# VRC: Visual Reward Consistency in GUI-Agent Reinforcement Learning

**Anonymous authors**
Paper under double-blind review

## Abstract

Visual reward models can classify recorded trajectories accurately yet provide weak feedback for policy learning. We examine this gap in graphical user interface (GUI) reinforcement learning using paired visual evidence on BrowserGym/MiniWoB. Four data-matched verifier objectives compare binary cross-entropy with invariance and failure-sensitivity losses. A separate policy experiment compares a BCE verifier's single-view reward with five-view soft-min scoring from the combined-loss verifier. Across three seeds, sensitivity supervision achieved the highest validation AUROC, while BCE had the lowest banner inconsistency. On 2,160 evaluation episodes across 24 task families, BCE reward achieved 33.06% success and the soft-min reward achieved 32.22%. Neither the original comparison nor a post-hoc crossed seed-task analysis resolved an improvement from soft-min. Most held-out successes came from one task family, and the soft-min reward showed weak separation between successful and failed policy trajectories. The policy experiment changes verifier training and aggregation together and excludes the sensitivity-only verifier. The observed result concerns this reward combination; the downstream value of sensitivity supervision remains open.

## 1 Introduction

Task-specific success checks are costly to extend when a GUI agent encounters new applications or instructions. Learned and prompted evaluators already support web-agent and device-control reinforcement learning in WebRL and DigiRL (Qi et al., 2025; Bai et al., 2024). Vision-language models (VLMs) also provide zero-shot rewards, success judgments, and preference feedback in control tasks (Rocamonde et al., 2024; Du et al., 2023; Wang et al., 2024). The question here is how a visual reward responds to the evidence available during policy learning.

Two properties matter for terminal verification. A harmless presentation change should preserve the score, whereas evidence of an unsuccessful execution should lower it. These requirements become harder when the screenshots omit decisive state or the verifier lacks an explicit task instruction. Active verification acquires additional observations (Dai et al., 2026; Shi et al., 2026), while task-adaptive rubrics derive completion criteria from an available task instruction (Xiong et al., 2026b). We instead study a fixed, passive interface comprising initial and terminal screenshots plus the action history.

Offline discrimination and downstream policy success test different aspects of reward quality. A classifier can separate constructed failures yet remain insensitive to errors encountered by its policy. Reward-model overoptimization provides a broader reason to evaluate an independent outcome rather than the optimized score alone (Gao et al., 2023). We test whether paired consistency and failure-sensitivity supervision produce useful reward differences beyond those learned from the same augmented data with binary cross-entropy (BCE).

The study uses BrowserGym's MiniWoB tasks, whose validators provide outcome labels independently of the learned scorer (Le Sellier de Chezelles et al., 2025; Shi et al., 2017). We first compare four training objectives on a fixed pool of successful trajectories, natural failures, banner views, and constructed negatives. We then train policies using either BCE single-view reward or combined-loss five-view soft-min reward. This second comparison tests two complete reward configurations rather than separate effects of training and aggregation.



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.

The experiments yield three findings. Explicit invariance did not improve banner consistency over augmented-data BCE, while sensitivity supervision improved validation discrimination. The combined-loss soft-min reward did not yield a resolved policy-success gain over BCE reward. Task-level outcomes and evidence-substitution diagnostics locate additional limits of the tested scorers, including concentrated held-out success and weak failure-score separation.

## 2 Related Work

**Visual rewards and success detection.** Vision-language reward methods differ in what they predict and how they obtain supervision. Zero-shot VLM rewards use image-language representations (Rocamonde et al., 2024), while SuccessVQA formulates task-conditioned success detection as visual question answering (Du et al., 2023). RL-VLM-F learns a reward model from VLM-generated image preferences (Wang et al., 2024). FuRL addresses misalignment between visual-language rewards and task objectives during policy learning (Fu et al., 2024). GVL estimates visual-trajectory progress through in-context temporal ordering (Ma et al., 2025). RoboReward learns task-conditioned, graded completion rewards from robotics videos, including counterfactual instruction labels and temporal truncations (Lee et al., 2026). These precedents motivate separating progress, preference, and binary success targets. Our target is terminal GUI success; truncated prefixes supply training negatives rather than a process-reward objective.

**GUI-agent learning and verification.** WebRL combines outcome-supervised reward learning with an evolving web-task curriculum (Qi et al., 2025); DigiRL uses automated evaluation for offline-to-online device-control learning (Bai et al., 2024). ProRe and Agentic Reward Modeling acquire additional interface evidence before judging task completion (Dai et al., 2026; Cui et al., 2026). I-RA checks post-execution system, application, and GUI state (Shi et al., 2026). GUI-PRA uses state-specific criteria and visual tools for process evaluation (Xiong et al., 2026a). AdaptRubric supplies category-level and instruction-specific completion criteria to a trajectory verifier (Xiong et al., 2026b). Our scorer has neither an independent textual goal nor active access to the environment. For evaluation, World of Bits and BrowserGym support controlled web tasks (Shi et al., 2017; Le Sellier de Chezelles et al., 2025). WebArena, VisualWebArena, OSWorld, and AndroidWorld cover richer web, visual, desktop, and mobile interactions (Zhou et al., 2024; Koh et al., 2024; Xie et al., 2024; Rawles et al., 2025). All reported experiments use MiniWoB and a text-observation policy.

**Reward overoptimization and aggregation.** Reward gaming concerns conflict between an optimized proxy and the intended return (Skalse et al., 2022); overoptimization studies examine how that conflict develops with optimization pressure (Gao et al., 2023). Incomplete verifiers can reward shortcut solutions (Helff et al., 2026), while reward shaping seeks to reduce exploitation of learned reward models (Fu et al., 2026). Constrained RLHF controls component rewards through adaptive constraints (Moskovitz et al., 2024). Reward-model ensembles instead combine distinct predictors, including minimum and uncertainty-penalized rewards (Coste et al., 2024). Shared errors can persist even across diverse ensemble members (Eisenstein et al., 2024). Chasing the Tail refines rubrics to distinguish strong responses in the high-reward region (Zhang et al., 2026). Our soft-min combines transformed inputs to one verifier, not independently trained models or constrained reward components. We report evidence sensitivity and task success separately; screenshot substitution alone does not identify an optimization-induced reward exploit.

**Consistency and multiple views.** Data-augmentation theory and AugMix establish precedents for learning stable predictions under valid image transformations (Chen et al., 2020; Hendrycks et al., 2020). MVR learns video-based reward shaping from multiple rendered camera viewpoints (Luo et al., 2026). VISTA trains GUI-grounding policies with target-preserving crops, whereas GUI-RCPO derives test-time rewards from spatial agreement among sampled predictions (Qiu et al., 2026; Du et al., 2026). Faithful mobile-agent training regularizes reasoning-action consistency and addresses low-variance reward groups (Hu et al., 2026). SCIR aligns internal reward judgments for self-rewarding language models (Zhou et al., 2025). In our experiment, invariance acts on a terminal verifier's logits, and sensitivity acts on labeled successful-failed evidence pairs. Camera views, screenshot transformations, separate models, and sampled predictions supply different forms of variation; their errors need not be independent.

# 3 METHOD

Figure 1 summarizes paired evidence construction, learned-reward policy optimization, and independent outcome evaluation. Miniature interfaces illustrate how the scorer's visual inputs differ across these stages.

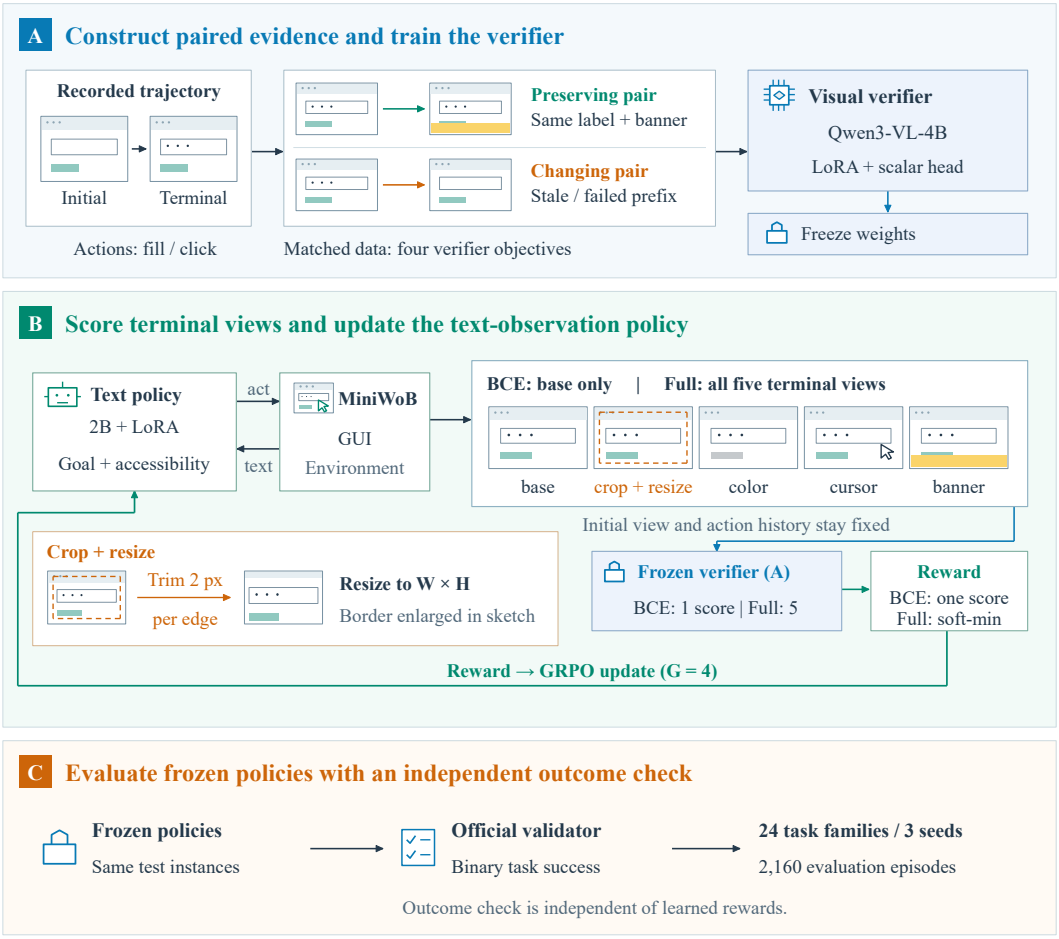


Figure 1: Method overview with schematic GUI views. (A) Four verifier objectives use matched paired evidence: banner pairs retain labels; stale-initial and unsuccessful-prefix records provide changing pairs. (B) A text-observation policy interacts with MiniWoB. Frozen verifiers produce either BCE single-view reward or Full five-view soft-min reward for GRPO. Crop removes 2 pixels per edge and resizes to the original dimensions; the illustrated border is exaggerated. (C) Frozen policies are evaluated on matched instances using the official validator, independently of learned rewards.

## 3.1 OBSERVATIONS, EVIDENCE, AND OUTCOMES

An episode has goal $x$, public observations $o_{0:T}$, actions $a_{0:T-1}$, screenshots $s_{0:T}$, and binary official outcome $y$. The policy receives $x$ and compact accessibility text containing element identifiers, roles, names, field values, and interaction state, without screenshot tensors. The verifier receives $e = (s_0, s_T, a_{0:T-1})$ and predicts $p_\phi(e) = \sigma(z_\phi(e))$. It has no separate textual goal input, so instruction interpretation depends on information visible in the screenshots or action history.

The official validator supplies data labels and the final evaluation outcome. Its scalar is excluded from policy prompts and reward-service payloads; the learned reward enters the policy advantage. The validator also checks outcomes before and after banner insertion during reward construction. S-

tandard success termination remains enabled, allowing episode length to reveal completion indirectly. Data construction and banner auditing therefore retain oracle access.

### 3.2 Paired evidence construction

We collect 480 successful training trajectories with a deterministic document object model (DOM) expert and 120 trajectories with the frozen policy. The latter include 48 successes and 72 genuine failures. A separate set of 120 expert successes supplies validation anchors on disjoint instances of the same 12 task families. The expert chooses actions from the rendered goal and DOM, not from the task validator. Each base example contains the initial screenshot, terminal screenshot, executed actions, and outcome label.

A **preserving pair** joins a base example to a banner view in a fixed bottom-right region. The operational validity check requires identical validator outcomes before and after insertion. A **changing pair** joins a successful source to a stale-initial or unsuccessful-prefix record. Stale-initial targets contain the initial screenshot twice and an empty action trace, labeled negative under the assumption that the initial state is incomplete. Prefix targets retain screenshots and actions with zero recorded validator reward, before the first success. Temporal truncation also supplies reward supervision in RoboReward (Lee et al., 2026); here it changes both visual and action evidence. These constructed negatives are distinct from the 72 natural policy failures.

The training set contains 2,229 examples: 600 base, 600 banner, 528 stale-initial, and 501 unsuccessful-prefix records. Validation contains 478 examples: 120 base, 120 banner, 120 stale-initial, and 118 prefix records. All validation negatives are constructed from successful expert trajectories; no naturally failed policy trajectories enter that set. Validation therefore measures discrimination on the paired construction distribution.

### 3.3 Data-matched verifier objectives

A scalar head reads the final attended hidden state of Qwen3-VL-4B-Instruct (Bai et al., 2025), adapted with low-rank adaptation (LoRA) (Hu et al., 2022). Every source and target receives BCE supervision,

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right].$$

Let $\mathcal{P}$ be preserving relations and $\mathcal{C}$ successful-source/changing-target relations in that batch. The auxiliary terms are

$$\mathcal{L}_{\text{inv}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}(z_i - z_j)^2, \qquad \mathcal{L}_{\text{sens}} = \frac{1}{|\mathcal{C}|}\sum_{(i,j)\in\mathcal{C}}\max\{0, 1-(z_i - z_j)\}.$$

A term is zero when its relation set is empty. We compare $\mathcal{L}_{\text{BCE}} + \lambda_p\mathcal{L}_{\text{inv}} + \lambda_c\mathcal{L}_{\text{sens}}$ for $(\lambda_p, \lambda_c) \in \{(0,0),(1,0),(0,1),(1,1)\}$. The corresponding conditions are BCE, + invariance, + sensitivity, and Full. All four use the same augmented pool and relation sampler, isolating auxiliary-loss effects under this training schedule. Unlike AugMix's posterior-distribution consistency (Hendrycks et al., 2020), the preserving loss matches scalar logits; the sensitivity loss separates labeled successful and unsuccessful records.

Each run uses 100 AdamW updates (Loshchilov & Hutter, 2019), with four sampled relations and eight separately encoded examples per update. The resulting 400 relation samples and 800 example exposures cover only part of the 1,629 available relations. LoRA rank is 16, learning rate is $2\times10^{-5}$, and seeds are 17, 29, and 43. Threshold selection maximizes validation true-positive rate subject to a false-acceptance rate of at most 0.10. The same validation examples determine the threshold and its reported metrics.

### 3.4 Multi-view reward aggregation

BCE reward uses one terminal view; the Full verifier scores $K = 5$ views: original, crop-resize jitter, reduced saturation, cursor overlay, and audited banner. Initial images and action traces remain fixed across views. Each view encodes two images, giving two image presentations per BCE reward and ten per five-view reward. Only banners appear in supervised preserving pairs; crop, color, and cursor changes are inference-time probes. These probes leave environment state fixed but can obscure information needed for judgment. Their evidence validity was not independently annotated. In contrast to rendered-camera-view reward shaping (Luo et al., 2026) or model ensembles (Coste et al., 2024), all five views share one verifier and one underlying screenshot.

For view scores $v_k = p_\phi(g_k(e))$, we consider their mean and the normalized soft minimum

$$R_\beta(e) = -\beta \log \left( \frac{1}{K} \sum_{k=1}^{K} \exp(-v_k/\beta) \right), \qquad \beta = 0.10.$$

This aggregation gives greater influence to lower-scoring views. In particular, with $v_{\min} = \min_k v_k$ and $\bar{v} = K^{-1} \sum_k v_k$,

$$v_{\min} \le R_\beta \le \min\{\bar{v},\ v_{\min} + \beta \log K\}, \qquad \frac{\partial R_\beta}{\partial v_j} = \frac{\exp(-v_j/\beta)}{\sum_k \exp(-v_k/\beta)}.$$

For dispersion small relative to $\beta$, the local expansion is $R_\beta \approx \bar{v} - \mathrm{Var}(v)/(2\beta)$. Soft-min therefore penalizes view disagreement, but leaves shared errors uncorrected. We evaluate it as a conservative scoring rule without assuming calibrated uncertainty or guaranteed reward correctness. Appendix E derives these identities.

### 3.5 Group-relative policy optimization

The policy is Qwen3-VL-2B-Instruct with a fresh rank-16 LoRA adapter (Bai et al., 2025; Hu et al., 2022). Group-relative policy optimization (GRPO) uses four completion scores $r_1, \ldots, r_4$ to define $A_i = (r_i - \bar{r})/(s_r + \epsilon)$ (Shao et al., 2024; Mroueh, 2025). We use PPO-style ratio clipping at 0.2 (Schulman et al., 2017) and a reference-policy Kullback-Leibler (KL) coefficient of 0.02. TRL's token-normalized DAPO loss setting supplies the normalization (Yu et al., 2025), with group reward scaling and temperature 1.0. This adopts a loss component rather than the complete DAPO system. Each run has 300 optimizer updates, one optimization iteration per generated batch, and four completions per generation batch. Per-device batch size and gradient accumulation are both one.

We attempted action-free visual reward, trajectory BCE reward, Full mean reward, and Full soft-min reward. At update 25, eligibility requires at least 20% nonzero logged rewards and mean within-group standard deviation of at least 0.01. The check uses generation-batch logs available through that update. Sigmoid scores make the nonzero condition weak; dispersion determined the observed failures. All three seeds must pass for a method to enter final evaluation, and failed seeds are not replaced.

Training uses tool-mediated interactions and a 256-token completion limit. Evaluation uses strict JSON actions, greedy decoding, a 96-token limit, and at most 12 actions per episode, identically across methods. Verifier training and serving also use different instruction strings. These training-serving differences may introduce reward mismatch. No crossed training-objective-by-aggregation experiment was performed.

## 4 Experimental Protocol

Policies train on 12 task families and are evaluated on those families plus 12 held-out families. Each method-seed cell contains ten deterministic instances per task, totaling 240 episodes per seed, 720 per method, and 2,160 across three methods. Instance keys match across methods within each seed. Success is the official binary outcome, independent of verifier threshold selection. The estimand

weights 24 task families and three training seeds equally, averaging ten instances within each cell. In-distribution (ID) and out-of-distribution (OOD) denote training and held-out MiniWoB task families, respectively.

The primary contrast specified before final evaluation is soft-min minus BCE. Our original analysis uses bootstrap resampling (Efron, 1979) with 10,000 replicates, drawing seeds and then task means within each drawn seed. We report percentile intervals from these replicates for the two policy contrasts. A paired randomization test independently flips signs of 72 seed-task effects. Holm correction covers the soft-min/BCE and BCE/base comparisons (Holm, 1979). Cell-level sign flips do not preserve dependence among tasks sharing a trained checkpoint, which can overstate independent training evidence.

Two post-hoc analyses examine that dependence. A crossed bootstrap uses 50,000 replicates, resampling seeds and tasks independently but applying a common task draw across sampled seeds. A seed-block sign test flips whole-seed effects and enumerates eight assignments. Its smallest nonzero two-sided p-value is 0.25. Resampling frequency does not increase the number of trained policies; three seeds provide limited power (Henderson et al., 2018; Colas et al., 2018; Agarwal et al., 2021). Task win/tie/loss counts and leave-one-task-out effects are additional descriptive checks. Appendix C specifies the sampling units and assumptions of both bootstrap procedures.

# 5 Results

## 5.1 Verifier discrimination and consistency

Table 1: Validation metrics, averaged across three seeds. AUROC and MFS are higher-is-better; FAR, FRR, ECE, and ICE are lower-is-better. Bold identifies the best mean. Threshold selection uses these same validation examples.

| Objective | AUROC | FAR | FRR | ECE | ICE | MFS |
|---|---|---|---|---|---|---|
| BCE | 0.9946 | 0.0658 | 0.0028 | 0.0604 | **0.0104** | 0.8488 |
| + invariance | 0.9841 | 0.0952 | 0.0319 | 0.1326 | 0.0166 | 0.6097 |
| + sensitivity | **0.9983** | **0.0462** | **0.0014** | **0.0542** | 0.0149 | **0.8898** |
| Full | 0.9896 | 0.0798 | 0.0083 | 0.0913 | 0.0191 | 0.7280 |

Table 1 reports area under the receiver operating characteristic curve (AUROC), false-acceptance rate (FAR), false-rejection rate (FRR), and expected calibration error (ECE) (Guo et al., 2017). Invariance consistency error (ICE) averages absolute probability differences on preserving pairs. The recorded minimal-flip sensitivity (MFS) metric averages the signed source-minus-target probability difference on changing pairs. These diagnostics use probabilities, whereas the auxiliary losses act on logits. Appendix C defines threshold selection and ten-bin ECE; Figure 2 shows seed variability.

Sensitivity-only achieved the highest mean AUROC and largest changing-pair separation. Adding invariance alone lowered mean AUROC relative to BCE and increased validation ICE. Full had weaker means than sensitivity-only on all six metrics. BCE already trained on banner examples with matched labels; the explicit invariance penalty supplied no observed improvement at the tested unit weight. All twelve thresholded runs satisfied the FAR constraint on the same validation set used for threshold selection.

## 5.2 Reward support and final policy success

The action-free and mean-reward conditions each failed the early dispersion gate in one seed, excluding both methods from final evaluation. Their final success rates were therefore not measured. Appendix F records the failed gates and the outcomes of eligible conditions.

Tables 2 and 3 report success rates and paired policy contrasts. Soft-min showed no resolved benefit over BCE under the original analysis or the seed-block check. The interval permits effects of several percentage points in either direction, leaving equivalence unresolved. ID and OOD contrasts were both negative; Appendix F reports their intervals.

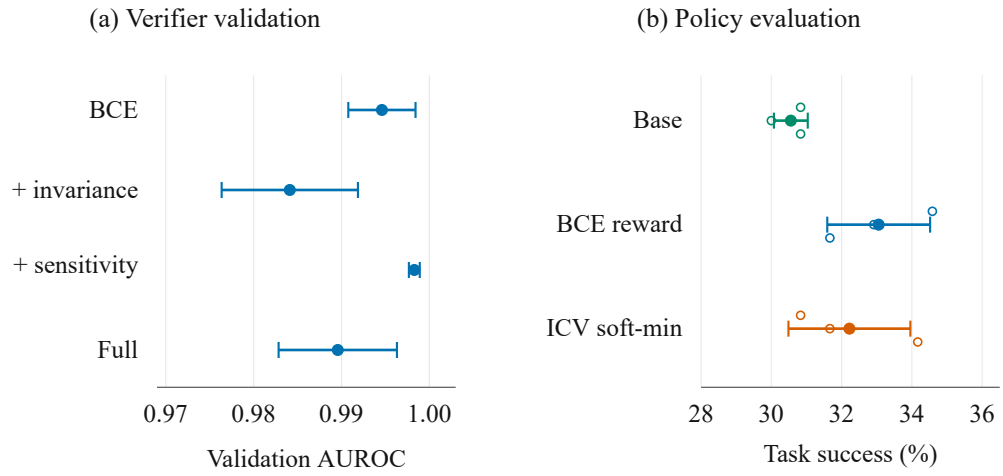


Figure 2: Verifier AUROC and final policy success. Filled points show means and error bars show sample standard deviations across three seeds, not confidence intervals. Small open circles in panel (b) show individual seed success rates, offset vertically for visibility. Panel (a) is a validation diagnostic; panel (b) evaluates different complete reward packages, not all four verifier ablations.

Table 2: Official success rates in percent. ID denotes training-task families on final instances; OOD denotes held-out task families. Standard deviation (SD) is across three seeds. Each method has 720 episodes.

| Policy | Overall | ID | OOD | Seed SD |
|---|---|---|---|---|
| Frozen base | 30.56 | 51.94 | 9.17 | 0.48 |
| BCE reward | **33.06** | **55.83** | **10.28** | 1.46 |
| ICV soft-min | 32.22 | 54.72 | 9.72 | 1.73 |

BCE exceeded the frozen base in every seed, but inference depended on the statistical unit. The cell-level test reached its adjusted significance threshold; the whole-seed test did not. The observed gain applies to this task panel, with substantial uncertainty across independent training runs.

Table 3: Policy contrasts. Effects and confidence intervals (CI) are in percentage points. CI uses the original nested seed-task bootstrap; Holm p uses the original cell-level sign test. Seed-block p is an explicitly post-hoc dependence check with only three seeds. Appendix F reports crossed-bootstrap intervals.

| Contrast | Effect | 95% CI | Holm p | Seed-block p |
|---|---|---|---|---|
| Soft-min minus BCE | -0.83 | [-4.58, 3.19] | 0.5973 | 0.75 |
| BCE minus base | +2.50 | [0.42, 5.28] | 0.0396 | 0.25 |

### 5.3 Task concentration and failure descriptors

Task-level differences were uneven (Figure 3). Relative to BCE, soft-min improved fewer families than it worsened; BCE showed the opposite pattern relative to the base. Removing any single family preserved each aggregate contrast's direction. Table 7 reports task win/tie/loss counts and leave-one-task-out ranges.

Held-out success was concentrated in click-collapsible, which every method solved on all instances. Seven held-out families had no successes under any method. Thus, the aggregate OOD rate mainly reflects one solved family rather than transfer across the selected task set.

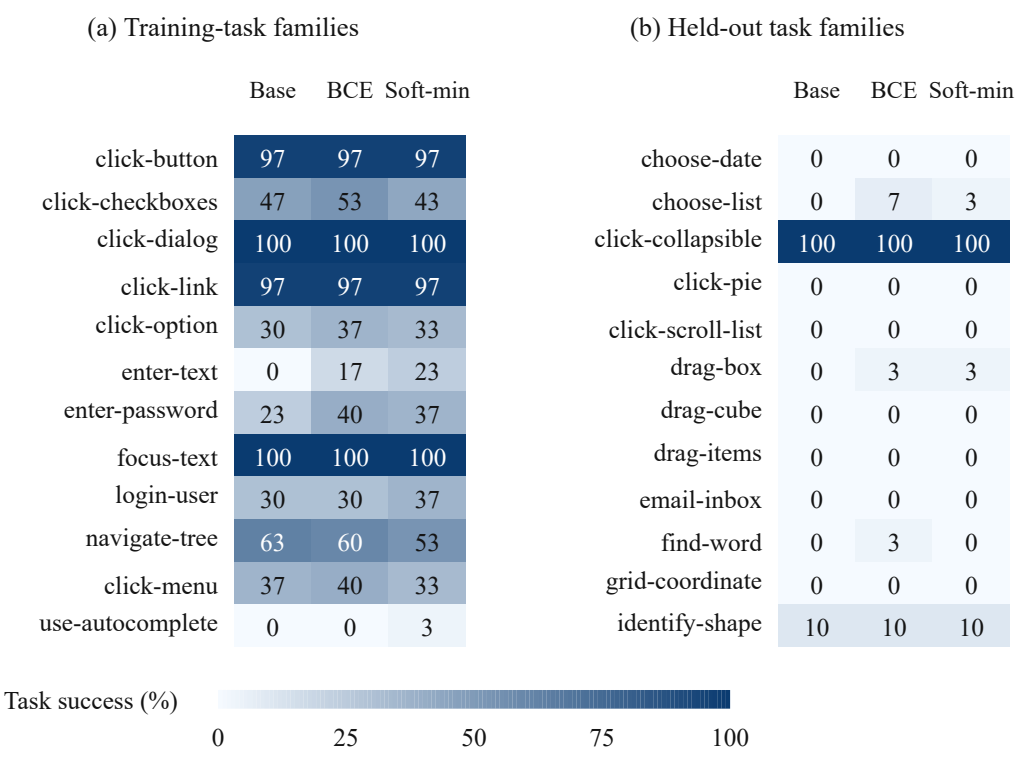


Figure 3: Post-hoc task success, pooled over three seeds and ten instances per seed. Cells show rounded percentages and share a 0-100 scale. Left: training-task families; right: held-out task families. Each cell has 30 episodes. Seven held-out families have zero successes for all methods, while click-collapsible accounts for most held-out successes.

Table 8 also counts failures reaching the 12-action horizon and failures containing at least three repetitions of a single action. These overlapping descriptors summarize behavior without distinguishing planning, observation, or parser errors.

### 5.4 Screenshot-substitution diagnostics

We rescore the 1,440 final trajectories from the two learned-reward policies under terminal-image substitutions. The initial-image test replaces the terminal screenshot with the initial screenshot; the previous-image test uses the preceding screenshot. The failure-swap test uses a terminal screenshot from a failed episode of the same task family. The original action trace is retained. A failure donor need not share the original goal or instance, and none of these tests executes a state-changing action. They measure response to inconsistent visual evidence rather than verified counterfactual task outcomes.

Table 4: Post-hoc score diagnostics on each method's own policy trajectories. Gap is mean absolute reward-outcome difference. Drops are original minus substituted reward; larger positive values indicate greater score reduction. Coverage is the fraction with an available failure donor.

| Reward | Gap | Initial drop | Previous drop | Failure drop | Coverage |
|---|---|---|---|---|---|
| BCE | 0.4778 | 0.0221 | 0.0100 | 0.1181 | 0.4550 |
| ICV soft-min | 0.4498 | 0.0119 | 0.0112 | 0.0147 | 0.4367 |

Soft-min showed a smaller absolute reward-outcome gap and a weaker response to failed-terminal substitution (Table 4). Its success-failure reward separation was also smaller (Table 9), consistent with a compressed outcome signal. However, each scorer evaluated trajectories generated by its

own policy. A common held-out trajectory bank is needed to separate scorer behavior from policy-induced distribution differences.

Table 9 reports soft-min's five-view score spread; BCE's multi-view spread was not measured. In this failure-heavy set, uniformly lower scores can reduce the absolute reward-outcome gap without improving ranking or policy success. The gap therefore complements, rather than replaces, discrimination and outcome measurements.

### 5.5 Resource measurements

The experiments used a 96 GB Blackwell GPU for policy training and a 48 GB RTX 4090 for verifier serving. Verifier training peaked at approximately 10.6 GiB. A 100-update verifier run plus 478 validation scores took about 115 seconds on Blackwell and 195 seconds on RTX 4090. Recorded segments of the six eligible policy runs averaged 35.8 minutes and totaled approximately 3.6 hours. Maximum recorded policy-side allocation was 5.59 GiB. These timers exclude concurrent verifier serving, failed conditions, collection, evaluation, and pre-resume work. End-to-end GPU cost and budget-matched throughput were not measured.

## 6 Discussion and Limitations

The verifier ablation separates augmentation from explicit pair constraints. Outcome-labeled banner examples already produced stable BCE predictions; adding logit matching did not improve that stability at the tested weight. Combining invariance with sensitivity also failed to retain sensitivity-only's discrimination. Loss weights and pair sampling may explain this interaction and warrant further testing. More broadly, reward-model ensembles reduce but do not always eliminate overoptimization (Coste et al., 2024). Agreement supplies limited protection when predictors share errors, including across ensemble members with different pretraining seeds (Eisenstein et al., 2024).

The policy result leaves the effects of verifier training and aggregation unresolved because both change in the comparison. Crossing BCE and Full checkpoints with single-view and soft-min scoring would distinguish those effects. Including sensitivity-only would test whether the strongest offline discriminator provides a better training signal. Those comparisons require matched policy initialization, rollout budgets, and reward-inference accounting.

The fixed evidence interface imposes a separate information limit. SuccessVQA and AdaptRubric explicitly condition judgments on task requirements (Du et al., 2023; Xiong et al., 2026b), whereas this verifier has no separate textual goal. Presentation changes also cannot recover application state absent from the screenshots. ProRe and IRA address that problem by acquiring additional evidence, at the cost of tool access and interaction (Dai et al., 2026; Shi et al., 2026). Goal-conditioned and active verifiers therefore test different capabilities from passive visual consistency.

Generalization remains limited by constructed validation negatives, one model family, and a small MiniWoB task panel with concentrated held-out success. A natural-failure bank, including high-reward mistakes, would test whether validation discrimination survives the policy's visited distribution (Lee et al., 2026; Zhang et al., 2026). Three training seeds leave modest effects uncertain (Henderson et al., 2018; Colas et al., 2018). The policy receives text observations, while banner auditing retains validator access; deployment without those interfaces remains untested. Prompt mismatches and contradictory action evidence in screenshot substitutions are additional sources of reward uncertainty.

## 7 Conclusion

This MiniWoB study evaluates paired verifier training and its use in policy rewards. Sensitivity supervision produced the strongest validation discrimination, while Full five-view soft-min scoring showed no resolved success gain over BCE single-view reward. Training and aggregation remain confounded, leaving sensitivity-only's downstream value open. Task-level outcomes and evidence-substitution diagnostics add information that validation AUROC and pooled policy success alone do not capture.

## Reproducibility Statement

The accompanying project contains instance manifests, collection and training scripts, model and optimizer configurations, per-seed predictions, policy episode records, and figure-generation code. Original aggregate results and the additional post-hoc analysis are stored separately. Appendices specify task partitions, metrics, implementation settings, and execution deviations.

## AI Use Statement

We used generative AI tools for writing assistance and polishing, literature retrieval and discovery, and research ideation and execution. These uses included revising manuscript text for clarity, identifying related work, planning and designing experiments, assisting with implementation and interpretation of results, and preparing figures. The authors remain responsible for the accuracy and integrity of the manuscript, including its claims, analyses, citations, and artifacts.

## References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. DigiRL: Training In-The-Wild Device-Control Agents with Autonomous Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 12461–12495, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1704ddd0bb89f159dfe609b32c889995-Abstract-Conference.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A Group-Theoretic Framework for Data Augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. URL https://www.jmlr.org/papers/v21/20-163.html.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments. *arXiv preprint arXiv:1806.08295*, 2018. URL https://arxiv.org/abs/1806.08295v2.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward Model Ensembles Help Mitigate Overoptimization. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/dda7f9378a210c25e470e19304cce85d-Abstract-Conference.html.

Chaoqun Cui, Jing Huang, Shijing Wang, Liming Zheng, Qingchao Kong, and Zhixiong Zeng. Agentic Reward Modeling: Verifying GUI Agent via Progressive Trajectory-Grounded Interaction. *arXiv preprint arXiv:2602.00575*, 2026. URL https://arxiv.org/abs/2602.00575.

Gaole Dai, Shiqi Jiang, Ting Cao, Yuqing Yang, Yuanchun Li, Rui Tan, Mo Li, and Lili Qiu. ProRe: A Proactive Reward System for GUI Agents via Reasoner-Actor Collaboration. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.21823v2.

Yong Du, Yuchen Yan, Fei Tang, Zhengxi Lu, Chang Zong, Weiming Lu, Shengpei Jiang, and Yongliang Shen. Test-Time Reinforcement Learning for GUI Grounding via Region Consistency. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 30593–30601, 2026. doi: 10.1609/aaai.v40i36.40314. URL https://ojs.aaai.org/index.php/AAAI/article/view/40314.

Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-Language Models as Success Detectors. In *Conference on Lifelong Learning Agents*, volume 232, pp. 120–136. PMLR, 2023. URL https://proceedings.mlr.press/v232/du23b.html.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or Herding? Reward Model Ensembles Mitigate but do not Eliminate Reward Hacking. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5u1GpUkKtG.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Qi Han, and Yuxia Wang. Reward Shaping to Mitigate Reward Hacking in RLHF. *arXiv preprint arXiv:2502.18770*, 2026. URL https://arxiv.org/abs/2502.18770v7. Version 7, September 2026.

Yuwei Fu, Haichao Zhang, Di Wu, Wei Xu, and Benoit Boulet. FuRL: Visual-Language Models as Fuzzy Rewards for Reinforcement Learning. In *International Conference on Machine Learning*, volume 235, pp. 14256–14274. PMLR, 2024. URL https://proceedings.mlr.press/v235/fu24j.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, volume 70, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Lukas Helff, Quentin Delfosse, David Steinmann, Ruben Härle, Hikaru Shindo, Patrick Schramowski, Wolfgang Stammer, Kristian Kersting, and Felix Friedrich. LLMs Gaming Verifiers: RLVR can Lead to Reward Hacking. In *ICLR Workshop on Logical Reasoning of Large Language Models*, 2026. URL https://arxiv.org/abs/2604.15149v1.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning That Matters. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11694. URL https://ojs.aaai.org/index.php/AAAI/article/view/11694.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.02781v2.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6 (2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Haowen Hu, Pengzhou Cheng, Zheng Wu, Lingzhong Dong, Gongshen Liu, and Zhuosheng Zhang. Faithful Mobile GUI Agents with Guided Advantage Estimator. *arXiv preprint arXiv:2605.01208*, 2026. URL https://arxiv.org/abs/2605.01208v1.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. In *Annual Meeting of the Association for Computational Linguistics*, pp. 881–905, 2024. doi: 10.18653/v1/2024.acl-long.50. URL https://aclanthology.org/2024.acl-long.50/.

Thibault Le Sellier de Chezelles, Maxime Gasse, Alexandre Lacoste, Massimo Caccia, Alexandre Drouin, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Graham Neubig, Quentin Cappart, Russ Salakhutdinov, and Nicolas Chapados. The BrowserGym Ecosystem for Web Agent Research. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=5298fKGmv3.

Tony Lee, Andrew Wagenmaker, Karl Pertsch, Percy Liang, Sergey Levine, and Chelsea Finn. RoboReward: General-Purpose Vision-Language Reward Models for Robotics. *arXiv preprint arXiv:2601.00675*, 2026. URL https://arxiv.org/abs/2601.00675v2.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lirui Luo, Guoxi Zhang, Hongming Xu, Yaodong Yang, Cong Fang, and Qing Li. MVR: Multi-view Video Reward Shaping for Reinforcement Learning. In *International Conference on Learning Representations*, 2026. URL https://mvr-rl.github.io/.

Yecheng Jason Ma, Joey Hejna, Ayzaan Wahid, Chuyuan Fu, Dhruv Shah, Jacky Liang, Zhuo Xu, Sean Kirmani, Peng Xu, Danny Driess, Ted Xiao, Jonathan Tompson, Osbert Bastani, Dinesh Jayaraman, Wenhao Yu, Tingnan Zhang, Dorsa Sadigh, and Fei Xia. Vision Language Models are In-Context Value Learners. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.04549.

Ted Moskovitz, Aaditya Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca Dragan, and Stephen McAleer. Confronting Reward Model Overoptimization with Constrained RLHF. In *International Conference on Learning Representations*, pp. 21998–22025, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5eee634cb9729b8bcc2ec9f2a46a74ae-Abstract-Conference.html.

Youssef Mroueh. Reinforcement Learning with Verifiable Rewards: GRPO's Effective Loss, Dynamics, and Success Amplification. *arXiv preprint arXiv:2503.06639*, 2025. URL https://arxiv.org/abs/2503.06639v4.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning. In *International Conference on Learning Representations*, pp. 79791–79821, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c66e1fcc9691aae706250638f36f681b-Abstract-Conference.html.

Xinyu Qiu, Yunzhu Zhang, Heng Jia, Shuheng Shen, Changhua Meng, and Linchao Zhu. VISTA: View-Consistent Self-Verified Training for GUI Grounding. *arXiv preprint arXiv:2606.14579*, 2026. URL https://arxiv.org/abs/2606.14579v1.

Chris Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents. In *International Conference on Learning Representations*, pp. 406–441, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/01a83bc2f2732a58e6aa731e659e7101-Abstract-Conference.html.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-Language Models are Zero-Shot Reward Models for Reinforcement Learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.12921v2.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300v3.

Chenrui Shi, Yuwei Wu, Yang Liu, Ruining Feng, Zirui Shang, Zhi Gao, Lifeng Fan, and Che Sun. Interactive Reward Agent: GUI Task Evaluation via Environment-State Verification. *arXiv preprint arXiv:2607.25904*, 2026. URL https://arxiv.org/abs/2607.25904v2.

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. World of Bits: An Open-Domain Platform for Web-Based Agents. In *International Conference on Machine Learning*, pp. 3135–3144. PMLR, 2017. URL https://proceedings.mlr.press/v70/shi17a.html.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and Characterizing Reward Gaming. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. RL-VLM-F: Reinforcement Learning from Vision Language Foundation Model Feedback. In *International Conference on Machine Learning*, volume 235, pp. 51484–51501. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24bn.html.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52040–52094, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Tao Xiong, Xavier Hu, Yurun Chen, Yuhang Liu, Changqiao Wu, Pengzhi Gao, Wei Liu, Jian Luan, and Shengyu Zhang. GUI-PRA: Process Reward Agent for GUI Tasks. *arXiv preprint arXiv:2509.23263*, 2026a. URL https://arxiv.org/abs/2509.23263v3. Version 3, August 2026.

Tao Xiong, Xavier Hu, Wenkai Wang, Qinzhuo Wu, Changqiao Wu, Pengzhi Gao, Wei Liu, Jian Luan, and Shengyu Zhang. Task-Adaptive Rubrics for GUI Reward Modeling. *arXiv preprint arXiv:2608.24174*, 2026b. URL https://arxiv.org/abs/2608.24174v1.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. Chasing the Tail: Effective Rubric-based Reward Modeling for Large Language Model Post-Training. In *International Conference on Learning Representations*, pp. 133430–133457, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d8183233dbb325a7e165909042a47e15-Abstract-Conference.html.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *International Conference on Learning Representations*, pp. 15585–15606, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

Xin Zhou, Yiwen Guo, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. Self-Consistency of the Internal Reward Models Improves Self-Rewarding Language Models. *arXiv preprint arXiv:2502.08922*, 2025. URL https://arxiv.org/abs/2502.08922v1.

## A Exact Task Partitions

Training/ID families are click-button, click-checkboxes, click-dialog, click-link, click-option, enter-text, enter-password, focus-text, login-user, navigate-tree, click-menu, and use-autocomplete. Held-out/OOD families are choose-date, choose-list, click-collapsible, click-pie, click-scroll-list, drag-box, drag-cube, drag-items, email-inbox, find-word, grid-coordinate, and identify-shape. Each final method-seed cell contains ten instances of each family. Training, verifier validation, and final evaluation use distinct instance scopes. Seeds and task assignments are recorded before final policy evaluation.

## B Implementation Details

The model identifiers are Qwen/Qwen3-VL-2B-Instruct for the policy and Qwen/Qwen3-VL-4B-Instruct for the verifier. Both use bfloat16 (BF16) backbones with LoRA rank 16, alpha 32, and dropout 0.05 on query, key, value, output, gate, up, and down projections. The verifier uses AdamW, learning rate $2 \times 10^{-5}$, and gradient norm clipping at one. The policy learning rate is $10^{-5}$. Policy checkpoint model cards record TRL 1.9.2, Transformers 5.14.1, and PyTorch 2.12.1+cu130; verifier adapter configurations record PEFT 0.20.0.

Accessibility observations expose BrowserGym element identifiers rather than Chrome node identifiers. Generic interactive elements inherit visible descendant text when their own accessible name is empty. Duplicate element identifiers are removed. Evaluation parses one JSON action at a time and executes only the first object if several are emitted. Available actions include click, double-click, fill, focus, hover, key press, scroll, selection, upload, and no-op. Parse and action failures remain in the episode records and count toward the final endpoint.

The verifier service receives initial and terminal images plus actions, without a separate textual goal or oracle scalar. For action-free scoring, actions are removed. Single-view BCE and multi-view ICV use seed-matched verifier checkpoints. The service runs separately from the policy trainer. Training and serving use different instruction strings; the reported checkpoints and scores reflect those respective interfaces.

## C Metrics and Statistical Details

Verifier threshold candidates are the observed probabilities together with 0, 0.5, and 1. Candidates satisfying FAR at most 0.10 are ordered by true-positive rate, then lower FAR, then lower threshold. AUROC is threshold-free. Ten-bin ECE is the occupancy-weighted absolute difference between mean predicted probability and positive-label frequency within each bin. ICE averages $|p_i - p_j|$ over banner pairs; MFS averages $p_i - p_j$ over successful-source/changing-target pairs. Neither diagnostic averages over the three additional inference-time pixel probes.

For methods $a, b$, let $d_{st}$ be the mean paired outcome difference over ten instances for seed $s$ and task $t$. The effect is $\hat{\Delta} = (3 \cdot 24)^{-1} \sum_{s,t} d_{st}$. The original bootstrap samples three seeds with replacement and, independently within each selected seed, 24 tasks. The added crossed bootstrap draws one common list of 24 tasks and applies it to all three sampled seeds. Both use percentile intervals; neither resamples individual instances inside a cell. Thus, within-cell instance uncertainty is not separately modeled.

The original randomization test flips individual $d_{st}$ signs. The added seed-block diagnostic enumerates $(\eta_1, \eta_2, \eta_3) \in \{-1, 1\}^3$ for $3^{-1} \sum_s \eta_s \bar{d}_s$. Its two-sided p-value is the fraction of eight assignments at least as extreme as the observed effect. This requires seed-effect sign exchangeability under its null and is presented as a sensitivity diagnostic, not a replacement confirmatory analysis. Original and post-hoc results remain separately labeled.

## D Execution Record and Diagnostic Scope

The run registry includes twelve verifier runs, twelve planned learned-reward policy attempts, and three base-policy evaluations. Method eligibility is fixed before final-oracle evaluation. Action-free

and mean-reward conditions each fail one seed's early gate and do not enter the final comparison. A stalled soft-min seed-17 run resumes from update 125, replaying updates 126-128; recorded wall segments do not include all pre-resume work. A few cumulative seed-17 success counters were visible after eligibility and analysis choices were fixed. Evaluation was therefore not operator-blinded, although the fixed episodes and comparison definitions were retained.

For substitution diagnostics, a banner difference mask is transferred onto the substituted image within the fixed bottom-right region. If dimensions differ, the bottom-right-aligned overlap is used, retaining the target dimensions and pixels outside that region. This repair concerns diagnostic image construction, not policy training. Failure swaps select an available failed terminal from the same task family; only covered cases contribute to the corresponding drop. Each scorer is evaluated on its own policy's records, so cross-method score differences are not paired common-input comparisons.

## E Soft-Min Bounds

Let $m = \min_k v_k$. Factoring out $\exp(-m/\beta)$ gives $R_\beta = m - \beta \log[K^{-1} \sum_k \exp(-(v_k - m)/\beta)]$. The bracket lies in $[1/K, 1]$, establishing $m \leq R_\beta \leq m + \beta \log K$. Jensen's inequality gives $K^{-1} \sum_k \exp(-v_k/\beta) \geq \exp(-\bar{v}/\beta)$ and hence $R_\beta \leq \bar{v}$. Differentiating the log-sum-exp yields the normalized negative-exponential weights in Section 3.4.

For $v_k = \bar{v} + \delta_k$ with $\sum_k \delta_k = 0$ and $\max_k |\delta_k|/\beta$ small, expanding the exponential and logarithm gives $R_\beta = \bar{v} - (2\beta K)^{-1} \sum_k \delta_k^2 + O(\max_k |\delta_k|^3/\beta^2)$. This is a local approximation, not a calibrated lower confidence bound. The identities explain aggregation behavior only; they make no assumption connecting a verifier probability to the true task outcome.

## F Detailed Results and Sensitivity Analyses

Tables 5–9 report seed-level outcomes and supplementary diagnostics, with post-hoc analyses identified in their captions. Each seed-level success rate in Table 5 uses 240 episodes.

Table 5: Overall policy success in percent for every evaluated seed.

| Policy | Seed 17 | Seed 29 | Seed 43 |
|---|---|---|---|
| Frozen base | 30.83 | 30.00 | 30.83 |
| BCE reward | 31.67 | 32.92 | 34.58 |
| ICV soft-min | 34.17 | 31.67 | 30.83 |

The action-free condition fails its update-25 gate in seed 29 with mean within-group reward standard deviation 0.00420. The mean-reward condition fails the same gate in seed 43 with 0.00496. Neither method enters the final comparison. Table 6 retains both bootstrap analyses: the original procedure samples tasks inside each sampled seed, whereas the post-hoc crossed procedure uses a common task draw across sampled seeds. Table 3 reports their associated original and seed-block test statistics without equating the underlying independence assumptions.

Table 6: Overall effects and 95% bootstrap confidence intervals in percentage points. Crossed intervals are post-hoc.

| Contrast | Effect | Original CI | Crossed CI |
|---|---|---|---|
| Soft-min minus BCE | -0.83 | [-4.58, 3.19] | [-4.86, 3.75] |
| BCE minus base | +2.50 | [0.42, 5.28] | [0.14, 5.97] |

For soft-min minus BCE, the descriptive ID effect is -1.11 percentage points with original interval [-8.33, 6.39]; the OOD effect is -0.56 points with interval [-1.67, 0.56]. These stratum intervals are descriptive rather than additional confirmatory endpoints. The original unadjusted two-sided randomization p-values are 0.5973403 for soft-min minus BCE and 0.0197980 for the BCE/base comparison; Table 3 gives Holm-adjusted values. The post-hoc whole-seed sign p-values are 0.75 and 0.25, respectively.

Table 7: Post-hoc task-level contrasts. Win/tie/loss counts use task means over all seeds. Leave-one-task-out (LOTO) ranges are in percentage points.

| Contrast | Wins | Ties | Losses | LOTO effect range |
|---|---|---|---|---|
| Soft-min minus BCE | 3 | 14 | 7 | [-1.16, -0.43] |
| BCE minus base | 8 | 15 | 1 | [1.88, 2.75] |

Table 8 summarizes raw episode counts. Horizon failures reach at least 12 actions. Repetition failures contain at least three actions with only one distinct action string. The two failure descriptors overlap and are not causal error categories. All methods obtain 30 successes on click-collapsible, accounting for 30 of 33, 30 of 37, and 30 of 35 OOD successes, respectively.

Table 8: Episode counts underlying the failure and OOD-concentration summaries. Each policy has 720 final episodes, of which 360 are OOD.

| Policy | Failures | Horizon | Repetition | OOD successes |
|---|---|---|---|---|
| Frozen base | 500 | 256 | 102 | 33 |
| BCE reward | 482 | 228 | 91 | 37 |
| ICV soft-min | 488 | 235 | 96 | 35 |

Table 9: Additional post-hoc reward diagnostics. Separation is mean reward on successful minus failed trajectories. BCE has no multi-view spread measurement because its conventional scorer uses one view.

| Reward | Success-failure separation | Five-view spread |
|---|---|---|
| BCE | 0.1421 | Not measured |
| ICV soft-min | 0.0187 | 0.0579 |

Full-precision outcome records, reward scores, and task-level means accompany the experiment artifacts.