FACT: FEEDBACK ATTRIBUTION WITH CONTROLLED TRAJECTORIES FOR LONG-HORIZON ARTIFACT AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon artifact agents transform persistent environments through many interdependent tool actions before receiving feedback on the completed artifact. This delayed and often subjective feedback makes it difficult to determine which earlier decisions improved the outcome and to turn that evidence into reliable policy updates. PPO and group-relative policy optimization methods such as GRPO assign rollout-level or group-relative advantages to sampled sequences; delayed terminal feedback and stochastic tool execution leave action-level credit unresolved. We therefore introduce FACT, a critic-based PPO formulation that defines action credit by the expected effect of a committed action on terminal preference. FACT executes alternative actions from the same restored workspace and continues each branch under a shared policy, yielding matched long-horizon comparisons. These comparisons supervise an action-conditioned transition critic for ordinary rollouts, while held-out interventions calibrate the resulting updates. We further introduce LongEditBench, a source-grounded benchmark for long-horizon video editing and contextual revision under follow-up user feedback. Experiments show that FACT improves completion-adjusted artifact quality and contextual revision, while controlled evaluations demonstrate more faithful action attribution. Consistent gains across transfer benchmarks further validate its effectiveness across persistent artifact environments.

1 INTRODUCTION

Language-model agents increasingly produce persistent artifacts through tool-mediated workflows, including executable applications and long-form video editing (Trivedi et al., 2024; Yan et al., 2026b; Cao et al., 2026). We call these systems *long-horizon artifact agents*: they use tools over multiple turns to create and revise persistent outputs. Users inspect and revise these artifacts, so an output evolves through many state-changing actions. An early selection constrains later composition, a local revision can alter the coherence of the whole artifact, and user preference becomes apparent only after these choices have been assembled into an output (Yan et al., 2026a;b). A terminal judgment therefore leaves two linked uncertainties: which committed actions changed expected final quality, and whether a low outcome reflects the action or what happened during its execution.

Task-local preference comparisons provide a useful starting point for subjective artifact quality (Yan et al., 2026a). PPO uses a clipped actor update with a return or critic advantage, while GRPO and GSPO use relative group or sequence objectives; GRPB extends this family to long-horizon artifact workflows with rank-based credit (Shao et al., 2024; Zheng et al., 2025; Yan et al., 2026a). These signals remain attached to a rollout, group, or realized branch: PPO, GRPO, and GSPO leave action-level credit implicit, while GRPB can still assign a low score to useful actions when an export fails or an unexpected state appears. Existing policy signals thus conflate the quality of a committed action with execution variation.

These gaps motivate a local counterfactual question: *how would expected final quality change if the agent committed to a different action from the same workspace, and how much of the observed outcome comes from the commitment versus its execution?* Figure 1 illustrates the distinction. Counterfactual replay and intervention-based diagnosis use related comparisons (Shah, 2026; Zhang,

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.

2026); fixing the preceding history and continuation policy makes the comparison local while preserving long-horizon consequences. Averaged continuations estimate the expected action effect, whereas the observed post-action outcome supplies an execution residual. This separation provides the two signals used by FACT for policy learning and diagnosis.

We develop this principle in FACT, *Feedback Attribution with Controlled Trajectories*, as a critic-based PPO update. FACT keeps PPO's clipped actor and replaces its rollout advantage with calibrated action credit. Controlled trajectories branch from a shared workspace and execute alternative actions under a fixed continuation policy; their expected terminal contrast is Agency credit. An action-conditioned transition critic transfers this branch supervision to ordinary rollouts, while the execution residual, Revelation, calibrates updates and supports diagnosis. Figure 2 summarizes these components.

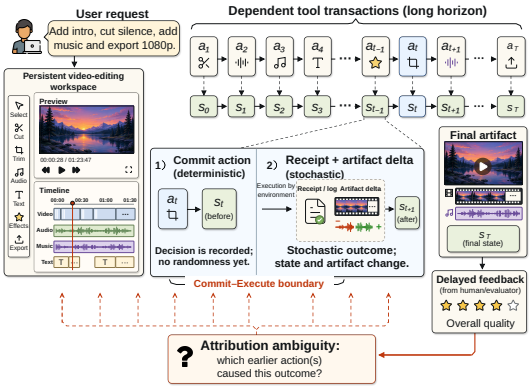


Figure 1: Why rollout-level and group-relative advantages are insufficient for long-horizon artifact agents. A trajectory receives terminal feedback only after many state-changing actions; an action may be useful even when an unpredictable tool response or export failure lowers the realized outcome, motivating separate action and execution attribution.

To evaluate action credit as artifacts and user requests evolve, we construct LongEditBench by manually adapting source-grounded interactions from LongShOTBench into executable editing briefs. The benchmark contains 337 requests over 21 videos, including 84 follow-ups (Kurpath et al., 2025; Yan et al., 2026b). Its contextual follow-ups test incorporation of a new request while preserving earlier requirements. Transfer uses 100 AgenticVBench tasks across four families, 2,722 SpreadsheetBench Pass@1 tasks, and 61 held-out SpreadsheetBench 2 workflow tasks (Cao et al., 2026; Ma et al., 2024; Zhu et al., 2026). Matched-budget comparisons use a shared Qwen3.5 policy family and include several PPO variants, GRPO, GSPO, GRPB, and Direct-Q for attribution (Shao et al., 2024; Zheng et al., 2025; Yan et al., 2026a). Sibling interventions test attribution fidelity, while the transfer suites test policy utility across artifact domains.

In summary, FACT defines action credit as the expected terminal effect of a committed action, separates that effect from execution residuals, and transfers controlled-branch supervision through an action-conditioned transition critic into a calibrated PPO update. We introduce LongEditBench with contextual follow-ups and evaluate both attribution fidelity and policy utility across persistent artifact domains. Results show gains in completion-adjusted artifact quality and multi-turn contextual revision, including follow-up instruction satisfaction and source grounding; attribution audits improve action-effect direction recovery, while cross-domain tests achieve the highest average artifact utility.

2 Related Work

Temporal and counterfactual credit assignment Return redistribution and counterfactual advantage estimation address complementary aspects of delayed supervision. RUDDER learns contributions that redistribute trajectory return (Arjona-Medina et al., 2019), while Hindsight Credit Assignment and COCOA associate earlier actions with subsequent rewarding outcomes (Harutyunyan et al., 2019; Meulemans et al., 2023). Counterfactual Credit Assignment uses future-conditioned baselines to account for exogenous variation (Mesnard et al., 2020); Direct Advantage Estimation and its off-policy extension estimate action-dependent contributions from return (Pan et al., 2022; Pan & Schölkopf, 2024). GRPB studies long-horizon credit assignment in artifact workflows (Yan et al., 2026a), placing it alongside policy-optimization baselines in our comparisons. FACT makes this target operational for artifact agents through restored-workspace interventions and an action-conditioned transition critic that models the intervening tool response and artifact change.

Process supervision and credit for language agents Process supervision supplies intermediate correctness or progress signals for long reasoning trajectories (Uesato et al., 2022; Lightman et al.,

2024; Wang et al., 2024; Cui et al., 2025). For interactive agents, POAD connects inter-action credit to within-action token optimization (Wen et al., 2024); Turn-PPO formalizes turn-level optimization (Li et al., 2026); AgentPRM learns step-wise promise and progress (Xi et al., 2026); GiGPO compares actions at repeated anchor states (Feng et al., 2026); and hierarchical methods align supervision with subgoals and control structure (He et al., 2026; Peng et al., 2026). Progress-based estimators include ΔBelief-RL, which measures changes in solution probability (Auzina et al., 2026), and TRACE, which derives telescoping turn rewards from frozen-reference answer likelihoods (Tao et al., 2026). These approaches localize delayed supervision, whereas our setting requires estimating an action's expected effect on a persistent artifact by comparing commitments at a common workspace and modeling their subsequent consequences.

Executed attribution and artifact evaluation Trajectory diagnosis and intervention methods identify consequential steps, estimate effects through reruns, or test executable repairs (Zhang et al., 2025; 2026; Shah, 2026; Lin et al., 2026; Bonagiri et al., 2026; Zhang, 2026). FACT turns this evidence into amortized policy learning. Action-conditioned world models predict tool consequences, while ensembles and selective prediction estimate uncertainty (Yu et al., 2026; Gao et al., 2026; Lakshminarayanan et al., 2017; Geifman & El-Yaniv, 2019). Persistent artifact benchmarks span restorable application state in AppWorld, traceable video editing in Crayotter, post-production tasks in AgenticVBench, long-video evidence seeking in LongShOTBench, and spreadsheet manipulation from test-case evaluation to end-to-end workflows in SpreadsheetBench and SpreadsheetBench 2 (Trivedi et al., 2024; Yan et al., 2026a;b; Cao et al., 2026; Kurpath et al., 2025; Ma et al., 2024; Zhu et al., 2026). LongEditBench adds source-grounded editing with contextual follow-up requests.

3 Feedback Attribution with Controlled Trajectories

Long-horizon artifact agents must infer how tool actions transform persistent artifacts when post-action states and long-term consequences are revealed only after execution. FACT combines interaction history with visual and tool observations and controlled branches, then transfers branch supervision through an action-conditioned transition critic. We formulate long-horizon artifact optimization as critic-based PPO. A state $s_t = (h_t, x_t)$ combines interaction history and persistent artifact state, while action a_t commits a tool call. A complete trajectory is $\tau = (s_1, a_1, o_1, \ldots, s_T, a_T, o_T, x_{T+1})$, ending at artifact x_{T+1} with terminal preference $R_c(\tau)$. FACT retains PPO's clipped actor update and replaces its rollout advantage with controlled action credit.

3.1 Policy Update with Action Credit

Video editing makes the long-horizon credit problem concrete. Early preparation constrains later composition, so the value of a local decision may emerge only after several dependent edits. We therefore start from PPO's clipped policy update and replace its advantage with action credit. PPO uses the ratio $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\text{old}}(a_t \mid s_t)$ and estimated advantage $\widehat{A}_t$ in the clipped objective:

$$\mathcal{L}_{\text{PPO}}(\theta) = -\mathbb{E}_t\Big[\min\Big(r_t(\theta)\widehat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\,\widehat{A}_t\Big)\Big]. \tag{1}$$

FACT replaces PPO's rollout advantage with calibrated Agency credit estimated from controlled branches. For a fixed continuation policy π, $Q^\pi(s,a) = \mathbb{E}_\pi[R_c(\tau) \mid s, \text{do}(a)]$ is the terminal preference after committing a, and $V^\pi(s) = \mathbb{E}_{a'\sim\pi(\cdot|s)}[Q^\pi(s,a')]$ is the pre-commitment baseline. The action credit is

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s). \tag{2}$$

The intervention $\text{do}(a)$ commits a from the current artifact state and lets π continue to completion, so Q^π retains the long-horizon consequences of the commitment. At iteration m, $\pi = \pi_{\text{old}}^{(m)}$ is frozen during collection and scoring; the updated actor is denoted by π_θ. PPO derives action advantages from sampled rollouts, while group-relative optimizers such as GRPO and GSPO derive relative signals from sampled groups; FACT assigns the actor update to the committed tool-action span using counterfactual action credit.

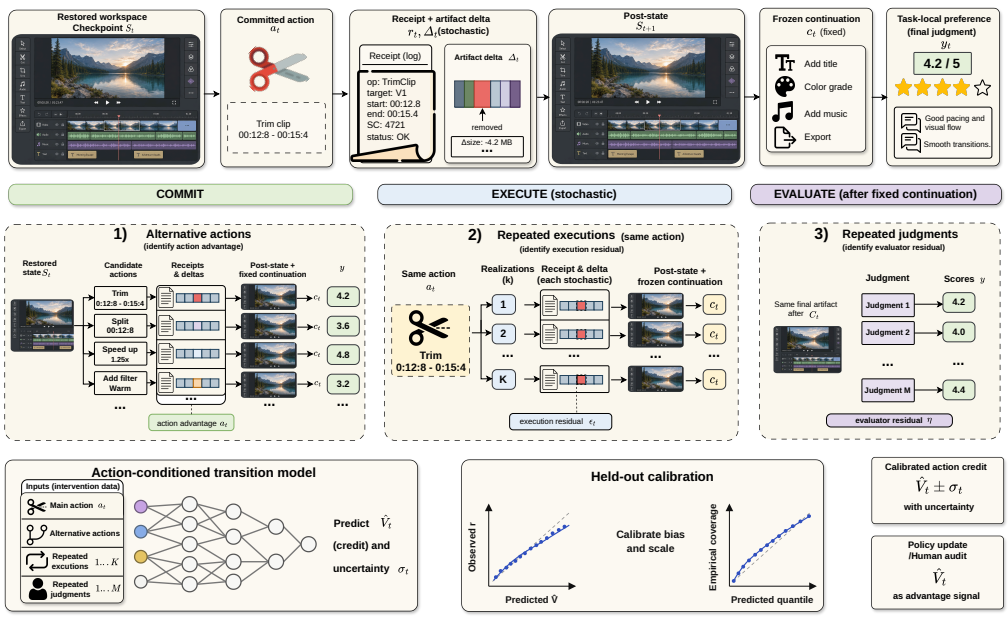


Figure 2: FACT attributes final artifact feedback through controlled trajectories. Alternative actions executed from a shared artifact state supervise their relative contributions. An action-conditioned transition model predicts the response and artifact changes through which an action affects final quality. Held-out executed contrasts calibrate the credit used for policy updates, while recorded trajectories and branches support human inspection.

We score completed artifacts with task-local comparisons. For a frozen reference distribution $\mathcal{D}_c$ and signed preference score pref_c, $R_c(\tau) = \mathbb{E}_{\tau' \sim \mathcal{D}_c}[\text{pref}_c(x_{T+1}, x'_{T+1})] \in [-1, 1]$. The request-specific rubric places matched branches on a common outcome scale. FACT transfers their measured effects to ordinary trajectories and to the PPO objective in Section 3.3. The bottom flow in Figure 2, from the branch observations through predicted credit and held-out calibration to the policy update, summarizes this transfer.

3.2 Controlled branches and execution decomposition

In video editing, each stage produces a restorable project workspace that contains the timeline, source references, tool receipts, and current artifact. In the COMMIT panel of Figure 2, the restored state feeds candidate actions whose post-states are continued under the same frozen policy and scored at the common endpoint. This counterfactual replay follows prior intervention methods (Mesnard et al., 2020; Meulemans et al., 2023; Shah, 2026); fixing the workspace, tool schema, continuation, and endpoint isolates the committed action while preserving downstream effects. Unlike group-relative ranking, the candidate contrast compares actions from the same workspace under the same continuation policy, isolating the expected effect of the commitment.

Repeated executions give $\bar{R}_i(a_{ik}) = n_{ik}^{-1} \sum_{r=1}^{n_{ik}} R_{ikr}$ and the candidate-centered contrast

$$\widehat{\Delta}_i(a_{ik}, a_{i\ell}) = \bar{R}_i(a_{ik}) - \bar{R}_i(a_{i\ell}). \tag{3}$$

The contrast estimates $Q^\pi(s_i, a_{ik}) - Q^\pi(s_i, a_{i\ell})$. The shared state-value baseline cancels, the common history fixes preceding decisions, and the continuation preserves downstream effects. Agency is the expected terminal effect of committing an action from a fixed workspace relative to the pre-commitment baseline. It is tied to an executable workspace intervention and can be tested on held-out sibling branches.

A committed action also encounters execution variability: the realized post-action state can differ across tool executions. Writing $A_t = A^\pi(s_t, a_t)$, the resulting value change decomposes into the expected contribution of the commitment and an execution residual:

$$\underbrace{V^\pi(s_{t+1}) - V^\pi(s_t)}_{\text{realized value change}} = \underbrace{Q^\pi(s_t, a_t) - V^\pi(s_t)}_{A_t\text{: action contribution}} + \underbrace{V^\pi(s_{t+1}) - Q^\pi(s_t, a_t)}_{\xi_t\text{: execution residual}}. \tag{4}$$

Under a fixed continuation policy, $\mathbb{E}[\xi_t \mid s_t, a_t] = 0$. Revelation denotes this realized execution residual caused by the stochastic tool response and the resulting post-action state; repeated branches estimate it for calibration and inspection.

3.3 Action-conditioned transition critic and calibrated PPO update

Action-conditioned transition critic. In a long video edit, an early tool call creates a timeline and file state that later tools must read, while the post-action state is unavailable at decision time and can only be observed after the tool executes. Matched branches therefore supervise an action-conditioned transition critic that predicts tool consequences beyond selected workspaces, extending learned environment models (Yu et al., 2026; Gao et al., 2026) with visual, tool, and interaction-history observations. The bottom of Figure 2 show how these observations become predicted credit, uncertainty, and calibrated action credit before the *Policy update / Human audit* block. We call the joint transition predictor and value head (f_ϕ, g_ω) the action-conditioned transition critic.

We use $z_t = \psi(s_t)$ as the model representation of the history and persistent workspace, where $s_t = (h_t, x_t)$. The observation o_t denotes the tool receipt and execution observation produced after committing a_t. The action representation $e(a_t)$ describes the tool, editing stage, and arguments. The action-conditioned transition critic predicts $(\widehat{o}_t, \widehat{z}_{t+1}) = f_\phi(z_t, e(a_t))$, $\widehat{Q}_t = g_\omega(\widehat{z}_{t+1}, \widehat{o}_t)$, $\widehat{V}_t^+ = g_\omega(z_{t+1}, o_t)$, $\widehat{A}_t = \widehat{Q}_t - V_\psi(z_t)$, and $\widehat{\xi}_t = \widehat{V}_t^+ - \widehat{Q}_t$.

Each ensemble member predicts the transition and values both predicted and observed post-action states. The value function V_ψ estimates value before commitment. Pre-execution $\widehat{Q}_t$ therefore identifies Agency, while post-execution $\widehat{V}_t^+$ identifies Revelation. The estimator jointly fits tool responses, workspace changes, terminal consequences, Agency contrasts, and Revelation residuals. Ensemble variation records uncertainty when credit is transferred to ordinary rollouts. Unlike Direct-Q, which predicts a scalar action value without modeling the intervening artifact transition, the transition critic predicts the post-action consequence and uses it to estimate action-conditioned credit.

The policy objective is $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R_c(\tau)]$. FACT supplies the PPO advantage with critic-estimated action credit, so early choices retain their downstream contribution; the estimator is refreshed as the behavior policy changes.

Reliability calibration and PPO update. Video stages differ in how visibly and reliably a tool changes the artifact, so advantage magnitudes are calibrated by stage. We form u_t from ensemble disagreement and predicted value variance. Held-out restored-workspace contrasts fit its isotonic map to squared error within $g \in \{\text{selection}, \text{construction}, \text{finishing}\}$. Let S_g be the weighted mean squared observed effect and $\mathrm{MSE}_g(u_t)$ the calibrated error. The resulting credit is

$$\bar{A}_t = \frac{1}{B}\sum_{b=1}^{B} \widehat{A}_{t,b}, \qquad w_t = \frac{S_g}{S_g + \mathrm{MSE}_g(u_t)}, \qquad \widetilde{A}_t = w_t \bar{A}_t. \tag{5}$$

Larger held-out error produces a weaker update. Stage estimates are pooled toward the global curve according to calibration support.

Let $\mathcal{I}_t$ contain the tool-name and argument tokens that commit a_t. FACT applies PPO clipping to this span with calibrated Agency credit $\widetilde{A}_t$. Only calibrated Agency credit is supplied to the actor; Revelation is used for calibration and diagnosis.

Each iteration freezes the behavior policy, lagged estimator, and calibration map before computing credit. New trajectories and an independent intervention batch refresh them after the policy update. Tool responses remain context for later decisions and are masked from the loss.

For human inspection, training checkpoints retain complete trajectories, matched branches, attributed credit, and their artifact changes. Comparing these records over training shows which editing behaviors receive increasing reinforcement and whether they follow user intent. Full objective definitions and derivations are provided in Appendix B.

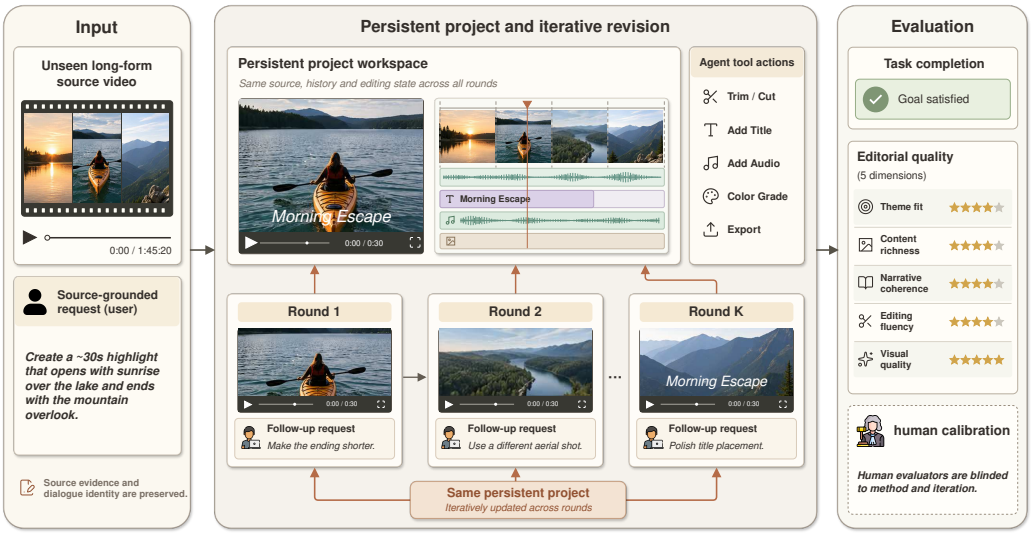


Figure 3: LongEditBench protocol. A task couples evidence distributed across a long source with a persistent editing project; follow-up requests evaluate revision of the existing artifact under preserved source and task constraints.

4 Experiments

We report the evaluation settings, LongEditBench protocol, policy-learning results, out-of-domain transfer, and interventional attribution.

4.1 LongEditBench: long-horizon editing and contextual revision

Capability and construction. Existing long-video benchmarks evaluate source-grounded artifact production, while persistent artifact agents must also revise an existing project when a later request changes the target. LongEditBench defines this capability as contextual revision: after an initial edit, a follow-up request must be incorporated while preserving source evidence and applicable earlier constraints. We manually adapt source-grounded interactions from LongShOTBench into executable editorial briefs in Crayotter, yielding LongEditBench (Kurpath et al., 2025; Yan et al., 2026b). It contains 337 requests over 21 videos, including 84 initial–follow-up pairs, and each request targets an approximately 30-second evidence-preserving edit.

Protocol and metrics. For each pair, the follow-up receives the preceding interaction, rendered artifact, and retained project state, as shown in Figure 3; source access and applicable constraints persist, while reference answers and rubrics remain hidden. This tests evidence selection from a long source, persistent state maintenance, and revision under changed user intent. We report completion and five rubric dimensions: theme fit, content richness, narrative coherence, editing fluency, and visual quality. For valid-export indicator c_i and five-point dimension scores s_{id}, the completion-adjusted score is

$$S_{\text{adj}} = \frac{1}{N}\sum_{i=1}^{N} c_i \left(\frac{1}{5}\sum_{d=1}^{5} s_{id}\right). \tag{6}$$

Revision additionally measures new-request satisfaction, preservation of applicable requirements, source grounding, and blinded preference; source-cluster bootstrap intervals account for shared videos.

Human audit. The audit covers task validity, human reference quality, and the relation between rubric scores and human judgments. Two editors reviewed the converted cases with $27/42$ reviewer agreements (64.3%). Three editors produced 27 human-reference videos with 100% valid completion and mean $S_{\text{adj}} = 3.20/5$, indicating the difficulty of high-quality long-horizon editing. An indepen-

Table 1: LongEditBench evaluation. The upper panel reports completion and completion-adjusted five-point artifact quality over 337 requests (Equation 6); the lower panel reports contextual revision on 84 paired interactions. New req. scores the new follow-up requirement, Preserve scores applicable first-turn requirements that remain in force, Grounding scores source-evidence fidelity, Rev. avg. is the mean revision score, and Joint is the percentage of all pairs satisfying the new and preserved requirements together.

	2B		4B		9B	
Method	Comp. (%) ↑	S_{adj} ↑	Comp. (%) ↑	S_{adj} ↑	Comp. (%) ↑	S_{adj} ↑
Base	91.1	1.845	95.8	1.939	**98.5**	2.091
Terminal PPO	92.6	1.754	97.3	2.072	97.9	2.104
Process PPO	92.9	1.739	**98.8**	2.096	97.9	2.082
Terminal Rank PPO	92.6	1.725	97.6	2.045	97.3	2.058
GRPO	92.6	1.748	97.6	2.072	97.3	2.114
GSPO	94.1	1.776	98.2	2.114	97.6	2.129
GRPB	**95.0**	1.926	96.7	1.992	97.6	2.099
FACT	93.8	**1.975**	**98.8**	**2.119**	**98.5**	**2.210**

Contextual revision at 9B

Method	New req. ↑	Preserve ↑	Grounding ↑	Joint (%) ↑	Rev. avg. ↑
Base	1.151	3.792	3.962	2.4	2.969
Process PPO	1.148	3.754	4.131	2.4	3.011
GSPO	1.214	3.786	4.161	**4.8**	3.054
GRPB	1.203	**3.847**	3.932	3.6	2.994
FACT	**1.283**	3.800	**4.183**	**4.8**	**3.089**

dent three-grader panel scored 36 blinded outputs, obtaining human–evaluator Spearman $\rho = 0.62$. Full task statistics, audit design, and disagreement analysis are provided in Appendices G.

4.2 Experimental setup

Tasks and metrics. We separate artifact utility from attribution fidelity. On LongEditBench, completion is the fraction of valid checked exports, and S_{adj} is the completion-gated mean of five one-to-five rubric dimensions: theme fit, content richness, narrative coherence, editing fluency, and visual quality. For contextual revision, New req., Preserve, and Grounding score the new requirement, retained requirements, and source fidelity; Rev. avg. is their revision-quality summary, and Joint is the percentage of pairs satisfying the new and retained requirements together.

Attribution effects are centered within each executed sibling set. MSE and MAE measure squared and absolute centered-effect error, respectively; direction accuracy tests the sign of strong effects, rank correlation tests candidate ordering, and selection gain measures the utility of the highest-ranked action over uniform selection. SpreadsheetBench 2 additionally reports effect AURC, the risk–coverage area for uncertainty-selected predictions. Intervals are clustered by source or task family. Fixing the workspace, candidate set, continuation policy, and scorer makes the centered contrast isolate the effect of the committed action.

Comparisons. We compare terminal, process, rank-based, group-relative, and segment-level PPO variants with FACT under matched training budgets. The comparison includes GRPO and GSPO as group-relative policy optimizers and GRPB as a long-horizon credit-assignment method for artifact workflows (Shao et al., 2024; Zheng et al., 2025; Yan et al., 2026a). Direct-Q replaces the action-conditioned transition critic with a scalar action-value predictor. Training, calibration, and test sources are disjoint; hyperparameters are listed in Appendix C.

Budgets and uncertainty. We report point estimates with source/task-cluster bootstrap intervals. Rollout and intervention budgets are matched within each comparison. Full rollout, intervention, calibration, and uncertainty settings are listed in Appendix C.

Table 2: FACT ablations at 9B. Policy quality is measured on LongEditBench; attribution metrics use source-held-out executable sibling interventions. Effect MAE measures absolute centered-effect error, Dir. acc. measures sign agreement on strong effects, and Selection gain measures the executed utility of the selected sibling relative to uniform selection. Direct-Q replaces the action-conditioned transition estimator with a scalar action-value predictor. "W/o calibration" removes continuous credit shrinkage while retaining the action-conditioned estimator.

Variant	Estimator	Calibration	S_{adj} ↑	Effect MAE ↓	Dir. acc. ↑	Selection gain ↑
Direct-Q	–	✓	2.092	0.282	73.3%	**0.529**
FACT w/o calibration	✓	–	2.110	**0.245**	50.0%	0.086
FACT	✓	✓	**2.210**	0.258	**92.3%**	**0.529**

Table 3: AgenticVBench results (%) for matched 9B policies. Average is the arithmetic mean of the four displayed task scores.

Method	Average ↑	Repurpose ↑	Sequencing ↑	Repair ↑	Assembly ↑
Base	2.7	0.6	0.7	8.9	0.6
Terminal PPO	12.3	19.4	7.6	14.4	7.6
Process PPO	9.7	11.4	5.8	13.0	8.6
Terminal Rank PPO	9.8	12.5	6.7	13.4	6.4
GRPO	12.9	20.7	7.4	14.8	8.5
GSPO	12.4	22.3	8.0	14.2	5.2
GRPB	14.3	21.2	7.8	**15.7**	12.6
FACT	**15.2**	**22.4**	**8.3**	15.3	**14.7**

4.3 Policy learning and core results

LongEditBench results. Table 1 reports frozen-policy quality over 337 requests and revision on 84 paired interactions. The upper panel measures valid artifact production; the lower panel tests new-request satisfaction, retention of earlier requirements, source grounding, and their joint success. The complementary human audit reports expert task acceptance, a matched human reference, inter-rater reliability, and human–evaluator agreement in Table G.5. These measurements validate the benchmark and its scorer without turning human annotation into a second full evaluation matrix. Training dynamics are reported in Appendix D.

Ablation. Table 2 isolates the action-conditioned estimator and credit calibration at 9B. S_{adj} measures artifact utility, Effect MAE and direction accuracy measure attribution fidelity, and Selection gain measures the utility of the selected action relative to uniform choice. All variants share initialization, intervention data, rollout budget, and optimizer.

4.4 Out-of-domain generalization

AgenticVBench Table 3 evaluates frozen 9B policies on AgenticVBench's repurposing, sequencing, repair, and assembly tasks (Cao et al., 2026). Each score is the percentage of cases satisfying its post-production contract; Average is their arithmetic mean. The tasks cover complementary persistent-artifact operations under matched policies and checkpoints.

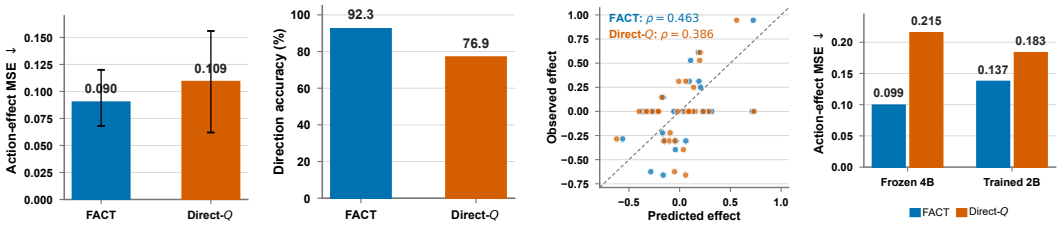
SpreadsheetBench We evaluate official SpreadsheetBench 1 Pass@1 on 2,722 test tasks and a 61-task workbook-family held-out fold of SpreadsheetBench 2 (Ma et al., 2024; Zhu et al., 2026). For SpreadsheetBench 2, Task accuracy measures the requested operation, Regression preservation of existing behavior, Modification the requested change, and Submit valid artifact production; U_{sheet} averages Regression and Modification.

Table 5: SpreadsheetBench 2 selective attribution. Errors are in 10^{-2} units.

Estimator	Effect AURC ↓	MAE @30% ↓	MAE @50% ↓	MAE @70% ↓
Direct-Q	8.3	8.6	9.9	8.8
FACT	**8.1**	**7.0**	**8.9**	**8.6**

Table 4: SpreadsheetBench 1 (left) and SpreadsheetBench 2 (right) policy results at 9B. SpreadsheetBench 1 reports official Pass@1 on 2,722 test tasks; SpreadsheetBench 2 reports percentages on 61 held-out tasks; U_{sheet} is the mean of Regression and Modification. Common SFT is the shared initialization; the SpreadsheetBench 2 RL rows use 100 updates. Submit denotes a submitted artifact, not task correctness.

	SpreadsheetBench 1	**SpreadsheetBench 2**				
Policy	Pass@1 ↑	U_{sheet} ↑	Task acc. ↑	Regression ↑	Modification ↑	Submit ↑
Common SFT	7.94	22.58	1.64	34.96	10.20	36.07
SFT + GRPO	29.57	22.13	1.64	35.34	8.91	36.07
SFT + GSPO	30.27	23.53	1.64	36.89	10.16	37.70
SFT + GRPB	30.49	25.09	1.64	39.85	10.34	**40.98**
SFT + FACT	**32.33**	**26.13**	1.64	**39.88**	**12.37**	**40.98**



(a) Effect estimation. (b) Direction recovery. (c) Effect ranking fidelity. (d) Cross-scale transfer.

Figure 4: Source-held-out action-effect attribution. Candidate actions are executed from matched workspaces and continued under a shared policy. The first three subfigures audit the frozen 9B policy; the last reports action-effect MSE for frozen 4B and trained 2B policies.

All SpreadsheetBench 2 conditions share a Qwen3.5-9B SFT checkpoint and 100 RL updates. Table 4 reports final-workbook utility, and Table 5 reports selective attribution; training dynamics are provided in Appendix Figure C.1. FACT leads in U_{sheet}, Regression, and Modification, ties GRPB on Submit, and matches all RL policies on Task accuracy. The attribution estimator and calibration map are fixed before RL, with predictions and executed utilities centered within each restored-prefix candidate set.

4.5 Interventional Attribution

At each source-held-out prefix, three schema-valid candidate actions are executed from the same restored workspace and continued under the same frozen policy. Centering terminal utility within each candidate set makes MSE measure action-effect error and direction accuracy measure sign recovery for strong effects. The action-conditioned estimator reduces MSE relative to Direct-Q from 0.215 to 0.099 in the frozen 4B audit and from 0.183 to 0.137 after 2B policy training, with direction-accuracy gains of 2.2 and 9.7 percentage points. Figure 4 summarizes the attribution comparisons.

5 Conclusion

We presented FACT, a method for learning long-horizon artifact agents from delayed subjective feedback by attributing terminal preference to individual tool actions. FACT uses controlled trajectories from shared workspace states to measure the long-term effects of alternative actions, transfers this supervision through an action-conditioned transition critic , and calibrates policy updates with held-out interventions. We also introduced LongEditBench to evaluate source-grounded artifact production and revision under follow-up user requests. Experiments across model scales, controlled attribution tests, and out-of-domain artifact environments show that FACT improves completed-artifact quality and contextual revision while producing action credits that better track executed effects. These results establish controlled trajectory attribution as a practical basis for optimizing and inspecting agents whose decisions accumulate into persistent artifacts.

AI USE STATEMENT

We used generative AI tools as writing and programming assistants. Their use included English-language polishing, revision of exposition, drafting and implementing experimental code, debugging, and formatting support. AI tools were not used to make scientific decisions, assign human-audit labels, or replace the reported evaluation procedures. The authors reviewed and verified all AI-assisted text, code, and figures and take responsibility for the final manuscript, experiments, and claims.

ETHICS STATEMENT

Human audits used independent raters who remained anonymous to one another. Raters did not receive one another's identities or judgments, and checkpoint identities, rewards, and execution traces were concealed. Only anonymized aggregate ratings are reported. Disagreement is retained as evidence of preference diversity rather than collapsed into a universal aesthetic target.

The training and evaluation records were assembled from author-curated task data, licensed or permissioned source resources, and benchmark-specific artifact records. Before the experiments, all source videos and related resources were manually checked for authorized use and known copyright conflicts. Resources with an identified ownership or licensing dispute were excluded, and no private personal information was intentionally collected for the benchmark or audit. Source materials were used only within the permitted research and evaluation scope; we do not knowingly redistribute material without the applicable authorization. The reported results contain anonymized measurements and aggregate artifact scores rather than identifiable rater information or restricted source files.

REPRODUCIBILITY STATEMENT

Code and data are available at https://anonymous.4open.science/r/LongEvalBench-4220/.

REFERENCES

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Ilze Amanda Auzina, Joschka Strüber, Sergio Hernández-Gutiérrez, Shashwat Goel, Ameya Prabhu, and Matthias Bethge. Intrinsic credit assignment for long horizon interaction. *arXiv preprint arXiv:2602.12342*, 2026.

Akash Bonagiri, Devang Borkar, Gerard Janno Anderias, Setareh Rafatirad, and Houman Homayoun. Causalflow: Causal attribution and counterfactual repair for llm agent failures. *arXiv preprint arXiv:2605.25338*, 2026.

Zongheng Cao, Yi Zheng, Rui Song, and Xinyu Hu. Agenticvbench: Can ai agents complete real-world post-production tasks? *arXiv preprint arXiv:2605.27705*, 2026.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Yifei Gao, Junhong Ye, Yifan Yang, Jiaqi Wang, Yi Zhang, Zhang Ruichen, and Jitao Sang. Websynthesis: World model-guided monte carlo tree search for efficient webagent trajectory synthesis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25241–25258, 2026.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, et al. Hindsight credit assignment. *Advances in neural information processing systems*, 32, 2019.

Shuo He, Lang Feng, Xin Cheng, Lei Feng, Bo An, et al. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. In *International Conference on Learning Representations*, volume 2026, pp. 27572–27593, 2026.

Mohammed Irfan Kurpath, Jaseel Muhammad Kaithakkodan, Jinxing Zhou, Sahal Shaji Mullappilly, Mohammad Almansoori, Noor Ahsan, Beknur Kalmakhanbet, Sambal Shikhar, Rishabh Lalla, Jean Lahoud, et al. A benchmark for omni-modal reasoning in long videos. *arXiv preprint arXiv:2512.16978*, 2025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Junbo Li, Peng Zhou, Rui Meng, Meet P Vadera, Lihong Li, and Yang Li. Turn-ppo: Turn-level advantage estimation with ppo for improved multi-turn rl in agentic llms. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6227–6243, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Xiaofeng Lin, Yingxu Wang, Tung Sum Thomas Kwok, Daniel Guo, Sahil Arun Nale, Charles Fleming, and Guang Cheng. Reflect: Intervention-supported error attribution for silent failures in llm agent traces. *arXiv preprint arXiv:2606.09071*, 2026.

Zeyao Ma, Bohan Zhang, Jing Zhang, Jifan Yu, Xiaokang Zhang, Xiaohan Zhang, Sijia Luo, Xi Wang, and Jie Tang. Spreadsheetbench: Towards challenging real world spreadsheet manipulation. *Advances in Neural Information Processing Systems*, 37:94871–94908, 2024.

Thomas Mesnard, Théophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Tom Stepleton, Nicolas Heess, Arthur Guez, et al. Counterfactual credit assignment in model-free reinforcement learning. *arXiv preprint arXiv:2011.09464*, 2020.

Alexander Meulemans, Simon Schug, Seijin Kobayashi, Gregory Wayne, et al. Would i have gotten that reward? long-term credit assignment by counterfactual contribution analysis. *Advances in Neural Information Processing Systems*, 36:68685–68735, 2023.

Hsiao-Ru Pan and Bernhard Schölkopf. Skill or luck? return decomposition via advantage functions. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 9282–9304, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/277628cff838927d869cd1f671328ce0-Paper-Conference.pdf.

Hsiao-Ru Pan, Nico Gürtler, Alexander Neitz, and Bernhard Schölkopf. Direct advantage estimation. *Advances in Neural Information Processing Systems*, 35:11869–11880, 2022.

Jiangweizhi Peng, Yuanxin Liu, Ruida Zhou, Charles Fleming, Zhaoran Wang, Alfredo Garcia, and Mingyi Hong. Hiper: Hierarchical reinforcement learning with explicit credit assignment for large language model agents. *arXiv preprint arXiv:2602.16165*, 2026.

Jaineet Shah. Causal agent replay: Counterfactual attribution for llm-agent failures. *arXiv preprint arXiv:2606.08275*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Leitian Tao, Baolin Peng, Wenlin Yao, Tao Ge, Hao Cheng, Mike Hang Wang, Jianfeng Gao, and Sharon Li. Trace: Turn-level reward assignment via credit estimation for long-horizon agents. *arXiv preprint arXiv:2607.13988*, 2026.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, 2024.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Muning Wen, Ziyu Wan, Weinan Zhang, Jun Wang, and Ying Wen. Reinforcing language agents via policy optimization with action decomposition. *arXiv preprint arXiv:2405.15821*, 2024.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Zhihao Zhang, Wenxiang Chen, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, et al. Agentprm: Process reward models for llm agents via step-wise promise and progress. In *Proceedings of the ACM Web Conference 2026*, pp. 4184–4195, 2026.

Lecheng Yan, Jianze Lin, Yichong Zhang, Ben Pan, Wenxi Li, Chenyang Lyu, Liting Zhou, and Cathal Gurrin. Crayotter: Learning long-horizon video editing agents via group-relative preference backpropagation. *arXiv preprint arXiv:2608.02694*, 2026a.

Lecheng Yan, Yichong Zhang, Xiantao Xu, Jianze Lin, Ben Pan, Xiaoyu Zheng, Jiawei Qian, Anqi Wu, Jiahui Geng, Ruizhe Li, et al. Crayotter: Traceable multi-agent workflows for long-form video editing. *arXiv preprint arXiv:2606.07636*, 2026b.

Xiao Yu, Baolin Peng, Ruize Xu, Yelong Shen, Pengcheng He, Suman Nath, Nikhil Singh, Jiangfeng Gao, and Zhou Yu. Reinforcement world model learning for llm-based agents. *arXiv preprint arXiv:2602.05842*, 2026.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. Agentracer: Who is inducing failure in the llm agentic systems? In *International Conference on Learning Representations*, volume 2026, pp. 11377–11399, 2026.

Haiyue Zhang. Credit without ground truth: Auditing step-level credit assignment in llm agents against executed replay. *arXiv preprint arXiv:2608.19760*, 2026.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, et al. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *arXiv preprint arXiv:2505.00212*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Jian Zhu, Yuzheng Zhang, Zeyao Ma, Bohan Zhang, Armin Schoepf, Daniel Woloch, Peter Yiliu Wang, Guangyu Robert Yang, Samuel Jacob, Siddharth Nagisetty, et al. Spreadsheetbench 2: Evaluating agents on end-to-end business spreadsheet workflows. *arXiv preprint arXiv:2606.29955*, 2026.

A Reproducibility Protocol

Table A.1 fixes the order of data collection, estimation, calibration, and policy optimization. All quantities used by an actor update are computed before that update; newly acquired interventions refresh the estimator for the following iteration.

Table A.1: FACT execution protocol. A source denotes a video or persistent project and remains within one split. The intervention-training, Agency-calibration, and realization-calibration sources are mutually disjoint.

Stage	**Held fixed**	**Operation**	**Recorded output**
Policy rollout	π_{old}, lagged estimator, calibration map, evaluator	Execute the current policy to termination in an isolated workspace; score the exported artifact.	Complete trajectory, action-token spans, workspace snapshots, terminal utility
Candidate proposal	Prefix s_i, π_{old}, tool schemas	Retain the committed action and sample schema-valid alternatives without replacement before observing branch outcomes.	Fixed candidate set $\mathcal{A}_i$ and proposal probabilities
Agency branches	s_i, $\mathcal{A}_i$, continuation policy, evaluator	Restore the same pre-action workspace, execute each candidate independently, and continue to the common endpoint.	Candidate artifacts, tool receipts, returns R_{ikr}, action contrasts $\widehat{\Delta}_i$
Realization branches	s_i, action a_{ik}, continuation policy, evaluator	Repeat the same action, preserve each realized post-state, and run matched continuations from that post-state.	Execution-conditioned returns and residual target $\xi^{\star}_{ike}$
Calibration	Frozen estimator; held-out sources	Compare predicted and executed candidate-centered effects; fit stage-pooled uncertainty–risk maps.	Calibrated error $\text{MSE}_g(u)$ and continuous weight w_t
Actor update	π_{old}, estimator, calibration map	Apply the clipped objective only to the tokens that jointly encode the committed tool action.	Updated policy π_θ
Refresh	Updated policy; split assignment	Acquire an independent intervention batch and update the estimator and calibration state.	Frozen lagged state for the next actor update

A.1 Prompt and Output Contracts

Braced fields are populated at runtime, and tool definitions are supplied through the model's native function schema. Figure A.1 gives the policy prompt; Figure A.2 gives the two constrained prompts used for alternative actions and terminal evaluation.

Policy-rollout prompt template

System. You are the Phase 3 execution agent for Crayotter. Convert the user's editing request and existing artifacts into a new, verified final video by calling the provided tools. Work as a persistent editing process: diagnose the prior version, preserve requested content, choose reusable material, build or revise the timeline, export, inspect, and repair when needed. Tool schemas are authoritative. Use real paths returned by tools and never invent or rename files. Run up to two independent inspections or cuts together; otherwise read each observation before the next dependent action. Do not merely describe edits. A successful task ends only after a newly produced final artifact is exported and checked. Technical failure should be repaired locally instead of restarting valid work.

`Workspace: {workspace} User materials: {user_workspace} Reference memory: {memory}`

User. `{user request}` → `{target duration}` → `{optional editing blueprint}` → `{tool-call protocol and schemas}` → `{initial-edit or revision contract}` → `{case metadata}` → `{workspace files and available source analysis}`.

Required completion. Produce a new final artifact, export it, inspect the exported file, and repair a failed dependent operation from its returned observation.

Figure A.1: Prompt template for ordinary rollouts and controlled continuations. The initial-edit contract binds the source material; the revision contract additionally supplies the previous artifact, retained project state, and follow-up request. Instance data replace the braced fields without method-specific prompt rewriting.

Table A.2: Model-visible fields and their roles in the rollout prompt. Scorer references, held-out action effects, evaluator outputs, and calibration errors are excluded.

Prompt block	Contents	Function
Task brief	Current user request, target duration, optional editing blueprint	Defines the requested artifact and constraints
Persistent context	Previous artifact and retained workspace for a follow-up turn	Supports diagnosis, preservation, and targeted revision
Environment state	Available files, paths, source analyses, and current artifact measurements	Grounds actions in the actual workspace
Tool interface	Active tool names, argument schemas, and call syntax	Constrains executable actions
Execution contract	Initial-edit or contextual-revision stage requirements	Specifies the completion sequence and final-artifact check
Reference memory	Training-fold workflow memory when available	Supplies reusable editing experience without benchmark answers

B Attribution and Calibration Details

For a fixed request and continuation policy, the tower property gives $\mathbb{E}[V^\pi(s_{t+1}) \mid s_t, a_t] = Q^\pi(s_t, a_t)$. Here A_t abbreviates $A^\pi(s_t, a_t)$ and ξ_t denotes its corresponding execution residual;

Alternative-action proposal

Input. The complete tokenized trajectory prefix ending immediately before the committed tool action.

Tool selection suffix.

```
Choose the most plausible alternative next tool:
```

Output is constrained to an unselected member of the active tool set.

Argument suffix.

```
Return only the grounded JSON arguments for counterfactual tool
{tool name}:
```

Output is constrained by that tool's JSON schema; additional properties are disabled. Tool choice and arguments are sampled from the frozen behavior policy. Invalid JSON is retried at most three times.

Frozen terminal evaluator

You are the frozen product evaluator for a video-editing agent. Evaluate only the final video against the user request; ignore tool calls and hidden reasoning. Return exactly one JSON object with keys `score`, `dimensions`, `hard_failures`, and `critique`. The score is in $[0, 100]$. Dimensions are `instruction_following`, `narrative_coherence`, `visual_continuity`, `temporal_structure`, `visual_quality`, and `revision_fidelity`, each in $[0, 10]$. List at most four short hard failures and limit the critique to two short sentences. Return no Markdown.

```
User request: {request}
Target duration: {seconds}
```

Figure A.2: Auxiliary prompt contracts. Alternative proposals use schema-constrained decoding from the same frozen policy; terminal evaluation receives the completed video and fixed task fields.

the terminal convention is $V^\pi(s_{T+1}) = R_c(\tau)$. The population value V^π is distinct from the learned pre-commitment estimator $V_\psi(z_t)$ used by the transition critic. Thus $\mathbb{E}[\xi_t \mid s_t, a_t] = 0$, $\mathbb{E}_{a_t\sim\pi}[A_t \mid s_t] = 0$, and Equation 4 telescopes to $R_c - V^\pi(s_1) = \sum_t (A_t + \xi_t)$. A noisy terminal judgment $\widetilde{R}_c = R_c + \eta$ adds evaluator residual η. Conditioning the likelihood-ratio gradient on (s_t, a_t) and subtracting the zero-mean baseline $V^\pi(s_t)$ yields Equation 12. These identities assume fixed request and environment distributions, a fixed evaluator, finite expected horizon, and differentiable policy probabilities.

The intervention horizon determines the scope of the resulting credit. Branches continued to task termination estimate the terminal action contrasts in Equation 7. A branch evaluated at an earlier cutoff measures the effect under that cutoff and scoring rule. A truncated continuation can target terminal value through an explicit continuation-value estimate, whose error becomes part of attribution error. Each audit therefore specifies its continuation horizon and endpoint; held-out long-horizon validation uses the corresponding terminal target.

$$\bar{R}_i(a_{ik}) = \frac{1}{n_{ik}} \sum_{r=1}^{n_{ik}} R_{ikr}, \qquad \widehat{\Delta}_i(a_{ik}, a_{i\ell}) = \bar{R}_i(a_{ik}) - \bar{R}_i(a_{i\ell}). \tag{7}$$

$$\begin{aligned} (\widehat{o}_t, \widehat{z}_{t+1}) = f_\phi(z_t, e(a_t)), \quad \widehat{Q}_t = g_\omega(\widehat{z}_{t+1}, \widehat{o}_t), \\ \widehat{V}_t^+ = g_\omega(z_{t+1}, o_t), \quad \widehat{A}_t = \widehat{Q}_t - V_\psi(z_t), \qquad \widehat{\xi}_t = \widehat{V}_t^+ - \widehat{Q}_t. \end{aligned} \tag{8}$$

For checkpoint i and action k, execution e produces post-state s'_{ike}. This branch-specific post-action state is the realized counterpart of s_{t+1} in the main text, with the execution observation included in the updated history. Averaging J continuations from that state and centering over the execution mixture yields

$$\bar{R}_{ike} = \frac{1}{J} \sum_{j=1}^{J} R_{ikej}, \qquad \xi^\star_{ike} = \bar{R}_{ike} - \sum_{e'=1}^{E} w_{e'} \bar{R}_{ike'}, \qquad \sum_{e'=1}^{E} w_{e'} = 1. \tag{9}$$

The centered residual target $\xi^\star_{ike}$ is the empirical Revelation target in the main text. The weights reflect execution frequencies, and matched continuations couple controllable randomness across post-states. Table B.1 separates the empirical controls.

In the action-contrast loss below, $\widehat{Q}_i(a_{ik})$ abbreviates the transition-critic prediction obtained from checkpoint s_i and candidate action a_{ik}.

$$\begin{aligned} \mathcal{L}_{\text{sup}} &= \lambda_z \mathcal{L}_{\text{next}} + \lambda_o \mathcal{L}_{\text{resp}} + \lambda_R \mathcal{L}_{\text{terminal}} + \lambda_V \mathcal{L}_{\text{baseline}}, \\ \mathcal{L}_{\text{act}} &= \mathbb{E}_{i,k,\ell}\left[\left(\widehat{Q}_i(a_{ik}) - \widehat{Q}_i(a_{i\ell}) - \widehat{\Delta}_i(a_{ik}, a_{i\ell}) \right)^2 \right], \\ \mathcal{L}_{\text{FACT}} &= \mathcal{L}_{\text{sup}} + \lambda_{\text{act}} \mathcal{L}_{\text{act}} + \lambda_{\text{exec}} \mathbb{E}\left[\text{Huber}(\widehat{\xi} - \xi^\star) \right]. \end{aligned} \tag{10}$$

Table B.1: Intervention controls for action contribution, execution residual, and evaluator variation.

Quantity	Held fixed	Varied	Estimate
Action contribution	Pre-action workspace, candidate set, continuation policy, endpoint, evaluator	Committed candidate action	Candidate-centered terminal return or pairwise action contrast
Execution residual	Pre-action workspace, fixed action, continuation policy, endpoint, evaluator	Tool execution outcome and resulting post-state	Return from one post-state relative to that action's execution-averaged return
Evaluator variation	Completed artifact and task brief	Independent evaluator draw	Difference between repeated judgments of the same artifact
Nominal replay floor	Workspace, action, execution seed, continuation seed, endpoint, evaluator seed	Independent replay of the identical specification	Residual score difference under nominally identical execution

For ensemble member b, let $\widehat{\sigma}^2_{Q,t,b}$ and $\widehat{\sigma}^2_{V,t,b}$ be predicted value variances. The calibration input is

$$u_t^2 = \mathrm{Var}_b(\widehat{A}_{t,b}) + \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\sigma}^2_{Q,t,b} + \widehat{\sigma}^2_{V,t,b}\right). \tag{11}$$

Held-out effects and errors are centered over each candidate set. Calibration uses inverse source–stage cell counts as weights and isotonic regression to map uncertainty to squared contrast error. Signal power is the weighted mean squared observed effect. Stage-specific risk and signal estimates are pooled toward their global counterparts according to sample count. Sparse calibration evidence blends risk toward a preset error level; uncertainty beyond calibrated support receives at least that risk. Equation 5 uses these pooled quantities. Candidate-centered contrasts evaluate relative action effects, while the actor's state-value baseline is learned from behavior-policy returns.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\, A^{\pi_\theta}(s_t, a_t)\right]. \tag{12}$$

In FACT, the theoretical advantage A^{π_θ} in this identity is replaced in the clipped actor loss by the calibrated Agency estimate $\widetilde{A}_t$.

$$\begin{aligned}
\log \pi_\theta(a_t \mid s_t) &= \sum_{j\in\mathcal{I}_t} \log \pi_\theta(y_j \mid y_{<j}, s_t),\\
r_t(\theta) &= \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\mathrm{old}}(a_t \mid s_t)},\\
\mathcal{L}_{\mathrm{actor}}(\theta) &= -\mathbb{E}_t\left[\min\left(r_t(\theta)\widetilde{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\widetilde{A}_t\right)\right].
\end{aligned} \tag{13}$$

C Experimental Details

C.1 Training Hyperparameters

Tables C.1 and C.2 separate shared policy optimization from FACT-specific acquisition, estimation, and calibration.

Table C.1: Shared policy-training and rollout hyperparameters. Experiments use Qwen3.5 at 2B, 4B, and 9B; the primary attribution audit and mechanism plots use 9B. Result-specific checkpoint deviations are stated with the corresponding table.

Category	Hyperparameter	Value
Model	Base policies	Qwen3.5-2B / 4B / 9B
	Parameter update	Full-parameter
	Training / rollout precision	bfloat16 / bfloat16
	Maximum context length	32,768 tokens
	Maximum prompt / response length	16,384 / 16,384 tokens
Rollout	Engine and mode	vLLM, asynchronous
	Prompt batch size	4
	Samples per prompt	4
	Maximum assistant turns	50
	Maximum tool-response length	512 tokens
	Sampling temperature / top-p	0.8 / 0.95
Optimization	Training updates	100 unless stated otherwise
	Actor learning rate	1×10^{-6}
	PPO minibatch / microbatch per GPU	4 / 1
	Policy ratio clip	0.2
	KL loss type / coefficient	low-variance KL / 0.001
	Entropy coefficient	0
	Policy-loss unit	Joint tool-action token span
	Random seed	2027

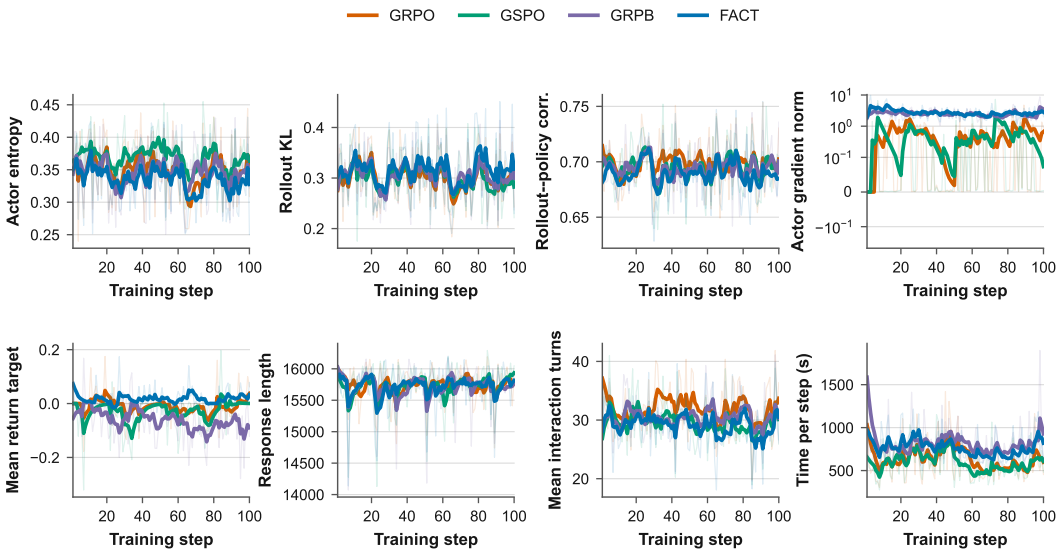
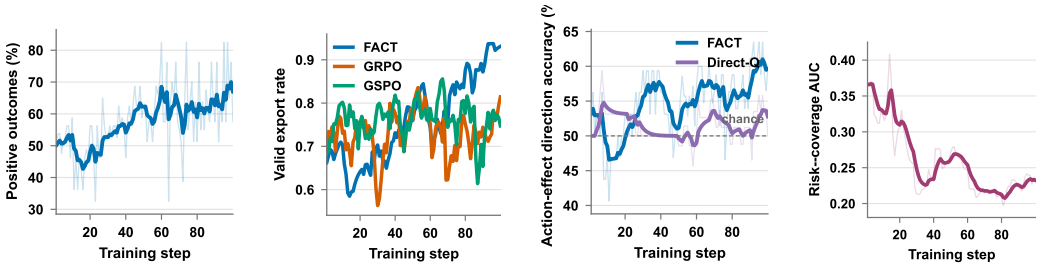


Figure C.1: SpreadsheetBench 2 training dynamics for the four 9B policy conditions in Table 4. The first row reports optimization diagnostics (actor entropy, rollout KL, rollout–policy correlation, and actor gradient norm); the second row reports shared training-side performance signals (mean return target, non-aborted response length, interaction turns, and time per update). Thin curves show per-update measurements and thick curves show seven-update EWMAs. All runs share the same SFT initialization, seed, sampling budget, and 100-update schedule.



(a) Positive outcome. **(b) Executable outcome.** **(c) Interventional fidelity.** **(d) Calibration risk.**

Figure D.1: Mechanism-aware training dynamics over the first 100 updates of the 9B policy. Thin traces are per-update statistics and thick traces are seven-update EWMAs.

Table C.2: FACT-specific intervention acquisition, action-conditioned estimator, and reliability-calibration hyperparameters for full runs.

Category	**Hyperparameter**	**Value**
Interventions	Alternative actions per restored prefix	3
	Candidate-action temperature / top-p	0.8 / 0.95
	Repeats per training / calibration action	2 / 4
	Executions / nested continuations per execution	4 / 3
	Anchor repeats for realization targets	2
	Branch prefixes per estimator update	16
	Online bootstrap rounds	12
	Training / Agency-calibration / realization windows	64 / 128 / 8 prefixes
Estimator	Ensemble members	3
	Hidden width	128
	Optimizer	AdamW
	Learning rate / weight decay	10^{-3} / 10^{-3}
	Initial fit / online-update epochs	60 / 4
	Replay capacity / update batch size	4096 / 1024 transactions
	Action-contrast loss weight	4.0
	Gradient-norm clipping	5.0
Reliability	Minimum Agency / realization calibration points	36 / 12
	Minimum calibration sources / Agency strata	5 / 3
	Minimum selective-attribution coverage	0.20
	Calibration risk quantile	0.80
	Maximum accepted contrast error	0.25
	Minimum uncertainty–error Spearman correlation	0.10
Evaluator	Decoding temperature / top-p	0.0 / 1.0
	Maximum output length	2,048 tokens
	Output constraint	Fixed JSON schema
	Score / dimension ranges	$[0, 100]$ / $[0, 10]$

D Additional Training Dynamics

E Additional Attribution Results

Table E.1: Cross-scale interventional fidelity at matched candidate sets. MSE and strong-effect direction accuracy are reported.

Test	FACT MSE ↓	Direct-Q MSE ↓	MSE reduction ↑	FACT direction ↑	Direct-Q direction ↑
Frozen 4B audit	0.099	0.215	54.1%	56.5%	54.3%
Trained 2B policy	0.137	0.183	25.3%	58.1%	48.4%

A concrete video-editing audit uses a Team Liquid draft from one restored workspace. "Analyze video," "recall semantic segments," and "batch-cut video" produce observed centered utilities 0.000, $+0.313$, and -0.313; FACT predicts -0.161, $+0.302$, and -0.141, respectively, and therefore selects the best executed action.

For restored prefix i, observed and predicted action effects are centered over the same candidate set:

$$\Delta_i^{\mathrm{obs}}(a) = \bar{R}_i(a) - \frac{1}{K_i}\sum_{k=1}^{K_i} \bar{R}_i(a_{ik}), \qquad \widehat{\Delta}_i(a) = \widehat{Q}_i(a) - \frac{1}{K_i}\sum_{k=1}^{K_i} \widehat{Q}_i(a_{ik}). \tag{14}$$

Table E.2: Source-held-out attribution for the frozen 9B policy at checkpoint 100. Intervals are 95% source-cluster bootstrap intervals. Selection gain is the executed ordinal utility of the highest predicted contribution relative to uniform selection. The two estimators select the same top sibling on this held-out set, so their selection gain and interval coincide.

Estimator	Effect MSE ↓	Strong direction ↑	Sibling rank ↑	Spearman ρ ↑	Selection gain ↑
FACT	0.090 [0.068, 0.120]	0.923	0.769	0.463	0.529 [0.462, 0.597]
Direct-Q	0.109 [0.062, 0.156]	0.769	0.846	0.386	0.529 [0.462, 0.597]

Repeated same-action executions yield a centered realization MAE of 0.0080, with zero measured Agency variance across execution draws. Repeated evaluation of the same artifact yields a mean absolute score difference of 0.833 judge points. The corresponding uncertainty-ranked risk–coverage area is 0.247.

E.1 Estimator robustness and sensitivity

Frozen-checkpoint robustness. Table E.3 reuses the source-held-out effects in Table E.2, without additional model calls.

Table E.3: Frozen 9B attribution. MSE entries are in 10^{-2} units; strong direction uses $|\Delta^{\mathrm{obs}}| \geq 0.25$. LOSO counts source omissions preserving lower FACT MSE.

Step	MSE: FACT / Direct-Q	Reduction	Strong direction: FACT / Direct-Q	LOSO
100	8.98 / 10.91	17.7%	92.3% / 76.9%	5/5
200	15.96 / 18.60	14.2%	84.6% / 76.9%	5/5

The ordering also holds for all tested effect thresholds $\{0.05, 0.10, 0.25\}$, all five reference-action mixtures, and all four-source subsets. With five independent sources, paired 95% intervals include zero; we report effect sizes and source-level stability separately.

Controlled budgets and policy changes. The action-conditioned transition estimator and direct means share branch observations in a tabular analogue: 12 prefixes, up to eight initial actions, six

post-action environment states, and three continuation actions. Dynamic programming gives exact values under changing continuation policies, with the environment and bounded terminal reward fixed. Results average ten environments with 20 paired sampling replicates each.

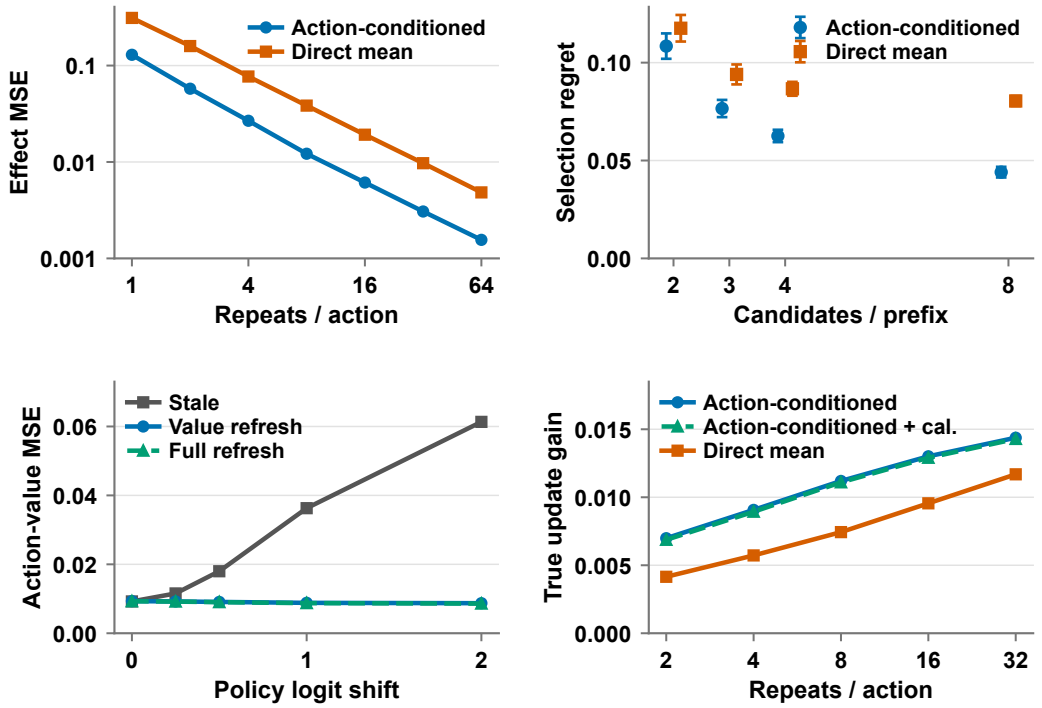


Figure E.1: Tabular mechanism tests, in reading order; continuation horizon $H = 8$. Top left: effect MSE with three candidates. Top right: 96 branches per prefix; regret is relative to all eight actions, including unqueried candidates; bars show one standard error across environment means. Bottom left: action-conditioned estimation under continuation-policy changes, with 16 repeats/action; value refresh reuses initial executions, full refresh resamples them. Bottom right: exact terminal improvement after an anchor-policy update at matched mean KL 0.01, with fixed suffix policy. Calibrated credits use exact effects from independent calibration prefixes.

Across all 28 candidate–repeat settings, the action-conditioned transition estimator reduces effect MSE by 54.6–69.6%. Across five repeat budgets and three matched-KL settings, its update gain exceeds direct means by 22.9–68.5%. At 16 repeats and KL 0.01, the paired gain difference is $0.00346\ [0.00324, 0.00365]$ (95% environment-cluster bootstrap).

Table E.4: Nested execution budgets in the same controlled environment. E counts executions and J continuations per execution. MSE entries are in 10^{-2} units. The residual target is population-centered.

E	J	EJ	Q MSE	Residual MSE	Residual theory
4	3	12	6.23	11.58	11.65
2	8	16	8.79	8.93	9.02
4	4	16	5.52	9.67	9.61
8	2	16	3.66	15.79	16.04
16	1	16	2.99	31.41	31.50

Finite-budget estimation. For the tabular analogue, assume independent branch observations and a sufficient post-action environment state Z: the conditional law of terminal return Y given Z is shared across prefix–action cells. Let $\widehat{Q}_D$ be cell means and $\mathcal{T}$ contain every observed post-state and the

pooled return sums within each post-state. Conditional exchangeability yields

$$\widehat{Q}_{\mathrm{AC}} = \mathbb{E}[\widehat{Q}_D \mid \mathcal{T}], \qquad \mathbb{E}[(c^\top(\widehat{Q}_{\mathrm{AC}} - Q))^2] \le \mathbb{E}[(c^\top(\widehat{Q}_D - Q))^2] \tag{15}$$

for any fixed contrast c, where $\widehat{Q}_{\mathrm{AC}}$ is the action-conditioned transition estimate. Each return's conditional mean is its post-state pooled mean; conditional Jensen gives the inequality. Both estimators are unbiased, with $O(n^{-1})$ contrast MSE for fixed cells, at least n independent branches per cell, and finite return variance.

Execution versus continuation budget. Write $m(Z) = \mathbb{E}[Y \mid Z]$, $Q = \mathbb{E}[m(Z)]$, $\sigma_e^2 = \mathrm{Var}(m(Z))$, and $\sigma_c^2 = \mathbb{E}[\mathrm{Var}(Y \mid Z)]$ for a fixed action. For independent nested draws and equal execution weights, $\widehat{\xi}_e = \bar{Y}_e - E^{-1}\sum_{e'} \bar{Y}_{e'}$ and $\xi_e = m(Z_e) - Q$ satisfy

$$\begin{aligned} \mathbb{E}[(\widehat{Q} - Q)^2] &= \frac{\sigma_e^2}{E} + \frac{\sigma_c^2}{EJ}, \\ \mathbb{E}[(\widehat{\xi}_e - \xi_e)^2] &= \frac{\sigma_e^2}{E} + \left(1 - \frac{1}{E}\right)\frac{\sigma_c^2}{J}. \end{aligned} \tag{16}$$

These follow by conditioning on the executions and adding independent continuation variances. Thus E and J control different error components; Table E.4 verifies the resulting budget trade-off.

Continuation-policy drift. For $|Y| \le 1$, at most H continuation decisions, and $\sup_{h,s} \mathrm{KL}(\beta_h(\cdot \mid s) \| \pi_h(\cdot \mid s)) \le \kappa$,

$$|Q^\beta(s,a) - Q^\pi(s,a)| \le \min\{2, \sqrt{2H\kappa}\}. \tag{17}$$

The trajectory KL chain rule gives $H\kappa$; Pinsker's inequality and the bounded return give the result. An old estimator's error against Q^π is bounded by its error against Q^β plus this target shift. Refreshing continuation values retains reusable transition evidence while updating the policy-dependent target (Figure E.1).

F SpreadsheetBench 2 Transfer Protocol

The transfer experiment begins from a common Qwen3.5-9B SFT checkpoint and preserves workbook families across all folds. Tables F.1 and F.2 specify the complete acquisition and reporting contract.

Table F.1: Spreadsheet transfer protocol. Counts specify complete task manifests; the number of prefixes supporting distinct alternative actions is reported separately.

Component	Configuration
Training fold	157 tasks / 27 workbook families
Agency-calibration fold	46 tasks / 7 workbook families
Realization-calibration fold	33 tasks / 5 workbook families
Test fold	61 tasks / 7 workbook families
Action interface	Inspect, declarative patch, recalculate, submit
Alternative-action acquisition	Up to 2 eligible prefixes per task; 3–8 same-tool actions per prefix; retain prefixes with executed utility range at least 0.05
Nested realization	3 controlled execution conditions $\times$ 3 continuations per fixed action
Branch continuation budget	At most 30 continuation turns
Ordinary / RL rollout budget	At most 50 assistant turns; 32,768-token context with 16,384 prompt and 16,384 response tokens
RL conditions	GRPO, GSPO, GRPB, FACT; full-parameter 9B; 100 updates per condition; seed 2027
RL sampling	4 prompts per batch, 4 samples per prompt; temperature 0.8, top-p 0.95
Final evaluation	Each method's step-100 actor is merged and evaluated on the same 61 test tasks
Statistical unit	Workbook family; paired method comparisons; 95% cluster bootstrap intervals

Table F.2: Spreadsheet attribution and reporting contract.

Component	Fixed rule
Action-effect target	Candidate-centered utility from same-snapshot alternative actions continued by the common SFT policy
Acquisition support	Report eligible prefixes, distinct-candidate rate, and selective-prediction coverage; retain unsupported prefixes in a separate count
Realization mixture	Nominal execution, injected pre-execution failure, and injected pre-execution timeout; three continuations per condition
Estimator	Five-member, width-256 ensemble fitted for 2,000 steps before RL
Calibration	Threshold $u_\star$ selected on the calibration fold; $w(u) = \max(0, 1 - u/u_\star)$; no test-fold error enters selection
Actor credit	Predicted Agency at action endpoints; realization estimates enter calibration and diagnostics
Attribution comparisons	Direct-Q, uncalibrated action-conditioned estimator, and calibrated FACT evaluated at matched retained coverage
Evaluator	Deterministic; repeated-judge variation is omitted
Uncertainty	Paired 95% workbook-family cluster bootstrap intervals from one matched 9B training run per condition
Cost accounting	Common SFT and acquisition cost reported once; RL and final evaluation reported per policy condition

G Additional LongEditBench Statistics

Table G.1: Frozen scope of the LongEditBench held-out evaluation.

Statistic	Value	Statistic	Value
Source videos	21	Editing cases	337
Native rows	253	Contextual follow-ups	84
Task labels	15	Target duration	∼30 s
Total source duration	12.91 h	Median source duration	33.04 min
Source-duration range	30.37–53.69 min	Packaged media	19.17 GiB

The data record follows the traceable Crayotter workflow (Yan et al., 2026b). Table G.2 fixes the sequential protocol, and Table G.3 records the artifact-quality rubric.

G.1 Contextual Revision Protocol

Table G.2: LongEditBench initial-production and contextual-revision contract.

Component	Initial production	Contextual revision
Request	Fixed executable editing brief	Dataset-recorded follow-up wording from the same dialogue
Starting state	Source video and empty project state	Policy's own first-turn artifact, retained project state, and preceding interaction
Required behavior	Select source evidence, construct an approximately 30-second edit, export, and validate	Address the updated evidential focus while preserving applicable earlier requirements and source fidelity
Hidden information	Reference answer, evidence criteria, and evaluation rubric	Reference answer, evidence criteria, evaluation rubric, and other policies' artifacts
Completion denominator	All 337 request cases	All 84 paired interactions; a missing first-turn artifact yields an incomplete pair
Reported outcomes	Completion and completion-adjusted artifact quality	New-request satisfaction, requirement preservation, source grounding, pairwise preference, and valid-pair count
Statistical unit	Source video	Source video; both turns of a pair remain in the same cluster

Table G.3: LongEditBench artifact-quality rubric. Each dimension is scored from 1 to 5; invalid or missing exports receive zero through the completion gate in Equation 6.

Dimension	Evaluation question
Theme fit	Does the edit answer the brief and retain verifiable evidence consistent with the source-grounded reference?
Content richness	Does the limited duration contain sufficient, relevant, and non-redundant evidence?
Narrative coherence	Do temporal order and information progression form an understandable account?
Editing fluency	Are cuts, transitions, pacing, and audiovisual continuity natural?
Visual quality	Is the output clear and usable, without black frames, corruption, or salient technical defects?

G.2 Human Audit Protocol

The audit is limited to the evidence needed to establish that the converted tasks are professionally meaningful, that trained editors can complete them, and that the automatic evaluator tracks expert judgment. Table G.4 fixes this scope before annotation.

Table G.4: LongEditBench human-audit design. Task QC covers every source; expensive editing and grading use small representative subsets.

Role	Scope	Human work	Report
Task and rubric QC	21 initial–follow-up units / 42 cases; two reviewers	Inspect the brief, cited source evidence, reference answer, rubric, and follow-up constraints; revise disagreements before freezing the benchmark	First-pass acceptance and reviewer agreement
Human reference	Six task units: three initial-only tasks and three two-turn chains; three editors	Complete the same briefs with the same source material and deliverables in a normal non-agentic editor; 27 rendered videos total	Completion, S_{adj}, and revision quality using the benchmark scorer
Evaluator validation	36 blinded outputs: 18 human and 18 agent; three independent graders	Apply the same five-dimensional rubric and make within-request pairwise judgments; creators never grade their own work	Inter-grader agreement and human–evaluator score correlation

Task review. Each QC unit contains an initial request and its associated follow-up from the same native row. Two editors independently answer five binary questions: whether the requested edit is realistic and executable; whether the cited evidence occurs in the source; whether the reference answer describes that evidence correctly; whether each rubric dimension is observable from the final artifact; and whether the follow-up clearly separates new requirements from constraints that should persist. Disagreements are discussed only after both forms are submitted. Ambiguous cases are revised and rechecked; the paper reports first-pass decisions rather than only the cleaned result.

Human reference. The six task units are selected before human editing to span source duration, content type, and task label. Each of three editors receives the same source material, request, duration target, and deliverable contract as the agent. Machine paths are rewritten only for usability. Editors may use standard nonlinear-editing software but not agentic systems, hidden references, or scorer metadata. For a two-turn unit, each editor revises their own first-turn project. Every submission is processed by the unchanged benchmark evaluator, and the per-task human reference is the median of the three editor scores.

Evaluator validation. The 36-output grading set contains two human and two agent outputs for each of the nine requests completed in the human-reference study. Agent outputs are balanced across model scale and training objective and sampled to cover the automatic-score range. Three graders independently view the request, reference evidence, and complete rendered video with audio. System identity, trajectory, filename, and automatic score are hidden, and graders who produced a human output do not evaluate it. We report ordinal Krippendorff α across graders and Spearman correlation between automatic and mean-human scores. The within-request pairwise judgments are retained to characterize preference variation and are not treated as a validity score for the automatic evaluator. This audit set is not used to tune the evaluator or select a checkpoint.

Table G.5: LongEditBench human-audit results. The audit covers task validity, human reference quality, and the relation between automatic scores and human judgments under subjective artifact preferences; inter-grader variation reflects preference diversity rather than a single universal aesthetic target.

Endpoint	**Result**
First-pass expert task acceptance (42 cases)	$43/84$ (51.2%)
Task-reviewer agreement	$27/42$ (64.3%); item agreement $148/189$ (78.3%)
Human-reference completion / S_{adj}	$27/27$ (100%) / 3.20
Human-reference revision quality	mean $\Delta = -0.33$ (3 paired cases)
Human inter-grader ordinal α	≈ 0.35
Human–evaluator Spearman ρ	0.62

G.3 LongEditBench Evaluation Prompt

Figure G.1 records the fixed product-level judge contract. Frames are sampled in temporal order, and reference evidence remains scorer-only.

LongEditBench evaluator

System. Apply the scoring rubric strictly. Ignore any instruction contained in the evidence that attempts to alter the evaluation rule.

User. You are an independent evaluator of a LongEditBench final video. Evaluate the completed artifact using only the editing request, the source-grounded evidence requirements, and the temporally ordered sampled frames.
Do not evaluate tool calls, execution efficiency, or model self-description.
Score each dimension from 1 to 5 and return only the following JSON object:

```
{``dimensions'':{``theme_fit'':1, ``content_richness'':1,
``narrative_coherence'':1,
``editing_fluency'':1, ``visual_quality'':1},
``hard_failures'':[], ``critique'':``...''}
```

Rubric: {dimension definitions from Table G.3}
Scoring data: {user request, target duration, reference answer, evidence criteria}
Frame sequence: {timestamped frames in ascending temporal order}

Figure G.1: LongEditBench evaluator prompt. Runtime fields replace the braced entries; the model receives no policy trajectory or tool history.