REWIND TO RETHINK: LOCALIZED REPAIR FOR MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reasoning models can continue to rely on an early judgment even after generating additional reasoning, allowing an initial error to persist into the final answer. We define step-to-step interventional support using offline Shapley attribution: preceding steps are retained or masked, and their marginal contributions to the mean log-likelihood of fixed subsequent text are averaged across coalitions. The resulting matrix reveals how an early judgment can continue to support later reasoning and the final answer, a pattern we call answer inheritance. Inspired by this observation, we propose BANKROLL, a training-free method that rolls back to regenerate an earlier judgment together with the reasoning built on it. Cashout monitors intermediate answers to determine when thinking stops. Refinance uses online local masking tests as a heuristic to select a rollback point, truncates the chain before that step, and samples alternative continuations. Writeoff retains the original chain if no rollback point is chosen, no continuation is selected, or the probe-score acceptance check fails. Experiments across five benchmarks and three multimodal models show that BANKROLL consistently improves task performance over standard chain-of-thought (CoT).

1 INTRODUCTION

Chain-of-thought (CoT) reasoning allows multimodal models to solve image–text tasks through intermediate observations and deductions (Wei et al., 2022; Zhang et al., 2024). These steps become part of the context for the rest of the response. A mistaken interpretation of an image or an invalid deduction can therefore become a premise for subsequent reasoning. Even as the chain grows, later steps may continue to rely on the same judgment and carry its error into the final answer (Liao et al., 2025).

Recent work maps sentence-level influence through masking and counterfactual resampling (Bogdan et al., 2025) and estimates dependencies among fixed reasoning units using linear structural equations (Wei et al., 2026). A step's contribution can depend on which other steps remain in context: masking one statement may have little effect when another repeats the same premise. We therefore analyze step-to-step support offline as a cooperative game (Shapley, 1953). Preceding steps are players, and each coalition retains a subset of them while masking the rest in place. With the model, image–question input, and original target text fixed, we score each coalition by the target's mean conditional log-likelihood over content tokens. Each predecessor's Shapley value is the weighted average of its marginal changes to this score across coalitions.

Figure 1 illustrates this support in a depth-map question. At T_{19}, the model describes Point A as lying in a shallow region. This judgment has positive Shapley contributions to multiple summary steps, including Y_6, which concludes that Point A is the shallowest location. The model ultimately selects Point A, whereas the reference answer is Point B. We call the persistence of support from an earlier judgment into later reasoning and the final answer *answer inheritance*.

Inspired by this observation from the offline Shapley analysis, we restart generation before an earlier judgment so that both the judgment and the reasoning built on it can change. Choosing where to restart matters: models can struggle to locate mathematical reasoning errors even with a reference solution (Srivatsa et al., 2025), while a supplied error location can facilitate correction (Tyen et al., 2024). At inference time, we use local sensitivity of the answer-probe score as a heuristic for choosing where to try revision. Controlling generation cost also involves two decisions: when to stop the

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.

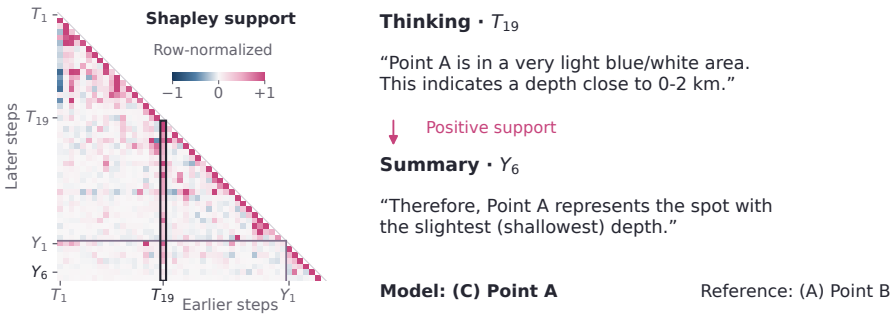


Figure 1: An earlier judgment supports an incorrect final answer.

original chain and how far to extend alternative continuations before selection. Prior work addresses the first decision by stopping reasoning when confidence in an intermediate trial answer reaches a threshold (Yang et al., 2026a). The second arises because completing every candidate would spend computation on continuations that are ultimately discarded.

We propose **BANKROLL**, a training-free method for frozen multimodal models that determines when to stop reasoning and where to restart it for local revision. **Cashout** uses intermediate answer probes to determine when thinking stops. Once thinking ends, naturally or through Cashout, **Refinance** independently masks candidate steps and uses the resulting score drops to select a rollback point. It samples short continuations from the prefix before that point and completes only the one with the highest available last-probe score, limiting the cost of exploring alternatives. **Writeoff** compares the score retained at selection with the original chain's last-probe score to decide whether to accept the completed repair or retain the original chain. Experiments across five benchmarks and three multimodal models demonstrate BANKROLL's effectiveness in improving performance on a range of multimodal tasks.

Our contributions are:

- We characterize answer inheritance through an offline cooperative game over reasoning steps. Shapley attribution traces how support from an earlier judgment can persist into later reasoning and the final answer.
- We propose BANKROLL, a training-free method for frozen multimodal models that controls when reasoning stops and selectively regenerates an earlier judgment together with the reasoning built on it.
- Evaluation across five benchmarks and three multimodal models shows that BANKROLL improves task scores over Standard CoT in every evaluated setting. The gains extend from multimodal reasoning to visual hallucination evaluation and text recognition.

2 Analysis and Motivation

We analyze answer inheritance offline by measuring how masking earlier steps changes the likelihood of the original later text. Following removal-based explanations (Covert et al., 2021), we specify the masking operation, the target score, and the rule for aggregating contributions.

For image–question input $x = (I, q)$, let $Z = (z_1, \ldots, z_n)$ denote the generated chain. Thinking steps are delimited by line breaks and summary steps by sentence boundaries. The summary begins immediately after the thinking block's closing marker. The model p_θ, input, and original target text remain fixed throughout the analysis.

For target z_t, the players are its preceding steps, indexed by $N_t = \{1, \ldots, t-1\}$. A coalition $S \subseteq N_t$ retains the original text of its members and masks all other predecessors in place:

$$C_t(S) = \big(x, \bar{z}_1(S), \ldots, \bar{z}_{t-1}(S)\big), \qquad \bar{z}_j(S) = \begin{cases} z_j, & j \in S, \\ (\rho)^{|z_j|}, & j \notin S, \end{cases} \tag{1}$$

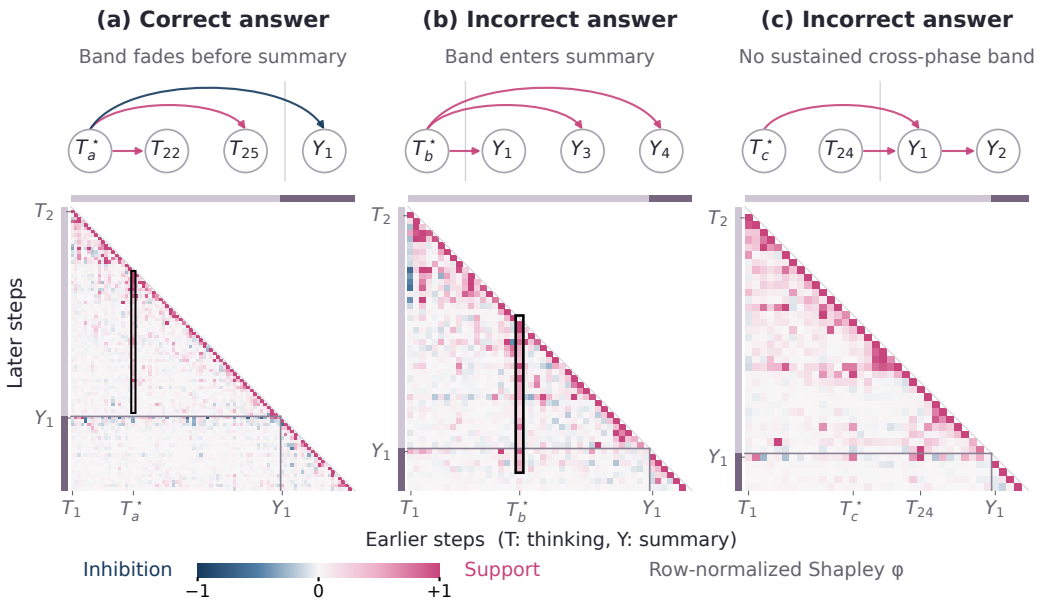


Figure 2: Step-to-step Shapley contributions in three example chains. Top graphs show selected contributions to fixed target text, with thinking (T) and summary (Y) nodes in generation order and schematic spacing. Equal-width arrows indicate support (rose) or inhibition (blue). Bottom matrices use row normalization, so absolute magnitudes cannot be compared across targets. Black boxes highlight selected bands, and phase marks indicate summary onset.

where ρ is a fixed replacement token and $|z_j|$ is the original token count. Replacement preserves each step's length and position. For a fixed, nonempty set W_t of content-token positions in z_t, the coalition value is the target's mean conditional log-likelihood:

$$v_t(S) = \frac{1}{|W_t|} \sum_{\ell \in W_t} \log p_\theta(z_{t,\ell} \mid C_t(S), z_{t,<\ell}) \,, \tag{2}$$

where $z_{t,\ell}$ is target token ℓ and $z_{t,<\ell}$ its original within-target prefix. No alternative target text is generated. The marginal contribution of source $i \in N_t \setminus S$ is the score change when its original text is restored:

$$\Delta_{t,i}(S) = v_t(S \cup \{i\}) - v_t(S). \tag{3}$$

A source's marginal contribution can vary with the other retained predecessors. We therefore average it across coalitions using Shapley weights (Shapley, 1953):

$$\phi_{t,i} = \sum_{S \subseteq N_t \setminus \{i\}} \frac{|S|!\,\big(|N_t| - |S| - 1\big)!}{|N_t|!}\, \Delta_{t,i}(S). \tag{4}$$

Positive values indicate that retaining source i raises the original target score on average across the weighted coalitions. Negative values indicate inhibition.

We estimate the Shapley values by permutation sampling (Mitchell et al., 2022). Let σ be a uniformly sampled permutation of N_t, and let $P_i(\sigma)$ contain the indices preceding i in that permutation. The permutation determines coalition membership, while steps in $C_t(S)$ retain their original order. The equivalent expectation and estimator from L_t independent draws are

$$\begin{aligned} \phi_{t,i} &= \mathbb{E}_\sigma[\Delta_{t,i}(P_i(\sigma))] \,, \\ \widehat{\phi}_{t,i} &= \frac{1}{L_t} \sum_{r=1}^{L_t} \Delta_{t,i}(P_i(\sigma_r)) \,. \end{aligned} \tag{5}$$

Exactly $|S|!(|N_t| - |S| - 1)!$ of the $|N_t|!$ permutations satisfy $P_i(\sigma) = S$, recovering the weight in Equation (4).

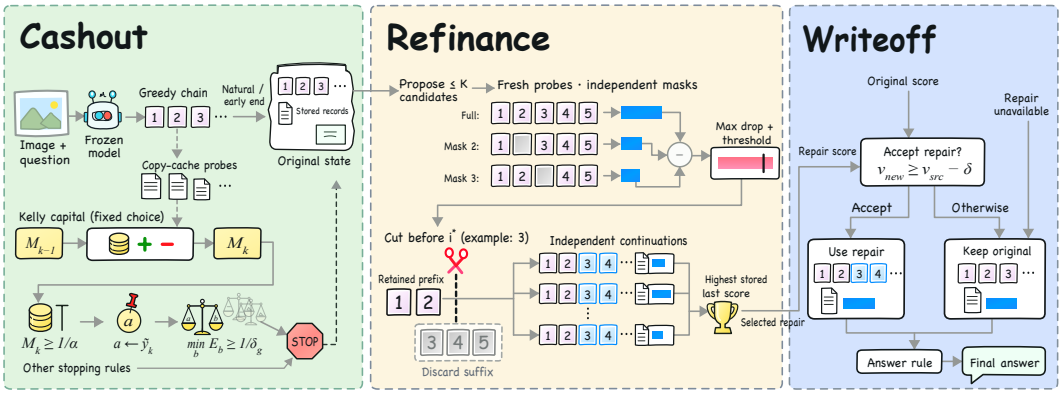


Figure 3: Overview of BANKROLL. Answer probes guide Cashout's stopping decisions. Refinance independently masks candidate steps to locate a rollback point and samples continuations from the retained prefix. The highest stored last-probe score determines selection. Writeoff accepts the repair if its criterion holds, otherwise retaining the original state.

The estimates $\widehat{\phi}_{t,i}$ form a lower-triangular matrix with target rows t and source columns $i < t$. The first step has no predecessors and contributes no entries. For visualization, a row with estimates available for every predecessor is normalized as

$$m_t = \max_{j<t} |\widehat{\phi}_{t,j}|, \qquad \widetilde{\phi}_{t,i} = \frac{\widehat{\phi}_{t,i}}{m_t + \epsilon}, \qquad \epsilon > 0. \tag{6}$$

The offset prevents division by zero. Unevaluated entries remain missing, not zero. Normalized magnitudes describe relative contributions within a target row and cannot be compared across targets.

Figure 2(b) shows repeated positive contributions from an early step to the original summary text. The highlighted band in (a) does not persist into the summary, while (c) shows no sustained cross-phase support band.

The persistent support in (b) motivates revisiting an earlier judgment together with the reasoning built on it. We restart generation from the prefix before a selected step, retaining the preceding context while allowing both the judgment and its subsequent reasoning to change.

3 Method: BANKROLL

BANKROLL uses answer probes on the frozen model p_θ to track predictions during generation and examine their sensitivity to earlier reasoning (Figure 3). Successive probes guide Cashout's stopping decision, while local masking tests guide Refinance's selection of a step to revisit once thinking ends. Continuations are sampled from the prefix before that step, and Writeoff compares the selected continuation with the original chain to decide which state supplies the final answer. Naturally completed and early-stopped chains undergo the same procedure, with at most one repair round per original chain.

3.1 Cashout

We generate the original chain greedily and query its current answer at each thinking-step boundary. A fixed suffix appended to a copy of the key–value cache closes thinking and elicits a probe, leaving the main generation context unchanged.

For fixed-choice tasks, normalizing candidate-letter logits yields the answer distribution π_k, prediction $\widetilde{y}_k = \arg\max_j \pi_k(j)$, and score $v_k = \max_j \pi_k(j)$. The initial distribution π_0 is read after prefilling. Within Cashout, the Cashout-Cert stopping rule combines capital accumulation with pairwise answer comparisons. It updates a capital process from consecutive probe distributions:

$$d_k = 1 - \sum_j \pi_k(j)\pi_{k-1}(j), \qquad M_0 = 1, \quad M_k = M_{k-1}[1 + \lambda(c - d_k)]. \tag{7}$$

The overlap between probe distributions makes d_k small when both favor the same letter with high probability. Here, $c - d_k$ is the unit return and λ is the betting scale in the Kelly-style capital update (Kelly, 1956; Ramdas et al., 2023). Preset $0 \le c < 1$ and $0 \le \lambda < 1/(1-c)$ keep each multiplier positive. For $\lambda > 0$, capital grows when $d_k < c$ and declines when $d_k > c$.

The accumulated capital determines when to fix a reference answer for subsequent comparisons. For preset $\alpha, \delta_{\mathrm{g}} \in (0, 1)$, the first time M_k crosses $1/\alpha$, Cashout-Cert fixes the current prediction as a. Pairwise counting starts at the next probe. For each alternative b, n_b counts predictions of either a or b, and x_b counts predictions of a. Both start at zero and continue without resetting if M_k later declines. The comparison scores are

$$E_b = 2^{n_b} \frac{B(x_b + \frac{1}{2}, n_b - x_b + \frac{1}{2})}{B(\frac{1}{2}, \frac{1}{2})}, \qquad G_k = \min_{b \neq a} E_b, \tag{8}$$

where B is the Beta function. Cashout-Cert stops thinking when $G_k \ge 1/\delta_{\mathrm{g}}$, requiring every pairwise score to reach the threshold. Each E_b measures pairwise imbalance in either direction, so this condition establishes neither a lead nor correctness for a. The parameters α and δ_{g} do not bound overall answer error.

Independently of Cashout-Cert, a saturation rule stops fixed-choice thinking after m_{sat} consecutive probes predict the same letter, each with $v_k \ge v_{\mathrm{sat}}$; the initial π_0 is excluded. Open-ended probes greedily generate short trial answers and score them by the geometric mean of their token probabilities (Yang et al., 2026a). Thinking stops when two consecutive answers match exactly after normalization.

When a stopping rule triggers, we append the fixed answering suffix and decode the answer. Both early-stopped and naturally completed states S store the complete chain, probe predictions $\widetilde{y}_k$ and scores v_k, and the answer $y(S)$. The last probe is the final one performed during generation and may precede chain completion.

3.2 Refinance

Given a generated state S, Refinance selects a thinking step to revisit based on how much masking it lowers the answer-probe score. The candidate set $\mathcal{C}$ contains at most K steps with nonzero reasoning-shift word counts, ranked by descending count and then earlier position.

A fresh probe on the input, complete thinking text, and fixed probe suffix supplies the baseline. These scores s are computed separately from generation-time scores v. Each candidate is masked independently in place by replacing its tokens one-for-one with the designated token, keeping all other thinking fixed. For fixed-choice tasks, every context scores the baseline letter $a^\star = \arg\max_a p_{\mathrm{full}}(a)$:

$$s_{\mathrm{full}} = p_{\mathrm{full}}(a^\star), \qquad s_{\mathrm{masked},i} = p_{\mathrm{masked},i}(a^\star). \tag{9}$$

Open-ended probes generate a separate trial answer in each context and score it by the geometric mean of its token probabilities. Answers can differ across contexts. We measure the score drop as

$$dv_i = s_{\mathrm{full}} - s_{\mathrm{masked},i}, \qquad i^\star = \arg\max_{i \in \mathcal{C}} dv_i. \tag{10}$$

For fixed-choice tasks, dv_i is the step's local marginal contribution to the baseline letter's score with all other thinking present. A positive drop indicates local dependence of the probe score on the step and can arise from either a correct or an incorrect judgment. The drop may be small when the same premise appears in other steps. For nonempty $\mathcal{C}$, we select the largest drop, break ties by earlier position, and require $dv_{i^\star} \ge \tau_{\mathrm{dv}} > 0$.

Revisiting the selected step allows its judgment and subsequent reasoning to change together. We truncate before $i^\star$, deleting the pivot and its entire suffix while retaining the input and preceding reasoning. Appendix F illustrates this operation on a chart question. From this prefix, we sample up to R independent candidate continuations using nucleus sampling (Holtzman et al., 2020). In the main experiments, $R = 3$, and each candidate is limited to three reasoning steps before selection. Fixed-choice continuations use step-end probes and saturation stopping, with the Cashout-Cert capital update and pairwise comparisons disabled. For candidate state S_r, let v_r denote its generation-time last-probe score at selection. We select

$$r^\star = \arg\max_{r:\, v_r \text{ available}} v_r, \qquad v_{\mathrm{new}} = v_{r^\star}, \tag{11}$$

breaking ties by generation order, without voting over answers. We then complete reasoning and answering in $S_{r^\star}$ while retaining the selection-time score v_{new} for Writeoff. Rejected candidates are discarded without generating complete continuations. The completed state enters Writeoff for comparison with the original state.

3.3 WRITEOFF

Refinance selects among newly sampled continuations. Writeoff decides whether to replace the retained original state S_{src} with the selected state. Let v_{src} denote the original chain's generation-time last-probe score, with v_{new} fixed at continuation selection. These scores may refer to different predicted answers.

When both scores are available, acceptance requires

$$v_{\text{new}} \geq v_{\text{src}} - \delta, \qquad \delta \geq 0. \tag{12}$$

The original chain sets the reference score, and acceptance permits a decrease of up to δ. The comparison does not verify answer correctness.

Acceptance adopts the entire state $S_{r^\star}$, including its chain, probe records, and associated answer. We retain S_{src} if the comparison fails or v_{src} is missing. The original state is also retained when no candidate is proposed, the pivot fails its threshold, or no continuation has a last-probe score.

3.4 FINAL ANSWER

The state S_{out} selected by Writeoff supplies the default prediction $\widehat{y} = y(S_{\text{out}})$. Open-ended tasks return this answer directly.

For fixed-choice tasks, we read the prediction $\widetilde{y}_{\text{out}}$ and score v_{out} from the stored record of the selected state's last probe during generation. We replace $\widehat{y}$ with $\widetilde{y}_{\text{out}}$ if both are available, $v_{\text{out}} \geq v_{\text{ovr}}$, and $\widetilde{y}_{\text{out}} \neq \widehat{y}$. Otherwise, we retain the default answer. This output rule leaves the selected state unchanged. Algorithm 1 gives the complete procedure, with one repair round detailed in Algorithm 2.

Algorithm 1 BANKROLL

Input: x, frozen p_θ, task type
Output: Final prediction $\widehat{y}$

$C \leftarrow x$; $H \leftarrow [\,]$
Initialize Cashout (Section 3.1)
For fixed-choice tasks, record π_0 in H
while thinking is unfinished **do**
  Extend C greedily to a step boundary or generation end
  if a thinking step ends **then**
    Probe a copy of C; append to H
    Update Cashout; **if** triggered, **break**
  end if
end while
Complete answering in C; extract y
$S_{\text{src}} \leftarrow (C, H, y)$
$S_{\text{out}} \leftarrow \text{REPAIR}(S_{\text{src}})$
$\widehat{y} \leftarrow y(S_{\text{out}})$
if fixed-choice task **then**
  $(\widetilde{y}_{\text{out}}, v_{\text{out}}) \leftarrow \text{LASTPROBE}(S_{\text{out}})$
  if both exist and $v_{\text{out}} \geq v_{\text{ovr}}$ **then**
    $\widehat{y} \leftarrow \widetilde{y}_{\text{out}}$
  end if
end if
return $\widehat{y}$

Algorithm 2 Refinance and Writeoff

Input: Original state S_{src}
Output: Retained or repaired state

$\mathcal{C} \leftarrow \text{PROPOSEPIVOTS}(S_{\text{src}}, K)$
if $\mathcal{C} = \varnothing$, **return** S_{src}
Freshly probe S_{src} for s_{full}
for all $i \in \mathcal{C}$ **do**
  Independently mask step i for $s_{\text{masked},i}$
  $dv_i \leftarrow s_{\text{full}} - s_{\text{masked},i}$
end for
$i^\star \leftarrow \arg\max_{i \in \mathcal{C}} dv_i$
if $dv_{i^\star} < \tau_{\text{dv}}$, **return** S_{src}
$P \leftarrow$ prefix strictly before step $i^\star$
$\{S_r\} \leftarrow \text{SAMPLECANDIDATES}(P, R)$
Read each stored last-probe score v_r
$\mathcal{V} \leftarrow \{r : v_r \text{ is available}\}$
if $\mathcal{V} = \varnothing$, **return** S_{src}
$r^\star \leftarrow \arg\max_{r \in \mathcal{V}} v_r$
$v_{\text{new}} \leftarrow v_{r^\star}$
Complete reasoning and answering in $S_{r^\star}$ without changing v_{new}
Read v_{src} from $\text{LASTPROBE}(S_{\text{src}})$
if v_{src} is missing, **return** S_{src}
if $v_{\text{new}} \geq v_{\text{src}} - \delta$, **return** $S_{r^\star}$
return S_{src}

SAMPLECANDIDATES generates up to R short candidate continuations from P using nucleus sampling (Section 3.2). LASTPROBE returns the prediction and score from a state's last stored generation-time probe; missing values remain unavailable.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We use Qwen3.5-9B (Qwen Team, 2026) as the primary model, with GLM-4.1V-9B-Thinking (GLM-V Team, 2025) and InternVL3.5-8B (Wang et al., 2025b) for additional evaluation.

Benchmarks. We evaluate multimodal understanding and reasoning on M^{3}CoT (Chen et al., 2024b), MMStar (Chen et al., 2024a), and MMMU-Pro (Yue et al., 2025), visual hallucination on Hallusion-Bench (Guan et al., 2024), and visual text recognition and parsing on CC-OCR (Yang et al., 2024). Table 1 reports results on all five benchmarks.

Baselines. We compare six training-free baselines. Standard CoT (Wei et al., 2022) generates a single reasoning chain, while Direct answer omits reasoning. Concise CoT adds an instruction to keep reasoning concise. Dynamic Early Exit in Reasoning (DEER) (Yang et al., 2026a), adapted to the same models and answer probes, stops reasoning when intermediate trial-answer confidence reaches a threshold. No-Think uses routing inspired by Think or Not? (Yong et al., 2025), accepting the initial direct answer when its confidence reaches a threshold and running Standard CoT otherwise. CoT-SC (5) applies self-consistency (SC) (Wang et al., 2023) by majority vote over the final answers of five independently sampled complete reasoning chains per example.

Table 1: Task scores and generated-token counts across five benchmarks and three multimodal models.

	M^{3}CoT		**MMStar**		**MMMU-Pro**		**HallusionBench**		**CC-OCR**	
Method	Score↑	T_{p} ↓	Score↑	T_{p} ↓	Score↑	T_{p} ↓	Score↑	T_{p} ↓	Score↑	T_{p} ↓
Qwen3.5-9B										
Standard CoT	81.06	1431	75.33	1768	54.62	3111	77.92	1656	77.59	1505
Direct answer	65.66	358	69.40	272	37.98	445	75.29	205	79.01	349
DEER†	80.16	970	77.40	1005	65.66	2066	80.23	1238	78.14	981
No-Think‡	81.15	1311	75.47	1596	54.97	2952	78.55	1571	77.38	1132
CoT-SC (5)	<u>82.36</u>	7701	<u>78.80</u>	9780	<u>67.98</u>	16256	**81.49**	9327	78.35	6769
Concise CoT	81.54	1319	76.67	1091	62.72	1713	68.66	984	<u>79.05</u>	1172
BANKROLL	**82.70**	1853	**79.00**	1796	**69.02**	4049	<u>81.28</u>	1941	**79.55**	641
GLM-4.1V-9B-Thinking										
Standard CoT	60.70	927	60.47	801	39.13	2067	75.18	442	61.89	1982
Direct answer	51.42	437	54.00	375	23.70	494	71.40	336	62.62	1271
DEER†	70.06	571	<u>66.93</u>	511	**49.94**	1468	74.97	312	59.83	1699
No-Think‡	66.91	792	64.47	661	41.39	1872	75.39	385	61.88	2973
CoT-SC (5)	63.76	4677	62.67	4023	47.46	10492	**75.92**	2355	58.09	9220
Concise CoT	<u>70.45</u>	738	61.40	678	43.24	1695	75.29	368	**63.27**	1320
BANKROLL	**71.14**	1118	**68.13**	901	<u>49.77</u>	2461	<u>75.50</u>	551	<u>63.23</u>	2776
InternVL3.5-8B										
Standard CoT	68.33	1444	66.67	1154	<u>55.43</u>	3003	74.45	1124	67.95	1471
Direct answer	66.65	214	63.60	155	42.02	310	70.24	81	<u>69.08</u>	351
DEER†	<u>70.53</u>	646	**69.80**	531	55.20	1445	73.40	689	68.06	1095
No-Think‡	62.08	923	66.73	767	44.45	2282	73.29	956	68.78	1297
CoT-SC (5)	69.63	5607	69.20	5506	54.28	14963	**76.55**	5015	67.85	5545
Concise CoT	66.13	1027	66.33	1070	52.77	2783	74.13	754	63.53	1209
BANKROLL	**73.34**	1276	<u>69.60</u>	1017	**55.90**	2410	<u>75.81</u>	1062	**69.68**	622

Notes. Our DEER† adaptation stops reasoning when intermediate trial-answer confidence reaches a threshold. No-Think‡ accepts the initial direct answer if its confidence reaches a threshold and runs Standard CoT otherwise. T_{p}: mean total generated tokens per example, including answers and probes but excluding prefill, rounded to integers. Bold and underlining mark the best and second-best scores per model and benchmark, respectively.

Metrics. We report accuracy (%) on the first four benchmarks and the official 0–100 score on CC-OCR (Yang et al., 2024). T_{th} counts thinking tokens in the final chain, including the retained prefix and excluding the answer. T_{p} counts total generated tokens, including answers and probes but

excluding prefill. For CoT-SC (5), both counts sum over all five sampled chains. Token counts are averaged over evaluation examples.

BANKROLL configuration. Each repair considers up to three candidate continuations, each limited to three reasoning steps before selection by its last answer-probe score. Both task scores and T_{p} in Table 1 use this configuration. T_{p} includes all tokens actually generated for the candidates and answer probes; rejected candidates are not expanded into complete continuations. Appendix C reports sensitivity to repair and stopping parameters.

4.2 Main Results

BANKROLL makes an earlier judgment revisable after subsequent reasoning has already built on it. This capability complements DEER's stopping and No-Think's initial routing. Table 1 shows the effectiveness of the combined framework: BANKROLL improves task scores over Standard CoT and No-Think in all 15 settings and over DEER in 13. On MMMU-Pro, gains over Standard CoT reach 14.40 and 10.64 percentage points for Qwen3.5-9B and GLM-4.1V-9B-Thinking. Improvements on HallusionBench and CC-OCR across all three models extend this benefit beyond the reasoning benchmarks.

The comparison with CoT-SC (5) highlights the value of exploring alternatives around a retained prefix. BANKROLL selects among short continuations and completes only the selected one, avoiding repeated prefix generation and full expansion of discarded candidates. It achieves higher accuracy in all nine reasoning comparisons while using fewer generated tokens in all 15 settings, including candidate and probe outputs. On M^{3}CoT with Qwen3.5-9B, accuracy reaches 82.70% with 1853 tokens, compared with 82.36% and 7701 tokens for CoT-SC (5) (Figure 4). Local revision can thus provide a competitive alternative to voting over complete chains. CoT-SC (5) retains higher HallusionBench accuracy by 0.21–0.74 percentage points.

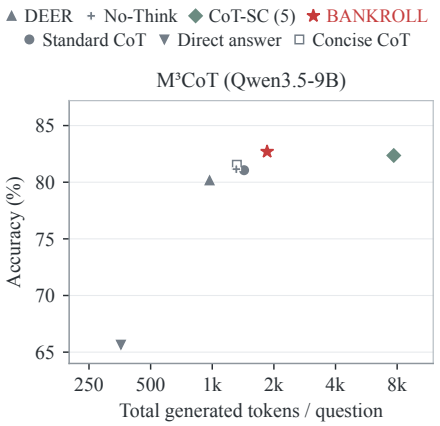


Figure 4: Accuracy–generation trade-off on M^{3}CoT (Qwen3.5-9B).

Relative to Standard CoT, the generation trade-off depends on the model. InternVL3.5-8B improves scores while reducing total generated tokens on all five benchmarks; on M^{3}CoT, accuracy increases from 68.33% to 73.34% as token use falls from 1444 to 1276. Qwen3.5-9B and GLM-4.1V-9B-Thinking instead use more tokens in nine of ten settings. BANKROLL's performance gains therefore occur with either lower or higher total generation than Standard CoT.

4.3 Core Ablations

The Qwen3.5-9B ablations in Table 2 show the accuracy benefit of Refinance's rollback and regeneration procedure. Removing Refinance saves tokens but lowers accuracy by 1.68 and 1.53 percentage points on M^{3}CoT and MMStar, respectively. Paired outcomes show that repair corrects some errors made by the configuration without Refinance while reversing few of its correct answers (Figure 5). In contrast, extending original and repair continuations by disabling Cashout retains probes but uses 25% and 47% more tokens, with accuracy rising on M^{3}CoT and falling on MMStar. Repair revisits the earlier judgment itself, complementing Cashout's control over further generation.

The remaining ablations connect this benefit to selective repair. Replacing reasoning-shift candidate nomination with random nomination while retaining masking and repair reduces accuracy. Uniform pivot sampling from the same $\mathcal{C}$ on the same repair chains, with truncation, continuation generation, and Writeoff unchanged, also lowers accuracy by 1.03 and 1.80 percentage points relative to maximum-dv selection. The corresponding T_{p} changes are below 4%. Masking additionally controls whether to regenerate: bypassing its tests roughly doubles token use without consistent accuracy gains. Removing Writeoff's score-based check for repair adoption while retaining other fallbacks,

Table 2: Core ablations of BANKROLL with Qwen3.5-9B.

Configuration	M^{3}CoT		MMStar	
	Score↑	T_p ↓	Score↑	T_p ↓
BANKROLL	82.70	1853	79.00	1796
w/o Cashout	82.92	2314	78.33	2633
w/o Cashout-Cert	81.75	1881	78.40	1872
w/o Refinance	81.02	1387	77.47	1408
Random candidate proposal	81.36	1951	77.00	1817
Random pivot selection	81.67	1861	77.20	1736
w/o masking tests	82.79	3542	78.60	3806
w/o acceptance check	82.53	1853	78.27	1796
w/o last-probe override	82.27	1853	78.13	1796

Notes. w/o: without. Results use the full evaluation sets.

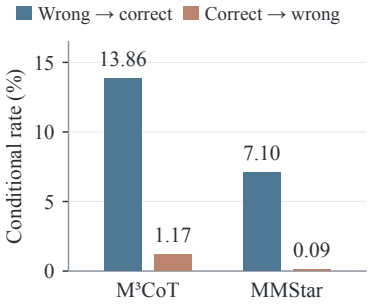


Figure 5: Correction and corruption rates of Refinance.

or the separate last-probe override for final-answer selection, lowers accuracy at unchanged token counts.

5 Related Work

5.1 Reasoning Attribution

ContextCite attributes generated statements via sparse linear surrogates of context ablations (Cohen-Wang et al., 2024). TokenSHAP and SalaMAnder apply Shapley attribution to prompt tokens and mathematical expressions in few-shot reasoning demonstrations, respectively (Horovicz & Goldshmidt, 2024; Xin et al., 2025). Thought Anchors measures sentence-to-sentence and answer influence via masking and counterfactual resampling (Bogdan et al., 2025). AttriCoT fits linear structural equations to log-probability changes, estimating dependencies among fixed generated units, including answers (Wei et al., 2026). Our offline analysis measures support in multimodal chains using Shapley-weighted marginal changes in fixed-target log-likelihood under length-preserving masking.

5.2 Inference Control

Early stopping uses DEER's intermediate-answer confidence (Yang et al., 2026a), Adaptive Think's answer entropy (Yong et al., 2025), uncertainty-keyword arrivals or conformal calibration (Xie et al., 2026), and sliding-window answer consistency in Dynasor's CoT probes (Fu et al., 2025). Thought dropout and policy optimization train multimodal models to reason selectively (Wang et al., 2025a).

Training-free Self-Refine revises outputs via self-feedback (Madaan et al., 2023); self-correction without external feedback can impair reasoning (Huang et al., 2024a). Earlier perturbations of initially correct CoTs make recovery less reliable (von Recum et al., 2026). Self-Backtracking trains models to revisit symbolic reasoning and planning states (Yang et al., 2026b); reward-guided early pruning selects first steps to continue (Liao et al., 2025). Attention–gradient scores guide StepFlow's information-flow interventions (Xu et al., 2026); CausalFlow repairs agent traces via counterfactual attribution and minimal edits (Bonagiri et al., 2026). OPERA combines attention-based logit penalties with token rollback and reselection in multimodal beam decoding (Huang et al., 2024b). BANKROLL combines early stopping and suffix regeneration on frozen multimodal models, using local masking to select rollback steps and answer-probe scores to select and accept continuations.

6 Conclusion

We analyze answer inheritance through Shapley-based interventional support between multimodal reasoning steps. An incorrect-answer example illustrates persistent support from an early judgment into the summary. Motivated by this observation, BANKROLL combines adaptive stopping with rollback to regenerate a selected judgment and its subsequent reasoning without training. BANKROLL improves task performance over Standard CoT across five benchmarks and three multimodal models.

Ethics Statement

Our experiments use publicly available datasets and involve no human participants or personal data. BANKROLL uses answer-probe scores from the same model that generates the answers to guide stopping and local repair. These scores do not verify answer correctness. Treating a high score or an accepted repair as verification may encourage overreliance on incorrect outputs. Applications in which errors affect people require validation of model outputs appropriate to the intended use.

Reproducibility Statement

Section 2 defines the offline attribution analysis. Section 3 and Algorithms 1 and 2 describe BANKROLL, including stopping, rollback, continuation selection, and final-answer handling. Section 4 and Appendix A describe the evaluated models, benchmarks, data splits, metrics, and baselines. Appendix B reports the shared hyperparameters, probe suffix, continuation sampling settings and seed rule, masking tokens, and token accounting. Appendix C reports hyperparameter sensitivity and distinguishes BANKROLL evaluations from Cashout replay on fixed generation traces. We plan to release our code after acceptance.

References

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought Anchors: Which LLM reasoning steps matter?, 2025. URL https://arxiv.org/abs/2506.19143.

Akash Bonagiri, Devang Borkar, Gerard Janno Anderias, Setareh Rafatirad, and Houman Homayoun. CausalFlow: Causal attribution and counterfactual repair for LLM agent failures, 2026. URL https://arxiv.org/abs/2605.25338.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Main Conference*, 2024b. arXiv:2405.16473.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2409.00729.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL https://www.jmlr.org/papers/v22/20-1316.html.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana Rosing, Ion Stoica, and Hao Zhang. Efficiently scaling LLM reasoning programs with Certaindex. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4751. URL https://papers.nips.cc/paper_files/paper/2025/hash/d037fd021c9aace128b8ce25001cdb6c-Abstract-Conference.html.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09751.

Miriam Horovicz and Roni Goldshmidt. TokenSHAP: Interpreting large language models with Monte Carlo Shapley value estimation. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 1–8, Miami, FL, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4science-1.1. URL https://aclanthology.org/2024.nlp4science-1.1/.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

J. L. Kelly, Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35(4): 917–926, 1956. doi: 10.1002/j.1538-7305.1956.tb03809.x.

Baohao Liao, Xinyi Chen, Sara Rajaee, Yuhui Xu, Christian Herold, Anders Søgaard, Maarten de Rijke, and Christof Monz. Lost at the beginning of reasoning, 2025. URL https://arxiv.org/abs/2506.22058.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling permutations for Shapley value estimation. *Journal of Machine Learning Research*, 23(43):1–46, 2022. URL https://jmlr.org/papers/v23/21-0439.html.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4), 2023. doi: 10.1214/23-STS894.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, number 28 in Annals of Mathematics Studies, pp. 307–317. Princeton University Press, 1953.

KV Aditya Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar. LLMs cannot spot math errors, even when allowed to peek into the solution. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://arxiv.org/abs/2509.01395.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.826.

Alexander von Recum, Leander Girrbach, and Zeynep Akata. Are reasoning LLMs robust to interventions on their chain-of-thought? In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.07470.

Jiaqi Wang, Kevin Qinghong Lin, James Cheng, and Mike Zheng Shou. Think or not? selective reasoning via reinforcement learning for vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-3712.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265v2.

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. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Dennis Wei, Yannis Belkhiter, Erik Miehling, and Radu Marinescu. Local causal attribution of chain-of-thought reasoning. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. arXiv:2606.21821.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yangxinyu Xie, Tao Wang, Soham Mallick, Yan Sun, Georgy Noarov, Mengxin Yu, Tanwi Mallick, and Edgar Dobriban. Statistical early stopping for reasoning models, 2026. URL https://arxiv.org/abs/2602.13935v3.

Yue Xin, Chen Shen, Shaotian Yan, Xiaosong Yuan, Yaoming Wang, Xiaofeng Zhang, Chenxi Huang, and Jieping Ye. SalaMAnder: Shapley-based mathematical expression attribution and metric for chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8558–8577, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.455.

Xiaoyu Xu, Yulan Pan, Xiaosong Yuan, Zhihong Shen, Minghao Su, Yuanhao Su, and Xiaofeng Zhang. Reasoning fails where step flow breaks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In *International Conference on Learning Representations*, 2026a.

Xiao-Wen Yang, Xuan-Yi Zhu, Ding-Chu Zhang, Wen-Da Wei, Jie-Jing Shao, Zhi Zhou, Lan-Zhe Guo, and Yu-Feng Li. Step back to leap forward: Self-Backtracking for symbolic reasoning and planning in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27657–27665, 2026b. doi: 10.1609/aaai.v40i33.39986. URL https://ojs.aaai.org/index.php/AAAI/article/view/39986.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, Lianwen Jin, and Junyang Lin. CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy, 2024. URL https://arxiv.org/abs/2412.02210.

Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. Think or not? Exploring thinking efficiency in large reasoning models via an information-theoretic lens. In *Advances in Neural Information Processing Systems*, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=y1pPWFVfvR.

APPENDIX A EXPERIMENTAL SETUP

Evaluation and scoring. We evaluate Qwen3.5-9B, GLM-4.1V-9B-Thinking, and InternVL3.5-8B on the splits in Table 3. Accuracy is averaged over individual questions; CC-OCR scores are averaged on the evaluator's 0–1 scale and multiplied by 100.

Table 3: Evaluation data and scoring.

Benchmark	Split	Questions	Scoring
M^3CoT	test	2,318	Exact option-letter match
MMStar	val	1,500	Exact option-letter match
MMMU-Pro	test	1,730	Exact option-letter match
HallusionBench	test	951	Exact Yes/No match
CC-OCR	test	7,058	Official score

Hyperparameter transfer. All rules and hyperparameters were determined during development using only Qwen3.5-9B. They were fixed before inference on M^3CoT test and transferred unchanged to the other four benchmarks and to GLM-4.1V-9B-Thinking and InternVL3.5-8B.

APPENDIX B IMPLEMENTATION DETAILS

Shared configuration. Qwen3.5-9B, GLM-4.1V-9B-Thinking, and InternVL3.5-8B use the same BANKROLL hyperparameters in Table 4, with no model-specific retuning for GLM or InternVL. Original chains use greedy decoding, with at most one repair round per chain.

Table 4: BANKROLL hyperparameters shared by all three models.

Parameter	Value	Parameter	Value
Pivot shortlist K	3	Continuations R	3
Masking threshold τ_{dv}	0.005	Steps before selection	3
Sampling temperature	1.0	Nucleus top-p	0.95
Acceptance tolerance δ	0.10	Override threshold v_{ovr}	0.5
Saturation score v_{sat}	0.99	Unchanged probes m_{sat}	5
Cashout unit return c	0.2	Betting scale λ	1.125
Capital threshold α	0.05	Pairwise gate δ_g	0.05

Probe suffix. The fixed probe suffix is `"\n</think>\n\nThe answer is "` in escaped notation, where `\n` denotes a newline and the suffix ends with one space. Each generation-time probe appends this suffix to a copy of the current key–value cache.

Saturation. Fixed-choice saturation requires five consecutive identical predictions, each scoring at least 0.99.

Runtime configuration. All three models run in bfloat16 on a single GPU and use seed $42 + \mathrm{CRC32}(q) + i$ for continuation i of question q. Online masking uses token ID 248046 for Qwen3.5-9B, token ID 82573 for GLM-4.1V-9B-Thinking, and token ID 83262 for InternVL3.5-8B.

Continuation and answer handling. Only the selected short continuation is completed. Writeoff retains its selection-time score, while final-answer override reads the last probe of the state retained by Writeoff. Fixed-choice chain answers use a directly read forced letter when available, followed by text parsing as a fallback.

Token accounting. T_{p} counts all tokens actually generated for the original chain, the selected continuation through completion, discarded short candidates, answers, and probes, including work later rejected by Writeoff. Counts are averaged per example. Prefilling and masking recomputation are excluded; offline Shapley analysis is accounted for separately.

Offline attribution configuration. For the mathematics-989 and geography-3180 chains in Figure 2, offline attribution uses `shap.PermutationExplainer` (SHAP 0.51.0) with seed 13. For each

target z_t, we set `max_evals` to $2(t-1)+1$ and average one uniformly sampled permutation with its reverse. This paired implementation uses $L_t = 2$ in Equation (5), with the two permutations coupled. Summing the per-target budgets over $t = 2, \ldots, n$ gives $n^2 - 1 = O(n^2)$ coalition-value evaluations for the full step-to-step attribution matrix, where n counts generated thinking and summary steps.

APPENDIX C HYPERPARAMETER SENSITIVITY

Table 5 examines BANKROLL's hyperparameter sensitivity on Qwen3.5-9B. Panel (a) evaluates final predictions under different Refinance, Writeoff, and final-answer settings; panel (b) replays Cashout on fixed full-generation traces without local revision. Here, R limits the number of candidate continuations, and s limits the reasoning steps generated per candidate before selection.

Table 5: Hyperparameter sensitivity on Qwen3.5-9B. Brackets give default values; bold rows report default results.

(a) Refinance, Writeoff, and final-answer settings

		M^{3}CoT		MMStar	
Parameter	Value	Acc.	T_{p}	Acc.	T_{p}
Default	—	**82.70**	**1853**	**79.00**	**1796**
Acceptance tolerance δ [0.10]	0.00	82.57	1853	78.33	1796
	0.05	82.40	1853	78.33	1796
	0.20	82.61	1853	78.27	1796
Candidate count and steps (R, s) [$(3, 3)$]	$(1, 3)$	80.24	1616	77.93	1612
	$(3, 1)$	81.79	1727	77.87	1699
	$(3, 5)$	81.88	1914	77.80	1841
	$(5, 3)$	82.18	2093	78.20	1977
	$(5, 5)$	82.27	2213	77.93	2072
Pivot shortlist K [3]	1	80.72	1666	77.60	1678
	2	81.62	1791	77.93	1748
Masking threshold τ_{dv} [0.005]	0.01	82.48	1768	78.13	1734
	0.02	82.31	1686	78.00	1644
Answer override v_{ovr} [0.5]	0.3	82.57	1853	78.40	1796
	0.7	82.53	1853	78.27	1796
Continuation sampling temperature [1.0]	0.6	81.62	1837	78.07	1767
	1.3	81.84	1862	78.07	1798

(b) Cashout replay on fixed traces

		M^{3}CoT			MMStar	
Parameter	Value	Stop (%)	Acc.	T_{th}	Acc.	T_{th}
Default	—	**29.8**	**80.11**	**1210.6**	**77.27**	**1361.7**
Saturation score v_{sat} [0.99]	0.97	36.8	79.94	1178.1	77.27	1307.2
	0.995	26.1	79.90	1224.7	77.27	1384.9
Unchanged probes m_{sat} [5]	3	57.0	79.94	1110.8	77.33	1221.3
	7	16.0	79.77	1253.9	77.33	1438.0

Note. Acc. is accuracy (%); Stop is stop coverage. Token counts are means per question: T_{p} includes all generated tokens, including probes and discarded continuations, and T_{th} counts thinking tokens. T_{p} is rounded to the nearest integer.

Refinance, Writeoff, and final answers. Using one continuation, a smaller pivot shortlist, or a higher masking threshold saves tokens at an accuracy cost. On both benchmarks, $(R, s) = (3, 3)$ achieves higher accuracy with fewer generated tokens than $(3, 5)$, $(5, 3)$, or $(5, 5)$. Acceptance and override sweeps leave token use unchanged, with accuracy drops from the default of at most 0.30 percentage points on M^{3}CoT and 0.73 on MMStar. Temperature changes yield larger maximum drops of 1.08 and 0.93 points, respectively. For the temperature sweep, continuations are resampled from stored rollback prefixes at top-$p = 0.95$; examples without repair retain their recorded outcomes.

Cashout saturation. Replay takes the probe argmax at the stopping point, or at the final probe if no rule fires, without final-answer decoding. Reducing m_{sat} from 5 to 3 saves 8.2% of thinking tokens

on M^{3}CoT with an accuracy loss of 0.17 percentage points. On MMStar, it saves 10.3% and answers one additional question correctly (about 0.07 percentage points).

Cashout-Cert. Defaults are $c = 0.2$, $\lambda = 1.125$, and $\alpha = \delta_g = 0.05$. We sweep $c \in \{0.10, 0.20, 0.30\}$ with coupled $\lambda = (c - 0.02)/[c(1 - c)]$, and separately sweep $\lambda \in \{0.5, 1.125\}$ and each of $\alpha, \delta_g \in \{0.01, 0.05, 0.10\}$. Both accuracies and M^{3}CoT stop coverage remain unchanged across these sweeps. Mean thinking-token counts stay near 1,211 on M^{3}CoT and 1,362 on MMStar.

APPENDIX D EARLY SUPPORT CAN PERSIST INTO THE SUMMARY

Figure 6 compares Shapley and leave-one-out (LOO) attribution offline on the chains from Figure 2. In geography-3180, T19's Shapley support band extends through later reasoning into the summary of a chain ending with an incorrect answer. In the correct mathematics-989 chain, T19's band fades before the summary. Both bands are less pronounced under LOO. Neither operator shows a pronounced long-range band in temporal-commonsense-18, which also ends incorrectly.

Both operators use the same fixed target text, model, image–question input, length-preserving masking, evaluated source–target pairs, and score v_t from Section 2. LOO masks source i while retaining all other predecessors:

$$d_{t,i}^{\mathrm{LOO}} = v_t(N_t) - v_t(N_t \setminus \{i\}), \qquad i \in N_t. \tag{D.1}$$

If those predecessors carry similar information, this full-context marginal can be weak. Shapley averages source i's marginal contributions across weighted predecessor coalitions, tracing support as the retained context varies.

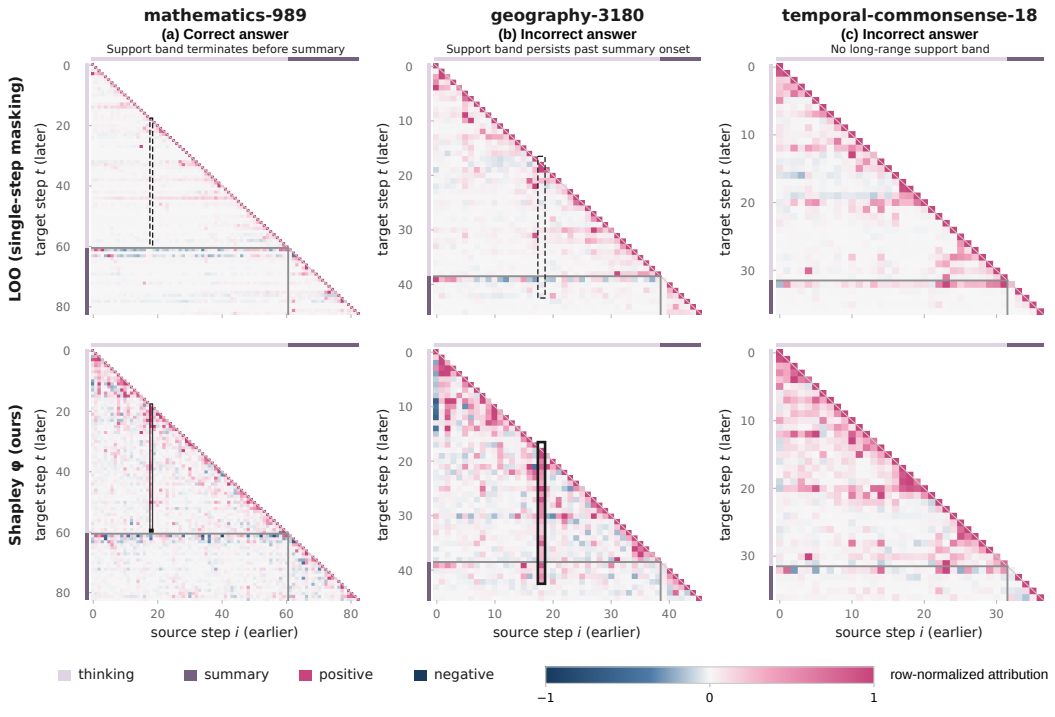


Figure 6: LOO (top) and Shapley (bottom) on three example chains. Rows are later targets and columns are earlier sources. Colors show support or inhibition, normalized independently within each row and operator. Absolute magnitudes cannot be compared across rows or operators. Gray lines mark summary onset. Solid outlines highlight Shapley support bands. Dashed outlines project the same regions onto LOO for reference.

Across all scored target rows containing T19, its median within-row rank is 5 with Shapley versus 26 with LOO in mathematics-989 (65 rows), and 2.5 versus 19.5 in geography-3180 (28 rows). These

ranks quantify the early sources' greater relative prominence under coalition averaging, while the bands trace how far their support persists along each chain. The geography case reveals answer inheritance: even after many intervening reasoning steps, the original summary can remain supported by an early judgment.

APPENDIX E FROM ATTRIBUTION TO LOCAL REPAIR

Persistent support for fixed later text motivates regenerating an earlier judgment together with the reasoning built on it (Section 2). Refinance uses local answer-probe sensitivity to select where regeneration begins (Section 3.2). We examine answer-level support on 120 Qwen3.5-9B M^{3}CoT test chains, randomly sampled within three outcome strata of 40 chains each: successful repairs, other repair outcomes, and untriggered chains, totaling 934 thinking steps. For each step, we compare its Shapley estimate $\widehat{\phi}_i^{\text{probe}}$ for the source answer probe (one pass, seed 13) with the full-context masking diagnostic

$$d_i^{\max} = \max_a p_{\text{full}}(a) - \max_a p_{\text{masked},i}(a), \tag{E.1}$$

where probabilities are normalized over answer letters and all other thinking remains fixed. This diagnostic uses each context's peak confidence; Refinance's dv_i scores the fixed unmasked baseline letter.

For a common value function, full-context masking gives a marginal contribution included in the Shapley average. The answer-probe diagnostics show frequent agreement in step ordering: among 109 original chains with at least three thinking steps, the median within-chain Spearman correlation is 0.531, with an interquartile range of 0.321–0.800 and 97 positive correlations.

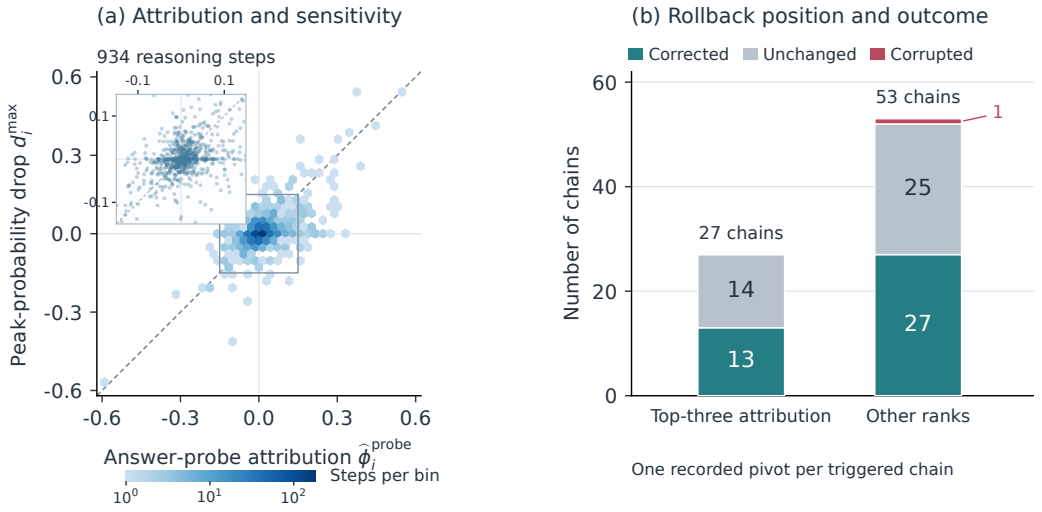


Figure 7: **Attribution, sensitivity, and repair outcomes.** (a) Counts of 934 steps in hexagonal bins (logarithmic color scale), with a central-region inset and an equality diagonal. (b) Final correctness transitions at 80 source-run rollback positions. Top-three refers to signed attribution rank within the original chain (competition ranks). Labels are outcome-stratified sample counts, not population repair rates; unchanged denotes unchanged correctness.

The recorded positions connect attribution ranks to observed repair outcomes (Figure 7b). Both rank groups contain corrections, while masking raises peak confidence at 71 of the 80 positions ($d_i^{\max} < 0$). Attribution and masking scores describe answer dependence; final correctness measures repair outcomes. We use these source-run positions for descriptive analysis, with the online selection rule evaluated by the controlled comparison below.

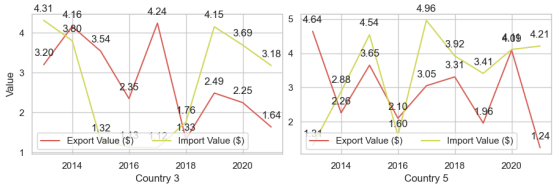
The random-pivot comparison keeps the same pivot candidate sets and repair procedure: up to three continuations per repair, each with at most three reasoning steps before selection. Only the continuation with the highest last-probe score is completed and passed to Writeoff. Under this shared protocol, maximum-dv selection exceeds uniform random pivot selection by 1.03 and 1.80 percentage

points on M^{3}CoT and MMStar (Section 4.3, Table 2). The complete repair procedure also corrects more answers than it corrupts relative to disabling Refinance (Figure 5), yielding a positive net benefit on both datasets.

APPENDIX F REOPENING AN INCORRECT VISUAL PREMISE

The rollback point determines which earlier judgments remain in the context for regeneration. Figure 8 illustrates this choice in Refinance (Section 3.2) on a chart question asking for the 2017 trade balances of Country 3 and Country 5. The correct option is B, $(3.12, -1.91)$.

(a) Visual question



What are the trade balances of Country 3 and Country 5 in 2017?

A (−1.91, 3.12) B (3.12, −1.91) ✓
C (−0.61, −1.91) D (−1.91, −0.61)

(b) Intervention outcomes

Intervention	Reasoning excerpts	Answer
Original	"Okay, so for 2017: Export = 4.24, Import = 1.76." "Trade Balance = 4.24 - 1.76 = +2.48."	**D** ×
Reflection prompt	PROMPT "Re-examine it carefully, correct it if needed, and then continue." "2017: Red=4.24, Yellow=1.76" "So Trade Balance = 4.24 - 1.76 = 2.48."	**C** ×
Resample ×5 + vote	B D D D D → majority D	**D** ×
BANKROLL (ours)	~~"Okay, so for 2017: Export = 4.24, Import = 1.76."~~ "Point 5 values: Export 4.24, Import 1.12." "So, Trade Balance for Country 3 in 2017 = 4.24 - 1.12 = 3.12."	**B** ✓

Figure 8: An incorrect chart reading before and after intervention. Selected excerpts show intermediate reasoning about Country 3. The answer column reports the recorded final outcomes. Strikethrough marks an excerpt from the discarded step. Refinance removes that entire step and all subsequent reasoning. Blue text shows the regenerated reading and calculation.

Reflection retains the premise. The original chain reads Country 3's import value as 1.76 and computes $4.24 - 1.76 = 2.48$. The arithmetic is consistent with the stated values, but the import reading is wrong. The original run ultimately answers D. A reflection prompt inserted after this reading asks the model to re-examine and correct its reasoning. The continuation again uses 1.76 and obtains 2.48. Its final answer changes to C, but the earlier visual judgment remains in both the context and the new calculation.

Voting selects among final answers. Five independently sampled full chains, none using the original reasoning prefix, yield $[\mathrm{B}, \mathrm{D}, \mathrm{D}, \mathrm{D}, \mathrm{D}]$. The correct answer B is generated, but majority voting returns D. Voting aggregates the final choices without revisiting the chart readings that produced them.

Rollback reopens the reading. Refinance selects the Country 3 analysis as its rollback point. It retains the input and preceding reasoning, then removes the selected step and its entire suffix. Generation resumes before the decision about Country 3's 2017 import value, allowing that value to be read again. The selected continuation reads the import value as 1.12 and the export value as 4.24, then computes $4.24 - 1.12 = 3.12$. The repaired run ultimately answers B. In this case, the repaired visual premise enters a newly generated calculation: the step containing the misreading and the reasoning built on it are revised together.

APPENDIX G AI USE STATEMENT

We used generative AI tools to assist with method implementation and translation. We also used these tools to summarize and analyze existing literature, explore research topics and identify research gaps, search for information, format references, and assist with manuscript typesetting and layout.

We did not use generative AI tools to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims; provide key ingredients for mathematical proofs or assist in writing proofs; propose or refine research hypotheses; design or provide feedback on research methodology or experiments; or clean or reformat datasets.

The authors take responsibility for the final content of this paper and its associated artifacts, including text and code produced with the assistance of generative AI.