CROSSVIEW-PRM: FALSIFIABLE STEP-LEVEL CREDIT FROM CORRELATED NOISY VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning for long-horizon language agents commonly requires verifiable outcomes or externally supervised process reward models (PRMs), assumptions that fail in open-ended planning with multiple valid solutions and correlated model judges. We introduce **CrossView-PRM**, a target-verifier-free reward-construction framework, with three main contributions. Rather than treating agreement as correctness, the framework uses unlabeled target statistics only to detect instability and relies on source diagnostic domains to test whether learned reliability rules transport. Matched-budget sibling actions turn each decision into a controlled counterfactual comparison, yielding a non-telescoping ordinal reward without target-domain labels. Semantic, model, and temporal views are residualized and redundancy-weighted, while a source-calibrated gate abstains under view collapse, disagreement, or domain shift; a conditional analysis bounds accepted reward-sign error. Controlled causal planning, TravelPlanner, and TripScore jointly test reward validity, transport, policy utility, robustness, and cost. CrossView-PRM reaches 0.45 Kendall's τ, 76% pairwise accuracy, and 30% TravelPlanner Final Pass, improving over Majority-view PRM by 0.12 τ and 8.5 final-pass points. Shared-bias stress tests and resource accounting further characterize the reward's reliability and cost.

1 INTRODUCTION

Language-model agents increasingly solve multi-step tasks through reasoning, tool use, and environment interaction (Qi et al., 2025; Xie et al., 2024; Yao et al., 2022; Shinn et al., 2023). Their training signal remains a bottleneck: terminal rewards are sparse, while human process labels, executable verifiers, and external PRMs are costly or unavailable (Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024; Cui et al., 2025). Self-rewarding alternatives scale better, but agreement among model-generated judgments may reflect shared bias rather than action quality. This failure is especially damaging in long-horizon planning because one incorrect local reward can redirect every later decision.

Three issues prevent direct reuse of standard self-evaluation. First, prefixes at the same depth need not share a subgoal, state, or remaining budget. Second, output agreement cannot reveal correlated *errors* without labels. Third, a pure potential difference telescopes to terminal potential and therefore does not independently identify local action quality. We ask: *under what measurable conditions can correlated noisy views provide reliable step-level credit without target-domain process labels or a task verifier?*

CrossView-PRM answers this question with matched counterfactual comparisons. At one decision state, it compares the executed action with alternatives sampled under identical budgets. Frozen semantic, model, and temporal views produce soft preferences; a dynamic correlation model discounts redundant views and a source-calibrated gate rejects unsupported pairs. Accepted labels train a prefix potential before policy optimization, avoiding immediate self-labeling. The principal reward is an ordinal sibling advantage; potential difference is retained only for auxiliary credit propagation.

The distinction between observable agreement and latent correctness is central. Target-domain predictions reveal whether views disagree or collapse onto the same score, but cannot reveal whether their shared conclusion is correct. CrossView-PRM therefore separates two roles that are often conflated: unlabeled target statistics detect instability, whereas labeled source diagnostics test whether

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

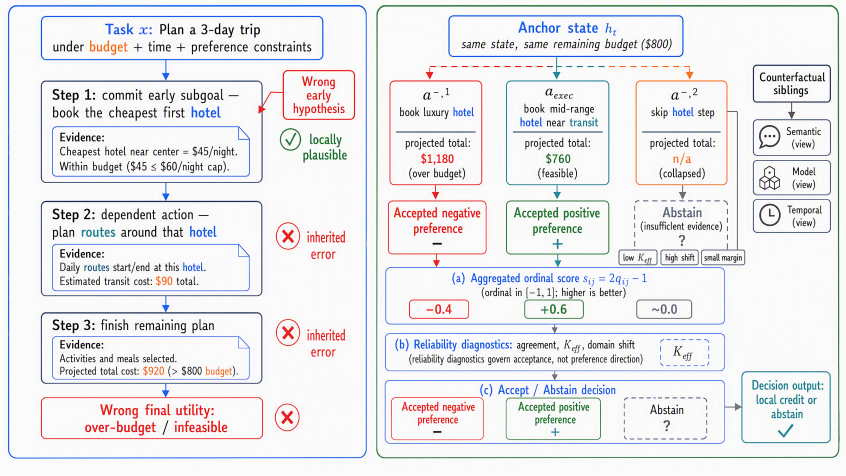

Figure 1: Motivation. A single reasoning chain can propagate an early, locally plausible error. CrossView-PRM restores the same decision state, compares the executed action with matched siblings, and either assigns ordinal credit from diverse views or abstains when evidence is unreliable.

a reliability rule transports. This separation makes the method falsifiable. Failed source-to-target transport disables learned weighting; insufficient target evidence triggers abstention; and a reward that improves local ranking but not downstream utility supports only the narrower reward-identification claim.

Our contributions are:

1. **Problem formulation.** We formalize target-verifier-free process supervision as local comparison among matched counterfactual actions, rather than unrestricted prefix ranking or absolute self-evaluation.
2. **Method.** We introduce correlation-aware cross-view aggregation, explicit abstention and transport gates, and a non-telescoping ordinal reward with frozen potential shaping.
3. **Analysis.** We derive a conditional reward-sign error bound that exposes view dependence, transport error, and potential approximation as separate failure terms.
4. **Evaluation.** We connect reward validity, cross-domain transport, downstream utility, mechanism robustness, and resource cost through five prespecified research questions.

Supplementary Appendix A–F contains expanded definitions, proofs, implementation details, complete tables, and the audit checklist.

2 Problem Formulation

For task $x \sim \mathcal{D}$, an agent observes history $h_t = (x, o_0, a_0, \ldots, o_t)$, samples $a_t \sim \pi_\theta(\cdot \mid h_t)$, and produces trajectory τ with latent utility $U(\tau, x) \in [0, 1]$. Target-domain reward construction and policy updates observe task text, histories, valid-action checks, and frozen auditor outputs, but not U, evaluator outputs, references, or process labels. Checkpoint selection separately uses validation reward; it is excluded from reward learning and policy gradients, and is not part of the verifier-free claim.

For frozen continuation policy π_c, define

$$V_U^{\pi_c}(h) = \mathbb{E}_{\tau \sim \pi_c(\cdot|h)}[U(\tau, x) \mid h].$$

At anchor h_t, let h_{t+1}^{+} be the executed successor and $h_{t+1}^{-,m}$ a sibling produced from the same restored state, transition randomness, and remaining budget. The unavailable local target is

$$Y_{t,m}^{\star} = \operatorname{sign}\left[V_U^{\pi_c}(h_{t+1}^{+}) - V_U^{\pi_c}(h_{t+1}^{-,m})\right].$$

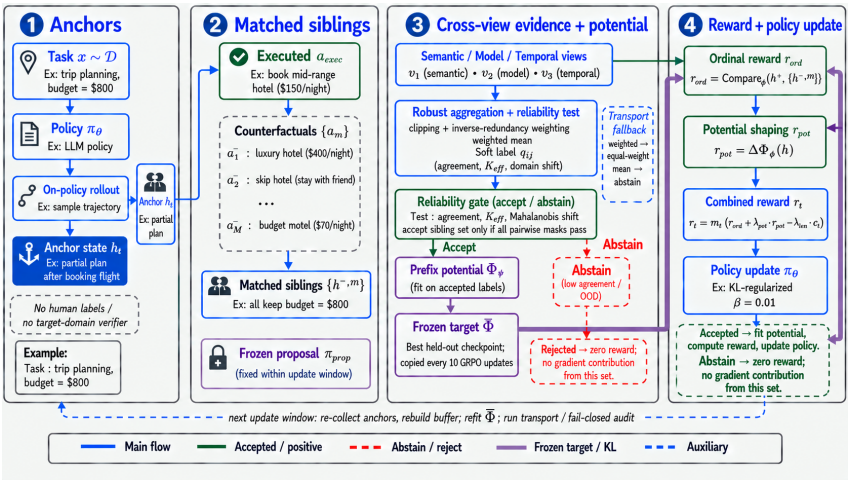

Figure 2: CrossView-PRM. Frozen views label matched sibling prefixes; screening precedes potential fitting; a frozen target potential supplies ordinal and auxiliary shaping rewards.

We seek a dense reward whose accepted ordinal signs agree with $Y^\star_{t,m}$, such that KL-regularized optimization improves $J_U(\theta) = \mathbb{E}_{x,\tau\sim\pi_\theta}[U(\tau, x)]$. The construction must be local, non-circular, correlation-aware, and fail-closed when reliability is unsupported.

These requirements induce three distinct statistical objects. A *pair-local* statistic determines whether the current sibling comparison is decisive; a *window-level* statistic detects whether nominally different views have become redundant; and a *source-to-target* test decides whether weights learned with oracle diagnostics remain admissible. Conflating these objects would let global agreement masquerade as pairwise confidence or let target data silently tune an oracle-dependent rule. Our algorithm keeps them separate and freezes every source-fitted component before target training.

3 METHOD

Overview and dependencies. Figure 2 gives the fixed dependency order. Module 1 constructs matched comparisons and view vectors; Module 2 produces a soft preference and acceptance mask; Module 3 learns a frozen potential and reward; and Module 4 updates the policy before creating the next window's frozen snapshots. Thus no trainable component produces the labels used for its own update. The principal innovation is the combination of *matched local counterfactuals* with *correlation-aware fail-closed aggregation*: siblings make a pair interpretable as action-level credit, while the gate decides whether multi-view evidence may use that credit at all.

Module 1: Matched counterfactual evidence. At optimization window b, frozen proposal and rollout snapshots generate $M = 4$ unique valid sibling actions and $L = 2$ equal-budget continuations per branch. Anchors with insufficient valid siblings are discarded. The default $K = 7$ views comprise two validated semantic renderings crossed with three frozen auditors (Qwen3.5-9B, Llama-3.1-8B-Instruct, and Gemma-2-9B-IT), plus the preceding Qwen policy checkpoint. Forward and reversed candidate orders are averaged:

$$v_k(h_i, h_j; x) = \tfrac{1}{2}[p_k^{A,B} + (1 - p_k^{B,A})].$$

Invalid outputs after one deterministic retry reject the pair; missing scores are never imputed.

Concretely, the simulator state immediately before the executed action is serialized and restored for every alternative. We sample at most $3M$ proposals at temperature 0.9, canonicalize them with a versioned action parser, remove exact duplicates and the executed action, and retain the first M admissible actions. Stochastic branches share transition and decoding seeds, remaining turns, and token ceilings. Continuations are evidence shown to auditors rather than independent views. This construction changes only the local action while holding the decision context fixed, which is necessary for interpreting the comparison as step-level credit.

Semantic renderings are generated before training and retained only if all numbers, entities, hard constraints, and tool schemas match the original task and every frozen auditor verifies bidirectional entailment. Each view receives the task, shared anchor, two successors, remaining budgets, and continuation sets in a fixed serialization. Auditor temperatures are fitted on source calibration data and then frozen. The temporal view uses the preceding-window checkpoint, never the policy being updated in the current window. The resulting complete K-view vector and shared nuisance metadata are the only inputs passed to Module 2.

Module 2: Correlation-aware aggregation and abstention. After clipping v_{nk} to $[0.05, 0.95]$, source-fitted ridge models remove observable nuisance effects (length, template, continuation budget, and task family). Let C_b be the Ledoit–Wolf residual correlation matrix from the preceding window ($C_0 = C_s$). Its effective view count and weights are

$$K_{\text{eff}}^{(b)} = \frac{\operatorname{tr}(C_b)^2}{\operatorname{tr}(C_b^2)}, \qquad w_k \propto \left(10^{-3} + \sum_{\ell \neq k} |(C_b)_{k\ell}|\right)^{-1}.$$

The soft label and disagreement are

$$q_{ij} = \sum_k w_k \tilde{v}_{k,ij}, \qquad d_{ij}^{(b)} = \sum_k w_k (\tilde{v}_{k,ij} - q_{ij})^2.$$

A source matrix C_s and target-window matrix C_b serve different roles: C_s calibrates the rule, whereas C_b uses only complete unlabeled view vectors from window $b-1$ and is frozen during window b. If fewer than 512 complete pairs are available, the next window is rejected. Consequently, policy updates cannot alter the weights used to score their own trajectories.

A frozen bge-m3 encoder and source PCA/covariance define Mahalanobis shift $d_{\text{shift}}(x)$. With u_{ij} indicating complete valid outputs,

$$m_{ij}^{(b)} = u_{ij}\, \mathbb{I}\Big[N_b \geq 512,\ K_{\text{eff}}^{(b)} \geq \kappa_K,\ d_{ij}^{(b)} \leq \kappa_V,\ d_{\text{shift}}(x) \leq \kappa_D,\ |q_{ij} - \tfrac{1}{2}| \geq \kappa_Q\Big].$$

Thresholds minimize selective error under at least 0.50 source coverage. Leave-one-domain-out transport enables weighting only when its aggregation gain has a positive 95% lower confidence bound and sign-error increase is at most 0.05; failure invokes equal weighting or full abstention (Lakshminarayanan et al., 2017; Geifman & El-Yaniv, 2017).

This gate is intentionally conservative. $K_{\text{eff}}^{(b)}$ detects window-wide collapse but does not certify correctness; $d_{ij}^{(b)}$ measures pair-specific disagreement; and d_{shift} restricts extrapolation. The margin test prevents nearly tied labels from entering potential training. A sibling set is accepted only if all four executed-versus-sibling pairs pass, so partially supported comparisons cannot leak into policy rewards. Module 2 therefore outputs exactly two objects to Module 3: the soft label q_{ij} and the immutable acceptance mask $m_{ij}^{(b)}$.

Module 3: Potential learning and reward construction. A Qwen3.5-4B scalar potential uses Bradley–Terry probability

$$p_\psi(h_i \succ h_j) = \sigma((\Phi_\psi(h_i) - \Phi_\psi(h_j))/\tau_\Phi)$$

and soft-label cross-entropy on accepted pairs. Semantic invariance, leave-one-view stability, and utility-preserving anti-hacking transformations regularize training (Appendix B.5). After task-disjoint validation, the target copy $\bar{\Phi}^{(b)}$ is frozen. For $\widehat{Y}_{t,m}^{(b)} = \operatorname{sign}[\bar{\Phi}^{(b)}(h_{t+1}^{+}) - \bar{\Phi}^{(b)}(h_{t+1}^{-,m})]$ and $m_t^{(b)} = \prod_m m_{+,m}^{(b)}$, the reward is

$$r_t = m_t^{(b)} \left[\frac{1}{M}\sum_{m=1}^{M} \widehat{Y}_{t,m}^{(b)} + \lambda_{\text{pot}}(\gamma \bar{\Phi}^{(b)}(h_{t+1}^{+}) - \bar{\Phi}^{(b)}(h_t)) - \lambda_{\text{len}} c_t\right].$$

Thus rejected sets receive no ordinal or shaping signal. We alternate potential fitting with ten GRPO updates, refresh snapshots only at window boundaries, and halt on repeated KL violation, non-positive diagnostic alignment, or reward–utility divergence.

For completeness, the fitted objective is

$$\mathcal{L}_\Phi = \mathbb{E}[\mathcal{L}_{\text{pair}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{hack}}\mathcal{L}_{\text{hack}}],$$

where $\mathcal{L}_{\text{pair}}$ is soft pairwise cross-entropy, $\mathcal{L}_{\text{inv}}$ preserves potential differences across equivalent task renderings, $\mathcal{L}_{\text{corr}}$ matches leave-one-view aggregates, and $\mathcal{L}_{\text{hack}}$ is invariant to utility-preserving verbosity and formatting changes. We use $\gamma = 1$, $\lambda_{\text{pot}} = 0.2$, and select all remaining coefficients on source/model-selection splits. Once selected, the potential is copied to $\bar{\Phi}^{(b)}$ and remains fixed while Module 4 consumes its reward; policy gradients never update the reward model.

Module 4: Alternating policy optimization. Each window executes seven ordered stages: collect on-policy anchors; generate matched siblings; elicit and screen cross-view preferences; fit the potential on a task-disjoint 90/10 split; freeze $\bar{\Phi}^{(b)}$; assign rewards; and run ten GRPO updates. Fourteen trajectories form one prompt group, observation tokens are masked, and standardized group returns define advantages. At the boundary, the policy is copied to the next proposal, rollout, and temporal-view snapshots before any new pair is scored. A window is skipped if effective view count or accepted coverage fails its threshold. Training retains the last valid checkpoint if policy KL exceeds 0.15 twice, proxy reward rises while diagnostic utility falls three times, or any non-finite value occurs.

Reliability analysis. For an accepted pair, let $Z_k = 2\tilde{v}_k - 1$, $S = \sum_k w_k Z_k$, $X = YS$, $\mu = \mathbb{E}[X \mid A] > 0$, and $\nu = \text{Var}(X \mid A)$.

Proposition 1. If the frozen potential reverses the aggregate sign with probability at most ϵ_Φ, then

$$\Pr(\widehat{Y} \neq Y \mid A) \leq \delta := \min\left\{1, \frac{\nu}{\nu + \mu^2} + \epsilon_\Phi\right\}.$$

For M accepted siblings, $\Pr[\text{sign}(\widehat{R}_t) \neq \text{sign}(R_t^\star) \mid m_t = 1] \leq \min\{1, \sum_m \delta_m\}$. The pair bound follows from Cantelli's inequality and the set bound from a union bound; no view-independence assumption is required. Positive covariance increases ν, making correlated failure explicit. Appendix C provides the proof and operational audit rule.

The operational audit replays target-window weights on labeled source folds to compute $\underline{\mu}$, $\overline{\nu}$, $\overline{\epsilon}_\Phi$, and $\overline{\delta}$. A non-positive $\underline{\mu}$, failed transport criterion, or held-out sign error above $\overline{\delta}$ triggers the registered fail-closed response. The audit concerns local signs under the frozen rollout policy, not J_U.

4 Experiments

4.1 Benchmarks and Baselines

CCP supplies 1,600 held-out sibling-pair tasks per seed across travel, scheduling, resource, and household domains. Target training uses 3,493 TripScore queries; selection uses 158 validation queries; final evaluation uses 1,000 synthetic and 219 real-user tasks. TravelPlanner uses its official 45/180/1,000 split and evaluators (Xie et al., 2024; Qu et al., 2025). All setting-matched methods use Qwen3.5-9B, 120 GRPO updates, 14 rollouts per prompt, and seeds $17, 42, 2026$. Baselines include no RL, self-consistency, single judge, Intuitor, RENT, Co-rewarding-I/II, Majority-view PRM, and potential-only PRM; outcome GRPO and PRISM are supervision-assisted references (Zhao et al., 2026; Prabhudesai et al., 2025; Zhang et al., 2026c; Ghimire et al., 2026).

CCP contains 3–8-decision deterministic planning problems with typed actions, hard constraints, soft preferences, and utilities normalized to $[0, 1]$. Exact oracle continuation values are obtained by enumeration below 10^6 remaining nodes and otherwise by 256 common-seed continuations. Oracle ties are retained as a separate class. Templates, entities, and seeds are disjoint among source diagnostics, target diagnostics, policy training, model selection, and sealed evaluation. Training workers cannot mount references, evaluator code, validation labels, or test identifiers.

We distinguish *policy-update budget*, which fixes the checkpoint, optimizer updates, prompt order, policy-token cap, and environment calls, from *reward-construction cost* (sibling, continuation, and auditor tokens). *Total token cost* is the realized sum of policy and reward-construction tokens, while *hardware cost* is realized H100 GPU-hours; we report both dimensions rather than aggregate their incompatible units. CrossView matches baselines only on the policy-update budget, not on either

Table 1: Main results, mean [95% CI]. TP is TravelPlanner; S/R are synthetic and real-user TripScore.

Method	CCP $\tau \uparrow$	Pair Acc.↑	Brier↓	TP Final↑	TripScore-S↑	TripScore-R↑
Qwen3.5-9B, no RL	–	–	–	6 [3,9]	.30 [-.20,.80]	.50 [-.10,1.00]
Self-consistency	.15 [.10,.20]	.59 [.56,.62]	.25 [.23,.27]	8 [5,12]	.50 [.10,.90]	.70 [.20,1.20]
Intuitor	.13 [.08,.18]	.57 [.54,.60]	.27 [.24,.29]	11 [7,15]	.70 [.30,1.00]	.90 [.50,1.30]
RENT	.12 [.07,.17]	.56 [.53,.59]	.28 [.25,.30]	10 [6,14]	.60 [.20,.90]	.80 [.40,1.20]
Co-rewarding-II	.31 [.25,.36]	.68 [.64,.71]	.21 [.19,.23]	18 [13,23]	1.10 [.80,1.40]	1.40 [1.00,1.70]
Majority-view PRM	.33 [.27,.39]	.70 [.66,.73]	.20 [.18,.22]	21.5 [16,27]	1.25 [1.00,1.50]	1.55 [1.20,1.90]
CrossView-PRM	**.45 [.40,.52]**	**.76 [.73,.80]**	**.17 [.14,.19]**	**30 [25,37]**	**1.58 [1.35,1.85]**	**1.92 [1.65,2.25]**
Outcome-only GRPO	–	–	–	41.5 [35,48]	1.80 [1.55,2.05]	2.28 [2.00,2.55]

realized cost dimension. Each run uses 16 H100 80GB GPUs and records wall time, GPU-hours, token category, environment calls, rejected anchors, and restarts. Full model revisions, optimization parameters, leakage checks, and hyperparameter grids are reported in Appendix D.

4.2 Evaluation Protocol

Each seed first calibrates nuisance regressors, correlation estimates, and acceptance thresholds on source CCP domains. A leave-one-domain-out audit then selects redundancy weighting, equal weighting, or complete abstention before target training. Training alternates twelve reward/policy windows. At every boundary, the audit records K_{eff}, coverage, CCP alignment, policy KL, reward scale, response length, reward–utility divergence, and the operational-audit statistics $(\mu, \nu, \epsilon_\Phi, \delta)$ together with the held-out sign error; complete audit statistics are reported in Appendix E. Checkpoint selection uses TripScore validation reward subject to coverage ≥ 0.50; each selected checkpoint is evaluated once on sealed TripScore and TravelPlanner tests. This is evaluator-assisted model selection, not verifier-free model selection; validation reward is excluded from reward construction and policy gradients.

Statistical analysis. The paired inferential unit is the evaluation task. For each of the three independently trained replicas, methods share task IDs and generation seeds. Reported intervals are two-level percentile bootstrap 95% CIs from 10,000 resamples, with replicas resampled at the outer level and tasks within replica at the inner level. Two-sided paired randomization tests are Holm-corrected across primary endpoints. We report all replica-level values and interpret the inferential results as paired task effects across these fixed training replicas. Metrics include Kendall's τ, pair accuracy, Brier score, selective risk/coverage, TravelPlanner pass rates, TripScore reward, and complete token/GPU accounting. Definitions, leakage controls, hyperparameters, and all secondary metrics are in Appendix D.

4.3 Overall Performance

Table 1 supports three primary observations.

- **Cross-view screening improves reward validity rather than coverage.** Against Majority-view PRM, CrossView raises Kendall's τ from 0.33 to 0.45, an absolute gain of 0.12 ([0.07,0.17], $p_{\text{Holm}} = .004$) and a 36.4% relative gain. Pair accuracy rises by 0.07, while Brier score falls from 0.20 to 0.17. Coverage decreases from 0.80 to 0.68, showing that the improvement comes from selective reliability rather than labeling more pairs.
- **Reward improvements transfer to planning utility.** TravelPlanner Final Pass rises from 21.5% to 30.0%, an 8.5-point or 39.5% relative improvement ([3.9,13.1], $p_{\text{Holm}} = .008$). TripScore-S/R increase by $0.33/0.37$, and all three seeds improve. The common direction across feasibility and preference-sensitive metrics links local ranking quality to downstream policy quality.
- **Verifier-free learning does not close the supervised gap.** Outcome-only GRPO reaches 41.5% Final Pass, 11.5 points above CrossView. CrossView's contribution is therefore reliable credit without a target verifier, not superiority to methods that consume terminal outcomes.

Table 2: Claim–evidence closure. LCB denotes a 95% lower confidence bound.

RQ	Test	Decision	Evidence
1	Reward validity	Pass	$\Delta\tau = +.12$, LCB $= .07$; accuracy $+.07$, Brier $-.03$, all corrected $p \leq .010$.
2	Transport	Pass	Four held-out gains $+.13$–$+.14$, all LCBs > 0; maximum sign-error increase $= .04 \leq .05$.
3	Policy utility	Pass	TP Final $+8.5$ points, LCB $= 3.9$; TripScore-S/R $+.33/+.37$, all three seeds improve.
4	Mechanism/robustness	Pass	Removing matching, weighting, or abstention changes τ by $-.19/-.07/-.10$; shared-bias false acceptance is .16 versus .38 for Majority PRM.
5	Cost characterization	Reported	Policy budget is matched; token/GPU h cost is $7.5\times$/2,500 per seed versus $4.5\times$/1,850. This is not an equal-resource comparison.

Table 3: Mechanism and transport tests. Ablation deltas are variant minus full CrossView; transport deltas are CrossView minus the best single auditor.

Test	Condition	$\Delta\tau$ [95% CI]	Δ TP Final [95% CI]	p_{Holm}
Ablation	Unmatched prefixes	-.19 [-.26,-.12]	-13.0 [-18.0,-8.0]	.004
Ablation	Equal/simple weighting	-.07 [-.11,-.03]	-6.5 [-10.5,-2.4]	.012
Ablation	No abstention	-.10 [-.15,-.05]	-8.0 [-12.6,-3.5]	.008
Transport	Travel held out	+.13 [+.07,+.19]	–	–
Transport	Schedule held out	+.14 [+.08,+.20]	–	–
Transport	Resource held out	+.14 [+.08,+.20]	–	–
Transport	Household held out	+.13 [+.06,+.19]	–	–

4.4 Mechanism, Transport, and Robustness

Matched credit and ordinal reward. Removing matched siblings causes the largest decline ($\Delta\tau = -.19$, ΔFinal $= -13.0$ points), followed by removing abstention $(-.10, -8.0)$ and redundancy weighting $(-.07, -6.5)$; all paired CIs exclude zero and corrected $p \leq .012$. Removing ordinal reward costs 10 final-pass points, versus 4 for potential shaping, confirming that the ordinal term is primary and potential shaping is auxiliary.

Transport and shared bias. In every held-out CCP domain, aggregation gains 0.13–0.14 τ; all lower confidence limits are positive and sign-error increase is at most 0.04, below the 0.05 gate. Under a shared misleading rationale, CrossView sacrifices 30 coverage points but limits accuracy degradation to 6 points and false acceptance to .16 [.11,.22], versus .38 [.32,.45] for Majority PRM. Candidate reversal and verbosity reduce CrossView accuracy by only 1 and 3 points, respectively, compared with 5 and 9 points for Majority PRM. These results show that abstention reacts to correlated failure instead of converting consensus into unconditional reward.

Resource trade-off. CrossView uses $7.5\times$ tokens and 2,500 H100 GPU h/seed versus $4.5\times$ and 1,850 for Majority PRM. Its $+8.5$-point Final-Pass gain ([3.9,13.1], $p_{\text{Holm}} = .008$) is measured under the shared policy-update budget, not equal token or hardware cost. RQ5 therefore reports a quality–cost trade-off rather than attributing the gain solely to aggregation or claiming end-to-end compute efficiency.

Calibration and selective risk. CrossView lowers Brier score from 0.20 to 0.17 and risk–coverage AUC from 0.24 to 0.14 relative to Majority-view PRM, while accepted coverage decreases from 0.80 to 0.68. Figure 3(b) shows that its selective error remains below both the majority and simple-mean aggregators across every registered operating point, rather than only at the chosen threshold. The

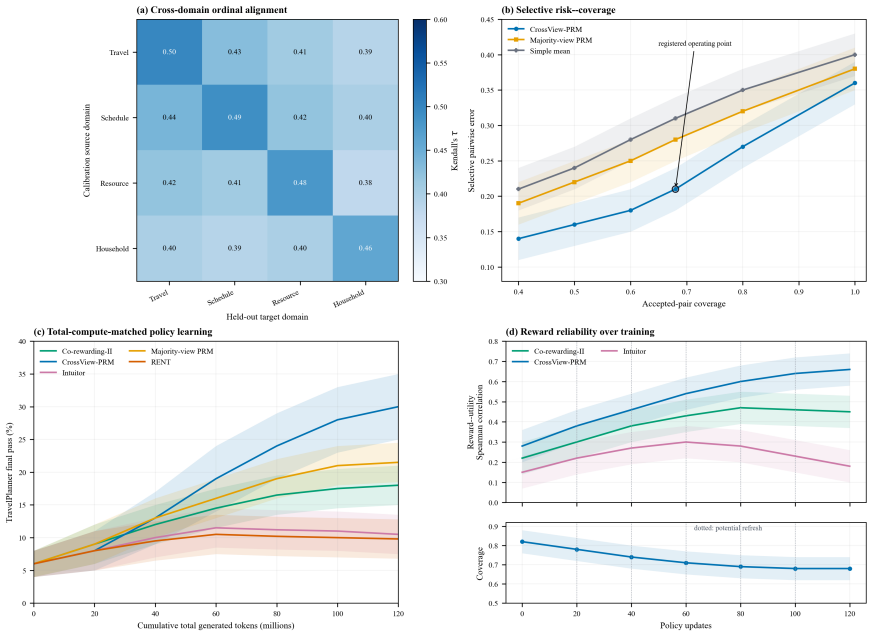

Figure 3: Mechanism diagnostics: held-out-domain alignment, selective risk–coverage, policy-budget-matched learning, and reward–utility alignment over updates. Bands are two-level bootstrap 95% CIs.

Table 4: Robustness changes from each method's unperturbed condition, in percentage points, except false-accept rate. Values are mean [95% CI].

Intervention and metric	Intuitor	Majority PRM	CrossView-PRM
Paraphrase: TP Final	-9 [-12,-6]	-6 [-8,-3]	**-3 [-4,-1]**
Candidate reversal: pair accuracy	-7 [-9,-4]	-5 [-7,-3]	**-1 [-2,0]**
Verbosity: pair accuracy	-12 [-16,-8]	-9 [-12,-6]	**-3 [-5,-2]**
Shared bias: pair accuracy	-18 [-24,-12]	-15 [-20,-10]	**-6 [-9,-4]**
Shared bias: false-accept rate	.42 [.36,.49]	.38 [.32,.45]	**.16 [.11,.22]**

0.10 absolute RC-AUC reduction indicates that the gate orders examples by realized risk: relaxing acceptance degrades gracefully instead of exposing a narrow, hand-selected operating point.

Training dynamics. Figure 3(c–d) connects reward behavior to policy learning. TravelPlanner Final Pass rises from approximately 6% before training to 30% after 120 updates, while Majority-view PRM saturates near 21.5%. Over the same windows, diagnostic reward–utility alignment increases from approximately 0.25 to 0.66, whereas accepted coverage decreases from 0.83 to 0.68. The opposing trends are consistent with progressively stricter selection as the policy moves away from the source-calibration distribution. Co-rewarding-II improves initially but plateaus below CrossView, and Intuitor's alignment declines after the midpoint, illustrating why proxy reward alone is insufficient for model selection.

Seed-level consistency. The three CrossView seeds obtain CCP Kendall's τ values $0.40/0.44/0.51$, compared with $0.28/0.33/0.38$ for Majority-view PRM. TravelPlanner Final Pass is $25/29/36\%$, compared with $17/21.5/26\%$. Thus every paired seed improves on both primary endpoints, and TripScore-S and TripScore-R show the same direction in all seeds. Complete replica-level values and paired randomization tests are provided in Appendix E.1.

Failure diagnostics. The full method records 0.3 [0,1] reward–utility divergence windows per seed. Removing abstention raises this to 4.5 [3,6], and updating the potential online rather than freezing it

also yields 4.5 [3,6]. The unmatched-prefix variant reaches 3.0 [2,4]. These failures separate three roles: matching defines a valid local contrast, freezing prevents a self-reinforcing label loop, and abstention limits policy updates after the evidence becomes unreliable. The fail-closed criteria stop or skip these windows before test evaluation; stopped checkpoints and rejected comparisons remain in the run ledger rather than being silently discarded.

5 Related Work

Process rewards and agent training. Process supervision localizes errors but usually requires human labels, executable verifiers, outcome-conditioned rollouts, or a stronger reward model (Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024; Khalifa et al., 2025; Zhang et al., 2026b;a). TravelPlanner exposes the gap between fluent generation and valid execution, while WebRL and Planner-R1 show the benefit of curricula, KL control, and dense constraint feedback (Xie et al., 2024; Qi et al., 2025; Zhu et al., 2025; Graves et al., 2017). We instead study target training without terminal verifiers or process annotations.

Intrinsic and self-generated feedback. Intuitor and RENT use certainty or entropy as intrinsic rewards but can favor confident errors (Zhao et al., 2026; Prabhudesai et al., 2025). Self-consistency and self-rewarding models aggregate sampled reasoning or model judgments (Wang et al., 2022; Yuan et al., 2024); Co-rewarding and PRISM add analogous inputs, delayed teachers, or a generated process model (Zhang et al., 2026c; Ghimire et al., 2026). CrossView differs by measuring residual dependence, testing transport on held-out domains, and assigning zero reward when evidence is unsupported.

Counterfactual credit and selective prediction. Potential shaping propagates delayed information but does not independently identify the better action at a state. Our matched siblings restore one anchor and equalize downstream budgets. The gate is related to ensemble uncertainty and selective prediction (Lakshminarayanan et al., 2017; Geifman & El-Yaniv, 2017), but separates unlabeled collapse statistics from oracle-backed transport tests; agreement is never treated as a correctness certificate.

6 Conclusion

CrossView-PRM constructs rewards and policy updates without target-domain evaluator outputs; checkpoint selection is evaluator-assisted. It improves CCP Kendall's τ from 0.33 to 0.45, TravelPlanner Final Pass from 21.5% to 30.0%, and shared-bias false acceptance from 0.38 to 0.16. Its $7.5\times$/2,500 token/GPU h cost exceeds Majority PRM's $4.5\times$/1,850, supporting a quality–cost trade-off rather than an end-to-end efficiency claim.

Reproducibility statement

The main text specifies the estimands, gate, data isolation, primary budgets, seeds, metrics, and tests. Appendix F records the complete protocol, and Appendix I lists the manifests, configurations, run ledger, and per-task records required to regenerate all tables and figures.

AI use statement

Generative AI tools were used to assist research ideation, language editing, and conversion of the authors' research notes into LaTeX. All AI-assisted text, mathematical claims, citations, and artifacts were reviewed by the authors, who take responsibility for the final content.

REFERENCES

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.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Mukesh Ghimire, Aosong Feng, Liwen You, Youzhi Luo, Fang Liu, and Xuan Zhu. Prism: A unified framework for post-training llms without verifiable rewards. *arXiv preprint arXiv:2601.04700*, 2026.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pp. 1311–1320. Pmlr, 2017.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. *arXiv preprint arXiv:2504.16828*, 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.

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.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*, 2025.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, et al. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 79791–79821, 2025.

Yincen Qu, Huan Xiao, Feng Li, Gregory Li, Hui Zhou, Xiangying Dai, and Xiaoru Dai. Tripscore: Benchmarking and rewarding real-world travel planning with fine-grained evaluation. *arXiv preprint arXiv:2510.09011*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents. *arXiv preprint arXiv:2402.01622*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Lingyin Zhang, Jun Gao, Xiaoxue Ren, and Ziqiang Cao. The bidirectional process reward model. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12564–12580, 2026a.

Zheng Zhang, Ziwei Shan, Kaitao Song, Yexin Li, and Kan Ren. Linking process to outcome: Conditional reward modeling for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 87995–88016, 2026b.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Xuan Li, Xiao Feng, Jiangchao Yao, Bo Han, et al. Co-rewarding: Stable self-supervised rl for eliciting reasoning in large language models. In *International Conference on Learning Representations*, volume 2026, pp. 82520–82558, 2026c.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. In *International Conference on Learning Representations*, volume 2026, pp. 2548–2581, 2026.

Siyu Zhu, Yanbin Jiang, Hejian Sang, Shao Tang, Qingquan Song, Biao He, Rohit Jain, Zhipeng Wang, and Alborz Geramifard. Planner-r1: Reward shaping enables efficient agentic rl with smaller llms. *arXiv preprint arXiv:2509.25779*, 2025.