordinate tolerance (sensitivity settings in the supplement); closed-loop metrics are paired task success with cluster bootstrap, step counts, and wall time. All arms decode greedily, so every action is a deterministic function of its rendered prompt, and the conditional second pass always replaces the proposal rather than selecting between samples—it cannot act as best-of-n.

Results

Selectivity Is Not Inherited from the Frozen Policy. On the desktop development split, the frozen policy shows no reliable preference for a task-relevant archived screenshot over the informative recent frame it would displace (recent-baseline construction details in the supplement): to the frozen policy the two are worth about the same, so profitable fidelity reallocation must be learned. Under the official multi-turn protocol the frozen replacement effect decays with the window and reverses at large B (supplement), while the learned interface keeps a positive selection effect across the budget axis.

Selective Use Within a Drift Envelope. Table 3 reports the pre-specified desktop DiD gate at $B{=}1$. Adapters are trained for matched steps; checkpoints follow the frozen selection rule. HGKV passes every gate: its selected checkpoint learns a positive DiD selection effect against the redundant-duplicate recent control, a re-scoring probe confirms the effect against the informative true-previous-frame control, and recent-only and wrong-history drift stay well inside the cap $\epsilon{=}0.02$; its no-history path is bitwise identical to the frozen policy under randomized nonzero adapter weights. The layer- and rank-matched ungated control also learns selectivity, confirming that the last-eight K/V position carries signal, but token gating adds a significant paired per-group advantage while holding drift further from the caps. Gating is what buys selectivity *inside* the envelope.

Fidelity Reallocation and Zero-Shot Transfer. The offline budget axis shows the same picture. Across the trained budgets $B \in \{1, 2, 4\}$, HGKV maintains positive selectivity—per-B DiD gates all pass with drift within the pre-specified caps—and the learned allocation improves over Recent-B at matched image counts, while keeping the full recent window in 19% of $B{=}4$ states. The ungated control again passes but with uniformly smaller selection effects, drifting toward the cap at the largest budget. Untrained-budget extrapolation ($B{=}8$) and the full per-budget tables are in the supplement.

Mobile closed loop (zero-shot). The 117-task mobile benchmark contributes no training or selection signal and is our primary test of *allocation*: its roster carries a memory-critical split deterministically partitioned via the benchmark's upstream metadata (multi-app vs. single-app), with zero manual filtering. The full-roster aggregates are secondary: over three runs at the primary budget $B{=}4$, CausalCache attains the highest overall success rate, 36.8%, against 30.2% for HGKV+Recent-B and 28.2% for the frozen summary-only baseline (Table 2), with the same sign in every round. The decomposition is clean: frozen Recent-B, HGKV+Recent-B, and the summary-only baseline are statistically indistinguishable (29.9/30.2/28.2%) — the behavioral counterpart of the drift cap — so the gain is carried by *which summarized events are promoted*, not by the adapter or by recency alone. Running the same selector on the fully frozen policy retains most of the margin (33.9%), and re-teaching it from frozen-policy-scored labels retains less (33.3%): the adapter contributes both at deployment and as the utility teacher. A recent-dose sweep of the frozen policy makes the last point explicit: across $B{=}0$–4 success is non-monotone with every paired contrast against $B{=}0$ crossing zero; promoting additional recent events alone buys nothing, so fidelity *allocation* rather than active visual budget *size* is the operative variable.

The primary closed-loop result is that split: cross-app tasks are constructed so that information from an earlier application is needed after the switch. The rule uses no model outcome, and candidates that do not truly require distant evidence only dilute the stratum toward zero, making the con-

Method	OSWorld-Verified (361)	MobileWorld (117)	MW memory-critical (62)	MW control (55)
Frozen, summary-only (B=0)	19.9	28.2	21.0	36.4
Frozen + Recent-B	32.7	29.9	22.6	38.2
HGKV + Recent-B	33.0	30.2	19.4	42.4
Frozen + selector	32.7	33.9	28.0	40.6
CausalCache (HGKV + selector)	**33.3**	**36.8**	**30.6**	**43.6**
frozen-taught selector variant	—	33.3	25.8	41.8

Table 2: Closed-loop success rates (%) at the primary deployment budget B=4; the offline budget sweep is in the supplement. OSWorld-Verified: official no-GDrive roster and official 15-step budget, one run per arm (a 30-step extended-horizon rerun of both closed-loop arms is reported in the text). MobileWorld: zero-shot means over three runs for all multi-run arms; the memory-critical column restricts to the 62 construction-defined cross-app tasks, where the allocation contrast is largest, and the control column to the 55 single-app tasks, where no restoration method should help. CausalCache deploys the HGKV adapter plus the budget-aware selector; Frozen + selector runs the same selector on the fully frozen policy, and the frozen-taught variant additionally trains the selector on frozen-policy-scored labels. High-fidelity arms beat summary-only memory on desktop but remain pairwise indistinguishable there. Full paired statistics and per-round results are in the supplementary document.

Adapter	DiD select	$\lvert A_r \rvert$	Wrong drift
HGKV	**+0.0224**	0.0088	0.0085
Ungated KV	+0.0177	0.0086	0.0071

Table 3: Desktop DiD evaluation on 94 development groups (B=1). DiD select is the target-relevant conditional margin ($A_s - A_r$); both adapters hold recent-window ($\lvert A_r \rvert$) and wrong-history drift within the cap 0.02 (intervals and all budgets in the supplement). The HGKV zero-budget path is exactly the frozen policy.

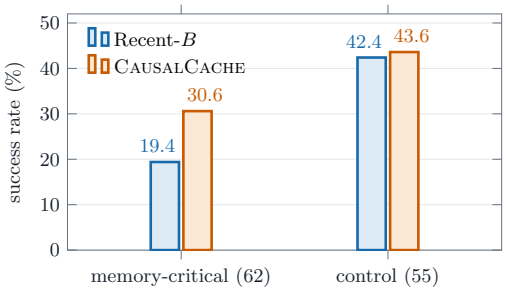

Figure 3: The core contrast at the primary budget: on construction-defined memory-critical tasks CausalCache gains +11.2pp over the same-budget recent allocation, while on matched single-app controls no effect is detectable (+1.2pp); the split-by-method interaction is +10.1pp (paired intervals in the supplement).

trast conservative (rule text and manifest hashes in the supplement). On the memory-critical stratum the same-budget contrast is decisive (Figure 3): CausalCache 30.6% vs. Recent-B 19.4%, a +11.2pp gain; on the controls no effect is detectable (43.6% vs. 42.4%); and the split-by-method interaction is +10.1pp (intervals in the supplement). The value of reallocation is conditional on distant visual dependency: the modest full-roster mean is composition, not absence of effect, since 47% of the roster is control tasks on which no high-fidelity restoration method should help.

As a complementary outcome-defined analysis, we call tasks that the frozen summary-only policy cannot complete within the official budget *memory-demanding*, and report $\Delta_{\text{long}} = P(\text{success} \mid \text{CausalCache}, \text{long}) - P(\text{success} \mid \text{baseline}, \text{long})$ on that stratum. Across 95 such mobile tasks, CausalCache gains 7.0pp over summary-only memory, while adding nothing on the 22 tasks the baseline already solves. These post-hoc strata are secondary to the construction-defined memory-critical split.

Desktop closed loop. On OSWorld-Verified (361 tasks, official no-GDrive roster and 15-step budget), all high-fidelity arms reach 32.7–33.3% (CausalCache 33.3%), each about +13pp over summary-only memory (19.9%, paired 95% CIs excluding zero), while same-budget allocations remain indistinguishable. The official step budget itself binds, however: 352 of 361 episodes exhaust the 15-step cap. Rerunning both arms with the cap doubled to 30 steps separates the allocations: Recent-4 rises to 42.4% (+9.5pp from the extra horizon) while CausalCache rises to 46.7% (+13.4pp), converting the extended horizon into +3.9pp more in-domain success and a +4.3pp margin at 30 steps versus +0.3 at 15. Under the official cap OSWorld therefore supports high-fidelity *availability*; conditional *allocation* shows on the horizon-extended desktop run and is tested zero-shot by the mobile memory-critical split. HGKV again shows no detectable behavioral difference on matched Recent-B allocations, mirroring the mobile finding. CausalCache adds a median of eight policy calls per task, about 7% of median end-to-end task time.

Ablations and Analyses

Policy-side per-B gates and the frozen recent-dose sweep appear in Results. Table 4 brackets the learned selector between allocation rules that share its budget and rendering exactly. Random exact-B and RGB-similarity restoration are indistinguishable from Recent-B: sparse exposure of archived pixels, or visual similarity to the current screen, is not utility. Inverting the learned score is significantly harmful at both budgets—a specificity control showing the

Allocation rule	B=2	B=4
Learned scorer	+.030	+.017
Random exact-B	+.006	+.008
RGB similarity	+.006	+.007
Inverted score	−.016	−.009
No recency feats	+.012	+.008

Table 4: Chooser ladder: log-likelihood margin over Recent-B at matched budgets with fresh full-set policy scoring. Only the learned ranking and its inversion separate from zero (episode-clustered intervals in the supplement).

scorer ranks real signal symmetrically rather than exploiting a rendering artifact. Retraining without the recency-membership features returns the selector to the noise band, so budget-context features carry substantial weight.

The budget is spent adaptively: at B=4 the learned scorer keeps the entire recent window on 19% of states and replaces all of it on 26%—a state-dependent k profile that neither random allocation nor RGB similarity reproduces. Although training supervises only the k=1 interface, these multi-slot compositions sustain the closed-loop gains of Table 2: the token-gated adaptation applies per restored frame and generalizes across replacement multiplicity.

Closed-loop budget sweep. Figure 4 extends the closed loop along the deployment budget axis with single-round mobile campaigns at $B \in \{1, 2, 8\}$, sharing every frozen weight with the primary setting. The allocation gain is positive at every budget and, pooled over $B \in \{1, 2, 4, 8\}$ with equal budget weight, reaches +4.9pp (task-paired bootstrap; interval in the supplement), while single-budget cells remain trend evidence. On the memory-critical subset the margin is largest at the trained budgets and narrows at the untrained B=8, where the eight-frame recent window already reaches back far enough to cover part of the split's evidence—further support that the operative variable is where the evidence sits relative to the window.

What the restored slots must carry. Two controls keep selection, budget, and the two-pass interface intact but replace each restored frame's pixels with text: the turn's verbatim recorded response, or an OCR transcription of the same screenshot (both run on the selector+frozen deployment, whose pixel version scores 28.0% on this split). On the memory-critical split verbatim text collapses to the Recent-4 level (18.0% vs. 19.4%; −10.4pp below pixels, task-paired p=0.006), ruling out a generic "reminder" effect; OCR recovers part of the level (24.2%) but its +4.8pp margin over Recent-4 does not separate from zero, while pixel restoration's +8.6pp does. Full tables in the supplement.

Deployment cost of the second pass. The conditional second pass fires on 81–87% of audited mobile steps; the selector costs under 40 ms. Instrumented across budgets, the added model time per step is 2.0/2.3/2.9 s at B=1/2/4—8.1/8.7/10.5% of the measured per-step wall time—and falls to roughly 5% at B=8, where the second pass fires on only 39% of steps. On desktop the median end-to-end task overhead is 7%. The re-decision's real-time cost thus peaks near one-tenth of wall time at the primary budget; complete intervals are in the supplement.

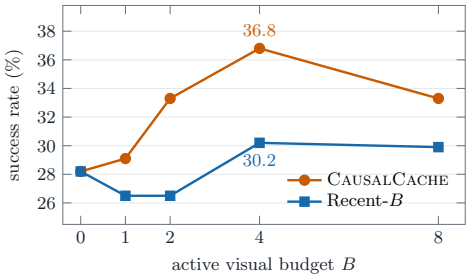

Figure 4: Closed-loop MobileWorld success along the budget axis (full roster). B=0/4: three-run means; other budgets single rounds (B=1/2 references: frozen dose curve; HGKV neutral on matched recent). Pooled over the sweep: +4.9pp (task-paired bootstrap; interval and full per-budget table in the supplement).

Limitations and Conclusion

The evidence assembles into three steps. First, the problem is fidelity allocation under a fixed visual budget: every event survives as text and only B regain pixels. Second, a frozen policy does not reliably separate a useful distant screenshot from the recent frame it would displace; HGKV learns to, inside a drift envelope that leaves matched-recent behavior untouched. Third, the learned selector reallocates profitably exactly where distant visual dependency exists: memory-critical tasks improve, controls show no detectable change. We control an active context budget, not archive capacity, and CausalCache reattaches rather than generates pixels. We claim only that some summarized events have greater conditional marginal utility than the recent realization they displace; pretraining contamination remains possible. In one sentence: CausalCache learns which summarized GUI events deserve to be seen again.

References

Boisvert, L.; Thakkar, M.; Gasse, M.; Caccia, M.; Le Sellier de Chezelles, T.; Cappart, Q.; Chapados, N.; Lacoste, A.; and Drouin, A. 2024. WorkArena++: Towards Compositional Planning and Reasoning-Based Common Knowledge Work Tasks. In *Advances in Neural Information Processing Systems*, volume 37.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT but Faster. In *The Eleventh International Conference on Learning Representations*.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. An Image Is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. In *Computer Vision – ECCV 2024*.

Cheng, K.; Sun, Q.; Chu, Y.; Xu, F.; YanTao, L.; Zhang, J.; and Wu, Z. 2024. SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9313–9332. Bangkok, Thailand: Association for Computational Linguistics.

Corallo, G.; and Papotti, P. 2024. FINCH: Prompt-Guided Key-Value Cache Compression for Large Language Models. *Transactions of the Association for Computational Linguistics*, 12: 1517–1532.

Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; and Su, Y. 2023. Mind2Web: Towards a Generalist Agent for the Web. In *Advances in Neural Information Processing Systems*, volume 36. Datasets and Benchmarks Track.

Drouin, A.; Gasse, M.; Caccia, M.; Laradji, I. H.; Del Verme, M.; Marty, T.; Vazquez, D.; Chapados, N.; and Lacoste, A. 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*, 11642–11662. PMLR.

Garg, D.; Caples, D.; Draguns, A.; Ravi, N.; Putta, P.; Garg, N.; Hebbar, P.; Joo, Y.; Gu, J.; London, C.; Schroeder de Witt, C.; and Motwani, S. 2025. REAL: Benchmarking Autonomous Agents on Deterministic Simulations of Real Websites. In *Advances in Neural Information Processing Systems*, volume 38.

Gutiérrez, B. J.; Shu, Y.; Gu, Y.; Yasunaga, M.; and Su, Y. 2024. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37.

Hong, W.; Wang, W.; Lv, Q.; Xu, J.; Yu, W.; Ji, J.; Wang, Y.; Wang, Z.; Dong, Y.; Ding, M.; and Tang, J. 2024. CogAgent: A Visual Language Model for GUI Agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14281–14290.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jiang, H.; Wu, Q.; Lin, C.-Y.; Yang, Y.; and Qiu, L. 2023. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13358–13376.

Jiang, H.; Wu, Q.; Luo, X.; Li, D.; Lin, C.-Y.; Yang, Y.; and Qiu, L. 2024. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1658–1677.

Koh, J. Y.; Lo, R.; Jang, L.; Duvvur, V.; Lim, M.; Huang, P.-Y.; Neubig, G.; Zhou, S.; Salakhutdinov, R.; and Fried, D. 2024. VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 881–905. Bangkok, Thailand: Association for Computational Linguistics.

Kong, Q.; Zhang, X.; Yang, Z.; Gao, N.; Liu, C.; Tong, P.; Cai, C.; Zhou, H.; Zhang, J.; Chen, L.; Liu, Z.; Hoi, S.; and Wang, Y. 2026. MobileWorld: Benchmarking Autonomous Mobile Agents in Agent-User Interactive and MCP-Augmented Environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6142–6167. San Diego, California, United States: Association for Computational Linguistics.

Lee, J.; Lee, Y.; Kim, J.; Kosiorek, A. R.; Choi, S.; and Teh, Y. W. 2019. Set Transformer: A Framework for Attention-Based Permutation-Invariant Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3744–3753. PMLR.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474.

Li, W.; Bishop, W.; Li, A.; Rawles, C.; Campbell-Ajala, F.; Tyamagundlu, D.; and Riva, O. 2024a. On the Effects of Data Scale on UI Control Agents. In *Advances in Neural Information Processing Systems*, volume 37.

Li, Y.; Huang, Y.; Yang, B.; Venkitesh, B.; Locatelli, A.; Ye, H.; Cai, T.; Lewis, P.; and Chen, D. 2024b. SnapKV: LLM Knows What You Are Looking for Before Generation. In *Advances in Neural Information Processing Systems*, volume 37.

Lin, K. Q.; Li, L.; Gao, D.; Yang, Z.; Wu, S.; Bai, Z.; Lei, S. W.; Wang, L.; and Shou, M. Z. 2025. ShowUI: One Vision-Language-Action Model for GUI Visual Agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19498–19508.

Liu, C.; Chen, L.; Zhou, H.; Zhang, X.; Kong, Q.; Tong, P.; Wang, W.; Yu, X.; Hoi, S.; and Wang, Y. 2026. What Memory Do GUI Agents Really Need? From Passive Records to Active Task-Driving States. arXiv:2606.31612.

Liu, Z.; Desai, A.; Liao, F.; Wang, W.; Xie, V.; Xu, Z.; Kyrillidis, A.; and Shrivastava, A. 2023. Scissorhands: Ex-

ploiting the Persistence of Importance Hypothesis for LLM KV Cache Compression at Test Time. In *Advances in Neural Information Processing Systems*, volume 36.

Liu, Z.; Yuan, J.; Jin, H.; Zhong, S.; Xu, Z.; Braverman, V.; Chen, B.; and Hu, X. 2024. KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 32332–32344. PMLR.

Lu, Q.; Shao, W.; Liu, Z.; Du, L.; Meng, F.; Li, B.; Chen, B.; Huang, S.; Zhang, K.; and Luo, P. 2025. GUIOdyssey: A Comprehensive Dataset for Cross-App GUI Navigation on Mobile Devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22404–22414.

Pan, Z.; Wu, Q.; Jiang, H.; Xia, M.; Luo, X.; Zhang, J.; Lin, Q.; Rühle, V.; Yang, Y.; Lin, C.-Y.; Zhao, H. V.; Qiu, L.; and Zhang, D. 2024. LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, 963–981.

Rawles, C.; Clinckemaillie, S.; Chang, Y.; Waltz, J.; Lau, G.; Fair, M.; Li, A.; Bishop, W. E.; Li, W.; Campbell-Ajala, F.; Toyama, D. K.; Berry, R. J.; Tyamagundlu, D.; Lillicrap, T. P.; and Riva, O. 2025. AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents. In *The Thirteenth International Conference on Learning Representations*.

Rawles, C.; Li, A.; Rodriguez, D.; Riva, O.; and Lillicrap, T. 2023. AndroidInTheWild: A Large-Scale Dataset for Android Device Control. In *Advances in Neural Information Processing Systems*, volume 36.

Sarch, G. H.; Jang, L.; Tarr, M. J.; Cohen, W. W.; Marino, K.; and Fragkiadaki, K. 2024. VLM Agents Generate Their Own Memories: Distilling Experience into Embodied Programs of Thought. In *Advances in Neural Information Processing Systems*, volume 37, 75942–75985.

Shapley, L. S. 1953. A Value for n-Person Games. In Kuhn, H. W.; and Tucker, A. W., eds., *Contributions to the Theory of Games II*, 307–317. Princeton, New Jersey: Princeton University Press.

Shi, Y.; Li, J.; Zhang, L.; Dongfang, Z.; Wu, B.; Tao, S.; Yan, Y.; Qin, C.; Liu, W.; Lin, Z.; Li, H.; Huang, Y.; Dai, S.; Hei, Y.; Ding, Y.; Li, X.; Wang, S.; Xu, C.; Liu, J.; Ma, X.; Zheng, Z.; Zhang, X.; Wang, B.; Yang, N.; Wu, J.; Tian, L.; Li, C.; and Hu, X. 2026. AndroTMem: From Interaction Trajectories to Anchored Memory in Long-Horizon GUI Agents. arXiv:2603.18429.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36.

Srivastava, S. S. 2026. Causal Intervention-Based Memory Selection for Long-Horizon LLM Agents. arXiv:2605.17641.

Xiao, G.; Tian, Y.; Chen, B.; Han, S.; and Lewis, M. 2024. Efficient Streaming Language Models with Attention Sinks. In *The Twelfth International Conference on Learning Representations*.

Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; Hua, T. J.; Cheng, Z.; Shin, D.; Lei, F.; Liu, Y.; Xu, Y.; Zhou, S.; Savarese, S.; Xiong, C.; Zhong, V.; and Yu, T. 2024. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. In *Advances in Neural Information Processing Systems*, volume 37. Datasets and Benchmarks Track.

Xu, H.; Zhang, X.; Liu, H.; Wang, J.; Zhu, Z.; Zhou, S.; Hu, X.; Gao, F.; Cao, J.; Wang, Z.; Chen, Z.; Liao, J.; Zheng, Q.; Zeng, J.; Xu, Z.; Bai, S.; Lin, J.; Zhou, J.; and Yan, M. 2026. Mobile-Agent-v3.5: Multi-Platform Fundamental GUI Agents. arXiv:2602.16855.

Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2025. VisionZip: Longer Is Better but Not Necessary in Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19792–19802.

Yao, S.; Chen, H.; Yang, J.; and Narasimhan, K. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In *Advances in Neural Information Processing Systems*, volume 35.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*.

Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Póczos, B.; Salakhutdinov, R.; and Smola, A. J. 2017. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30, 3391–3401.

Zeng, Z.; Hua, H.; Zou, B.; Cai, M.; Feris, R.; and Luo, J. 2026. MementoGUI: Learning Agentic Multimodal Memory Control for Long-Horizon GUI Agents. arXiv:2605.18652.

Zhang, Z.; Sheng, Y.; Zhou, T.; Chen, T.; Zheng, L.; Cai, R.; Song, Z.; Tian, Y.; Ré, C.; Barrett, C.; Wang, Z.; and Chen, B. 2023. H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36.

Zhao, A.; Huang, D.; Xu, Q.; Lin, M.; Liu, Y.-J.; and Huang, G. 2024. ExpeL: LLM Agents Are Experiential Learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19632–19642.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. MemoryBank: Enhancing Large Language Models with Long-Term Memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19724–19731.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; Alon, U.; and Neubig, G. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *The Twelfth International Conference on Learning Representations*.