RTWorld: Benchmarking Real-Time Game Playing Agent for Open-Ended Tasks in Digital Environment

Anonymous authors
Paper under double-blind review

Abstract

Generalist visual agents must perceive, plan, and act while the digital world continues to change, yet most evaluations implicitly let the world wait for model inference. This convention obscures whether an agent can turn a correct decision into a timely action. We introduce RTWorld, a benchmark for real-time game-playing agents built from 31 open-ended browser games with pixel-only observations, mixed keyboard and pointer control, and verifiable native outcomes. A fail-closed runtime preserves the live world clock, timestamps the full observation–inference–action path, rejects invalid or stale plans, binds each episode to one model replica, and records replayable hash-verified evidence. We evaluate 13 vision-language systems using common seeds and complement the main benchmark with controlled-delay, pause/live, delay-shape, and temporal-context interventions. Our results show that static model quality, response availability, and closed-loop action freshness are distinct factors: model rankings can change when inference no longer pauses the environment, and faster responses do not by themselves guarantee useful control. RTWorld provides a reproducible foundation for studying agents that must reason and act before the opportunity has passed.

1 Introduction

Generalist agents are increasingly expected to operate software through the same visual interface available to people. Games are a revealing testbed for this goal: perception, memory, planning, and control must remain coordinated over long horizons, while scores, deaths, deadlines, and changing world state make failures observable. Existing platforms and learned players have driven progress from Atari and 3D navigation to open-ended worlds (Bellemare et al., 2013; Beattie et al., 2016; Mnih et al., 2015; Fan et al., 2022; Wang et al., 2023a). Recent multimodal benchmarks broaden the range of games addressable through a common model interface (Zhang et al., 2025; Ouyang et al., 2026; Paglieri et al., 2025). Yet their usual turn-based abstraction omits a basic property of deployed interaction: the environment does not wait for the model to finish thinking.

This omission changes the capability being measured. An observation can be accurate when captured and obsolete when its action arrives; a sound plan can miss a platform, steer toward an object that has moved, or continue after the episode has changed state. Pausing during inference erases these failures and can favor a slower policy that would act too late in the live world. Server latency alone is also insufficient: capture, encoding, queueing, validation, and scheduling determine the age of the command that reaches the environment. Although real-time reinforcement learning treats delay as part of the controlled process (Travnik et al., 2018; Ramstedt & Pal, 2019), this principle is not yet a first-class contract for generalist vision-language agent evaluation.

A faithful benchmark must therefore combine consequentially changing tasks with end-to-end timing and auditable failure semantics. It must record when a frame was captured, when inference started and ended, whether the returned plan still belongs to the active interaction epoch, and how old each command is at dispatch. It must also separate policy failure from serving or evaluator failure. Without replica identity, immutable traces, synchronized video, and explicit exclusions, apparent gains can arise from hidden pauses, stale actions, mid-episode failover, or incomplete evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

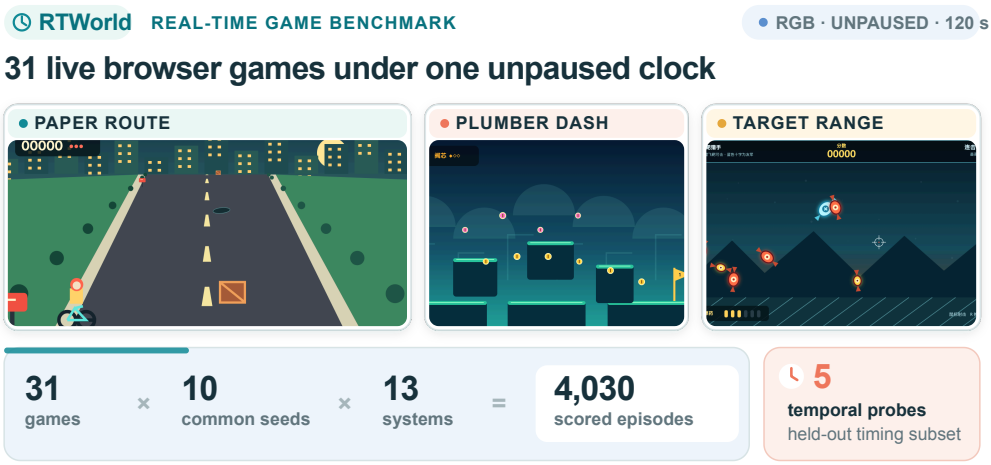

Figure 1: RTWorld at a glance. Arcade32 retains 31 source-audited browser games whose state changes consequentially without input, spanning mixed keyboard and pointer control. The repeated benchmark evaluates 13 deployed systems on ten common seeds per game; five games form the assigned-delay intervention subset. Statistics and category counts are derived from the frozen game catalog, accepted model roster, and released cohort manifests.

We introduce RTWorld, a benchmark and execution framework designed around these requirements. Its Arcade32 suite retains 31 source-audited browser games spanning platforming, driving, shooting, navigation, reaction, planning, and continuous pointer control. Agents receive only RGB frames and public timing metadata and emit a versioned action language; DOM state, game objects, structured scores, and future frames remain hidden. The world clock advances through capture, inference, admission, and dispatch. Schema, interruption-epoch, and source-age checks prevent invalid or stale plans from controlling the game. Each episode is bound to one model replica and yields hash-linked inputs, outputs, traces, results, and video. Latency and freshness thus become observable properties of the evaluated system rather than experimental noise.

Our evaluation separates broad capability measurement from causal timing diagnosis. We compare 13 deployed vision-language systems on all games using ten common seeds, then study a frozen temporal subset with assigned response delays, pause/live pairs, alternative delay shapes, and controlled visual context. The strongest static player need not dominate across timing conditions, and pausing can materially alter measured capability. The relevant quantity is the interaction among response availability, stale-plan admission, and active control, motivating a joint quality–timing profile rather than a hardware-dependent speed leaderboard.

Our contributions are:

- 31 source-audited browser games under a common pixel-only interface;
- a fail-closed continuous-time runtime with end-to-end freshness, stale-response rejection, replica affinity, and replayable evidence; and
- a common-seed 13-system benchmark plus controlled interventions that isolate where real-time competence is gained or lost.

The full comparison covers 13 systems; primary five-game interventions cover three Qwen systems with one exploratory MiniCPM replication. Dustline is implemented but not evaluated here. RTWorld measures the deployed agent loop, not hardware-normalized throughput or general computer-use safety.



Figure 2: The 31 evaluated Arcade32 environments. Every tile is a real RGB frame captured from the audited browser game; models receive this same pixel-only interface rather than DOM or privileged game state. Accent marks identify the five-game assigned-delay intervention subset.

2 Related Work

Game environments and learned players. ALE, Gym, DeepMind Lab, ViZDoom, PySC2, NetHack, TextWorld, and Procgen made game interaction reproducible across arcade, 3D, strategy, text, and procedurally generated settings (Bellemare et al., 2013; Brockman et al., 2016; Beattie et al., 2016; Kempka et al., 2016; Vinyals et al., 2017; Kuttler et al., 2020; Cote et al., 2019; Cobbe et al., 2020). Specialized agents reached landmark results in Atari, Go, StarCraft, Dota, and world-model control (Mnih et al., 2013; 2015; Silver et al., 2016; Vinyals et al., 2019; OpenAI et al., 2019; Hafner et al., 2025). Minecraft and multi-world agents broadened this agenda toward open-ended, instructed behavior (Fan et al., 2022; Baker et al., 2022; Wang et al., 2023a; Lifshitz et al., 2023; Wang et al., 2023b; Li et al., 2025; SIMA Team et al., 2024; 2025; Bakhtin et al., 2022).

Generalist game-agent benchmarks. Recent work tests language and multimodal agents across live games, navigation, long demonstrations, physical reasoning, and verifiable tasks (Paglieri et al., 2025; Hu et al., 2025a;b; Ahn et al., 2025; Park et al., 2025; Magne et al., 2025; Wang et al., 2025b; Zheng et al., 2025b;a; Zhang et al., 2026; Waytowich et al., 2024; Nasir et al., 2024; Ruoss et al., 2025; Taesiri et al., 2025; Tong et al., 2025; Wang et al., 2025a; Xie et al., 2025; Xu et al., 2025; Sun et al., 2025; Tan et al., 2025; Zhao et al., 2023). VideoGameBench directly exposes real-time latency and supplies a paused variant; GameWorld emphasizes standardized browser tasks and state-verifiable scoring (Zhang et al., 2025; Ouyang et al., 2026). RTWorld instead measures the closed-loop timing mechanism with common seeds, assigned response delays, pause/live pairs, stale-plan admission, and immutable episode evidence.

Computer-use agents. WebShop, Mind2Web, WebArena, VisualWebArena, WorkArena, BrowserGym, WebGym, OSWorld, and OSWorld 2.0 progressively expand grounded web and desktop control (Yao et al., 2022; Deng et al., 2023; Zhou et al., 2024; Koh et al., 2024; Drouin et al., 2024; Le Sellier De Chezelles et al., 2024; Bai et al., 2026; Xie et al., 2024; Yuan et al., 2026). SeeAct, AssistGUI, Cradle, ALFWorld, and OSWorld-MCP study visual grounding, desktop automation, embodied interaction, and tool invocation (Zheng et al., 2024; Gao et al., 2023; Tan et al., 2024; Shridhar et al., 2021; Jia et al., 2025).

Agent scaffolds, perception, and timing. ReAct, Reflexion, and Toolformer motivate structured reasoning, memory, and tool use (Yao et al., 2023; Shinn et al., 2023; Schick et al., 2023); visual

Table 1: Default contracts of related benchmarks. ✓: required; ∘: supported or studied, but not required throughout; —: absent. RGB is native rendered control; Dose is assigned response delay; Pair is common-seed pause/live; Gate is admission-time stale rejection; Audit is a released replay trajectory or stronger evidence.

Benchmark	RGB	Live	Dose	Pair	Gate	Audit
OSWorld 2.x (Yuan et al., 2026)	✓	∘	—	—	—	∘
GameWorld (Ouyang et al., 2026)	✓	∘	—	—	—	∘
VideoGameBench Full v3 (Zhang et al., 2025)	✓	✓	—	—	—	∘
BrowserGym / WebArena (Le Sellier De Chezelles et al., 2024; Zhou et al., 2024)	∘	—	—	—	—	∘
RTWorld	✓	✓	✓	✓	✓	✓

diagnostic studies caution that fluent multimodal outputs can mask basic perception failures (Tong et al., 2024; Rahmanzadehgervi et al., 2025; Al-Tahan et al., 2024; Guan et al., 2024). Real-time RL treats delay as part of the controlled process, whereas MLPerf measures inference systems without downstream control consequences (Travnik et al., 2018; Ramstedt & Pal, 2019; Reddi et al., 2020). Table 1 summarizes the closest benchmark contracts.

3 Benchmark Design

3.1 Environment families

Arcade32 contains 31 original Canvas games spanning platforming, driving, shooting, reaction, planning, and continuous control. Dustline is a browser first-person combat task with a write-only relative-pointer action transport. Four deterministic mini-games expose explicit 10–60 Hz timing requirements for controlled latency studies. All scored episodes end on death or at a hard 120-second timeout, at which point the native score is frozen.

The name Arcade32 is retained for path compatibility with the original 32-game collection. Before evaluation, we removed one input-driven puzzle whose board did not change autonomously; the retained real-time catalog therefore contains 31 games. Every retained game passed a source-level audit requiring no-input evolution to alter survival, reachability, score, or a time-bounded opportunity during active play. Dustline specifies and tests the environment and transport contract, but this paper does not report model results on Dustline; all reported VLM scores are from Arcade32 or the five-game Arcade32 temporal subset.

Arcade32 is intentionally heterogeneous rather than a collection of visual reskins: its catalog includes platforming, driving, fixed- and free-flight shooting, landing, maze navigation, reaction, puzzle, and continuous pointer tasks. Each game has a source-reviewed control manifest and three disclosed reporting thresholds. Thresholds are joined only after an episode; they are not included in policy context. These are designer-set ordinal anchors frozen before model outcomes, not fitted human percentiles or learned normalization constants. We therefore retain every native score and use threshold normalization only for within-game aggregation.

3.2 Observation and action contract

The policy receives only rendered RGB frames and public timing metadata. It does not receive DOM state, game objects, or structured score, terminal, or evaluator channels; human-visible HUD pixels remain in the RGB frame. Arcade actions use three primitives—normalized pointer movement, click, and bounded-duration key presses—under a versioned per-game allowlist. Invalid, late, wrong-epoch, and stale outputs fail closed; inputs are released on replacement, interruption, and shutdown.

3.3 Unpaused execution and deployment affinity

Rendering, physics, hazards, and terminal deadlines continue during capture, inference, validation, and dispatch. Each episode acquires an exclusive lease on one declared model replica. The deployment and server-configuration hashes must remain stable until the terminal record, trace, result shard, and release record are durable. Mid-episode failover invalidates the attempt.

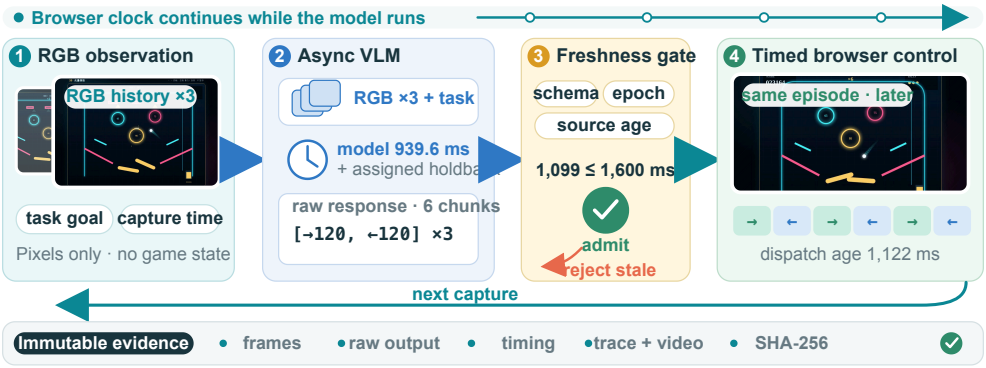

Figure 3: The fail-closed RTWorld runtime on a continuously advancing world clock. The overlay binds one accepted Vector Pinball response (seed 4007) to its measured latency, admission age, and action program; the later frame is from the same audited episode. Assigned holdback precedes schema, epoch, and source-age checks. Fresh plans become bounded action chunks, stale plans are rejected, and every stage enters immutable episode evidence.

3.4 Evidence contract

Every accepted attempt stores the task, prompt, action schema, control manifest, observation candidate, deployment binding, lossless model-input archive, raw event trace, terminal result, synchronized video, and hashes. A completion marker is written last. Diagnostic, candidate-search, and scored cohorts occupy different roots and cannot be pooled.

4 Execution Architecture

The evaluator separates observation capture, model planning, action admission, dispatch, and evidence publication. Capture continues while a request is in flight. The strict planner permits at most one request in flight; a response carries the source frame sequence, capture timestamp, and interruption epoch. The executor checks schema, epoch, and source age when admitting a returned plan, then recomputes and records age at each chunk dispatch without a second maximum-age gate. Admitting a replacement cancels the old plan, advances the epoch, and releases held keys and pointer buttons; shutdown disables admission and releases all inputs.

Each scored episode obtains an exclusive lease from a direct model-replica adapter. The lease records the served model, tokenizer hash, deployment ID, server configuration, and supported wire protocol. Requests carry replica affinity, and pre- and post-episode probes must agree. Load-balancer health alone is therefore insufficient: a cohort is accepted only after authenticated multimodal inference and stable episode-level identity.

Artifacts are first written to staging, synchronized, copied into an immutable attempt directory, and individually SHA-256 hashed. The completion record is the final write. A separate lease-release record remains outside the immutable attempt so that failed cleanup cannot be hidden by a successful result. The cohort index hashes its manifest, task plan, aggregate, report, and quarantine record. Repeated-cohort analysis revalidates these raw hashes instead of trusting a precomputed success flag.

5 Metrics

For game g and reporting threshold $\tau_{g,3}$, we report native score s_g and raw threshold-normalized score $n_g = s_g/\tau_{g,3}$, while retaining the T3/T6/T9 attainment indicators separately. The primary benchmark index clips n_g to $[0, 3]$ so a single high-scale game cannot dominate; raw values remain available. We do not average native scores across games. For each action plan we report schema admission, admitted and applied actions, canceled actions, and stale rejections. Plan occupancy is the sum of all dispatched action durations divided by observed live episode time, clipped to one.

Active-control occupancy uses the same definition but excludes `press([], duration)`, which is an explicit live-time wait. Dispatch event coverage is reported separately so pointer-heavy control is not assigned zero coverage by construction; stall ratio is one minus plan occupancy.

For freshness threshold τ, fresh-control coverage $F(\tau)$ is the fraction of live episode time occupied by dispatched actions whose source observation age at dispatch is at most τ. We report F at 100, 250, 500, 1000, and 2000 ms, plus the corresponding fraction of instantaneous dispatch events so pointer-heavy tasks are not assigned zero coverage by construction.

Latency percentiles are recomputed from raw completed events pooled across accepted attempts and therefore estimate a completed-event-weighted conditional distribution; terminal in-flight calls can be right-censored. Per-run percentiles are never averaged. We additionally report equal-episode summaries and suppress dispatch tails with fewer than 100 events. On deterministic tasks, a latency curve is summarized by trapezoidal AUC over the frozen base delays. Task score and a deadline-aware real-time score remain separate, so a policy cannot compensate for unsafe timing with a high task score.

For model m and game g, let $Q_{m,g}$ be its bounded score when the world is paused only during inference. The primary assigned-availability stress summary integrates the policy-specific score profile through 1000 ms over $u = d/1000$ as $\mathrm{LevelAUC}_{\leq 1\mathrm{s}} = \int_0^1 Y(u)du$. The corresponding LossAUC is $\int_0^1 [Y(0) - Y(u)]du$ and is signed: negative values denote average improvement over zero holdback. Full-grid AUC through 2000 ms remains a secondary deployed-system-plus-safety-gate estimand because the 1400-ms maximum-plan-age gate mechanically blocks the two largest assignments. The gate can still reject a sub-threshold assignment when model latency plus holdback exceeds 1400 ms. An outcome-unseen frozen reference-controller gate conditionally enables a cross-fitted factorization of paused competence and normalized temporal load. Because that all-games gate fails, we suppress the factorization rather than choosing a scale post hoc; realized action age and freshness remain post-treatment diagnostics.

6 Experimental Protocol

6.1 Controlled latency matrix

We evaluate oracle, pixel-only scripted, random, and no-op policies on all four deadline tasks at 0, 40, 80, 120, 250, and 500 ms base delay. Each cell uses 20 seeds. Policies paired by task, seed, and base delay receive the same jitter and spike schedule: uniform 20 ms jitter, 5% spike probability, and 500 ms spike magnitude. The oracle reads privileged task state and is only a controller upper bound; it is never compared as a deployable visual agent.

6.2 Strict visual-language evaluation

Vision-language systems use the same action grammar, game-specific prompt, token cap, observation candidates, episode duration, and direct-replica lease contract. Evaluated checkpoints first enter a score-blind interface and breadth screen, then run all 31 games on ten common held-out seeds. The runner injects a versioned 32-bit seed before game JavaScript executes; this controls declared game randomness but not wall-clock scheduling or bit-exact replay. A 31-game coverage round is only an infrastructure gate.

Candidate selection and confirmatory evaluation use disjoint seeds. Diagnostics, search, ablations, and confirmatory rounds occupy separate roots. A round is accepted only when all 31 attempts are strict-valid and evidence-complete, all leases and videos are complete, and infrastructure failures, policy-invalid attempts, admitted stale responses, pauses, and quarantine entries are all zero. Infrastructure-excluded rounds are retained and reported; policy-invalid rounds are not silently replaced.

Table 11 reports exact checkpoints, serving precision, and tensor parallelism; corresponding model cards are cited for Qwen2.5-VL, Qwen3.5, Qwen3.8, Qwen3-VL, Gemma-4, InternVL3.5, MiniCPM-V-4.5, Ovis2.5, and Mistral Small 3.1 (Qwen Team, 2025; 2026a;b;c; Google DeepMind, 2026; OpenGVLab, 2025; OpenBMB, 2025; AIDC-AI, 2025; Mistral AI, 2025). The abbreviation "Qwen3-VL-235B" denotes `Qwen/Qwen3-VL-235B-A22B-Instruct-FP8`; weights are FP8 and

runtime is BF16. All accepted systems run on NVIDIA H100 80GB GPUs; TP in the table is also the GPU count per replica. Every scored episode holds one exclusive direct-replica lease, with no concurrent scored episode and at most one planner request in flight. A real multimodal request warms and verifies each service before scoring. Candidate manifests send a 2048-token request cap and the deployments advertise the same generation ceiling. The separate 256-token `arcade-functions-v1` field is the tokenizer-derived bound required to serialize any legal action, not the request cap; all scored deployments satisfy both bounds. The artifact records the exact serving image/flags, configuration hash, prompt/image token counts when exposed, and request count per episode. Because stacks and TP allocations differ, latency rows describe deployed systems rather than hardware-normalized architecture effects; we make no cost or throughput claim.

For browser-task lower controls, no-op and versioned uniform-random policies run all 31 games on ten common held-out base seeds under the same unpaused terminal and evidence contracts. Five frozen task-aware pixel controllers run the five-game ablation subset on its five seeds. These controllers receive only RGB and public controls; they are reported as transparent references, not as learned agents or privileged upper bounds.

6.3 Paused and controlled-delay interventions

On the five frozen temporal games, we pair strict unpaused episodes with a separate paused-during-inference protocol under common models and seeds. The paused track is intentionally non-deployable and is never pooled into the main benchmark. We report paired score gaps, rank agreement, and bootstrap rank-reversal probabilities.

Before VLM outcomes, an oracle and independent frozen RGB controller calibrate six delays on disjoint seeds. The optional RET_g@2000 scale is enabled only if every game passes frozen headroom, endpoint-degradation, and cross-controller agreement gates. Failure retains uniform LevelAUC/LossAUC and forbids per-game scale selection.

For assigned-delay response curves, a model-independent schedule holds each complete response for an additional 0, 250, 500, 1000, 1500, or 2000 ms before admission, while the world and active action chunk continue. The three Qwen anchors were frozen across sub-second natural-latency tiers before response outcomes were inspected. Because holdback begins only after generation, natural model latency is not treated as exogenous. Delay order follows a deterministic rotation within each seed block and reverses between blocks. This limits gross order imbalance but is not random assignment; estimates are frozen-schedule assigned-delay contrasts. The ten blocks are not exactly position-balanced: mean within-block positions range from 2.1 to 2.9 across the six delays. Delay versus planned absolute execution index has Pearson $r = 0$ and Spearman $\rho = -0.0023$ in each 300-cell root. Adding game and seed fixed effects plus a linear position term preserves every raw contrast sign; the largest absolute adjustment is 0.0002, 0.0035, and 0.0065 for the 7B, 32B, and Qwen3.8 anchors. These are drift-sensitivity checks: the bootstrap quantifies seed variation, planned index is only a time proxy, and nonlinear drift or carryover remains possible rather than randomized away. A separate experiment compares IID and burst-clustered permutations of the same pre-generated delay multiset; the deterministic delay suite is separate and not pooled. The full curve includes the deployed 1400-ms maximum-plan-age safety gate. Consequently, the 1500 and 2000 ms assignments are system-level policy endpoints at which positive inference latency makes admission impossible; they do not isolate task-dynamic sensitivity. The primary LevelAUC and signed LossAUC stop at 1000 ms, avoiding holdbacks that alone exceed the gate but not removing model-latency–gate interactions; full-grid values are secondary safety-gate estimands reported with admission, rejection, and freshness.

6.4 Compatibility and statistics

Pooling requires identical non-model source, prompt, schema, control, candidate, server, and evidence contracts. Per-game reports retain every native score and T3/T6/T9 outcome. Bootstrap intervals resample complete seed blocks; paired comparisons use common seeds. The controlled and main paired panels provide only ten seed blocks, and the temporal-context and OFAT studies only five; their intervals quantify conditional seed variation for the fixed games and serving stack, have coarse small-cluster support, and are not population-level coverage guarantees. Ten repetitions do not characterize extreme tails.

Table 2: Main 31-game results over ten common seeds (310 accepted episodes per system). Score is the macro clipped threshold-normalized mean with a seed-block bootstrap 95% interval. Latencies pool completed events; dispatch P95 is suppressed when fewer than 100 actions were dispatched.

Model	Positive games	Score [95% CI]	Model P50	Model P95	Dispatch P95
Qwen3.8-27B	21.3	0.367 [0.329, 0.407]	490.9	851.9	1850.4
Qwen2.5-VL-32B	16.6	0.118 [0.100, 0.139]	297.0	818.2	1980.0
Qwen2.5-VL-7B	12.4	0.072 [0.061, 0.083]	130.4	184.8	580.2
Qwen2.5-VL-3B	7.6	0.056 [0.044, 0.075]	117.3	199.1	510.0
Qwen3.5-4B	16.2	0.259 [0.231, 0.288]	143.4	188.0	868.0
Qwen3.5-27B	20.8	**0.440 [0.411, 0.468]**	414.8	848.6	1669.3
Qwen3-VL-235B	19.7	0.255 [0.217, 0.292]	299.6	684.6	1606.7
Qwen2.5-VL-72B	20.0	0.290 [0.257, 0.328]	531.2	744.5	1399.6
Gemma-4-31B	4.0	0.027 [0.022, 0.031]	49657.2	50124.5	–
InternVL3.5-8B	14.3	0.206 [0.169, 0.243]	433.8	600.5	851.0
MiniCPM-V-4.5	14.1	0.214 [0.192, 0.246]	131.7	233.0	616.8
Ovis2.5-9B	15.2	0.206 [0.164, 0.245]	182.1	361.8	711.0
Mistral-Small-3.1-24B	14.0	0.202 [0.181, 0.225]	286.8	941.9	2053.2

Table 3: Paired paused-versus-live gaps on five temporal games. Scores are macro clipped T3-normalized means over common seed blocks. The gap is paused minus strict unpaused with a paired-bootstrap 95% interval. No model pair exhibits a simultaneous-band-supported rank reversal.

Model	Unpaused	Paused	Gap [95% CI]
Qwen2.5-VL-7B	0.040	0.043	0.003 [-0.003, 0.010]
Qwen2.5-VL-32B	0.057	0.085	0.027 [0.000, 0.072]
Qwen3.8-27B	0.349	0.550	0.201 [0.032, 0.368]

6.5 Outcome-unseen frozen ablations

The primary temporal–context design uses eight corners from frames $\{1, 3\}$, image boxes $\{320 \times 240, 512 \times 320\}$, and cadence $\{150, 300\}$ ms, plus the canonical three-frame, 480×360, 200 ms center. Maximum plan age is fixed at 1400 ms throughout. Each of three models uses the same five games and common held-out seeds 7301–7305 (225 episodes per model; 675 total). Complete seed blocks are the bootstrap units, and the model–configuration interaction is a paired difference in differences. No resource-cost model or Pareto criterion is defined, so this is a design surface rather than an efficiency frontier. The model-independent execution order rotates configurations by seed block and alternates direction to limit warmup/time drift; it is balanced but not randomized, so we report controlled assigned-configuration contrasts rather than a randomized causal estimand.

A secondary one-factor-at-a-time study retains 800 and 2000 ms plan-age variants without pooling them into the factorial. Natural latency, dispatch age, and freshness remain descriptive; the online protocol never installs a response already stale at admission.

7 Results

Table 2 is the main repeated-seed result; the full event counts and all intervention tables appear in Appendix A. The 13 systems span Qwen2.5-VL, Qwen3-VL, Qwen3.5, Qwen3.8, Gemma, InternVL, MiniCPM-V, Ovis, and Mistral Small (Qwen Team, 2025; 2026c;a;b; Google DeepMind, 2026; OpenGVLab, 2025; OpenBMB, 2025; AIDC-AI, 2025; Mistral AI, 2025). Qwen3.5-27B ranks first at 0.440 [0.411, 0.468], followed by Qwen3.8-27B at 0.367 [0.329, 0.407]; the independent-seed random baseline is 0.197 [0.181, 0.217]. The leader is unchanged under cap 1, T3/T6/T9, all 62 ±20% T3 perturbations, and 30 of 31 leave-one-game-out rankings (minimum Spearman agreement 0.907). Table latency is event-pooled; for example, Qwen2.5-VL-32B has pooled P95 818.2 ms but an equal-episode mean P95 of 2309.9 ms, reported with event counts in the appendix.

At 1000 ms holdback, response admission falls to 92.5%, 10.9%, and 0.2% for the 7B, 32B, and Qwen3.8 anchors, and fresh active-control coverage is zero. The 1400-ms gate makes 1500/2000-ms responses structurally ineligible, so the primary AUC stops at 1000 ms but remains a deployment-system total effect. Pausing raises Qwen3.8 by 0.201, yet its paired-t interval [-0.005, 0.407] and exact sign-flip $p = 0.0547$ warrant a suggestive, not confirmatory, claim; no simultaneous-band rank

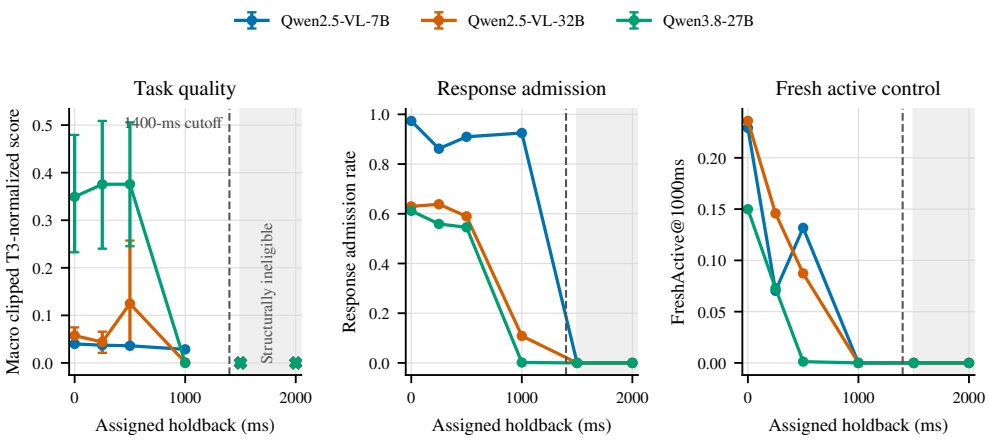

Figure 4: Policy-specific assigned-availability stress profile under the fixed 1400-ms source-age gate at plan admission. Left: score curves connect only the eligible assignments through 1000 ms; the 1500 and 2000 ms endpoints are marked separately because the holdback alone exceeds the gate. Center: response admission by assigned holdback. Right: fresh active-control coverage at 1000 ms; FreshActive is a post-hoc occupancy diagnostic, not an additional dispatch gate. Error bars resample complete ten-seed blocks; request events are measurements, not independent statistical samples.

reversal is supported. The OFAT plan-age shifts, all burst-minus-IID intervals, and the cross-model MiniCPM replication are also non-significant. Eight of 24 temporal-context interactions exclude zero without multiplicity correction; all cells and robustness checks are reported in the appendix.

8 Analysis

We distinguish malformed actions, stale responses, coverage stalls, infrastructure failures, and legal actions without progress. The immutable case in Figure 11 was selected by a frame-blind census after aggregate metrics were frozen and is descriptive, not a substitute for repeated seeds. Because all five reference-normalization gates fail, we suppress that factorization and report unconditional LevelAUC/LossAUC. Controlled delay estimates later response availability under the fixed scheduler, not a causal effect of model size or naturally varying service latency.

9 Limitations and Broader Impact

Results are conditional on ten seeds, the serving stack, prompt, tokenizer, and observation policy; they do not characterize extreme latency tails. Primary interventions cover three Qwen systems plus one exploratory MiniCPM replication, the temporal-context schedule is balanced but not randomized, Dustline is unevaluated, and interface-gate failures remain unscored rather than zero. The suite measures digital rather than physical control. Releases must remove credentials and await source/asset redistribution clearance; seeded game randomness does not control GPU, scheduling, or capture timing.

10 Conclusion

RTWorld treats timing as deployed interactive behavior. Its unpaused, assigned-delay, paired-pause, and hash-bound protocols identify the latency–staleness–control chain without attributing natural service latency causally. The evidence is deployment-conditioned, with no supported rank reversal or burst-ordering effect; broader cross-family replication remains essential.

REFERENCES

Jaewoo Ahn, Junseo Kim, Heeseung Yun, Jaehyeon Son, Dongmin Park, Jaewoong Cho, and Gunhee Kim. FlashAdventure: A benchmark for GUI agents solving full story arcs in diverse adventure games. In *Proceedings of EMNLP*, 2025.

AIDC-AI. Ovis2.5-9B model card. `https://huggingface.co/AIDC-AI/Ovis2.5-9B`, 2025. Accessed 2026-09-24.

Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. UniBench: Visual reasoning requires rethinking vision-language beyond scaling. *arXiv preprint arXiv:2408.04810*, 2024.

Hao Bai, Alexey Taymanov, Tong Zhang, Aviral Kumar, and Spencer Whitehead. WebGym: Scaling training environments for visual web agents with realistic tasks. *arXiv preprint arXiv:2601.02439*, 2026.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. In *Advances in Neural Information Processing Systems*, 2022.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, et al. DeepMind Lab. *arXiv preprint arXiv:1612.03801*, 2016.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Marc-Alexandre Cote, Akos Kadar, Xingdi Yuan, Ben Kybartas, Tavian Barnes, et al. TextWorld: A learning environment for text-based games. In *Computer Games Workshop*, 2019.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*, 2023.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Advances in Neural Information Processing Systems*, 2022.

Difei Gao, Lei Ji, Zechen Bai, Mingyu Ouyang, Peiran Li, et al. AssistGUI: Task-oriented desktop graphical user interface automation. *arXiv preprint arXiv:2312.13108*, 2023.

Google DeepMind. Gemma-4-31B model card. `https://huggingface.co/google/gemma-4-31B`, 2026. Accessed 2026-09-23.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *Nature*, 640:647–653, 2025.

Lanxiang Hu, Mingjia Huo, Yuxuan Zhang, Haoyang Yu, Eric P. Xing, Ion Stoica, Tajana Rosing, Haojian Jin, and Hao Zhang. lmgame-Bench: How good are LLMs at playing games? *arXiv preprint arXiv:2505.15146*, 2025a.

Lanxiang Hu, Qiyu Li, Anze Xie, Nan Jiang, Ion Stoica, Haojian Jin, and Hao Zhang. GameArena: Evaluating LLM reasoning through live computer games. *arXiv preprint arXiv:2412.06394*, 2025b.

Hongrui Jia, Jitong Liao, Xi Zhang, Haiyang Xu, Tianbao Xie, et al. OSWorld-MCP: Benchmarking MCP tool invocation in computer-use agents. *arXiv preprint arXiv:2510.24563*, 2025.

Michal Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaskowski. ViZDoom: A doom-based AI research platform for visual reinforcement learning. In *IEEE Conference on Computational Intelligence and Games*, 2016.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Heinrich Kuttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktaschel. The NetHack learning environment. *arXiv preprint arXiv:2006.13760*, 2020.

Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, et al. The BrowserGym ecosystem for web agent research. *arXiv preprint arXiv:2412.05467*, 2024.

Muyao Li, Zihao Wang, Kaichen He, Xiaojian Ma, and Yitao Liang. JARVIS-VLA: Post-training large-scale vision language models to play visual games with keyboards and mouse. *arXiv preprint arXiv:2503.16365*, 2025.

Shalev Lifshitz, Keiran Paster, Harris Chan, Jimmy Ba, and Sheila McIlraith. STEVE-1: A generative model for text-to-behavior in Minecraft. In *Advances in Neural Information Processing Systems*, 2023.

Loic Magne, Anas Awadalla, Guanzhi Wang, Yinzhen Xu, Joshua Belofsky, et al. NitroGen: An open foundation model for generalist gaming agents. *arXiv preprint*, 2025.

Mistral AI. Mistral-Small-3.1-24B-Instruct-2503 model card. `https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503`, 2025. Accessed 2026-09-24.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Muhammad Umair Nasir, Steven James, and Julian Togelius. GameTraversalBenchmark: Evaluating planning abilities of large language models through traversing 2d game maps. *arXiv preprint arXiv:2410.07765*, 2024.

OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

OpenBMB. MiniCPM-V-4.5 model card. `https://huggingface.co/openbmb/MiniCPM-V-4_5`, 2025. Accessed 2026-09-23.

OpenGVLab. InternVL3.5-8B model card. `https://huggingface.co/OpenGVLab/InternVL3_5-8B`, 2025. Accessed 2026-09-23.

Mingyu Ouyang, Siyuan Hu, Kevin Qinghong Lin, Hwee Tou Ng, and Mike Zheng Shou. Gameworld: Towards standardized and verifiable evaluation of multimodal game agents. *arXiv preprint arXiv:2604.07429*, 2026. Accessed 2026-09-26.

Davide Paglieri, Bartlomiej Cupial, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, et al. BALROG: Benchmarking agentic LLM and VLM reasoning on games. In *International Conference on Learning Representations*, 2025.

Dongmin Park, Minkyu Kim, Beongjun Choi, Junhyuck Kim, Keon Lee, et al. Orak: A foundational benchmark for training and evaluating LLM agents on diverse video games. *arXiv preprint*, 2025.

Qwen Team. Qwen2.5-VL model collection. `https://huggingface.co/collections/Qwen/qwen25-vl`, 2025. Model cards; accessed 2026-09-23.

Qwen Team. Qwen3.5 model collection. `https://huggingface.co/collections/Qwen/qwen35`, 2026a. Model cards; accessed 2026-09-23.

Qwen Team. Qwen3.8-27B model card. `https://huggingface.co/Qwen/Qwen3.8-27B`, 2026b. Accessed 2026-09-23.

Qwen Team. Qwen3-VL-235B-A22B-Instruct-FP8 model card. `https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct-FP8`, 2026c. Accessed 2026-09-23.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*, 2025.

Simon Ramstedt and Chris Pal. Real-time reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vijay Janapa Reddi et al. Mlperf inference benchmark. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture*, pp. 446–459, 2020.

Anian Ruoss, Fabio Pardo, Harris Chan, Bonnie Li, Volodymyr Mnih, and Tim Genewein. LMAct: A benchmark for in-context imitation learning with long multimodal demonstrations. *arXiv preprint arXiv:2412.01441*, 2025.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, et al. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

SIMA Team, Maria Abi Raad, Arun Ahuja, Catarina Barros, Frederic Besse, et al. Scaling instructable agents across many simulated worlds. *arXiv preprint arXiv:2404.10179*, 2024.

SIMA Team, Adrian Bolton, Alexander Lerchner, Alexandra Cordell, et al. SIMA 2: A generalist embodied agent for virtual worlds. *arXiv preprint arXiv:2512.04797*, 2025.

Haochen Sun, Shuwen Zhang, Lujie Niu, Lei Ren, Hao Xu, et al. Collab-Overcooked: Benchmarking and evaluating large language models as collaborative agents. *arXiv preprint arXiv:2502.20073*, 2025.

Mohammad Reza Taesiri, Abhijay Ghildyal, Saman Zadtootaghaj, Nabajeet Barman, and Cor-Paul Bezemer. VideoGameQA-Bench: Evaluating vision-language models for video game quality assurance. In *NeurIPS Datasets and Benchmarks Track*, 2025.

Weihao Tan, Wentao Zhang, Xinrun Xu, Haochong Xia, Ziluo Ding, Boyu Li, Bohan Zhou, Junpeng Yue, Jiechuan Jiang, Yewen Li, et al. Cradle: Empowering foundation agents towards general computer control. *arXiv preprint arXiv:2403.03186*, 2024.

Weihao Tan, Xiangyang Li, Yunhao Fang, Heyuan Yao, Shi Yan, et al. Lumine: An open recipe for building generalist agents in 3d open worlds. *arXiv preprint arXiv:2511.08892*, 2025.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, et al. Game-RL: Synthesizing multimodal verifiable game data to boost VLMs' general reasoning. *arXiv preprint arXiv:2505.13886*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. *arXiv preprint arXiv:2401.06209*, 2024.

Jaden B. Travnik, Kory W. Mathewson, Patrick M. Pilarski, and Richard S. Sutton. Reactive reinforcement learning in asynchronous environments. *Frontiers in Robotics and AI*, 5:79, 2018.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander S. Vezhnevets, et al. PySC2: StarCraft II learning environment. *arXiv preprint arXiv:1708.04782*, 2017.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michael Mathieu, Andrew Dudzik, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575:350–354, 2019.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Xinyu Wang, Bohan Zhuang, and Qi Wu. Are large vision language models good game players? *arXiv preprint arXiv:2503.02358*, 2025a.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, et al. JARVIS-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv preprint arXiv:2311.05997*, 2023b.

Zihao Wang, Xujing Li, Yining Ye, Junjie Fang, Haoming Wang, et al. Game-TARS: Pretrained foundation models for scalable generalist multimodal game agents. *arXiv preprint*, 2025b.

Nicholas R. Waytowich, Devin White, MD Sunbeam, and Vinicius G. Goecks. Atari-GPT: Benchmarking multimodal large language models as low-level policies in atari games. *arXiv preprint arXiv:2408.15950*, 2024.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Tuan Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, 2024.

Yunfei Xie, Yinsong Ma, Shiyi Lan, Alan Yuille, Junfei Xiao, and Chen Wei. Play to generalize: Learning to reason through game play. *arXiv preprint arXiv:2506.08011*, 2025.

Xinrun Xu, Pi Bu, Ye Wang, Borje F. Karlsson, Ziming Wang, et al. DeepPHY: Benchmarking agentic VLMs on physical reasoning. *arXiv preprint arXiv:2508.05405*, 2025.

Shunyu Yao, Howard Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, 2022.

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

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, et al. OSWorld 2.0: Benchmarking computer use agents on long-horizon real-world tasks. *arXiv preprint arXiv:2606.29537*, 2026. Version 2; benchmark release osworld-v2.1; accessed 2026-09-26.

Alex L. Zhang, Thomas L. Griffiths, Karthik R. Narasimhan, and Ofir Press. Videogamebench: Can vision-language models complete popular video games? *arXiv preprint arXiv:2505.18134*, 2025. arXiv:2505.18134v3, revised 2026-05-14; accessed 2026-09-24.

Kuan Zhang, Dongchen Liu, Qiyue Zhao, Jinkun Hou, Xinran Zhang, et al. GameVerse: Can vision-language models learn from video-based reflection? *arXiv preprint arXiv:2603.06656*, 2026.

Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, et al. See and think: Embodied agent in virtual environment. *arXiv preprint arXiv:2311.15209*, 2023.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. GPT-4V(ision) is a generalist web agent, if grounded. In *International Conference on Machine Learning*, 2024.

Xiangxi Zheng, Linjie Li, Zhengyuan Yang, Ping Yu, Alex Jinpeng Wang, et al. V-MAGE: A game evaluation framework for assessing vision-centric capabilities in multimodal large language models. *arXiv preprint arXiv:2504.06148*, 2025a.

Xinyue Zheng, Haowei Lin, Kaichen He, Zihao Wang, Qiang Fu, et al. MCU: An evaluation framework for open-ended game agents. In *International Conference on Machine Learning*, 2025b.

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

A Complete Tables and Additional Figures

Reproducibility statement. The artifact records source/protocol versions, deterministic seeds, commands, configuration hashes, immutable attempts, raw latency events, and fail-closed validators. Traces and videos will be released after license clearance. Plans were locally timestamped and hash-frozen, not externally preregistered. Controlled and paused runners differ only by one exact, manually reviewed allowlisted hash pair; all others are rejected, without claiming machine-verified semantic equivalence.

Executor state machine and clock semantics. Figure 3 summarizes this state machine. Each loop capture records an RGB frame and source timestamp. A cadence tick starts one in-flight request; the worker timestamps model completion, applies assigned post-response holdback, and queues the response. The executor drains that queue and admits only a schema-valid, same-epoch response whose source age is within `max_plan_age_ms`. Admission releases held inputs, cancels the old plan remainder, advances the epoch, and installs replacement chunks. At each chunk boundary, held keys are released before the next chunk is applied. Source age is recomputed and recorded at every dispatch, but the maximum-age rejection guard is applied at admission, not again for later chunks. Raw evaluator timestamps for capture, request, response, holdback, dispatch, and finalization use Python `time.monotonic_ns()` in one process clock domain. The paired paused protocol stops only the browser page clock; Python's monotonic clock and the model request continue. Derived active-time fields subtract the tracked pause interval. During finalization, the `finally` path disables admission and releases all inputs; the planner is then joined and remaining responses are drained. Results and artifact hashes are durably published before `COMPLETE.json`, the final write inside the immutable attempt. Lease release is then written to a sibling operational `LEASE_RELEASE.json`.

Table 4: Repeated strict Arcade32 results. Values are means across complete common-seed rounds; latency percentiles are pooled from raw events. Dispatch latency is conditional on an action being dispatched, so its event count may differ from the model-latency count; pooled dispatch P95 is suppressed when fewer than 100 dispatch events are available. Episode-balanced sensitivity is reported separately.

(a) Outcome quality

Model	Rounds	Positive [95% CI]	Clipped norm. [95% CI]
Qwen3.8-27B	10	21.300 [20.700, 21.900]	0.367 [0.329, 0.407]
Qwen2.5-VL-32B	10	16.600 [15.600, 17.600]	0.118 [0.100, 0.139]
Qwen2.5-VL-7B	10	12.400 [11.700, 13.000]	0.072 [0.061, 0.083]
Qwen2.5-VL-3B	10	7.600 [6.900, 8.300]	0.056 [0.044, 0.075]
Qwen3.5-4B	10	16.200 [15.300, 16.900]	0.259 [0.231, 0.288]
Qwen3.5-27B	10	20.800 [20.100, 21.500]	0.440 [0.411, 0.468]
Qwen3-VL-235B	10	19.700 [18.700, 20.600]	0.255 [0.217, 0.292]
Qwen2.5-VL-72B	10	20.000 [19.300, 20.700]	0.290 [0.257, 0.328]
Gemma-4-31B	10	4.000 [4.000, 4.000]	0.027 [0.022, 0.031]
InternVL3.5-8B	10	14.300 [13.700, 14.900]	0.206 [0.169, 0.243]
MiniCPM-V-4.5	10	14.100 [13.500, 14.800]	0.214 [0.192, 0.246]
Ovis2.5-9B	10	15.200 [14.500, 15.900]	0.206 [0.164, 0.245]
Mistral-Small-3.1-24B	10	14.000 [13.300, 14.600]	0.202 [0.181, 0.225]

(b) System latency

Model	Events (model/dispatch)	Model P50 (ms)	Model P95 (ms)	Dispatch P95 (ms)
Qwen3.8-27B	23023/35224	490.9	851.9	1850.4
Qwen2.5-VL-32B	22014/26274	297.0	818.2	1980.0
Qwen2.5-VL-7B	65544/47263	130.4	184.8	580.2
Qwen2.5-VL-3B	51464/27853	117.3	199.1	510.0
Qwen3.5-4B	47156/36636	143.4	188.0	868.0
Qwen3.5-27B	26758/37021	414.8	848.6	1669.3
Qwen3-VL-235B	35913/38616	299.6	684.6	1606.7
Qwen2.5-VL-72B	27353/56133	531.2	744.5	1399.6
Gemma-4-31B	657/8	49657.2	50124.5	–
InternVL3.5-8B	37290/32556	433.8	600.5	851.0
MiniCPM-V-4.5	79465/68470	131.7	233.0	616.8
Ovis2.5-9B	47210/52767	182.1	361.8	711.0
Mistral-Small-3.1-24B	39708/40550	286.8	941.9	2053.2

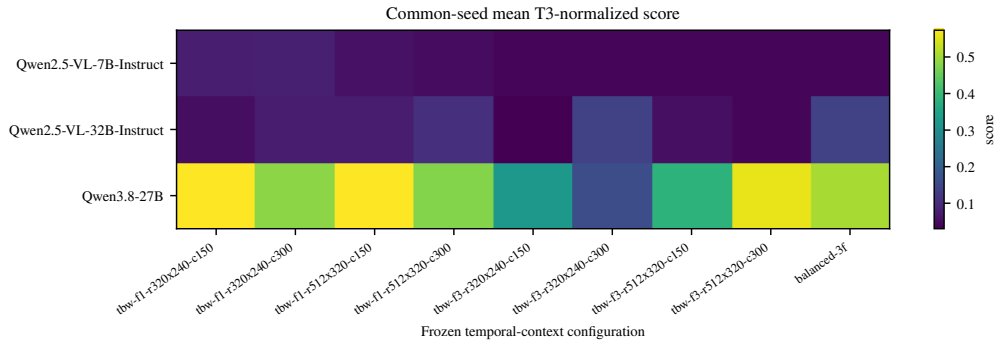

Figure 5: Three-model temporal-context design surface over eight corners and the canonical center. Values average the five games within each common seed block; natural serving latency is not a causal exposure.

Table 5: Policy-specific assigned-availability stress results on five temporal games under the fixed 1400-ms source-age gate at plan admission. Score and AUC brackets are 95% paired seed-block bootstrap intervals. LevelAUC (0–1 s) and signed LossAUC (0–1 s) are the primary stress summaries; negative LossAUC denotes improvement over zero holdback. Full-grid AUCs integrate over $u = d/2000$ and are secondary deployed-system-plus-safety-gate summaries. The 1-s cells can still hit the gate after adding model latency; 1500/2000-ms responses are ineligible by construction. Admit is response admission; Plan/Active are all-plan and non-wait control occupancy; FreshActive is a post-hoc active-control occupancy diagnostic whose dispatch age is at most 1000 ms, not an additional execution gate; AgeReject is max-plan-age rejection.

(a) Stress scores

Model	Score@0	Level AUC (0–1 s)	Loss AUC (0–1 s)	Level AUC (0–2 s)	Loss AUC (0–2 s)
Qwen2.5-VL-7B	0.040 [0.037, 0.044]	0.035 [0.034, 0.036]	0.005 [0.002, 0.009]	0.021 [0.020, 0.022]	0.019 [0.016, 0.023]
Qwen2.5-VL-32B	0.057 [0.038, 0.075]	0.065 [0.032, 0.117]	-0.008 [-0.056, 0.028]	0.033 [0.016, 0.059]	0.025 [-0.001, 0.047]
Qwen3.8-27B	0.349 [0.233, 0.479]	0.278 [0.220, 0.336]	0.071 [-0.048, 0.189]	0.139 [0.110, 0.168]	0.210 [0.094, 0.333]

(b) Mechanism diagnostics

Model	Admit@1s	Plan@1s	Active@1s	FreshActive@1s	AgeReject@1.5s
Qwen2.5-VL-7B	92.5%	5.4%	4.0%	0.0%	97.8%
Qwen2.5-VL-32B	10.9%	8.4%	6.5%	0.0%	74.9%
Qwen3.8-27B	0.2%	0.5%	0.5%	0.0%	92.0%

Table 6: Small-cluster sensitivity from frozen ten-seed evidence. Panel (a) reports paired pause gaps and between-model gap interactions. Panel (b) reports per-seed primary assigned-delay AUCs over 0–1000 ms. Student-t intervals and delete-one-seed-block (LOO) ranges use seed blocks as the sampling unit. Exact two-sided sign-flip tests are shown only for signed contrasts; they are not meaningful for nonnegative LevelAUC.

(a) Pause gaps and interactions

Estimand	Estimate	Paired-t 95% CI	Exact p	LOO range
Qwen2.5-VL-7B pause gap	0.0030	[-0.0053, 0.0113]	0.6250	[0.0000, 0.0051]
Qwen2.5-VL-32B pause gap	0.0273	[-0.0213, 0.0758]	0.1621	[0.0064, 0.0317]
Qwen3.8-27B pause gap	0.2013	[-0.0047, 0.4073]	0.0547	[0.1592, 0.2512]
Qwen3.8-27B − Qwen2.5-VL-7B	0.1983	[-0.0080, 0.4046]	0.0586	[0.1558, 0.2461]
Qwen3.8-27B − Qwen2.5-VL-32B	0.1740	[-0.0158, 0.3638]	0.0684	[0.1284, 0.2195]

(b) Primary assigned-delay AUC sensitivity

Model	Quantity	Estimate	Student-t 95% CI	Exact p	LOO range
Qwen2.5-VL-7B	LevelAUC	0.0350	[0.0334, 0.0365]	–	[0.0345, 0.0353]
Qwen2.5-VL-7B	LossAUC	0.0050	[0.0011, 0.0090]	0.0020	[0.0037, 0.0056]
Qwen2.5-VL-32B	LevelAUC	0.0652	[0.0098, 0.1207]	–	[0.0417, 0.0715]
Qwen2.5-VL-32B	LossAUC	-0.0078	[-0.0601, 0.0446]	0.8574	[-0.0153, 0.0129]
Qwen3.8-27B	LevelAUC	0.2785	[0.2078, 0.3491]	–	[0.2620, 0.2939]
Qwen3.8-27B	LossAUC	0.0707	[-0.0751, 0.2166]	0.2930	[0.0334, 0.1095]

Exact p values enumerate all $2^{10} = 1024$ sign assignments, retain zero differences, and count ties as extreme. LevelAUC is nonnegative, so a sign-exchangeability null about zero is not meaningful and its exact-p entries are omitted. Recomputed primary and full-grid AUC means match the published controlled-delay table at displayed precision; full-grid means also match the frozen aggregates numerically. All checks are descriptive and unadjusted.

Table 7: Exploratory cross-family replication on MiniCPM-V-4.5 using the frozen five-game, ten-seed design. Intervals resample or model complete seed blocks; exact tests are descriptive and unadjusted.

Estimand	Mean	Bootstrap 95% CI	Paired-t 95% CI	Exact p
$\text{LevelAUC}_{0--1000}$	0.0355	[0.0315, 0.0406]	[0.0299, 0.0412]	–
$\text{LossAUC}_{0--1000}$	0.0055	[-0.0032, 0.0147]	[-0.0054, 0.0163]	0.2812
Paused − zero holdback	-0.0037	[-0.0102, 0.0021]	[-0.0112, 0.0039]	0.2734

LevelAUC and signed LossAUC integrate the 0, 250, 500, and 1000-ms assigned-holdback curve after five-game macro-averaging. The pause gap is paused minus the paired zero-holdback score. LevelAUC has no sign-flip test because a nonnegative level is not sign-exchangeable about zero. This panel is exploratory cross-family replication only.

Table 8: Environment families and their roles in the evaluation.

Family	Tasks	Rate	Primary purpose
Arcade32	31	display-rate	Broad RGB control across 31 distinct game mechanics and mixed keyboard/pointer interfaces.
Dustline	1	display-rate	Tactical first-person control with a strict write-only relative-pointer transport.
Deadline tasks	4	10–60 Hz	Deterministic latency cliffs, replay, and common-random-number system characterization.

Table 9: Source-reviewed task taxonomy for all 31 retained Arcade32 games. "Temporal" marks membership in the five-game assigned-delay subset. Regime labels summarize dominant mechanics; every game still evolves without input and uses the common 120-s wall-clock horizon.

Regime	Games	Dominant control / horizon	Temporal subset
Platform and navigation	plumber-dash, road-frog, tower-hop, gold-runner, barrel-climb, cube-hopper	discrete keys with hazards/checkpoints	–
Shooting and combat	jungle-strike, alien-ranks, asteroid-patrol, star-swarm, bug-centipede, sky-defender, tank-arena, dojo-duel	sustained multi-key control	asteroid-patrol
Reactive aiming/interception	target-range, missile-guard, mole-combo, paddle-duel, brick-blitz	pointer or rapid positional correction	target-range
Continuous dynamics	sunset-racer, alpine-slalom, lunar-touchdown, paper-route, sky-flap	steering/thrust with short-lived opportunities	paper-route, sky-flap
Grid and planning under a clock	block-stack, bomb-grid, gem-chain, snake-rush, maze-munch	discrete plans while hazards/timers advance	block-stack
Physics and mixed arcade	vector-pinball, bubble-rescue	timed key sequences and collision dynamics	–

Table 10: Blind action baselines on Arcade32 (10 common seeds per game). The primary normalized metric is clipped per game to $[0, 3]$ before macro averaging; intervals resample complete 31-game seed blocks.

Policy	Strict/evidence	Positive rate [95% CI]	Clipped norm. [95% CI]
No-op	310/310	0.129 [0.129, 0.129]	0.032 [0.029, 0.036]
Uniform random	310/310	0.690 [0.671, 0.710]	0.197 [0.181, 0.217]
Random − no-op	310 pairs	–	0.165 [0.149, 0.184]

Table 11: Exact checkpoints for accepted Arcade32 systems. A dash revision means the registry did not preserve a public revision; the accepted run still binds model, tokenizer, and server-configuration hashes. FP8/BF16 distinguishes checkpoint weights from serving dtype. All systems use `request_max_tokens`=2048 and `server_max_new_tokens`=2048; `wire_action_token_bound`=256 is the distinct wire-format action bound.

Reported name	Repository/checkpoint	Revision	Weights/runtime	TP	Cells
Qwen3.8-27B	Qwen/Qwen3.8-27B	–	bfloat16	2	310
Qwen2.5-VL-32B	Qwen/Qwen2.5-VL-32B-Instruct	–	bfloat16	2	310
Qwen2.5-VL-7B	Qwen/Qwen2.5-VL-7B-Instruct	–	bfloat16	1	310
Qwen2.5-VL-3B	Qwen/Qwen2.5-VL-3B-Instruct	–	bfloat16	1	310
Qwen3.5-4B	Qwen/Qwen3.5-4B	–	bfloat16	1	310
Qwen3.5-27B	Qwen/Qwen3.5-27B	–	bfloat16	1	310
Qwen3-VL-235B	Qwen/Qwen3-VL-235B-A22B-Instruct-FP8	–	FP8/bfloat16	4	310
Qwen2.5-VL-72B	Qwen/Qwen2.5-VL-72B-Instruct	–	bfloat16	4	310
Gemma-4-31B	google/gemma-4-31B	–	bfloat16	1	310
InternVL3.5-8B	OpenGVLab/InternVL3_5-8B	9bb6a56a	bfloat16	1	310
MiniCPM-V-4.5	openbmb/MiniCPM-V-4_5	daef484c	bfloat16	1	310
Ovis2.5-9B	AIDC-AI/Ovis2.5-9B	d73b2283	safetensors/bfloat16	1	310
Mistral-Small-3.1-24B	mistralai/Mistral-Small-3.1-24B-Instruct-2503	68faf511	safetensors/bfloat16	1	310

Table 12: Score-blind interface exclusions. Diagnostic completions are not benchmark episodes or zero scores. Formal scoring starts only after at least one schema-valid admitted action. Diagnostics use the same `request_max_tokens`=2048, `server_max_new_tokens`=2048, and `wire_action_token_bound`=256 contract as accepted systems.

Candidate	Repository/checkpoint	Responses	Valid	Admitted	Gate reason
Phi-4-multimodal-instruct	microsoft/Phi-4-multimodal-instruct	12	0	0	No schema-valid action
SmolVLM2-2.2B-Instruct	HuggingFaceTB/SmolVLM2-2.2B-Instruct	4	0	0	All responses length-truncated
Muse-Glimmer-30B	meta-models/Muse-Glimmer-30B	17	0	0	No schema-valid action
LFM2-VL-1.6B	LiquidAI/LFM2-VL-1.6B	12	0	0	No schema-valid action
Idefics3-8B	HuggingFaceM4/Idefics3-8B-Llama3	31	0	0	No schema-valid action
Molmo2-4B	allenai/Molmo2-4B	3	0	0	All responses length-truncated
LLaVA-OneVision-Qwen2-7B	llava-hf/llava-onevision-qwen2-7b-ov-hf	3	0	0	No schema-valid action
Pixtral-12B	mistralai/Pixtral-12B-2409	41	0	0	No schema-valid action
DeepSeek-VL2-Small	deepseek-ai/deepseek-vl2-small	19	0	0	No schema-valid action
GLM-4.1V-9B-Thinking	zai-org/GLM-4.1V-9B-Thinking	8	0	0	No schema-valid action

Table 13: Score-definition sensitivity on the frozen 13-model Arcade32 panel (4,030 existing strict-valid episodes; no new rollouts). Entries are metric value with tie-aware rank in parentheses. Positive is the mean number of positive-score games per 31-game seed block; threshold columns are episode attainment rates; rank percentile is threshold-free and computed within each game–seed cell.

Model	Cap 3	Cap 1	Positive	T3	T6	T9	Rank pct.
Qwen3.5-27B	0.440 (1.0)	0.274 (1.0)	20.800 (2.0)	0.113 (1.0)	0.052 (1.0)	0.032 (1.0)	0.671 (1.0)
Qwen3.8-27B	0.367 (2.0)	0.238 (2.0)	21.300 (1.0)	0.106 (2.0)	0.026 (2.5)	0.003 (4.0)	0.619 (3.0)
Qwen2.5-VL-72B	0.290 (3.0)	0.216 (3.0)	20.000 (3.0)	0.074 (3.0)	0.003 (9.0)	0.000 (9.0)	0.622 (2.0)
Qwen3.5-4B	0.259 (4.0)	0.162 (5.0)	16.200 (6.0)	0.071 (4.0)	0.026 (2.5)	0.019 (2.0)	0.490 (7.0)
Qwen3-VL-235B	0.255 (5.0)	0.189 (4.0)	19.700 (4.0)	0.061 (5.0)	0.006 (7.0)	0.000 (9.0)	0.579 (4.0)
MiniCPM-V-4.5	0.214 (6.0)	0.140 (8.0)	14.100 (9.0)	0.048 (8.0)	0.006 (7.0)	0.000 (9.0)	0.474 (9.0)
Ovis2.5-9B	0.206 (7.0)	0.141 (7.0)	15.200 (7.0)	0.048 (8.0)	0.016 (4.0)	0.016 (3.0)	0.511 (5.0)
InternVL3.5-8B	0.206 (8.0)	0.144 (6.0)	14.300 (8.0)	0.055 (6.0)	0.006 (7.0)	0.000 (9.0)	0.486 (8.0)
Mistral-Small-3.1-24B	0.202 (9.0)	0.137 (9.0)	14.000 (10.0)	0.048 (8.0)	0.010 (5.0)	0.000 (9.0)	0.463 (10.0)
Qwen2.5-VL-32B	0.118 (10.0)	0.112 (10.0)	16.600 (5.0)	0.010 (10.0)	0.000 (11.5)	0.000 (9.0)	0.499 (6.0)
Qwen2.5-VL-7B	0.072 (11.0)	0.072 (11.0)	12.400 (11.0)	0.003 (11.5)	0.000 (11.5)	0.000 (9.0)	0.417 (11.0)
Qwen2.5-VL-3B	0.056 (12.0)	0.051 (12.0)	7.600 (12.0)	0.003 (11.5)	0.000 (11.5)	0.000 (9.0)	0.366 (12.0)
Gemma-4-31B	0.027 (13.0)	0.027 (13.0)	4.000 (13.0)	0.000 (13.0)	0.000 (11.5)	0.000 (9.0)	0.303 (13.0)

Table 14: Ranking robustness relative to the primary cap-3 T3-normalized ranking. LOO omits one complete game at a time; OAT perturbs one game's T3 by ±20%; global scaling multiplies all T3 values by 0.8–1.2. These are deterministic post-hoc stress tests, not empirical threshold calibration.

Stress test	Scenarios	Min. ρ	Max. rank shift	Top-1 retained
Leave-one-game-out	31	0.907	5.0	96.8%
One-at-a-time T3 ±20%	62	0.978	2.0	100.0%
Global T3 scaling	5	0.989	1.0	100.0%

Table 15: Fresh plan occupancy versus active-control occupancy at a 500-ms freshness threshold, recomputed from 4,030 frozen traces. Active control excludes `press([], duration)` waits. Dispatch columns are fractions of all dispatch events that are fresh and fresh-active, respectively. Entries are percent seed-block means with 95% bootstrap intervals.

Model	Plan occ.	Active occ.	Dispatch	Active dispatch
Gemma-4-31B	0.0 [0.0,0.0]%	0.0 [0.0,0.0]%	0.0 [0.0,0.0]%	0.0 [0.0,0.0]%
InternVL3.5-8B	0.9 [0.7,1.1]%	0.9 [0.7,1.1]%	18.4 [17.0,19.8]%	11.1 [9.8,12.4]%
MiniCPM-V-4.5	13.8 [11.5,16.2]%	13.8 [11.4,16.2]%	89.7 [87.2,92.0]%	66.1 [64.8,67.5]%
Mistral-Small-3.1-24B	39.1 [37.9,40.3]%	13.1 [12.2,13.9]%	33.3 [32.0,34.6]%	10.3 [9.4,11.3]%
Ovis2.5-9B	58.6 [57.7,59.4]%	58.6 [57.7,59.5]%	77.2 [75.3,79.2]%	58.3 [57.3,59.9]%
Qwen2.5-VL-32B	8.9 [7.8,9.9]%	7.0 [6.0,7.9]%	13.1 [10.8,15.5]%	11.2 [9.2,13.2]%
Qwen2.5-VL-3B	10.4 [9.2,11.7]%	10.4 [9.2,11.7]%	73.8 [71.4,76.5]%	51.4 [49.6,53.4]%
Qwen2.5-VL-72B	10.0 [8.8,11.0]%	7.4 [6.5,8.1]%	16.9 [16.3,17.5]%	8.6 [7.9,9.2]%
Qwen2.5-VL-7B	19.5 [17.3,21.8]%	13.4 [11.8,14.8]%	89.1 [88.0,90.3]%	38.1 [37.1,39.2]%
Qwen3-VL-235B	8.1 [7.2,9.3]%	7.0 [6.0,8.1]%	27.6 [26.2,29.1]%	22.8 [21.4,24.2]%
Qwen3.5-27B	44.4 [42.7,46.1]%	35.7 [34.7,36.8]%	39.4 [38.2,40.8]%	34.0 [33.1,35.0]%
Qwen3.5-4B	36.8 [35.0,38.7]%	36.4 [34.2,38.5]%	85.9 [84.5,87.3]%	63.2 [61.4,65.1]%
Qwen3.8-27B	1.3 [0.9,1.6]%	0.3 [0.2,0.4]%	1.6 [1.1,2.1]%	0.7 [0.4,1.0]%

Table 16: Episode-balanced latency sensitivity from the same 4,030 frozen traces, split into model and dispatch panels. Each episode contributes one within-episode quantile before seed-block averaging; equal-game columns first average eligible episodes within each game and then weight games equally. Brackets are 95% bootstrap intervals over ten complete 31-game seed blocks. Events and episodes with events are explicit. Terminal-inflight counts are completed model responses whose completion timestamp follows the recorded terminal observation. True right-censored request counts are not recoverable from these traces because the frozen schema has no auditable request-submitted event, so none are estimated. A dash suppresses dispatch P95 when there are fewer than 100 events, 20 episodes with events, or five seed blocks.

(a) Model latency

				P50 (ms)		P95 (ms)	
Model	Events	Eps.	Term.-inflight	Ep. mean 95% CI	Equal-game	Ep. mean 95% CI	Equal-game
Gemma-4-31B	657	310	309	49395.0 [48984.7,49736.1]	49395.0	49395.3 [48971.2,49731.3]	49395.3
InternVL3.5-8B	37290	310	269	1031.9 [855.7,1205.8]	1031.9	1973.1 [1663.7,2235.3]	1973.1
MiniCPM-V-4.5	79465	310	223	364.9 [233.3,496.4]	364.9	386.0 [256.0,515.7]	386.0
Mistral-Small-3.1-24B	39708	310	262	441.3 [435.9,446.7]	441.3	532.6 [521.4,543.3]	532.6
Ovis2.5-9B	47210	310	241	888.7 [714.7,1058.7]	888.7	920.5 [747.6,1086.2]	920.5
Qwen2.5-VL-32B	22014	310	273	1685.3 [1140.5,2318.2]	1685.3	2309.9 [1675.8,3034.3]	2309.9
Qwen2.5-VL-3B	51464	310	221	1057.4 [961.8,1170.9]	1057.4	1635.0 [1539.0,1719.8]	1635.0
Qwen2.5-VL-72B	27353	310	254	515.8 [509.2,522.4]	515.8	563.3 [558.2,567.7]	563.3
Qwen2.5-VL-7B	65544	310	209	140.3 [139.8,140.8]	140.3	162.5 [162.0,163.0]	162.5
Qwen3-VL-235B	35913	310	261	367.1 [360.3,374.6]	367.1	428.2 [423.4,432.7]	428.2
Qwen3.5-27B	26758	310	264	482.4 [469.0,496.2]	482.4	580.1 [568.7,593.0]	580.1
Qwen3.5-4B	47156	310	226	148.9 [146.9,150.8]	148.9	398.6 [296.5,502.5]	398.6
Qwen3.8-27B	23023	310	255	532.8 [521.3,543.9]	532.8	1040.5 [856.0,1233.6]	1040.5

(b) Dispatch latency

			P50 (ms)		P95 (ms)	
Model	Events	Eps.	Ep. mean 95% CI	Equal-game	Ep. mean 95% CI	Equal-game
Gemma-4-31B	8	8	905.1 [844.2,970.3]	884.9	–	–
InternVL3.5-8B	32556	300	592.6 [584.4,601.6]	594.6	654.2 [638.4,672.6]	655.2
MiniCPM-V-4.5	68470	305	324.3 [292.1,356.0]	324.0	400.3 [354.8,445.2]	399.3
Mistral-Small-3.1-24B	40550	310	745.5 [737.4,753.0]	745.5	992.1 [979.5,1005.1]	992.1
Ovis2.5-9B	52767	286	362.1 [356.4,368.1]	369.2	461.1 [452.7,468.9]	471.4
Qwen2.5-VL-32B	26274	277	781.1 [767.7,794.4]	834.2	1113.1 [1079.8,1155.0]	1176.7
Qwen2.5-VL-3B	27853	245	367.1 [364.3,369.6]	366.4	446.4 [438.5,453.7]	443.4
Qwen2.5-VL-72B	56133	310	810.5 [802.6,819.9]	810.5	1091.1 [1076.0,1103.6]	1091.1
Qwen2.5-VL-7B	47263	302	388.4 [385.5,391.3]	386.2	472.9 [466.6,480.1]	469.7
Qwen3-VL-235B	38616	310	684.8 [674.3,695.5]	684.8	926.6 [912.7,939.6]	926.6
Qwen3.5-27B	37021	310	688.2 [676.9,699.4]	688.2	939.0 [926.1,952.6]	939.0
Qwen3.5-4B	36636	310	408.0 [401.5,414.8]	408.0	535.9 [517.7,553.6]	535.9
Qwen3.8-27B	35224	310	924.5 [906.9,941.9]	924.5	1300.3 [1276.3,1326.4]	1300.3

Table 17: Sensitivity of the deterministic assigned-holdback contrasts to planned execution order. Adjusted regressions include game and seed fixed effects and a linear within-seed position term; intervals resample complete seed blocks. This is a frozen-schedule drift sensitivity check.

Model	Pearson r	Spearman ρ	Max. contrast shift
Qwen2.5-VL-7B	0.0000	-0.0023	0.0002
Qwen2.5-VL-32B	0.0000	-0.0023	0.0035
Qwen3.8-27B	0.0000	-0.0023	0.0065

Across each 300-cell plan, assigned delay versus absolute order has Pearson $r = 0.0000$ and Spearman $\rho = -0.0023$. The ten seed blocks are not exactly position-balanced; nonlinear drift and carry-over remain possible.

Table 18: Latency-robustness AUC on deterministic deadline-controlled tasks. Oracle is privileged and is reported only as a controller upper bound.

Game	noop	oracle	pixel	random
lane_intercept	0.284	0.314	0.313	0.232
projectile_dodge	0.002	0.003	0.003	0.001
pulse_parry	0.000	0.019	0.019	0.009
signal_tap	0.000	0.921	0.921	0.292

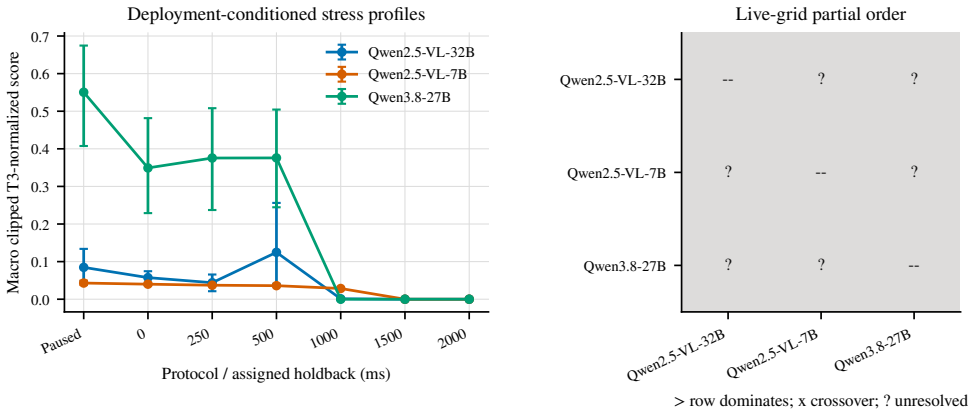

Figure 6: Deployment-conditioned temporal model ordering. Left: each deployment's common-seed stress profile from paused evaluation through increasing assigned post-response holdback. Right: the live-grid partial order using simultaneous seed-block bootstrap bands; point-estimate crossings do not qualify.

Table 19: Paired common-seed differences with 95% paired-bootstrap intervals. Differences are left minus right; positive values favor the left model. The 78 model-pair rows contain two exploratory intervals each and are not multiplicity adjusted.

Left	Right	Seeds	Δ positive [95% CI]	Δ clipped norm. [95% CI]
Qwen3.8-27B	Qwen2.5-VL-32B	10	4.700 [3.300, 6.200]	0.249 [0.218, 0.282]
Qwen3.8-27B	Qwen2.5-VL-7B	10	8.900 [8.200, 9.700]	0.295 [0.255, 0.336]
Qwen3.8-27B	Qwen2.5-VL-3B	10	13.700 [12.600, 14.800]	0.311 [0.274, 0.350]
Qwen3.8-27B	Qwen3.5-4B	10	5.100 [4.400, 5.900]	0.108 [0.065, 0.155]
Qwen3.8-27B	Qwen3.5-27B	10	0.500 [-0.400, 1.400]	-0.073 [-0.104, -0.043]
Qwen3.8-27B	Qwen3-VL-235B	10	1.600 [0.300, 3.000]	0.112 [0.062, 0.165]
Qwen3.8-27B	Qwen2.5-VL-72B	10	1.300 [0.300, 2.300]	0.077 [0.023, 0.135]
Qwen3.8-27B	Gemma-4-31B	10	17.300 [16.700, 17.900]	0.341 [0.304, 0.378]
Qwen3.8-27B	InternVL3.5-8B	10	7.000 [6.400, 7.500]	0.161 [0.116, 0.210]
Qwen3.8-27B	MiniCPM-V-4.5	10	7.200 [6.200, 8.100]	0.153 [0.116, 0.193]
Qwen3.8-27B	Ovis2.5-9B	10	6.100 [5.000, 7.200]	0.161 [0.104, 0.220]
Qwen3.8-27B	Mistral-Small-3.1-24B	10	7.300 [6.600, 7.900]	0.165 [0.122, 0.217]
Qwen2.5-VL-32B	Qwen2.5-VL-7B	10	4.200 [2.700, 5.500]	0.046 [0.022, 0.071]
Qwen2.5-VL-32B	Qwen2.5-VL-3B	10	9.000 [7.800, 10.300]	0.062 [0.049, 0.074]
Qwen2.5-VL-32B	Qwen3.5-4B	10	0.400 [-0.900, 1.800]	-0.141 [-0.172, -0.110]
Qwen2.5-VL-32B	Qwen3.5-27B	10	-4.200 [-5.500, -2.800]	-0.321 [-0.345, -0.299]
Qwen2.5-VL-32B	Qwen3-VL-235B	10	-3.100 [-4.200, -1.900]	-0.136 [-0.172, -0.102]
Qwen2.5-VL-32B	Qwen2.5-VL-72B	10	-3.400 [-4.500, -2.300]	-0.172 [-0.215, -0.133]
Qwen2.5-VL-32B	Gemma-4-31B	10	12.600 [11.600, 13.600]	0.092 [0.074, 0.110]
Qwen2.5-VL-32B	InternVL3.5-8B	10	2.300 [0.900, 3.600]	-0.087 [-0.133, -0.044]
Qwen2.5-VL-32B	MiniCPM-V-4.5	10	2.500 [1.400, 3.600]	-0.096 [-0.130, -0.063]
Qwen2.5-VL-32B	Ovis2.5-9B	10	1.400 [0.100, 2.800]	-0.088 [-0.130, -0.045]
Qwen2.5-VL-32B	Mistral-Small-3.1-24B	10	2.600 [1.400, 3.500]	-0.083 [-0.106, -0.059]
Qwen2.5-VL-7B	Qwen2.5-VL-3B	10	4.800 [3.800, 5.900]	0.016 [-0.005, 0.035]
Qwen2.5-VL-7B	Qwen3.5-4B	10	-3.800 [-4.900, -2.900]	-0.187 [-0.223, -0.153]
Qwen2.5-VL-7B	Qwen3.5-27B	10	-8.400 [-9.500, -7.400]	-0.368 [-0.395, -0.337]
Qwen2.5-VL-7B	Qwen3-VL-235B	10	-7.300 [-8.600, -5.900]	-0.183 [-0.224, -0.139]
Qwen2.5-VL-7B	Qwen2.5-VL-72B	10	-7.600 [-8.600, -6.600]	-0.218 [-0.260, -0.179]
Qwen2.5-VL-7B	Gemma-4-31B	10	8.400 [7.700, 9.100]	0.046 [0.035, 0.057]
Qwen2.5-VL-7B	InternVL3.5-8B	10	-1.900 [-2.700, -1.100]	-0.134 [-0.165, -0.102]
Qwen2.5-VL-7B	MiniCPM-V-4.5	10	-1.700 [-2.800, -0.700]	-0.142 [-0.182, -0.112]
Qwen2.5-VL-7B	Ovis2.5-9B	10	-2.800 [-3.800, -1.800]	-0.134 [-0.181, -0.085]
Qwen2.5-VL-7B	Mistral-Small-3.1-24B	10	-1.600 [-2.400, -0.700]	-0.129 [-0.156, -0.103]
Qwen2.5-VL-3B	Qwen3.5-4B	10	-8.600 [-9.500, -7.500]	-0.203 [-0.233, -0.174]
Qwen2.5-VL-3B	Qwen3.5-27B	10	-13.200 [-14.400, -12.000]	-0.384 [-0.409, -0.359]
Qwen2.5-VL-3B	Qwen3-VL-235B	10	-12.100 [-13.000, -11.000]	-0.199 [-0.239, -0.160]
Qwen2.5-VL-3B	Qwen2.5-VL-72B	10	-12.400 [-12.900, -11.900]	-0.234 [-0.279, -0.196]
Qwen2.5-VL-3B	Gemma-4-31B	10	3.600 [2.900, 4.300]	0.030 [0.019, 0.046]
Qwen2.5-VL-3B	InternVL3.5-8B	10	-6.700 [-7.800, -5.600]	-0.150 [-0.188, -0.114]

Table 19 (continued).

Left	Right	Seeds	Δ positive [95% CI]	Δ clipped norm. [95% CI]
Qwen2.5-VL-3B	MiniCPM-V-4.5	10	-6.500 [-7.500, -5.500]	-0.158 [-0.192, -0.126]
Qwen2.5-VL-3B	Ovis2.5-9B	10	-7.600 [-8.300, -7.000]	-0.150 [-0.185, -0.113]
Qwen2.5-VL-3B	Mistral-Small-3.1-24B	10	-6.400 [-7.400, -5.400]	-0.145 [-0.161, -0.130]
Qwen3.5-4B	Qwen3.5-27B	10	-4.600 [-5.700, -3.600]	-0.181 [-0.220, -0.138]
Qwen3.5-4B	Qwen3-VL-235B	10	-3.500 [-4.800, -2.300]	0.004 [-0.032, 0.043]
Qwen3.5-4B	Qwen2.5-VL-72B	10	-3.800 [-5.000, -2.800]	-0.031 [-0.067, 0.000]
Qwen3.5-4B	Gemma-4-31B	10	12.200 [11.400, 12.900]	0.232 [0.204, 0.261]
Qwen3.5-4B	InternVL3.5-8B	10	1.900 [0.700, 2.900]	0.053 [0.008, 0.096]
Qwen3.5-4B	MiniCPM-V-4.5	10	2.100 [1.200, 2.900]	0.045 [0.013, 0.075]
Qwen3.5-4B	Ovis2.5-9B	10	1.000 [-0.200, 2.100]	0.053 [0.011, 0.093]
Qwen3.5-4B	Mistral-Small-3.1-24B	10	2.200 [1.400, 3.100]	0.057 [0.035, 0.079]
Qwen3.5-27B	Qwen3-VL-235B	10	1.100 [-0.200, 2.500]	0.185 [0.143, 0.227]
Qwen3.5-27B	Qwen2.5-VL-72B	10	0.800 [-0.200, 1.900]	0.150 [0.105, 0.195]
Qwen3.5-27B	Gemma-4-31B	10	16.800 [16.100, 17.600]	0.413 [0.386, 0.440]
Qwen3.5-27B	InternVL3.5-8B	10	6.500 [5.800, 7.300]	0.234 [0.191, 0.278]
Qwen3.5-27B	MiniCPM-V-4.5	10	6.700 [5.500, 7.800]	0.226 [0.185, 0.267]
Qwen3.5-27B	Ovis2.5-9B	10	5.600 [4.800, 6.400]	0.234 [0.183, 0.289]
Qwen3.5-27B	Mistral-Small-3.1-24B	10	6.800 [5.700, 8.000]	0.238 [0.202, 0.278]
Qwen3-VL-235B	Qwen2.5-VL-72B	10	-0.300 [-1.400, 0.800]	-0.035 [-0.081, 0.004]
Qwen3-VL-235B	Gemma-4-31B	10	15.700 [14.700, 16.600]	0.228 [0.189, 0.266]
Qwen3-VL-235B	InternVL3.5-8B	10	5.400 [3.900, 6.800]	0.049 [-0.017, 0.110]
Qwen3-VL-235B	MiniCPM-V-4.5	10	5.600 [4.700, 6.500]	0.041 [-0.002, 0.082]
Qwen3-VL-235B	Ovis2.5-9B	10	4.500 [3.300, 5.600]	0.049 [-0.009, 0.111]
Qwen3-VL-235B	Mistral-Small-3.1-24B	10	5.700 [4.300, 7.100]	0.053 [0.010, 0.097]
Qwen2.5-VL-72B	Gemma-4-31B	10	16.000 [15.300, 16.700]	0.264 [0.227, 0.303]
Qwen2.5-VL-72B	InternVL3.5-8B	10	5.700 [4.800, 6.600]	0.084 [0.028, 0.141]
Qwen2.5-VL-72B	MiniCPM-V-4.5	10	5.900 [4.700, 7.100]	0.076 [0.040, 0.119]
Qwen2.5-VL-72B	Ovis2.5-9B	10	4.800 [4.000, 5.600]	0.084 [0.036, 0.147]
Qwen2.5-VL-72B	Mistral-Small-3.1-24B	10	6.000 [5.100, 7.000]	0.089 [0.055, 0.129]
Gemma-4-31B	InternVL3.5-8B	10	-10.300 [-10.900, -9.700]	-0.179 [-0.216, -0.143]
Gemma-4-31B	MiniCPM-V-4.5	10	-10.100 [-10.800, -9.500]	-0.187 [-0.219, -0.164]
Gemma-4-31B	Ovis2.5-9B	10	-11.200 [-11.900, -10.500]	-0.179 [-0.217, -0.139]
Gemma-4-31B	Mistral-Small-3.1-24B	10	-10.000 [-10.600, -9.400]	-0.175 [-0.197, -0.155]
InternVL3.5-8B	MiniCPM-V-4.5	10	0.200 [-1.000, 1.300]	-0.008 [-0.059, 0.038]
InternVL3.5-8B	Ovis2.5-9B	10	-0.900 [-1.800, 0.000]	0.000 [-0.053, 0.055]
InternVL3.5-8B	Mistral-Small-3.1-24B	10	0.300 [-0.500, 1.100]	0.004 [-0.032, 0.046]
MiniCPM-V-4.5	Ovis2.5-9B	10	-1.100 [-2.100, -0.100]	0.008 [-0.024, 0.040]
MiniCPM-V-4.5	Mistral-Small-3.1-24B	10	0.100 [-0.700, 0.900]	0.012 [-0.018, 0.047]
Ovis2.5-9B	Mistral-Small-3.1-24B	10	1.200 [0.100, 2.400]	0.004 [-0.026, 0.036]

Table 20: Independent reference-controller calibration (600 strict-valid episodes). RET is the area, in milliseconds, under no-op-floor-normalized retention through 2000 ms. All five outcome-unseen, hash-frozen per-game gates failed, so reference normalization and the downstream factorization are disabled; assigned-delay LevelAUC/LossAUC remain the uniform analysis.

Game	Oracle RET	RGB RET	RGB/oracle [95% CI]	Gate
Sky Flap	375	375	1.00 –	Fail
Paper Route	563	1864	3.31 [2.59, 4.01]	Fail
Block Stack	1958	1798	0.92 [0.90, 0.93]	Fail
Target Range	453	1625	3.58 [0.00, 3.72]	Fail
Asteroid Patrol	1915	1855	0.97 [0.92, 1.02]	Fail

Table 21: Non-Qwen intervention-panel disposition. Accepted matched-live evidence is kept distinct from the paused policy-invalid attempts; the latter are not replaced or pooled.

Model	Matched live	Live clipped score	Paused attempts	Disposition
Gemma-4-31B	50/50 accepted	0.000	2/50	invalid; excluded

Table 22: Qwen3.8-27B one-factor-at-a-time sensitivity on five games and five common seeds (225 strict-valid cells). Δ compares each setting with the canonical center using seed-block intervals. Fixed candidate-major execution makes these descriptive assigned-configuration contrasts, not randomized causal effects.

Configuration	Clipped norm.	Δ [95% CI]	Model P50	Dispatch P95
Canonical center	0.509	–	575	1341
Frames = 1	0.395	-0.114 [-0.322, 0.172]	291	1070
Frames = 2	0.426	-0.083 [-0.343, 0.136]	400	1170
Resolution 320×240	0.215	-0.294 [-0.485, -0.103]	383	1117
Resolution 512×320	0.584	0.075 [-0.155, 0.316]	450	1351
Cadence = 150 ms	0.439	-0.070 [-0.404, 0.244]	578	1405
Cadence = 300 ms	0.415	-0.094 [-0.337, 0.150]	572	1372
Max age = 800 ms	0.537	0.028 [-0.217, 0.348]	527	1166
Max age = 2000 ms	0.475	-0.034 [-0.377, 0.310]	573	1389

Table 23: Frozen-schedule assigned-profile contrasts at an outcome-unseen, hash-frozen 250 ms nominal mean. Score is the macro clipped T3-normalized score over five common seeds; differences are paired against constant delay. Rare-spike IID and burst share the complete nominal multiset, but trajectory-dependent realized prefixes are not guaranteed to have matched marginals.

Model	Assigned profile	Score [95% CI]	Δ vs. constant [95% CI]
Qwen2.5-VL-7B	Constant	0.037 [0.035, 0.038]	0.000 [0.000, 0.000]
Qwen2.5-VL-7B	Symmetric IID	0.035 [0.034, 0.036]	-0.002 [-0.002, -0.001]
Qwen2.5-VL-7B	Rare spike IID	0.036 [0.036, 0.037]	0.000 [-0.001, 0.000]
Qwen2.5-VL-7B	Rare spike burst	0.036 [0.035, 0.038]	0.000 [0.000, 0.000]
Qwen2.5-VL-32B	Constant	0.012 [0.000, 0.036]	0.000 [0.000, 0.000]
Qwen2.5-VL-32B	Symmetric IID	0.027 [0.000, 0.056]	0.015 [0.000, 0.044]
Qwen2.5-VL-32B	Rare spike IID	0.003 [0.000, 0.010]	-0.009 [-0.036, 0.010]
Qwen2.5-VL-32B	Rare spike burst	0.029 [0.002, 0.057]	0.017 [-0.021, 0.055]
Qwen3.8-27B	Constant	0.477 [0.335, 0.617]	0.000 [0.000, 0.000]
Qwen3.8-27B	Symmetric IID	0.523 [0.353, 0.672]	0.046 [-0.113, 0.215]
Qwen3.8-27B	Rare spike IID	0.265 [0.097, 0.480]	-0.212 [-0.434, 0.044]
Qwen3.8-27B	Rare spike burst	0.397 [0.189, 0.606]	-0.080 [-0.258, 0.078]

Model	Burst − rare-spike IID [95% CI]
Qwen2.5-VL-7B	0.000 [-0.001, 0.001]
Qwen2.5-VL-32B	0.026 [-0.005, 0.057]
Qwen3.8-27B	0.132 [-0.179, 0.376]

Table 24: Deployment-conditioned temporal ordering. Relations require a 95% simultaneous seed-block bootstrap band across the full condition grid; point-estimate crossings are insufficient.

Model pair	Live grid	Paused + live
Qwen2.5-VL-32B / Qwen2.5-VL-7B	Unresolved	Unresolved
Qwen2.5-VL-32B / Qwen3.8-27B	Unresolved	Unresolved
Qwen2.5-VL-7B / Qwen3.8-27B	Unresolved	Unresolved

Model	Configuration	Macro clipped T3-normalized score	95% CI
Qwen2.5-VL-7B-Instruct	tbw-f1-r320x240-c150	0.075	[0.051, 0.100]
Qwen2.5-VL-7B-Instruct	tbw-f1-r320x240-c300	0.078	[0.037, 0.139]
Qwen2.5-VL-7B-Instruct	tbw-f1-r512x320-c150	0.056	[0.039, 0.078]
Qwen2.5-VL-7B-Instruct	tbw-f1-r512x320-c300	0.047	[0.036, 0.068]
Qwen2.5-VL-7B-Instruct	tbw-f3-r320x240-c150	0.037	[0.036, 0.038]
Qwen2.5-VL-7B-Instruct	tbw-f3-r320x240-c300	0.041	[0.037, 0.044]
Qwen2.5-VL-7B-Instruct	tbw-f3-r512x320-c150	0.037	[0.036, 0.038]
Qwen2.5-VL-7B-Instruct	tbw-f3-r512x320-c300	0.041	[0.037, 0.045]
Qwen2.5-VL-7B-Instruct	balanced-3f	0.039	[0.037, 0.041]
Qwen2.5-VL-32B-Instruct	tbw-f1-r320x240-c150	0.050	[0.035, 0.069]
Qwen2.5-VL-32B-Instruct	tbw-f1-r320x240-c300	0.072	[0.044, 0.101]
Qwen2.5-VL-32B-Instruct	tbw-f1-r512x320-c150	0.075	[0.058, 0.087]
Qwen2.5-VL-32B-Instruct	tbw-f1-r512x320-c300	0.105	[0.063, 0.157]
Qwen2.5-VL-32B-Instruct	tbw-f3-r320x240-c150	0.030	[0.005, 0.072]
Qwen2.5-VL-32B-Instruct	tbw-f3-r320x240-c300	0.138	[0.003, 0.375]
Qwen2.5-VL-32B-Instruct	tbw-f3-r512x320-c150	0.053	[0.017, 0.095]
Qwen2.5-VL-32B-Instruct	tbw-f3-r512x320-c300	0.037	[0.003, 0.071]
Qwen2.5-VL-32B-Instruct	balanced-3f	0.136	[0.049, 0.250]
Qwen3.8-27B	tbw-f1-r320x240-c150	0.573	[0.470, 0.667]
Qwen3.8-27B	tbw-f1-r320x240-c300	0.478	[0.283, 0.609]
Qwen3.8-27B	tbw-f1-r512x320-c150	0.573	[0.486, 0.622]
Qwen3.8-27B	tbw-f1-r512x320-c300	0.472	[0.373, 0.532]
Qwen3.8-27B	tbw-f3-r320x240-c150	0.319	[0.091, 0.547]
Qwen3.8-27B	tbw-f3-r320x240-c300	0.159	[0.097, 0.229]
Qwen3.8-27B	tbw-f3-r512x320-c150	0.380	[0.291, 0.507]
Qwen3.8-27B	tbw-f3-r512x320-c300	0.555	[0.331, 0.680]
Qwen3.8-27B	balanced-3f	0.501	[0.298, 0.711]

Table 25: Three-model temporal-context design surface. Scores are clipped per game to $[0, 3]$ before macro aggregation. Intervals resample complete common-seed blocks. Natural service latency is descriptive only.

Model contrast	Configuration	Difference-in-differences [95% CI]
Qwen2.5-VL-7B-Instruct – Qwen2.5-VL-32B-Instruct	tbw-f1-r320x240-c150	0.121 [0.043, 0.210]
Qwen2.5-VL-7B-Instruct – Qwen2.5-VL-32B-Instruct	tbw-f1-r320x240-c300	0.103 [-0.021, 0.229]
Qwen2.5-VL-7B-Instruct – Qwen2.5-VL-32B-Instruct	tbw-f1-r512x320-c150	0.078 [-0.002, 0.209]
Qwen2.5-VL-7B-Instruct – Qwen2.5-VL-32B-Instruct	tbw-f1-r512x320-c300	0.039 [-0.016, 0.119]
Qwen2.5-VL-7B-Instruct – Qwen2.5-VL-32B-Instruct	tbw-f3-r320x240-c150	0.104 [0.014, 0.224]
Qwen2.5-VL-7B-Instruct – Qwen2.5-VL-32B-Instruct	tbw-f3-r320x240-c300	0.000 [-0.261, 0.220]
Qwen2.5-VL-7B-Instruct – Qwen2.5-VL-32B-Instruct	tbw-f3-r512x320-c150	0.081 [0.008, 0.156]
Qwen2.5-VL-7B-Instruct – Qwen2.5-VL-32B-Instruct	tbw-f3-r512x320-c300	0.100 [0.012, 0.222]

Table 26: Model–configuration interactions relative to each model's canonical center. Each row is $(\text{left}_{cfg} - \text{left}_{center}) - (\text{right}_{cfg} - \text{right}_{center})$. Intervals resample the five complete common-seed blocks and are individual, unadjusted 95% intervals; all 24 prespecified contrasts are reported across the three panels.

Model contrast	Configuration	Difference-in-differences [95% CI]
Qwen2.5-VL-7B-Instruct – Qwen3.8-27B	tbw-f1-r320x240-c150	-0.035 [-0.254, 0.225]
Qwen2.5-VL-7B-Instruct – Qwen3.8-27B	tbw-f1-r320x240-c300	0.063 [-0.108, 0.232]
Qwen2.5-VL-7B-Instruct – Qwen3.8-27B	tbw-f1-r512x320-c150	-0.055 [-0.273, 0.151]
Qwen2.5-VL-7B-Instruct – Qwen3.8-27B	tbw-f1-r512x320-c300	0.038 [-0.190, 0.255]
Qwen2.5-VL-7B-Instruct – Qwen3.8-27B	tbw-f3-r320x240-c150	0.181 [0.113, 0.255]
Qwen2.5-VL-7B-Instruct – Qwen3.8-27B	tbw-f3-r320x240-c300	0.344 [0.101, 0.586]
Qwen2.5-VL-7B-Instruct – Qwen3.8-27B	tbw-f3-r512x320-c150	0.119 [-0.187, 0.362]
Qwen2.5-VL-7B-Instruct – Qwen3.8-27B	tbw-f3-r512x320-c300	-0.052 [-0.220, 0.116]

Table 27: Model–configuration interactions relative to each model's canonical center. Each row is $(\text{left}_{cfg} - \text{left}_{center}) - (\text{right}_{cfg} - \text{right}_{center})$. Intervals resample the five complete common-seed blocks and are individual, unadjusted 95% intervals; all 24 prespecified contrasts are reported across the three panels.

Model contrast	Configuration	Difference-in-differences [95% CI]
Qwen2.5-VL-32B-Instruct – Qwen3.8-27B	tbw-f1-r320x240-c150	-0.157 [-0.388, 0.116]
Qwen2.5-VL-32B-Instruct – Qwen3.8-27B	tbw-f1-r320x240-c300	-0.040 [-0.213, 0.163]
Qwen2.5-VL-32B-Instruct – Qwen3.8-27B	tbw-f1-r512x320-c150	-0.133 [-0.312, 0.032]
Qwen2.5-VL-32B-Instruct – Qwen3.8-27B	tbw-f1-r512x320-c300	-0.001 [-0.180, 0.161]
Qwen2.5-VL-32B-Instruct – Qwen3.8-27B	tbw-f3-r320x240-c150	0.077 [-0.072, 0.222]
Qwen2.5-VL-32B-Instruct – Qwen3.8-27B	tbw-f3-r320x240-c300	0.344 [0.097, 0.591]
Qwen2.5-VL-32B-Instruct – Qwen3.8-27B	tbw-f3-r512x320-c150	0.038 [-0.228, 0.251]
Qwen2.5-VL-32B-Instruct – Qwen3.8-27B	tbw-f3-r512x320-c300	-0.153 [-0.291, -0.009]

Table 28: Model–configuration interactions relative to each model's canonical center. Each row is $(\text{left}_{cfg} - \text{left}_{center}) - (\text{right}_{cfg} - \text{right}_{center})$. Intervals resample the five complete common-seed blocks and are individual, unadjusted 95% intervals; all 24 prespecified contrasts are reported across the three panels.

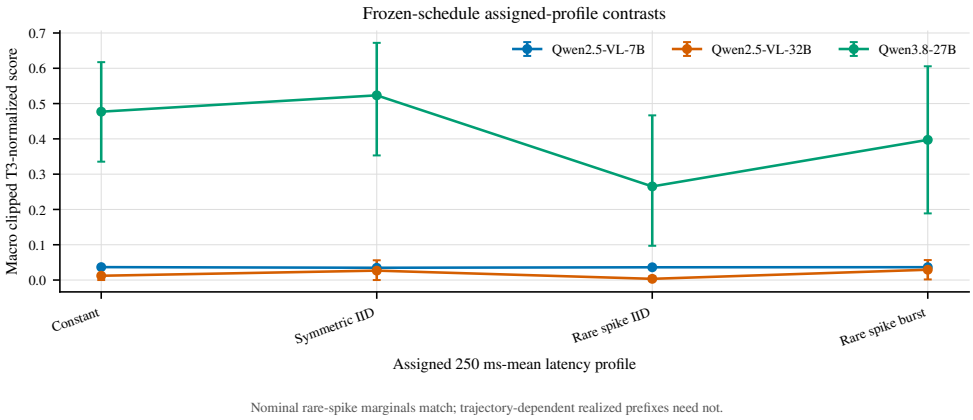

Figure 7: Latency-process risk under equal nominal mean delay. Constant, rare-spike IID, and burst-clustered profiles are assigned from frozen schedules. The IID/burst pair has the same complete delay multiset; estimates are frozen-schedule assigned-profile contrasts because closed-loop trajectories may consume different schedule prefixes.

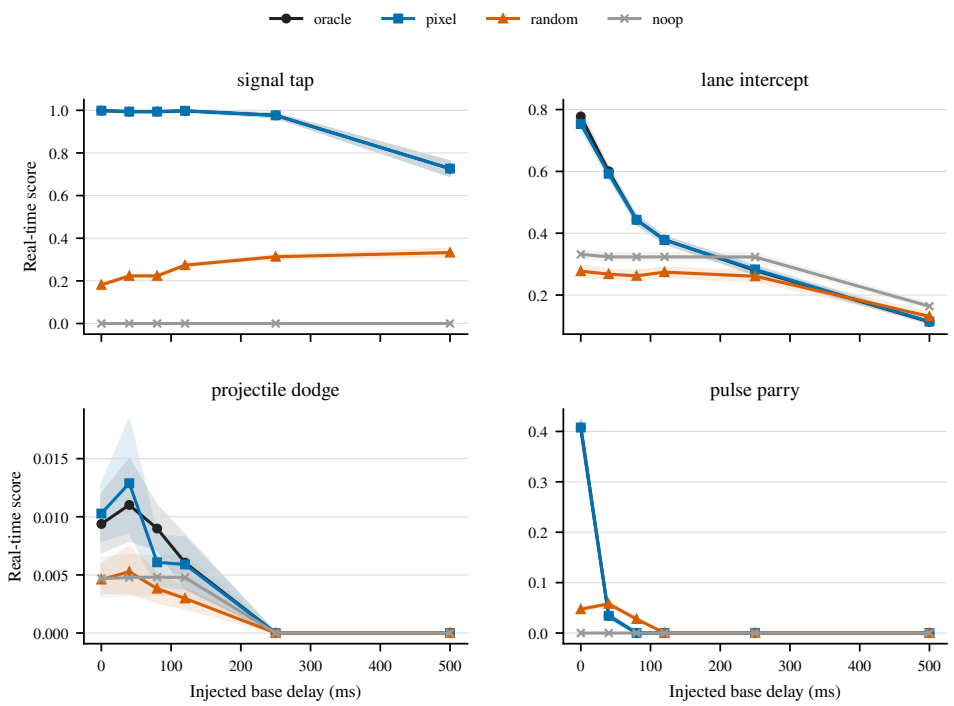

Figure 8: Controlled-delay real-time score curves. Shading gives deterministic seed-bootstrap 95% intervals over the 20 common seeds in each cell. The oracle is privileged and serves only as a controller upper bound.

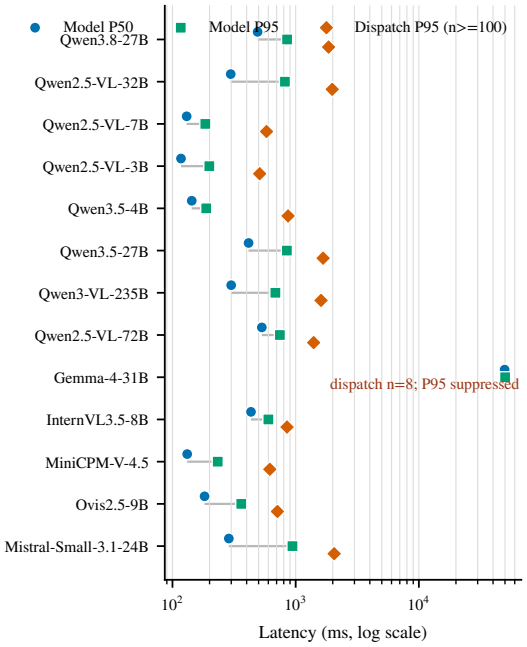

Figure 9: Completed-event-weighted model and capture-to-dispatch latency for accepted common-seed cohorts on a logarithmic axis. Dispatch latency includes capture, validation, scheduling, and action-horizon effects beyond model service time; tails with fewer than 100 dispatches are suppressed. The appendix reports equal-episode summaries, and seed blocks—not events—define statistical uncertainty.

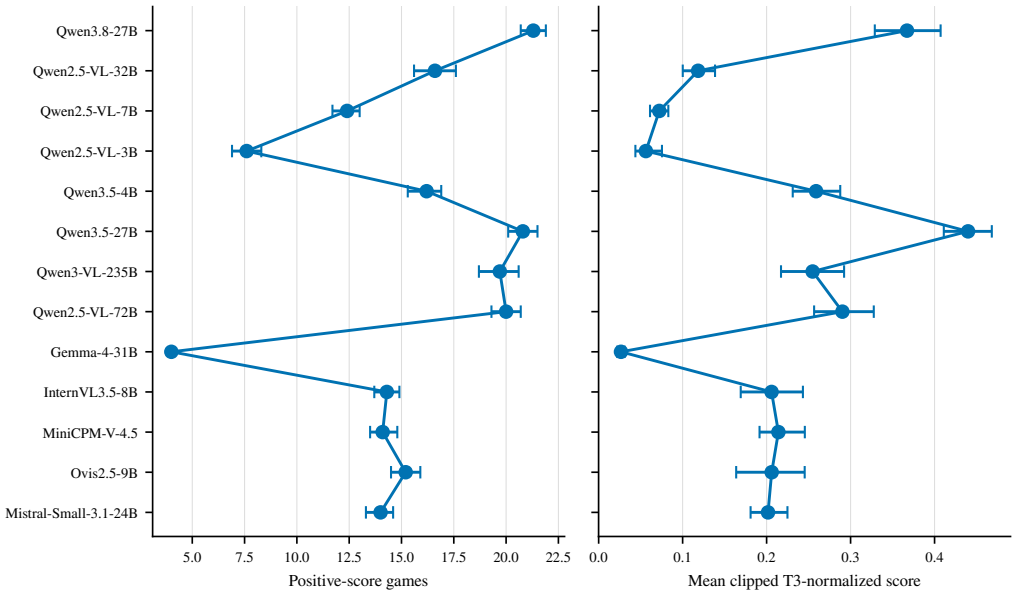

Figure 10: Repeated-seed Arcade32 quality with paired protocol conditions. Error bars are round-bootstrap 95% intervals.



Figure 11: Cross-model trajectory comparison on Vector Pinball (common seed 4007), using the same score-frozen selection and normalized video progress as Figures 12–13. Qwen3.8-27B is the only one of 13 formal models to score on this cell and executes coordinated alternating flipper commands. InternVL3.5-8B produces mostly rejected outputs, whereas MiniCPM-V-4.5 is schema-valid but repeatedly issues legal actions without sustained progress.

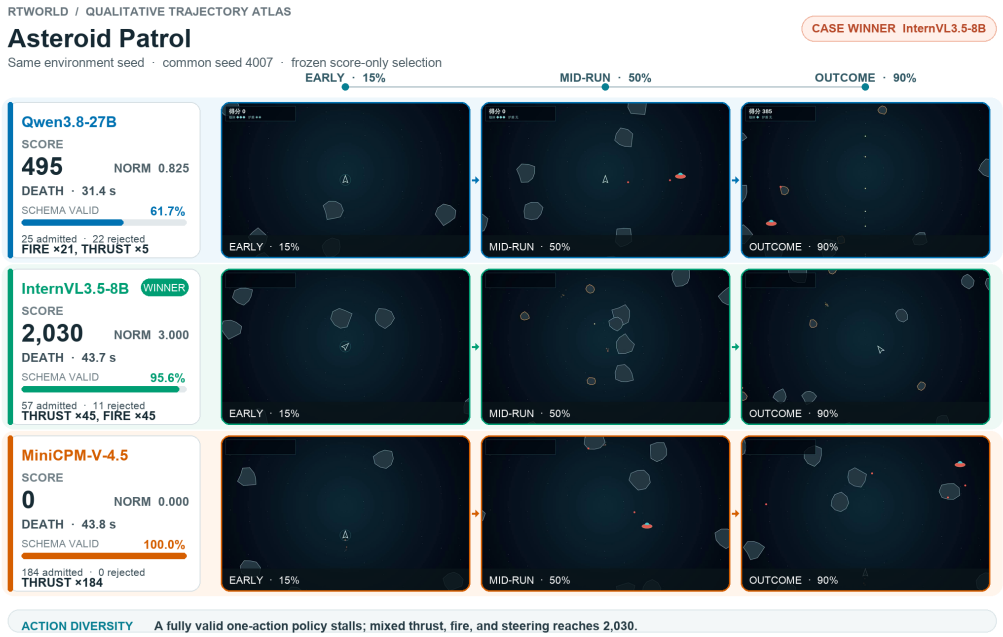

Figure 12: Cross-model trajectory comparison on Asteroid Patrol (common seed 4007), with frames sampled at 15%, 50%, and 90% of each rendered video. InternVL3.5-8B sustains mixed thrust, fire, and steering to score 2030; Qwen3.8-27B scores 495 despite substantially more rejected responses; MiniCPM-V-4.5 emits 184 schema-valid forward commands but scores zero. The case separates interface validity from closed-loop action diversity and progress; selection used frozen scores before frames were inspected.

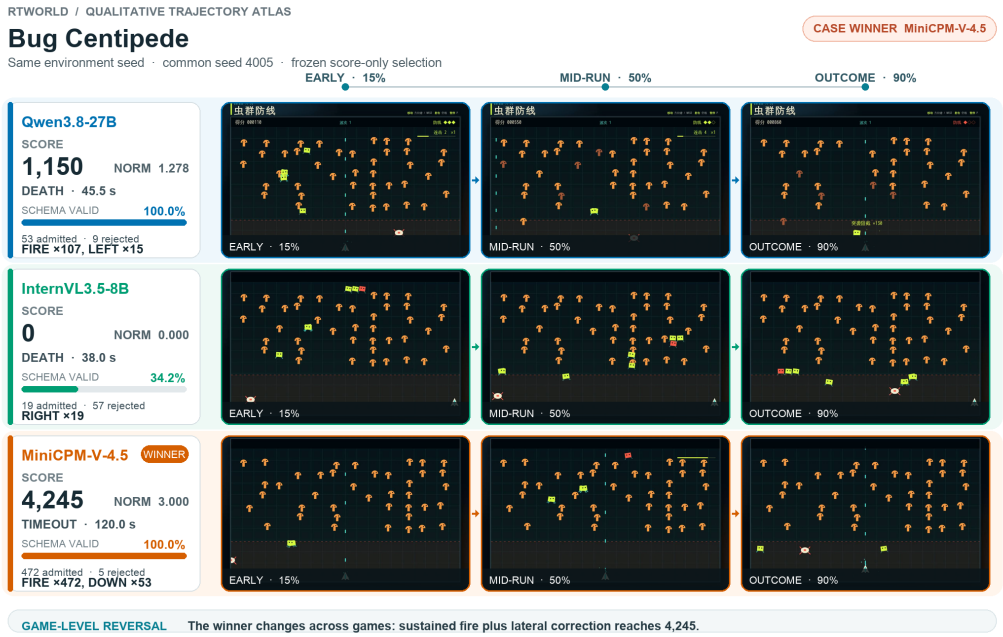

Figure 13: Cross-model trajectory comparison on Bug Centipede (common seed 4005). MiniCPM-V-4.5 reaches 4245 with sustained fire and lateral movement, Qwen3.8-27B reaches 1150, and InternVL3.5-8B scores zero while only 34.2% of responses are schema-valid. Together with Figure 12, this game-level reversal shows why a single qualitative example cannot stand in for the repeated-seed ranking. Frames and displayed statistics are bound to strict-valid immutable evidence.