EVIDENCEFLOW: REPAYING VISUAL EVIDENCE DEBT FOR TEST-TIME MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time scaling can improve multimodal reasoning without updating model parameters, but longer reasoning may not provide visual evidence required by later steps. Existing visual refocusing methods replay image tokens based on relevance to the current reasoning state, without examining whether realized visual support matches that state. We study the resulting mismatch between the current visual-demand distribution and the realized visual-support distribution. We define this mismatch as *Visual Evidence Debt*, quantified as one minus the overlapping probability mass of the two distributions. Its equivalent positive-residual form localizes visual tokens whose demand share exceeds their realized-support share. We therefore propose EVIDENCEFLOW, a training-free framework that performs marginal debt repayment under a fixed replay budget via a saturating monotone submodular objective. Experiments and controlled analyses across multiple MLLMs and benchmarks show improved reasoning accuracy over standard thinking and a more favorable accuracy–compute trade-off than the evaluated replay baselines. Our code is available in the supplementary material.

1 INTRODUCTION

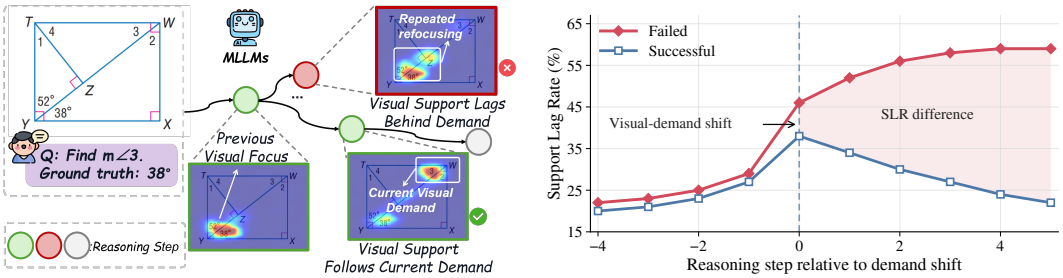
Test-time scaling extends inference-time computation to improve the reasoning of large language models without updating their parameters (Snell et al., 2024; Muennighoff et al., 2025). For multimodal reasoning, however, generating more text does not provide new visual evidence. Image tokens remain near the beginning of the context while the model-generated reasoning history continues to grow. This effect becomes more pronounced as reasoning proceeds. Consequently, later reasoning steps may rely more on previous textual conclusions and pay less attention to the image evidence required by the current step (Chu et al., 2025; Yang et al., 2025a; Ghosal et al., 2026).

Recent studies attempt to maintain visual grounding by allowing MLLMs to revisit visual information during reasoning. Qwen-LookAgain introduces visual-token copying and routing during vision–text reflection (Chu et al., 2025), while Look-Back trains MLLMs to refocus on visual information during later reasoning steps (Yang et al., 2025a). VisRef (Ghosal et al., 2026) instead provides a training-free selector based on relevance and diversity. These results establish visual revisitation as an effective complement to text-only test-time scaling.

However, existing refocusing methods primarily select visual tokens according to text–visual relevance to the current reasoning state. Relevance identifies visual evidence related to the current reasoning state, but does not determine which currently required evidence remains under-supported after accounting for visual support already realized along the reasoning trajectory. Such cases can arise naturally during multi-step reasoning. As shown in Figure 1a, visual support may remain concentrated on salient or previously revisited regions after the reasoning focus shifts to new objects, relations, or spatial details. The model may therefore reason about one region while receiving stronger visual support from another, a phenomenon we refer to as *"Saying is not seeing."*

To diagnose this behavior, we introduce the *Support Lag Rate* (SLR), which measures how often a visually correct statement is generated while the strongest visual support remains outside the corresponding evidence region. The diagnostic protocol and formal definition are provided in Section 2. We collect trajectories generated by Qwen3-VL-8B-Thinking on 400 questions sampled evenly from MathVista and RealWorldQA (Lu et al., 2024; xAI, 2024). The trajectories are aligned at detected

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.



(a) Demand–support mismatch.

(b) SLR around visual-demand shifts.

Figure 1: Demand–support mismatch during multimodal reasoning. (a) After a reasoning-focus shift, relevance-based refocusing may remain concentrated on a previously attended region, causing support to lag behind the current demand. (b) Trajectories are aligned at detected visual-demand shifts ($t = 0$). A higher post-shift SLR is observed for failed trajectories, indicating that their strongest support more frequently remains outside the required evidence region.

visual-demand shifts, and SLR is computed at each relative reasoning round. As shown in Figure 1b, failed trajectories exhibit a higher post-shift SLR than successful trajectories, revealing an observational association between persistent support lag and reasoning failure.

Motivated by this observation, we propose EVIDENCEFLOW, a training-free framework that treats visual replay as budgeted debt repayment. At each reasoning boundary, EVIDENCEFLOW estimates the current visual demand and the realized visual support. It defines the positive residual between them as *Visual Evidence Debt* and selects visual tokens according to their marginal debt repayment under a fixed replay budget. The selection objective is monotone submodular, and fixed-budget greedy selection has a theoretical approximation guarantee. Across multiple MLLMs and benchmarks, improved reasoning accuracy is obtained over standard thinking, together with a more favorable accuracy–compute trade-off than the evaluated replay baselines.

Our main contributions are summarized as follows:

- We formulate visual grounding during multimodal reasoning as a demand–support accounting problem, distinguishing current visual demand from history-aware realized support. We introduce SLR to diagnose support lag and Visual Evidence Debt to quantify and localize the remaining mismatch.
- We propose EVIDENCEFLOW, a training-free framework that selects and replays visual tokens through marginal debt repayment under a fixed visual-token budget. We formulate token selection as a monotone submodular optimization problem with a theoretical approximation guarantee for fixed-budget greedy selection.
- We validate the demand–support formulation through demand localization, support masking, and selector comparisons under a shared prescribed replay plan, and demonstrate improved accuracy across MLLMs and benchmarks.

2 PRELIMINARIES

Test-time multimodal reasoning and visual replay. Given an image I and a question q, a frozen MLLM with parameters θ encodes the image into N visual tokens $\mathcal{V} = \{v_i\}_{i=1}^{N}$ and generates reasoning rounds $z_1, \ldots, z_T$, followed by an answer. A reasoning round denotes the text generated between two candidate replay boundaries. Let $\mathcal{H}_t$ denote the interleaved prefix through z_t, including the original input, generated reasoning, and previously replayed visual tokens.

At the boundary following z_t, a replay policy may select $S_t \subseteq \{1, \ldots, N\}$ and insert $\mathcal{V}_{S_t}$ before generation is continued. Given a maximum replay ratio $\rho \in [0, 1]$, the per-replay budget is $B = \lfloor \rho N \rfloor$, with $|S_t| \leq B$. If replay is triggered, the next reasoning round is generated as

$$z_{t+1} \sim p_\theta(\cdot \mid \mathcal{H}_t, \mathcal{V}_{S_t}), \tag{1}$$

whereas no additional visual tokens are inserted otherwise. Source-token identities are retained across replay occurrences to support history-aware aggregation.

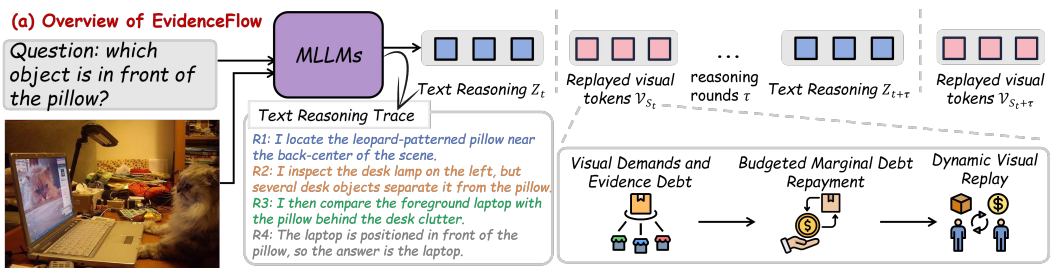


Figure 2: Overview of EVIDENCEFLOW. At each reasoning boundary, active visual demand and history-aware realized support are estimated, visual evidence debt is computed, and tokens with the largest marginal debt repayment are replayed under a fixed budget.

Diagnosing reasoning–support lag. For diagnostic evaluation, reasoning rounds are divided into clauses following Section 3.1. An external VLM judge is used to identify image-dependent claims, assess their consistency with the image, and localize the required evidence regions. Clauses without localizable evidence are excluded. The complete protocol is provided in Appendix B.3.

For dataset d, let $\mathcal{P}_d$ denote the retained region-groundable clauses. For each $(t, j) \in \mathcal{P}_d$, let $f_{t,j} \in \{0, 1\}$ indicate whether the visual claim is consistent with the image, and let $\mathcal{R}^*_{t,j} \subseteq \{1, \ldots, N\}$ contain the source visual tokens overlapping its evidence regions. Given the realized-support distribution $P_{t,j}$ defined in Equation 8, support localization is measured by

$$g_{t,j} = \mathbf{1}\left[\arg\max_i P_{t,ji} \in \mathcal{R}^*_{t,j}\right]. \tag{2}$$

For question q in dataset d, let $N^{d,q}_{a,b}$ denote the number of clauses satisfying $f_{t,j} = a$ and $g_{t,j} = b$. SLR is first computed within each question and then averaged across valid questions:

$$\mathrm{SLR}_d = \frac{1}{|\mathcal{Q}_d|} \sum_{q \in \mathcal{Q}_d} \frac{N^{d,q}_{1,0}}{N^{d,q}_{1,1} + N^{d,q}_{1,0}}. \tag{3}$$

SLR measures the fraction of visually correct claims whose strongest visual support lies outside the required evidence region. A lower SLR indicates stronger reasoning–support alignment. SLR is therefore used as a conditional support-localization diagnostic, rather than as evidence that support alignment causes visual correctness.

3 METHOD

Under the setup in Section 2, EVIDENCEFLOW operates at the boundary between consecutive reasoning rounds. After generating z_t, it (i) extracts the visual demands active in the current reasoning state, (ii) estimates the realized visual support and its mismatch with each demand, and (iii) replays the visual tokens with the largest marginal debt repayment before generating z_{t+1}. Figure 2 summarizes this process. All signals come from the frozen model's current forward pass; no additional training, external verifier, or parameter update is required.

3.1 ACTIVE VISUAL DEMANDS AND VISUAL EVIDENCE DEBT

A single reasoning round may contain several visual requirements. The mismatch is therefore modeled at the clause level by separately estimating the visual evidence required by each clause, its realized visual support, and the mismatch between them. We refer to these quantities as active visual demand, realized visual support, and visual evidence debt, respectively.

Active visual demands. We split the newly generated reasoning round z_t into clauses using punctuation and conjunction boundaries. Let $\mathcal{K}_t = \{1, \ldots, M_t\}$ index the resulting clauses and let $\mathcal{W}_{t,j}$ denote the textual token positions associated with clause j. At an intermediate decoder layer ℓ, we represent its active visual demand by

$$r_{t,j} = \frac{1}{|\mathcal{W}_{t,j}|} \sum_{u \in \mathcal{W}_{t,j}} h^{(\ell)}_u, \tag{4}$$

where $h_u^{(\ell)}$ is the hidden state at textual position u. The nonnegative compatibility between demand j and visual token i is

$$R_{t,ji} = \left[\cos\left(r_{t,j}, v_i^{(\ell)}\right)\right]_+, \tag{5}$$

where $[x]_+ = \max(x, 0)$. Unlike a single query–image relevance score, $\{R_{t,ji}\}_{j=1}^{M_t}$ preserves the multiple visual requirements that can coexist within one reasoning round. We normalize these compatibilities into a visual-demand distribution

$$Q_{t,ji} = \frac{R_{t,ji} + \epsilon/N}{\sum_{k=1}^{N} R_{t,jk} + \epsilon}, \qquad \sum_{i=1}^{N} Q_{t,ji} = 1. \tag{6}$$

The additive smoothing keeps the distribution well-defined when all compatibility scores are zero and preserves normalization, yielding the distribution vector $Q_{t,j} = (Q_{t,j1}, \ldots, Q_{t,jN})$.

Realized visual support. Relevance indicates which evidence is semantically related to a demand, but not how visual contribution is distributed across source tokens in the current forward pass. Let $\mathcal{O}_t(i)$ be the set of occurrences of visual token i in $\mathcal{H}_t$, including its original occurrence and any previous replay. We aggregate the normalized attention from clause j to these occurrences as

$$A_{t,ji} = \frac{1}{|\mathcal{W}_{t,j}|} \sum_{u \in \mathcal{W}_{t,j}} \sum_{o \in \mathcal{O}_t(i)} \widetilde{a}_{uo}^{(\ell)}. \tag{7}$$

Attention mass alone can overestimate tokens whose value vectors contribute little to the residual stream. We therefore modulate it by normalized value magnitude,

$$P_{t,ji} = \frac{A_{t,ji}\,\nu_i + \epsilon/N}{\sum_{k=1}^{N} A_{t,jk}\,\nu_k + \epsilon}, \qquad \nu_i = \frac{\|W_V v_i^{(\ell)}\|_2}{\max_k \|W_V v_k^{(\ell)}\|_2 + \epsilon}. \tag{8}$$

Replay occurrences are mapped to their source-token identities before aggregation, making $P_{t,ji}$ history-aware. Additive smoothing ensures $\sum_i P_{t,ji} = 1$, and the resulting distribution is denoted by $P_{t,j} = (P_{t,j1}, \ldots, P_{t,jN})$.

Definition 1 (Visual Evidence Debt)**.** For reasoning round t and active demand $j \in \mathcal{K}_t$, define its visual evidence debt as

$$D_{t,j} = 1 - \sum_{i=1}^{N} \min(Q_{t,ji}, P_{t,ji})\,. \tag{9}$$

Here, $D_{t,j} \in [0, 1]$ is one minus the probability mass shared by the current visual-demand and realized visual-support distributions. For comparisons across reasoning rounds with different numbers of clauses, we use the mean debt

$$\overline{D}_t = \frac{1}{|\mathcal{K}_t|} \sum_{j \in \mathcal{K}_t} D_{t,j}. \tag{10}$$

Because $Q_{t,j}$ and $P_{t,j}$ are normalized distributions, visual evidence debt admits the equivalent token-level residual and total-variation forms

$$E_{t,ji} = [Q_{t,ji} - P_{t,ji}]_+\,, \qquad D_{t,j} = \sum_{i=1}^{N} E_{t,ji} = \frac{1}{2}\,\|Q_{t,j} - P_{t,j}\|_1\,. \tag{11}$$

The residual $E_{t,ji}$ identifies source tokens for which the demand share exceeds the realized-support share. The first equality follows from $a - \min(a, b) = [a - b]_+$ and $\sum_i Q_{t,ji} = 1$; the proof is provided in Appendix D. Accordingly, Visual Evidence Debt measures demand–support allocation mismatch rather than the absolute magnitude of visual reliance. Since $Q_{t,j}$ and $P_{t,j}$ are normalized but derived from different model-internal signals, their difference is interpreted as an operational mismatch under the fixed estimators in Equations 6 and 8, rather than an absolutely calibrated amount of missing visual evidence. A large debt therefore indicates that much of the current visual demand remains unmatched by realized support.

3.2 Budgeted Marginal Debt Repayment

The token-level residual $E_{t,ji}$ identifies visual tokens whose demand share exceeds their realized-support share. We use the residual as an eligibility signal and the corresponding demand mass $Q_{t,ji}$ to measure the potential value of revisiting an under-supported token.

For a replay set $S \subseteq \{1, \ldots, N\}$, we define the normalized debt-repayment surrogate:

$$F_t(S) = \frac{1}{|\mathcal{K}_t|} \sum_{j \in \mathcal{K}_t} \min\left\{ D_{t,j}, \sum_{i \in S} Q_{t,ji}\, \mathbf{1}[E_{t,ji} > 0] \right\}, \tag{12}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Positive residuals determine eligibility, while $Q_{t,ji}$ weights the potential value of revisiting an under-supported token according to its current demand. Thus, $F_t(S)$ is a debt-repayment surrogate rather than a literal estimate of residual reduction. Using $E_{t,ji}$ directly would make the cap redundant because $\sum_i E_{t,ji} = D_{t,j}$, reducing the objective to residual ranking. The cap limits repeated coverage of the same clause, while averaging keeps $F_t(S) \in [0, 1]$ and thresholds comparable across rounds.

The marginal repayment obtained by adding token i to a partial selection S is

$$\Delta_t(i \mid S) = F_t(S \cup \{i\}) - F_t(S). \tag{13}$$

Starting from $S_t^{(0)} = \varnothing$, we repeatedly select the token with the largest marginal repayment:

$$i_b^\star = \arg\max_{i \notin S_t^{(b-1)}} \Delta_t(i \mid S_t^{(b-1)}), \qquad S_t^{(b)} = S_t^{(b-1)} \cup \{i_b^\star\}. \tag{14}$$

After each selection, the marginal repayments are recomputed to account for the debt already covered by the selected tokens.

Proposition 1 (Monotone Submodularity of Debt Repayment)**.** *The objective $F_t(S)$ in Equation 12 is monotone submodular. Under the cardinality constraint $|S| \le B$, greedy selection run to the full budget satisfies*

$$F_t(S_t^{\text{greedy}}) \ge (1 - 1/e) \max_{|S| \le B} F_t(S). \tag{15}$$

The proof follows the classical cardinality-constrained submodular result of Nemhauser et al. (1978) and is provided in Appendix E.

In practice, selection stops when $|S_t| = B$ or the largest remaining marginal repayment falls below λ_{tok}. The approximation guarantee in Proposition 1 applies to greedy selection run to the full budget B and not to this early-stopping variant.

3.3 Dynamic Visual Replay

Token selection specifies *what* is replayed, while the final component determines *whether* replay is required at the current reasoning boundary. Let S_t denote the adaptively selected set, and let

$$\Gamma_t = F_t(S_t) = \sum_{b=1}^{|S_t|} \Delta_t(i_b^\star \mid S_t^{(b-1)}) \tag{16}$$

denote its cumulative repayment, where $\Gamma_t \in [0, 1]$ by construction. Replay is triggered only if

$$\Gamma_t \ge \delta_{\text{rep}}, \tag{17}$$

where the round-level threshold δ_{rep} is used to suppress interventions with insufficient total repayment. This decision is separated from token-level early stopping: whether another token is added to S_t is controlled by λ_{tok}, whereas replay of the selected set is determined by δ_{rep}. If no token reaches λ_{tok}, $S_t = \varnothing$ and replay is not triggered.

When the condition in Equation 17 is satisfied, the selected tokens are restored to their spatial order, and generation is continued as in Equation 1. Otherwise, no visual tokens are inserted. At the next reasoning boundary, the support distribution is recomputed from $\mathcal{H}_{t+1}$, thereby incorporating replay occurrences into the history-aware estimate. Token priority is reduced after demand–support alignment is restored, but the token may be replayed if it is required by a later demand.

Table 1: Accuracy (%) across three MLLMs and five benchmarks. † marks results reported in VISREF (Ghosal et al., 2026). Best and second-best results are marked in dark red and light red. Δ denotes the average gain over Standard Thinking on benchmarks with reported results.

Model	Method	MathVista	MathVision	MM-Star	RealWorldQA	Geo3K	Δ
Qwen3-VL-8B Thinking	Standard Thinking	73.8	53.3	66.1	55.2	51.6	–
	Textual Self-Reflection	74.4	53.9	66.4	56.2	52.4	+0.7
	VISREF†	77.1	56.6	69.1	59.1	–	+3.4
	EVIDENCEFLOW	**79.0**	**60.5**	**71.2**	**62.4**	**60.1**	**+6.6**
InternVL3.5-8B	Standard Thinking	68.5	39.1	57.5	44.3	50.9	–
	Textual Self-Reflection	73.5	40.5	58.0	45.2	51.6	+1.7
	VISREF†	79.3	44.6	63.1	47.2	–	+6.2
	EVIDENCEFLOW	**81.1**	**48.7**	**65.4**	**50.1**	**57.1**	**+8.5**
SAIL-VL2-8B	Standard Thinking	72.8	30.3	47.4	56.9	48.1	–
	Textual Self-Reflection	74.0	31.6	49.2	58.4	48.9	+1.3
	VISREF†	78.2	37.3	55.3	61.2	–	+6.2
	EVIDENCEFLOW	**79.6**	**42.1**	**58.0**	**63.4**	**54.1**	**+8.3**

4 EXPERIMENTS

Our experiments test the central hypothesis of this work: *multimodal reasoning can fail when realized visual support does not match the current visual demand.* We ask whether EVIDENCEFLOW improves multimodal reasoning across models and benchmarks (RQ1), improves reasoning–support alignment (RQ2), and achieves a better accuracy–visual-compute trade-off through debt-aware replay (RQ3). These RQs cover effectiveness, alignment, and efficiency.

4.1 EXPERIMENTAL SETUP

Models and benchmarks. We use Qwen3-VL-8B-Thinking (Bai et al., 2025) as the primary backbone and evaluate transfer to InternVL3.5-8B (Wang et al., 2025) and SAIL-VL2-8B (Yin et al., 2025). Our evaluation covers MathVista (testmini) (Lu et al., 2024), MathVision (Wang et al., 2024), MM-Star (Chen et al., 2024), RealWorldQA (xAI, 2024), and Geometry3K (Geo3K) (Lu et al., 2021). All models remain frozen and are evaluated in their native thinking mode. Further details are provided in Appendix B.

Baselines. We compare EVIDENCEFLOW with Standard Thinking, Textual Self-Reflection, Random Replay, Relevance-only Replay, VISREF, and Full-image Replay. Following VISREF (Ghosal et al., 2026), Standard Thinking uses native reasoning without intervention, whereas Textual Self-Reflection uses the continuation cue introduced by s1 (Muennighoff et al., 2025). As no public implementation of VISREF is available, its published results are reported as external references; Appendix B.7 compares its token selector under the same prescribed replay boundaries and per-event token counts.

Metrics. We primarily report task accuracy, replayed visual tokens, trigger rate, and end-to-end latency. Mechanistic analyses additionally report VSA, SLR, UPR, RSC, and PRD. Metric definitions are provided in Appendices B.2 and B.3.

4.2 DOES EVIDENCEFLOW IMPROVE MULTIMODAL REASONING?

Table 1 evaluates RQ1 across three MLLMs and five benchmarks. Replay policies are compared under shared maximum per-replay visual-token budgets in Sections 4.4 and 5.1.

Finding 1. EVIDENCEFLOW achieves the highest accuracy across all three models and five benchmarks. Across these settings, its average gains over Standard Thinking and Textual Self-Reflection are 7.8 and 6.5 percentage points, respectively. The improvement is particularly pronounced on MathVision, where EVIDENCEFLOW outperforms Standard Thinking by up to 11.9 percentage points. A gain of up to 8.5 percentage points is also observed on Geo3K.

Table 2: Replay efficiency on Qwen3-VL-8B-Thinking at $\rho = 30\%$. Accuracy is macro-averaged over five benchmarks. The realized replay ratio is averaged over triggered events, and the trigger rate over eligible reasoning boundaries. Matched selector controls share the same fixed replay schedule and per-event budget.

Method	Accuracy (%) ↑	Realized replay ratio (%) ↓	Trigger rate (%) ↓	Latency (s) ↓
Standard Thinking	60.0	–	–	**8.1**
Textual Self-Reflection	60.7	–	–	9.4
Random Replay	61.0	30.0	30.0	9.1
Relevance-only Replay	62.8	30.0	29.4	9.1
Full-image Replay	65.5	100.0	28.0	14.8
Fixed-schedule selector controls				
Matched VISREF Selector	63.6	30.0	29.6	9.2
Matched Gated Demand	64.0	30.0	29.7	9.2
EVIDENCEFLOW (fixed)	65.1	30.0	30.0	9.3
EVIDENCEFLOW (adaptive)	**66.6**	**18.4**	**18.6**	8.8

4.3 DOES EVIDENCEFLOW IMPROVE REASONING–SUPPORT ALIGNMENT?

RQ2 evaluates whether EVIDENCEFLOW improves the alignment between the model's current reasoning and its realized visual support. Following the diagnostic protocol in Section 2, this step-level analysis is conducted using Qwen3-VL-8B-Thinking, with Standard Thinking, Relevance-only Replay, and EVIDENCEFLOW compared under the same setup.

Evaluation protocol. The analysis is conducted on 200 questions randomly sampled from each of MathVista, MathVision, and RealWorldQA. All methods are evaluated on the same 600 questions. We use `gpt-5.2-2025-12-11` as the external VLM judge. VSA and SLR are first computed within each question and then averaged across questions, so that each question contributes equally. The retained region-groundable clause counts are reported in Appendix B.3.

Alignment metrics. For each method, the realized-support distribution $P_{t,j}$ is computed using Equation 8. For a region-groundable clause (t, j), we measure visual support assigned to its evidence regions as $s_{t,j} = \sum_{i \in \mathcal{R}^*_{t,j}} P_{t,ji}$. Let $\mathcal{P}_{d,q}$ denote the retained clauses of question q. Visual Support Alignment (VSA) on dataset d is then defined as

$$\mathrm{VSA}_d = \frac{1}{|\mathcal{Q}_d|} \sum_{q \in \mathcal{Q}_d} \frac{1}{|\mathcal{P}_{d,q}|} \sum_{(t,j) \in \mathcal{P}_{d,q}} s_{t,j}. \tag{18}$$

A larger VSA indicates that more realized visual support is assigned to the regions required by the current reasoning clause. SLR is computed using Equation 3. The results for both metrics are summarized in Table 3.

Additional diagnostic metrics. Appendix B.3 reports VSA and SLR separately for MathVista, MathVision, and RealWorldQA. It additionally reports the Unfaithful Perception Rate, Reasoning–Support Consistency, and Perception–Reasoning Dissociation Rate, together with a human validation of the external VLM judge.

Table 3: Reasoning–support alignment for Qwen3-VL-8B-Thinking. VSA and SLR are question-averaged and macro-averaged across MathVista, MathVision, and RealWorldQA.

Method	**Avg. VSA ↑**	**Avg. SLR ↓**
Standard Thinking	0.454	30.8%
Relevance-only Replay	0.496	25.7%
EVIDENCEFLOW	**0.573**	**15.4%**

Finding 2. As shown in Table 3, EVIDENCEFLOW achieves the highest average VSA and the lowest average SLR across the three benchmarks. Compared with Relevance-only Replay, it increases average VSA from 0.496 to 0.573 and reduces average SLR from 25.7% to 15.4%. These results show that debt-aware replay better aligns realized visual support with the evidence required by the current reasoning step.

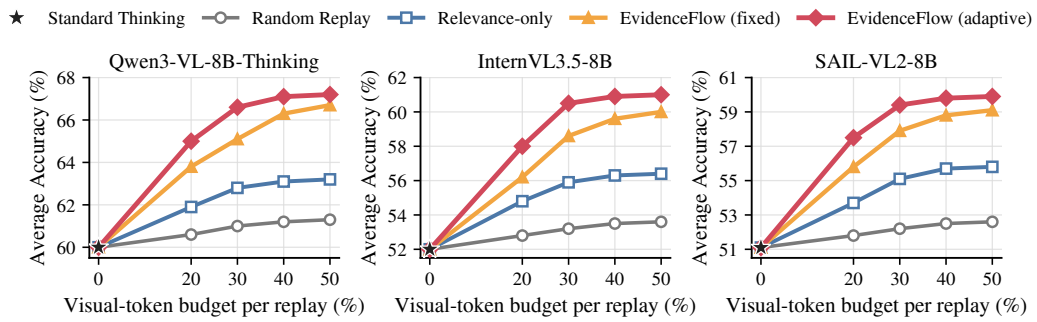


Figure 3: Accuracy under different maximum per-replay visual-token budgets, macro-averaged over MathVista, MathVision, MM-Star, RealWorldQA, and Geo3K. Standard Thinking is shown at $\rho = 0$, and model-specific vertical ranges are used for readability.

Table 4: Component ablation on Qwen3-VL-8B-Thinking at $\rho = 30\%$, with accuracy macro-averaged over five benchmarks, replay ratio over triggered events, and trigger rate over eligible reasoning boundaries.

Variant	Accuracy (%) ↑	Realized replay ratio (%)	Trigger rate (%) ↓
Full EVIDENCEFLOW	**66.6**	18.4	**18.6**
w/o replay-history tracking	63.8	27.1	27.6
w/o saturation	64.4	24.8	23.9
w/o repayment trigger	65.8	16.6	30.0

4.4 DOES DEBT-AWARE REPLAY USE VISUAL COMPUTE MORE EFFICIENTLY?

RQ3 examines whether debt-aware replay uses visual computation more efficiently. The per-replay visual-token budget ρ is varied over $\{20\%, 30\%, 40\%, 50\%\}$ while the textual decoding budget is fixed. For an input represented by $|\mathcal{V}|$ visual tokens, at most $\lfloor \rho|\mathcal{V}| \rfloor$ tokens can be selected at each replay. Random Replay, Relevance-only Replay, and fixed-budget EVIDENCEFLOW use the specified budget at each triggered event. Adaptive EVIDENCEFLOW may select fewer tokens through marginal-gain early stopping and suppress low-repayment events through debt-aware triggering.

Figure 3 reports accuracy macro-averaged over the five benchmarks. The results at $\rho = 30\%$ correspond to those reported for EVIDENCEFLOW in Table 1. Table 2 further reports realized visual computation and latency on Qwen3-VL-8B-Thinking at $\rho = 30\%$. Full-image Replay is included as a compute-unmatched reference.

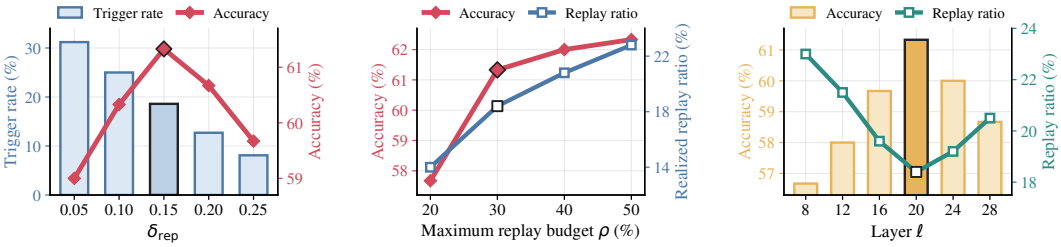
Finding 3. Debt-aware replay yields a more favorable accuracy–compute trade-off than uniformly allocated replay. At $\rho = 30\%$, adaptive EVIDENCEFLOW achieves average accuracies of 66.6%, 60.5%, and 59.4% on Qwen3-VL-8B-Thinking, InternVL3.5-8B, and SAIL-VL2-8B, respectively. On Qwen3-VL-8B-Thinking, the mean replay ratio per triggered event is reduced from 30.0% to 18.4%, a 38.7% reduction, while end-to-end latency is reduced from 9.3 to 8.8 seconds relative to fixed-budget EVIDENCEFLOW.

5 UNDERSTANDING EVIDENCEFLOW

5.1 COMPONENT ABLATION

Three components are ablated on Qwen3-VL-8B-Thinking at $\rho = 30\%$: replay-history tracking, saturation, and the repayment trigger.

Finding 4. Replay-history tracking yields the largest accuracy gain while reducing replay size and frequency, whereas saturation limits redundant demand coverage. Removing the repayment trigger slightly lowers the per-event replay ratio but raises the trigger rate from 18.6% to 30.0%, confirming that debt-aware activation suppresses unnecessary interventions. These components jointly improve token selection and compute allocation.



(a) Round-level replay threshold. (b) Maximum replay budget. (c) Representation layer.

Figure 4: Sensitivity analysis on the Geo3K validation split using Qwen3-VL-8B-Thinking. Accuracy and realized visual computation are reported for the repayment threshold δ_{rep}, replay budget ρ, and representation layer ℓ.

5.2 Sensitivity Analysis

All hyperparameters are selected using Qwen3-VL-8B-Thinking on the Geo3K validation split; the five reported test sets are used only for final evaluation, and the selected settings are frozen across all backbones. Sensitivity is evaluated for the round-level repayment threshold δ_{rep}, the maximum per-replay visual-token budget ρ, and the representation layer ℓ, with one factor varied at a time. The budget is fixed at $\rho = 30\%$ except in the budget sweep. The realized replay ratio is also reported to characterize the accuracy–compute trade-off.

Finding 5. Results identify a stable operating region. A small δ_{rep} triggers replay excessively, whereas a large value suppresses intervention. Setting $\delta_{\text{rep}} = 0.15$ achieves the highest validation accuracy of 61.33% with an 18.6% trigger rate. Increasing ρ from 20% to 30% improves accuracy from 57.67% to 61.33%, while larger budgets yield marginal gains and increase the realized replay ratio. The default is therefore set to $\rho = 30\%$, at which the ratio is 18.4%. Layer $\ell = 20$ achieves the highest validation accuracy and the lowest replay ratio among evaluated representations.

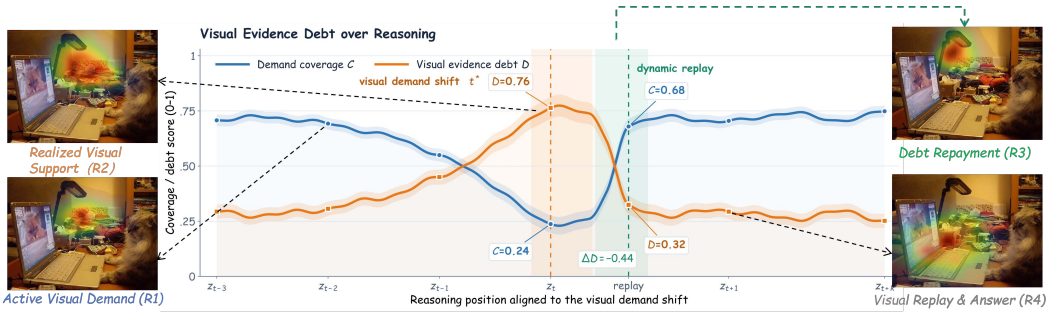


Figure 5: Representative debt-aware replay trajectory. Mean debt $\bar{D}_t$ and overlap coverage $\bar{C}_t = 1 - \bar{D}_t$ are shown around a replay event. Lower debt is observed in the subsequent reasoning round.

5.3 Qualitative Analysis

Representative trajectories are visualized to examine when replay is activated and which visual regions are selected. At each reasoning round, the mean visual evidence debt $\overline{D}_t$ is shown together with the replay decision and the selected visual tokens mapped to their original image regions.

6 Conclusion

We introduce Visual Evidence Debt to quantify the mismatch between current visual demand and realized support, with its positive-residual form localizing under-supported source tokens. Based on this formulation, EvidenceFlow selectively replays tokens through budgeted marginal debt repayment. Experiments across three MLLMs and five benchmarks demonstrate improved accuracy, visual-compute efficiency, and reasoning–support alignment, indicating that effective multimodal test-time scaling depends on how visual compute is allocated to current reasoning demands.

AI USE STATEMENT

Generative AI tools were not used for developing the method, conducting experiments, generating code, or preparing the manuscript. An external vision-language model was used only as an automated evaluator in the analysis experiments. All experimental settings, prompts, results, and manuscript contents were designed, verified, and finalized by the authors.

REFERENCES

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025. URL https://arxiv.org/abs/2503.04697.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

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*, 2024. URL https://arxiv.org/abs/2403.20330.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen Look Again: Guiding vision-language reasoning models to re-attention visual information. *arXiv preprint arXiv:2505.23558*, 2025. URL https://arxiv.org/abs/2505.23558.

Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. Don't look only once: Towards multimodal interactive reasoning with selective visual revisitation. *arXiv preprint arXiv:2505.18842*, 2025. URL https://arxiv.org/abs/2505.18842.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. Does thinking more always help? understanding test-time scaling in reasoning models. *arXiv preprint arXiv:2506.04210*, 2025. URL https://arxiv.org/abs/2506.04210.

Soumya Suvra Ghosal, Youngeun Kim, Zhuowei Li, Ritwick Chaudhry, Linghan Xu, Hongjing Zhang, Jakub Zablocki, Yifan Xing, and Qin Zhang. VisRef: Visual refocusing while thinking improves test-time scaling in multi-modal large reasoning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33404–33414, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ghosal_VisRef_Visual_Refocusing_while_Thinking_Improves_Test-Time_Scaling_in_Multi-Modal_CVPR_2026_paper.html.

Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. Don't overthink it: Preferring shorter thinking chains for improved LLM reasoning. *arXiv preprint arXiv:2505.17813*, 2025. URL https://arxiv.org/abs/2505.17813.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025a. URL https://arxiv.org/abs/2503.06749.

Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Junjie Hu, and Yong Jae Lee. VisualToolAgent (VisTA): A reinforcement learning framework for visual tool selection. *arXiv preprint arXiv:2505.20289*, 2025b. URL https://arxiv.org/abs/2505.20289.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6774–6786. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.528. URL https://aclanthology.org/2021.acl-long.528/.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02255.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. URL https://arxiv.org/abs/2501.19393.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

Alex Su, Haozhe Wang, Weimin Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025. URL https://arxiv.org/abs/2505.15966.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *The Thirty-Eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=QWTCcxMpPA.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. 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, pp. 24824–24837, 2022. URL https://arxiv.org/abs/2201.11903.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in LLMs. *arXiv preprint arXiv:2502.07266*, 2025. URL https://arxiv.org/abs/2502.07266.

xAI. RealWorldQA: A benchmark for real-world understanding. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. URL https://arxiv.org/abs/2411.10440.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in MLLM reasoning. *arXiv preprint arXiv:2507.03019*, 2025a. URL https://arxiv.org/abs/2507.03019.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-OneVision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025b. URL https://arxiv.org/abs/2503.10615.

Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, Wenzhuo Liu, Xiao Liang, Shuicheng Yan, and Chao Feng. SAIL-VL2 technical report. *arXiv preprint arXiv:2509.14033*, 2025. URL https://arxiv.org/abs/2509.14033.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. *arXiv preprint arXiv:2502.17422*, 2025a. URL https://arxiv.org/abs/2502.17422.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms. *arXiv preprint arXiv:2505.15436*, 2025b.

Yuyao Zhang, Zefeng Chen, Haizhai Zhang, Guorui Chen, Chang Yang, and Zhijiang Li. DVR: Dynamic visual re-reading for multimodal large language models via sub-question feedback. *Neurocomputing*, 700:134448, 2026. doi: 10.1016/j.neucom.2026.134448.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. URL https://arxiv.org/abs/2302.00923.

Appendix

Contents

A Related Work **14**

B Implementation and Reproducibility **14**

B.1 Implementation Details 14
B.2 Evaluation Protocol 15
B.3 Reasoning–Support Alignment Evaluation 16
B.4 Causal Validation of Debt-Localized Visual Evidence 19
B.5 Validation of Active Visual Demand 20
B.6 Causal Validation of Realized Visual Support 21
B.7 Matched Token-Selection and Realized Repayment Analysis 22

C Further Discussion **25**

C.1 Temporal role of visual replay 25

D Equivalent Forms of Visual Evidence Debt **25**

E Proof of Proposition 1 **26**

A Related Work

Multimodal reasoning models. Chain-of-thought prompting was first established for language models and was subsequently extended to multimodal inputs (Wei et al., 2022; Zhang et al., 2023). Longer and more structured visual reasoning has recently been obtained through supervised trajectories, staged decoding, and reinforcement learning, as in LLaVA-CoT, Vision-R1, and R1-OneVision (Xu et al., 2024; Huang et al., 2025a; Yang et al., 2025b). These developments build on broader progress in reinforcement-trained reasoning models (DeepSeek-AI et al., 2025). This progress has improved multimodal reasoning. However, access to the image evidence required by later steps is not ensured by additional textual reasoning alone.

Test-time scaling and textual reflection. Additional inference computation can be allocated to longer or multiple reasoning trajectories through test-time scaling (Snell et al., 2024). A trajectory is extended with a textual continuation cue by the budget-forcing strategy in s1, whereas reasoning length is learned in L1 (Muennighoff et al., 2025; Aggarwal & Welleck, 2025). Diminishing or non-monotonic returns have also been observed as reasoning chains become longer (Ghosal et al., 2025; Wu et al., 2025; Hassid et al., 2025). Following VisRef (Ghosal et al., 2026), *Standard Thinking* denotes native reasoning without intervention, and *Textual Self-Reflection* denotes the s1-style text-only continuation baseline. Neither setting is introduced in this work.

Visual refocusing in multimodal reasoning. Visual evidence can be revisited through explicit token copying or learned implicit refocusing (Chu et al., 2025; Yang et al., 2025a). Region retrieval and external visual tools have also been used during extended reasoning (Chung et al., 2025; Huang et al., 2025b; Su et al., 2025; Zhang et al., 2025b). In training-free settings, model-internal attention or gradient signals have been used to recover small visual details (Zhang et al., 2025a). Recent approaches further couple reasoning with adaptive visual perception or selective visual revisitation during inference (Zhang et al., 2025b; Chung et al., 2025). DVR dynamically re-encodes the image according to intermediate sub-questions (Zhang et al., 2026), while VisRef reinserts a relevance–diversity subset using a determinantal point process (Ghosal et al., 2026; Kulesza & Taskar, 2012). These methods establish the benefit of revisiting or reallocating visual information during reasoning, but do not explicitly account for the residual mismatch between the visual evidence currently required and the visual support already realized along the reasoning trajectory. In contrast, EvidenceFlow represents current visual demand and history-aware realized support jointly, and prioritizes their remaining positive residual.

B Implementation and Reproducibility

Implementation details and evaluation protocols are provided first. The subsequent analyses independently validate active visual demand, realized visual support, and debt-aware marginal repayment. Code, prompts, and evaluation scripts will be released upon publication.

B.1 Implementation Details

We use the official checkpoints and processors of Qwen3-VL-8B-Thinking, InternVL3.5-8B, and SAIL-VL2-8B, with all parameters frozen and images processed at each checkpoint's native resolution. All methods on the same backbone and dataset share the same chat template, decoding configuration, textual budget, stopping criteria, random seed, and answer parser. Greedy decoding is used with a limit of 8,192 generated text tokens, including reasoning and the final answer. For replay-based methods, replay decisions are evaluated at the first $K_{\text{cand}} = 10$ eligible reasoning boundaries. Generation then continues without further replay until the native end-of-sequence token or the text-token limit. Outputs without a valid parsed answer are counted as incorrect.

Each non-empty newline-delimited step defines a reasoning round; consecutive blank lines do not create additional rounds. Clauses are split at sentence-ending punctuation, semicolons, and standalone `and`, `but`, `while`, or `whereas`, excluding delimiters within mathematical expressions and decimal points. Empty fragments are discarded, and clause spans are mapped to text-token positions.

The main evaluation uses MathVista `testmini` (1,000 examples), MathVision (304), MM-Star (1,500), RealWorldQA `test` (765), and Geo3K `test` (601). Unless otherwise specified, we set

$\rho = 0.30$, $B = \lfloor \rho N \rfloor$, $\lambda_{\text{tok}} = 0.004$, $\delta_{\text{rep}} = 0.15$, $\ell = 20$, and $\epsilon = 10^{-6}$, selected on the Geo3K validation split and fixed thereafter.

Visual-token replay implementation. Cached input-level visual embeddings are replayed in source spatial order without rerunning the vision encoder. The prefix KV cache is retained, and only replayed tokens and subsequent text are processed incrementally. Replayed tokens are assigned fresh continuation positions after the current prefix rather than reusing their original sequence positions. For Qwen3-VL, the next scalar continuation indices are used on all three M-RoPE axes; for InternVL3.5 and SAIL-VL2, the native one-dimensional RoPE indices are advanced sequentially. The cache position is advanced by the number of replayed tokens, and subsequent text continues immediately after the last replay position. No image delimiters or other special tokens are reinserted during replay; only the cached projected visual embeddings are appended.

For demand and support estimation, $h_u^{(\ell)}$ and $v_i^{(\ell)}$ are taken from the input to layer-ℓ self-attention after its pre-attention RMSNorm. No additional feature normalization is applied except the ℓ_2 normalization implicit in cosine similarity. The same pre-attention RMSNorm output is used as the input to the value projection $W_V^{(\ell)}$ when computing ν_i. Layer indices are one-based throughout the paper, so $\ell = 20$ corresponds to decoder module index 19 in the implementation. Source-token identities are tracked separately from positional indices for history-aware support aggregation.

Realized-support extraction. The attention weight $a_{uo}^{(\ell,h)}$ is taken after causal masking and softmax, so inaccessible positions have zero weight. We retain and uniformly average all query heads:

$$\bar{a}_{uo}^{(\ell)} = \frac{1}{H_\ell^{\text{q}}} \sum_{h=1}^{H_\ell^{\text{q}}} a_{uo}^{(\ell,h)}. \tag{19}$$

Only original and replayed visual-token positions are retained. Their attention weights are normalized before occurrences of the same source token are summed:

$$\widetilde{a}_{uo}^{(\ell)} = \frac{\bar{a}_{uo}^{(\ell)}}{\sum_{o' \in \mathcal{O}_t} \bar{a}_{uo'}^{(\ell)} + \epsilon}, \qquad \mathcal{O}_t = \bigcup_{i=1}^{N} \mathcal{O}_t(i). \tag{20}$$

In Equation 8, $W_V^{(\ell)}$ denotes the model's complete value projection, including all key–value heads. This definition also applies to grouped-query attention, where several query heads share one key–value head. The value magnitude ν_i is computed from the original occurrence of source token i and shared across its replayed occurrences, while history dependence is captured by their context-dependent attention. Thus, $P_{t,j}$ provides a history-aware source-level attribution without explicitly modeling occurrence-specific variation in value magnitude, and is treated as an attribution proxy rather than an exact decomposition of visual contributions. Normalization over the N source-token identities is performed after attention aggregation and value gating.

B.2 Evaluation Protocol

Shared-budget comparisons. All internally evaluated methods share the same maximum reasoning-round and text-token limits, image encodings, and decoding settings. Budgeted visual-replay methods additionally share the candidate-boundary rule and maximum per-replay token budget B. Their activation policies and token selectors are varied, so realized replay totals and generated text lengths may differ. Full-image Replay is evaluated separately as a compute-unmatched reference. Unless otherwise specified, uncertainty is estimated using 10,000 paired bootstrap resamples. Questions are resampled within each dataset for masking and accuracy comparisons. The same resampled question indices are used across compared methods, and dataset-level estimates are macro-averaged within each bootstrap replicate. RDR and correlation estimates use question-level clustered resampling to preserve dependencies among replay events. Reported uncertainties are expressed as $\pm 1.96\,\text{SE}_{\text{boot}}$, where SE_{boot} is the standard deviation of the bootstrap estimates, yielding approximate 95% normal-based confidence intervals.

Baseline evaluation. Standard Thinking uses the native reasoning trace without additional text or visual tokens. Textual Self-Reflection appends the textual cue `Wait, think more` when the initial trace attempts to terminate and then resumes decoding under the same overall text-token limit.

Random Replay samples B source visual tokens uniformly without replacement. Relevance-only Replay selects the B tokens with the largest mean demand score $|\mathcal{K}_t|^{-1}\sum_j Q_{t,ji}$. Fixed EvidenceFlow runs the marginal selector to B tokens, whereas adaptive EvidenceFlow applies both λ_{tok} and δ_{rep}. For fixed-rate replay, the first $K_{\text{cand}} = 10$ eligible reasoning boundaries are indexed from 1, and replay is scheduled at boundaries $\{2, 5, 8\}$, corresponding to a nominal trigger rate of 30%. The same deterministic schedule is shared by all fixed-schedule selector controls, independently of model-internal relevance, demand, or support signals. At each reached replay boundary, each method selects exactly $B = \lfloor 0.3N \rfloor$ source visual tokens, with adaptive triggering and token-level early stopping disabled. Scheduled events that are not reached before termination, or at which no valid candidate remains, are skipped. The realized trigger rate is computed over candidate boundaries actually reached.

Ablation implementation. Without replay-history tracking, attention normalization and aggregation are restricted to original visual occurrences, and the resulting support scores are normalized over source tokens. Replayed tokens remain in the model context and are excluded only from support estimation.

Without saturation, the objective is replaced by

$$F_t^{\text{add}}(S) = \frac{1}{|\mathcal{K}_t|}\sum_{j\in\mathcal{K}_t}\sum_{i\in S} Q_{t,ji}\mathbf{1}[E_{t,ji} > 0].$$

The same budget, λ_{tok}, and δ_{rep} are retained, with marginal gains and cumulative repayment computed using F_t^{add}.

Without the repayment trigger, the activation rule is replaced by the fixed-rate schedule above. The saturating selector and token-level early stopping are retained. All other settings are unchanged.

Efficiency metrics. The realized replay ratio is $|S_t|/N$, averaged over replay events with $|S_t| > 0$. Trigger rate is the fraction of candidate replay boundaries at which visual tokens are inserted. All latency measurements are conducted on a single NVIDIA A800 40GB GPU with batch size 1 and bfloat16 precision, using PyTorch 2.6.0, Transformers 4.57.1, and CUDA 12.4. For support estimation, post-softmax attention probabilities at layer ℓ are captured directly from the attention module without an additional model forward pass.

Timing starts immediately before the initial multimodal model forward, after processor outputs have been transferred to the GPU, and ends after generation of the final output token. It therefore includes vision encoding, initial multimodal prefill, autoregressive decoding, signal extraction, token selection, and visual replay, while excluding disk I/O and CPU-side image preprocessing. CUDA execution is synchronized before and after each measurement and elapsed time is recorded using wall-clock timing. Five untimed warm-up runs are performed for each model–method configuration. Each example is then measured three times; the median latency of the three runs is computed per example and subsequently averaged across examples.

B.3 Reasoning–Support Alignment Evaluation

Sampling and generation. Two hundred questions are randomly sampled from each of MathVista, MathVision, and RealWorldQA, resulting in 600 questions. Standard Thinking, Relevance-only Replay, and EvidenceFlow are evaluated on the same questions using Qwen3-VL-8B-Thinking. All methods use the same decoding configuration, and the two replay-based methods share the same maximum per-replay visual-token budget.

Each generated reasoning trajectory is divided into clauses following the segmentation procedure in Section 3.1. An evaluation instance consists of the image, question, target clause, and reasoning prefix ending at that clause.

External VLM judge. The external VLM judge is instantiated as `gpt-5.2-2025-12-11`. For each evaluation instance, the judge first determines whether the target clause contains an image-dependent claim. If so, it assesses whether the claim is consistent with the image and returns the smallest set of image regions needed to verify it. A clause containing multiple visual claims is considered faithful only when all of them are supported by the image.

The judge is not given the method identity, subsequent reasoning, final answer, or model-internal signals. Deterministic decoding is used, and the judge is queried once for each clause. Responses must follow a fixed JSON format. A query is repeated only when its output cannot be parsed.

Judge prompt. The following prompt is used for all methods and datasets:

```
You are evaluating one clause from a multimodal reasoning
trajectory.

You are given:
1. an image;
2. a question about the image;
3. the reasoning prefix up to and including the target clause;
4. the target clause.

Complete the following tasks.

1. Determine whether the target clause contains a claim that
   can be directly verified from the image.
2. If it does, determine whether every image-dependent claim
   in the clause is supported by the image.
3. Return the smallest set of rectangular image regions
   needed to verify the claim.
4. If the claim depends on the entire image or cannot be
   localized reliably, mark it as non-localizable.

Do not evaluate the final answer or infer information from
later reasoning steps. Use only the provided image, question,
reasoning prefix, and target clause.

Return only the following JSON object:

{
  "image_dependent": true or false,
  "faithful": true, false, or null,
  "localizable": true or false,
  "evidence_regions": [
    [x_min, y_min, x_max, y_max]
  ]
}

Coordinates must be normalized to the range [0, 1000].
Return an empty region list when the clause is not
image-dependent or cannot be localized.
```

Mapping evidence regions to visual tokens. The returned bounding boxes are mapped to the source visual-token grid. A source token belongs to $\mathcal{R}^*_{t,j}$ if its corresponding image patch overlaps at least one returned evidence region. Clauses marked as non-localizable are excluded from the regional alignment metrics. The retained clause counts are reported in Table 5 for transparency, while VSA and SLR are aggregated at the question level.

For every retained clause, the realized-support distribution $P_{t,j}$ is computed from the logged model signals using Equation 8. These internal signals are not provided to the external judge.

Additional diagnostic metrics. For each dataset d, let $N^d_{a,b}$ denote the number of retained clauses satisfying $f_{t,j} = a$ and $g_{t,j} = b$, where $f_{t,j}$ indicates whether the visual claim is faithful and $g_{t,j}$ indicates whether the strongest realized support lies in the required evidence region.

Table 5: Complete reasoning–support diagnostics on Qwen3-VL-8B-Thinking. Clause counts are reported for transparency; VSA and SLR are aggregated at the question level. Adj. VSA additionally controls for evidence-region size. Best and second-best results are marked in dark red and light red.

Method	Dataset	$\lvert\mathcal{P}_d\rvert$	VSA ↑	Adj. VSA ↑	SLR ↓	UPR ↓	RSC ↑	PRD ↓
Standard Thinking	MathVista	816	0.456	0.337	29.4%	14.7%	64.4%	11.0%
	MathVision	1,038	0.416	0.270	36.4%	18.9%	56.0%	15.5%
	RealWorldQA	574	0.490	0.393	26.6%	12.2%	67.4%	9.6%
Relevance-only Replay	MathVista	835	0.498	0.388	25.0%	11.9%	69.9%	7.8%
	MathVision	1,064	0.459	0.324	30.8%	15.7%	62.9%	10.9%
	RealWorldQA	588	0.531	0.442	21.3%	9.9%	73.8%	7.0%
EVIDENCEFLOW	MathVista	828	**0.575**	**0.482**	**14.7%**	**7.6%**	**81.8%**	**4.1%**
	MathVision	1,056	**0.540**	**0.425**	**19.4%**	**10.4%**	**76.0%**	**6.1%**
	RealWorldQA	580	**0.604**	**0.529**	**12.1%**	**6.7%**	**84.5%**	**3.9%**

The Unfaithful Perception Rate (UPR) is defined as

$$\mathrm{UPR}_d = \frac{N^d_{0,0} + N^d_{0,1}}{|\mathcal{P}_d|}. \tag{21}$$

UPR measures the fraction of region-groundable visual clauses that are inconsistent with the image.

Reasoning–Support Consistency (RSC) is defined as

$$\mathrm{RSC}_d = \frac{N^d_{1,1}}{N^d_{1,1} + N^d_{1,0} + N^d_{0,1}}. \tag{22}$$

RSC measures how often a faithful visual statement and correctly localized support occur together, excluding clauses for which both are incorrect.

The Perception–Reasoning Dissociation Rate (PRD) is defined as

$$\mathrm{PRD}_d = \frac{N^d_{0,1}}{N^d_{1,1} + N^d_{0,1}}. \tag{23}$$

PRD measures how often the model receives its strongest visual support from the required region but still produces an incorrect visual statement.

Region-size robustness. Because larger evidence regions can mechanically receive more support mass, we additionally report a region-size-adjusted VSA. For clause (t, j), let $a_{t,j} = |\mathcal{R}^*_{t,j}|/N$ denote the fraction of visual tokens covered by its evidence region, and define

$$\widetilde{s}_{t,j} = \frac{s_{t,j} - a_{t,j}}{1 - a_{t,j}}. \tag{24}$$

The adjusted score is aggregated at the question level following the same protocol as VSA.

Per-dataset results. All diagnostic metrics are reported separately for the three datasets in Table 5. MathVision contains the largest number of region-groundable clauses and exhibits the weakest reasoning–support alignment among the three datasets. The overall trend is consistent across datasets. Across all datasets, EVIDENCEFLOW improves both VSA and size-adjusted VSA, while also improving RSC and reducing SLR, UPR, and PRD.

Human validation of the VLM judge. The judge is validated using a stratified sample of 300 clause-level outputs, with 100 sampled from each dataset. The sample is balanced across the three methods and includes both faithful and unfaithful predictions. This procedure is used to verify the reliability of the external judge rather than to manually relabel the complete evaluation set.

Two human evaluators independently inspect each sampled output. They verify whether the clause-type decision is correct, whether the assigned faithfulness label agrees with the image, and whether the returned regions contain sufficient visual evidence to assess the clause. The evaluators receive

Table 6: Human validation of the external VLM judge. Agreement on clause identification and faithfulness labels is reported together with the acceptance rate of localized evidence regions. The overall row is macro-averaged over three datasets; # Clauses gives the number of validated clauses.

Dataset	# Clauses	Clause-Type Agreement ↑	Faithfulness Agreement ↑	Region Acceptance ↑
MathVista	100	95.0%	92.0%	89.0%
MathVision	100	93.0%	90.0%	85.0%
RealWorldQA	100	96.0%	94.0%	90.0%
Overall	**300**	**94.7%**	**92.0%**	**88.0%**

the same image, question, reasoning prefix, and target clause as the judge, but are not shown the method identity, subsequent reasoning, final answer, or model internals.

The mean agreement between the judge and the human evaluators is reported in Table 6. Agreement is obtained on 93.3% of all binary verification decisions, with a Cohen's κ of 0.84 (Cohen, 1960). These results support the use of the external VLM judge for clause-level faithfulness and regional evidence evaluation.

Demand-shift alignment for Figure 1b. The introductory analysis uses 200 MathVista and 200 RealWorldQA questions. A boundary is marked as a visual-demand shift when the evidence-token set of the first region-groundable clause after the boundary is disjoint from that of the preceding region-groundable clause. Trajectories are aligned at the first such boundary and grouped as successful or failed according to final-answer correctness. SLR is then computed independently at each relative reasoning step. This analysis uses the same blinded judge and source-token mapping as the alignment evaluation above.

Since the successful–failed comparison is observational, we additionally examine the pattern separately by benchmark and reasoning-length group. The higher post-shift SLR of failed trajectories remains qualitatively consistent across these subsets. This analysis is used to check that the observed trend is not driven solely by benchmark or trajectory length.

B.4 Causal Validation of Debt-Localized Visual Evidence

We further examine whether the token-level debt residual identifies visual evidence that is required by the current reasoning process but remains under-supported. For each clause (t, j), the debt residual is

$$E_{t,ji} = [Q_{t,ji} - P_{t,ji}]_+, \tag{25}$$

where a larger value indicates that the demand allocation of token i exceeds its realized-support allocation.

A fixed subset of 600 region-groundable clauses is sampled from Standard Thinking trajectories generated by Qwen3-VL-8B-Thinking. One clause is sampled per question, with 200 questions from each of MathVista, MathVision, and RealWorldQA. The same subset is used in Appendix B.6. NLL increases and final-answer accuracy drops are computed per dataset and then macro-averaged. For each clause, 20% of the source visual tokens are masked according to four criteria:

- **Top Debt**: tokens with the largest $E_{t,ji}$ values;
- **Top Demand**: tokens with the largest $Q_{t,ji}$ values;
- **Top Support**: tokens with the largest $P_{t,ji}$ values;
- **Random**: uniformly sampled tokens under the same budget.

Selection scores are computed from the unmasked trajectory. All occurrences of selected source tokens are masked by setting attention logits from causally subsequent queries to these positions to $-\infty$ at every decoder layer. Sequence order and positional indices are preserved. The text-token

Table 7: Causal validation of debt-localized visual evidence at a 20% masking ratio. All strategies use the same clauses and masking budget as the corresponding 20% setting in Figure 6. Values are reported with 95% paired-bootstrap confidence intervals.

Selection Strategy	# Clauses	Target-Clause NLL Increase ↑	Final-Answer Accuracy Drop (pp) ↑
Random	600	0.27 ± 0.04	1.2 ± 0.9
Top Support ($P_{t,j}$)	600	0.72 ± 0.06	5.7 ± 1.7
Top Demand ($Q_{t,j}$)	600	0.49 ± 0.05	2.5 ± 1.3
Top Debt ($E_{t,j}$)	600	0.66 ± 0.05	5.2 ± 1.6

Table 8: Localization of active visual demand on region-groundable clauses. DLS measures the probability mass assigned to judge-identified evidence regions.

Demand signal	MathVista	MathVision	RealWorldQA	Avg.
Uniform	0.18	0.16	0.20	0.18
Attention Relevance	0.39	0.35	0.43	0.39
$Q_{t,j}$	**0.53**	**0.49**	**0.57**	**0.53**

prefix preceding the target clause is held fixed across policies and recomputed under the mask to rebuild its KV cache.

Target-clause NLL is averaged over textual tokens under teacher forcing. Final-answer accuracy is evaluated in a separate branch by freely continuing from the same pre-clause prefix, without teacher-forcing the target clause. The mask remains active throughout both evaluations. Each result is compared with its corresponding unmasked reference under the same scoring or decoding protocol. If the debt residual localizes demanded but under-supported evidence, Top-Debt masking should have a larger effect than Top-Demand masking, whereas Top-Support masking is expected to have the strongest effect because $P_{t,j}$ characterizes visual evidence already realized by the model.

As shown in Table 7, the largest degradation is observed under Top-Support masking, supporting $P_{t,j}$ as a proxy for realized visual reliance. Under paired resampling, Top-Debt exceeds Top-Demand by 0.17 ± 0.05 in target-clause NLL increase and by 2.7 ± 1.6 percentage points in final-answer accuracy drop, suggesting that the residual identifies prediction-relevant evidence beyond demand-only ranking. The utility of replaying such evidence is evaluated separately through the matched replay controls in Appendix B.7.

B.5 Validation of Active Visual Demand

The evidence regions returned by the blinded external judge are also used to validate whether $Q_{t,j}$ localizes the visual evidence required by the current clause. For a normalized demand signal $\pi_{t,j}$, the Demand Localization Score (DLS) on dataset d is defined as

$$\mathrm{DLS}_d(\pi) = \frac{1}{|\mathcal{P}_d|} \sum_{(t,j)\in\mathcal{P}_d} \sum_{i\in\mathcal{R}^*_{t,j}} \pi_{t,ji}, \tag{26}$$

where $\mathcal{P}_d$ contains the retained region-groundable clauses and $\mathcal{R}^*_{t,j}$ contains their judge-identified evidence tokens. A higher score indicates that more probability mass is assigned to the required evidence regions.

Two controls are considered. Uniform assigns $1/N$ probability to every visual token. Attention Relevance is obtained by normalizing the mean clause-to-image attention over the original visual tokens at layer ℓ, without using the semantic compatibility in Equation 5. Scores are computed separately for the three evaluated methods and then macro-averaged.

As shown in Table 8, $Q_{t,j}$ assigns the largest probability mass to the required evidence regions on all three datasets. Its average DLS reaches 0.53, compared with 0.39 for Attention Relevance and 0.18 for Uniform. These results provide independent evidence that $Q_{t,j}$ captures the visual evidence required by the current reasoning clause.

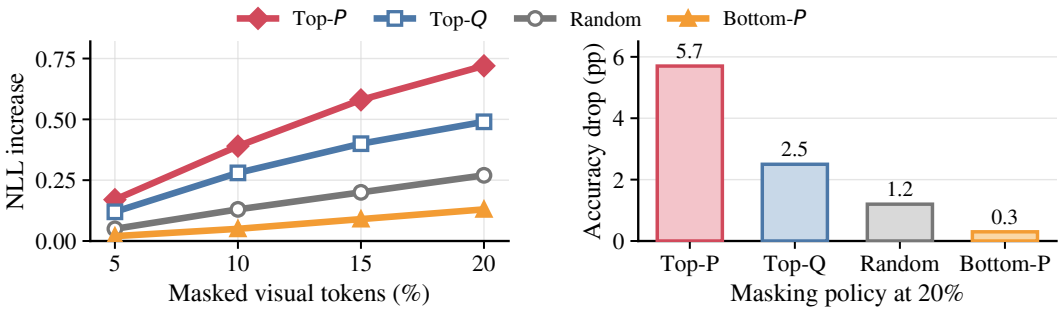


Figure 6: Causal validation of realized visual support on Qwen3-VL-8B-Thinking. Target-clause NLL increase is reported across masking ratios, and final-answer accuracy degradation is reported at 20% masking. The same fixed 600-clause subset is used as in Table 7.

Visual-dependence sanity check. Since $Q_{t,j}$ is normalized over source visual tokens, it characterizes the relative allocation of visual demand rather than the absolute magnitude of visual dependence. We therefore examine whether non-image-dependent clauses spuriously contribute to replay. Using the image-dependence labels already produced by the blinded external judge, clauses are partitioned into image-dependent and non-image-dependent groups. For each clause, we compute its visual evidence debt $D_{t,j}$ and the clause-level repayment contributed by the selected replay set,

$$\phi_{t,j}(S_t) = \min\left\{ D_{t,j}, \sum_{i \in S_t} Q_{t,ji} \mathbf{1}[E_{t,ji} > 0] \right\}.$$

Image-dependent clauses exhibit a mean debt of 0.284 and a mean repayment contribution of 0.091, compared with 0.176 and 0.023 for non-image-dependent clauses, respectively. The substantially lower debt and repayment contribution of non-image-dependent clauses indicate that the normalization of $Q_{t,j}$ does not systematically cause such clauses to dominate replay decisions in practice.

B.6 Causal Validation of Realized Visual Support

Because $P_{t,j}$ is an attention–value attribution proxy, a causal masking experiment is conducted to independently test whether tokens assigned higher support more strongly affect the model's prediction. We use the same fixed 600-clause evaluation subset as in Appendix B.4, with 200 clauses from each of MathVista, MathVision, and RealWorldQA. For every clause, $P_{t,j}$ and $Q_{t,j}$ are computed from the unmasked trajectory.

Visual tokens are masked according to four policies. Top-P and Bottom-P select tokens with the largest and smallest realized support, respectively; Top-Q selects tokens with the largest current demand; and Random provides a uniform control. The masking, prefix-recomputation, and continuation protocols follow Appendix B.4, using the same fixed 600-clause subset.

The target-clause NLL is averaged over textual tokens under teacher forcing. Its masking-induced increase is defined as

$$\Delta\mathrm{NLL}_{t,j} = \mathrm{NLL}_{t,j}^{\mathrm{masked}} - \mathrm{NLL}_{t,j}^{\mathrm{original}}. \tag{27}$$

A larger $\Delta\mathrm{NLL}_{t,j}$ indicates that the masked visual tokens contribute more strongly to prediction of the target clause. Final-answer accuracy is additionally evaluated by continuing generation from the same intervened prefix.

As shown in Figure 6, Top-P masking consistently causes the largest degradation. At 20% masking, the target-clause NLL and final-answer accuracy degradations are 0.72 and 5.7 percentage points, respectively, compared with 0.49 and 2.5 percentage points under Top-Q. Bottom-P masking reduces accuracy by only 0.3 percentage points. These results support $P_{t,j}$ as a proxy for visual reliance on Standard Thinking trajectories, where only original visual occurrences are present. We separately examine the history-aware setting after replay below.

Table 9: Realized-support validation after visual replay. All estimators use the same replay-containing prefixes and a 20% source-token masking budget. Reported ± values denote the half-width of the 95% paired-bootstrap confidence interval.

Support Estimator	# Questions	Target-Clause NLL Increase ↑
Random	428	0.29 ± 0.04
Original-only Support	428	0.55 ± 0.05
History-aware Support	428	0.71 ± 0.06
Occurrence-specific Support	428	0.73 ± 0.06

Replay-history support validation. The preceding masking experiment evaluates $P_{t,j}$ on Standard Thinking trajectories, where only original visual occurrences are present. We further evaluate realized-support attribution after visual replay using trajectories generated by adaptive EVIDENCEFLOW. For each question, we retain at most one target clause: the first region-groundable visual clause following the first replay event. This yields 428 questions in total, including 149 from MathVista, 136 from MathVision, and 143 from RealWorldQA.

All support estimators are computed from the same unmasked prefix containing the recorded replay history. We compare Random masking, Original-only Support that ignores replay occurrences, the History-aware Support in Equation 8, and an Occurrence-specific Support variant. For the latter, the source-shared term $A_{t,ji}\nu_i$ is replaced by

$$\widetilde{s}^{\mathrm{occ}}_{t,ji} = \frac{1}{|W_{t,j}|} \sum_{u \in W_{t,j}} \sum_{o \in \mathcal{O}_t(i)} \widetilde{a}^{(\ell)}_{uo} \nu_o, \tag{28}$$

where ν_o is computed separately for each visual occurrence. The resulting scores are normalized over the N source-token identities using the same smoothing as Equation 8.

For all policies, 20% of source visual tokens are masked. When a source token is selected, its original occurrence and all replayed occurrences in the recorded prefix are masked jointly. The generated text, replay sequence, and positional indices are held fixed, and the prefix KV cache is rebuilt under the mask before computing teacher-forced target-clause NLL. Results are macro-averaged across the three datasets, with uncertainty estimated by the same paired-bootstrap protocol described in Appendix B.2.

History-aware Support produces a substantially larger masking effect than Original-only Support, with a paired NLL-increase difference of 0.16 ± 0.05, showing that aggregating replay occurrences provides additional information about realized visual reliance. It also exceeds Random by 0.42 ± 0.06. Occurrence-specific Support yields only a small additional increase of 0.02 ± 0.03 over the source-shared History-aware estimator. Thus, explicitly modeling occurrence-specific value magnitudes provides limited additional attribution benefit in this setting, while the source-shared estimator retains most of the observed history-aware masking effect.

B.7 MATCHED TOKEN-SELECTION AND REALIZED REPAYMENT ANALYSIS

A comparison under a shared replay plan is conducted to isolate the effect of the token-selection objective. For each input x, adaptive EVIDENCEFLOW is first executed and its replay plan is recorded as

$$\Pi(x) = \{(b_r, k_r)\}_{r=1}^{R_x}, \tag{29}$$

where b_r denotes the ordinal reasoning boundary of replay event r, and k_r is the corresponding number of replayed visual tokens. Each control follows the recorded plan, selecting exactly k_r tokens at every scheduled boundary it reaches. Events after termination are skipped. Decoding settings and the text-token limit are shared. Thus, prescribed replay boundaries and per-event token counts are controlled, while realized replay totals, text lengths, and latency may differ. Accuracy comparisons are therefore conditional on the replay plan recorded from EVIDENCEFLOW.

Matched selectors. Random selection samples k_r source tokens uniformly without replacement. Relevance-only selection uses the clause-averaged demand score. Relevance–diversity selection

augments this score with a generic DPP-based diversity term. A matched VISREF selector additionally implements the published reasoning-conditioned DPP criterion of Ghosal et al. (2026) under the same recorded replay schedule and selected-set sizes. Residual-only selection uses the clause-averaged positive residual. Gated-demand selection uses the token-wise weight in Equation 12, but applies independent Top-k_r ranking without clause-level saturation or marginal-gain recomputation:

$$s_{t,i}^{\mathrm{rel}} = \frac{1}{|K_t|} \sum_{j \in K_t} Q_{t,ji}, \tag{30}$$

$$s_{t,i}^{\mathrm{res}} = \frac{1}{|K_t|} \sum_{j \in K_t} E_{t,ji}, \tag{31}$$

$$s_{t,i}^{\mathrm{gate}} = \frac{1}{|K_t|} \sum_{j \in K_t} Q_{t,ji}\, \mathbf{1}[E_{t,ji} > 0]. \tag{32}$$

For the generic relevance–diversity control, a positive semidefinite DPP kernel is constructed as

$$L_t^{\mathrm{rel-div}} = \mathrm{diag}(s_t^{\mathrm{rel}}) K_t \,\mathrm{diag}(s_t^{\mathrm{rel}}), \qquad (K_t)_{ii'} = \left\langle \hat{v}_i^{(\ell)}, \hat{v}_{i'}^{(\ell)} \right\rangle, \tag{33}$$

where $\hat{v}_i^{(\ell)}$ is the normalized representation of visual token i. Its replay set is approximately selected by greedy MAP inference:

$$S_t^{\mathrm{rel-div}} \approx \arg\max_{|S|=k_r} \log\det\left((L_t^{\mathrm{rel-div}})_S + \epsilon I\right). \tag{34}$$

For the matched VISREF selector, we follow its published reasoning-conditioned DPP criterion. Let $\mathcal{W}_t$ denote the textual positions of the current reasoning state. We construct

$$M_t = \sum_{u \in \mathcal{W}_t} h_u^{(\ell)} \left(h_u^{(\ell)}\right)^\top, \qquad \left(L_t^{\mathrm{VisRef}}\right)_{ii'} = \left(v_i^{(\ell)}\right)^\top M_t v_{i'}^{(\ell)}. \tag{35}$$

To handle rank-deficient kernels, a small positive diagonal regularizer is added, and the replay set is selected by

$$S_t^{\mathrm{VisRef}} \approx \arg\max_{|S|=k_r} \log\det\left((L_t^{\mathrm{VisRef}})_S + 10^{-6} I_{|S|}\right). \tag{36}$$

Greedy selection is continued until exactly k_r tokens are selected, with ties resolved by source-token index. This regularized implementation retains the published reasoning-conditioned kernel while keeping the prescribed replay boundaries and token counts fixed.

For each independently ranked control $m \in \{\mathrm{rel}, \mathrm{res}, \mathrm{gate}\}$, the replay set is

$$S_t^m = \mathrm{TopK}_{k_r}\left(\{s_{t,i}^m\}_{i=1}^N\right) \tag{37}$$

at boundary b_r.

The generic relevance–diversity control tests whether redundancy reduction alone explains the gain over relevance-only replay, while the matched VISREF selector provides a stronger reasoning-conditioned relevance–diversity comparison. The remaining selectors provide a progressive decomposition of Equation 12.

Method-agnostic fixed-schedule control. To test whether the selector advantage depends on the replay plan generated by EVIDENCEFLOW, Relevance-only, Matched VISREF, Matched Gated Demand, and the EVIDENCEFLOW selector are additionally evaluated under the common fixed schedule $\{2, 5, 8\}$ with $B = \lfloor 0.3N \rfloor$ tokens per reached boundary. Adaptive triggering and token-level early stopping are disabled, so only the token-selection rule is varied.

As reported in Table 2, EVIDENCEFLOW achieves 65.1% macro-average accuracy, compared with 64.0% for Matched Gated Demand, 63.6% for the Matched VISREF selector, and 62.8% for Relevance-only Replay. The corresponding paired improvements are 1.1 ± 0.5, 1.5 ± 0.6, and 2.3 ± 0.7 percentage points, respectively. The same ordering therefore holds without an EVIDENCEFLOW-generated replay plan. The Matched VISREF result remains a selector-only comparison and does not reproduce its native adaptive stopping mechanism.

Realized post-replay debt reduction. To examine whether the pre-replay objective predicts an actual reduction in subsequent demand–support mismatch, we conduct a controlled branching analysis using Qwen3-VL-8B-Thinking on the same 600 questions used in Section 4.3. At each replay event recorded by adaptive EVIDENCEFLOW, all branches start from the same pre-replay prefix $\mathcal{H}_t$. A no-replay branch is retained as the counterfactual reference. Each matched selector replays exactly k_r tokens at the same boundary and then generates one additional reasoning round using the same decoding configuration.

For selector m, realized debt reduction is defined as

$$\mathrm{RDR}_t^{(m)} = \overline{D}_{t+1}^{\text{no-replay}} - \overline{D}_{t+1}^{(m)}. \tag{38}$$

A positive value indicates that replay produces a lower next-round Visual Evidence Debt than continuing from the same prefix without replay. RDR captures the overall downstream effect of replay, which may include both improved visual support and changes in subsequent reasoning states. It is therefore interpreted as an end-to-end mismatch reduction measure rather than a pure token-level causal estimate of support recovery.

Because all matched selectors use the same prefix, replay boundary, and selected-set size, differences in $\mathrm{RDR}_t^{(m)}$ reflect the downstream effect of token selection. Full-image Replay is excluded from this matched analysis because it reinserts all N visual tokens; it remains a compute-unmatched reference in Table 2.

Fixed-target replay validation. To distinguish improved support for a fixed visual requirement from changes in generated content, the same branching setup is additionally evaluated using teacher-forced target-clause NLL. At each recorded event, the first visually groundable clause in the no-replay next-round continuation is considered as a target. Its visual correctness is verified from the image before replay-branch outcomes are inspected. Events without a verified target are excluded from this analysis.

The target clause and any preceding text in that continuation are held fixed across branches. Each selector uses only the shared pre-replay state to select exactly k_r tokens; neither the target text nor its annotations are used for selection. Any intervening text is teacher-forced identically before scoring the target clause. Fixed-target NLL reduction is defined as

$$\Delta\mathrm{NLL}_t^{(m)} = \mathrm{NLL}(c_t^\star \mid \mathcal{H}_t, u_t^\star) - \mathrm{NLL}\Big(c_t^\star \mid \mathcal{H}_t, \mathrm{Replay}(S_t^{(m)}), u_t^\star\Big), \tag{39}$$

where $u_t^\star$ denotes the fixed intervening text, which may be empty, and NLL is averaged over target-clause tokens. A positive value indicates that replay improves the likelihood of the same image-verified clause relative to no replay. Because targets originate from the no-replay continuation, this diagnostic is conditional on those reference targets.

The filtering retains 547 questions and 1,126 replay events. Event-level reductions are first averaged within each question. For each bootstrap replicate, questions are resampled within each dataset with selector pairing preserved, after which the dataset means are macro-averaged. Uncertainty is estimated using paired question-level clustered bootstrap resampling.

Matched Relevance–Diversity consistently improves over Matched Relevance, confirming the benefit of redundancy reduction. The matched VISREF selector further improves the relevance–diversity control through its reasoning-conditioned criterion. Matched Gated Demand also improves over the direct residual-magnitude control across all three backbones and remains the strongest independently ranked control.

Beyond token-wise gating, EVIDENCEFLOW improves macro-average accuracy over Matched Gated Demand by 2.7 ± 0.9 percentage points on Qwen3-VL-8B-Thinking, 2.6 ± 1.1 points on InternVL3.5-8B, and 2.8 ± 1.1 points on SAIL-VL2-8B. The consistent margins across all three backbones support the additional contribution of clause-level saturation and marginal-gain recomputation under the same prescribed replay plan. EVIDENCEFLOW also increases RDR from 0.056 under Matched Gated Demand to 0.078, corresponding to a paired improvement of 0.022 ± 0.008 under question-level clustered bootstrap resampling.

On fixed reference targets, EVIDENCEFLOW achieves an NLL reduction of 0.050 ± 0.007, compared with 0.036 ± 0.006 for Matched Gated Demand and 0.030 ± 0.006 for Matched Residual.

Table 10: Matched selector comparison under a shared replay plan. Accuracy is macro-averaged over five benchmarks. RDR and fixed-target ΔNLL are evaluated on Qwen3-VL-8B-Thinking; the latter uses 547 questions and 1,126 replay events. Values after ± denote approximate 95% paired-bootstrap confidence intervals.

Selector	Qwen3-VL-8B Accuracy (%)	InternVL3.5-8B Accuracy (%)	SAIL-VL2-8B Accuracy (%)	RDR ↑	Fixed-Target ΔNLL ↑
Matched Random	61.8	54.7	53.4	0.012	0.010 ± 0.004
Matched Relevance	62.5	55.8	54.9	0.027	0.019 ± 0.005
Matched Relevance–Diversity	62.9	56.4	55.4	0.033	0.023 ± 0.005
Matched VISREF Selector	63.5	57.1	56.1	0.042	0.027 ± 0.006
Matched Residual	63.2	56.8	55.8	0.039	0.030 ± 0.006
Matched Gated Demand	63.9	57.9	56.6	0.056	0.036 ± 0.006
EVIDENCEFLOW	**66.6**	**60.5**	**59.4**	**0.078**	0.050 ± 0.007

The corresponding paired improvements are 0.014 ± 0.005 and 0.020 ± 0.006, respectively. Unlike free-generation RDR, these comparisons hold the target text fixed and therefore separate improved likelihood of the same image-verified target from changes in generated content. They evaluate the predictive utility of the selected replay tokens without assuming that $F_t(S)$ is calibrated to actual NLL reduction. Accordingly, $Q_{t,j} - P_{t,j}$ remains an operational mismatch between the two estimators rather than an externally calibrated probabilistic discrepancy.

Across replay events, the predicted repayment surrogate $\Gamma_t = F_t(S_t)$ is positively associated with realized debt reduction (Spearman's $\rho = 0.59 \pm 0.07$ under question-level clustered bootstrap resampling). Since replay timing and selected-set sizes are matched, and the realized-repayment branches start from identical prefixes, these results indicate that $F_t(S)$ is informative of the subsequent downstream reduction in demand–support mismatch rather than merely favoring additional visual computation. This empirical association complements the guarantee in Proposition 1, which applies to optimization of the pre-replay objective $F_t(S)$ rather than to subsequent transformer dynamics.

C FURTHER DISCUSSION

C.1 TEMPORAL ROLE OF VISUAL REPLAY

The demand distribution $Q_{t,j}$ is estimated from the hidden states of z_t after it is generated. Thus, EVIDENCEFLOW neither predicts unseen visual demand in z_{t+1} nor alters the already generated z_t. Instead, it diagnoses visual requirements in the current reasoning state whose demand share exceeds their realized-support share at the replay boundary. Since z_{t+1} is generated conditionally on $\mathcal{H}_t$, it may verify, refine, or build upon the intermediate conclusions in z_t. Replay therefore acts as a prospective intervention based on a retrospective diagnosis, providing renewed visual support before generation continues. Accordingly, replay supports verification and continuation of the current reasoning state rather than anticipating the complete visual demand of the next round.

D EQUIVALENT FORMS OF VISUAL EVIDENCE DEBT

The equivalent residual and total-variation forms used in Section 3.1 are proved below. For any $a, b \geq 0$,

$$a - \min(a, b) = [a - b]_+. \tag{40}$$

Because $Q_{t,j}$ is a probability distribution, $\sum_i Q_{t,ji} = 1$. Therefore, the overlap definition in Equation 9 satisfies

$$\begin{aligned} D_{t,j} &= \sum_{i=1}^{N} Q_{t,ji} - \sum_{i=1}^{N} \min(Q_{t,ji}, P_{t,ji}) \\ &= \sum_{i=1}^{N} [Q_{t,ji} - \min(Q_{t,ji}, P_{t,ji})] \\ &= \sum_{i=1}^{N} [Q_{t,ji} - P_{t,ji}]_+ = \sum_{i=1}^{N} E_{t,ji}. \end{aligned} \tag{41}$$

The third equality follows from Equation 40. This establishes the positive residual interpretation of visual evidence debt.

For the total-variation form, the identity $|a - b| = a + b - 2\min(a, b)$ is used. Because both $Q_{t,j}$ and $P_{t,j}$ are probability distributions,

$$\begin{aligned} \frac{1}{2} \|Q_{t,j} - P_{t,j}\|_1 &= \frac{1}{2} \sum_{i=1}^{N} |Q_{t,ji} - P_{t,ji}| \\ &= \frac{1}{2} \sum_{i=1}^{N} [Q_{t,ji} + P_{t,ji} - 2\min(Q_{t,ji}, P_{t,ji})] \\ &= \frac{1}{2} \left[1 + 1 - 2 \sum_{i=1}^{N} \min(Q_{t,ji}, P_{t,ji}) \right] \\ &= 1 - \sum_{i=1}^{N} \min(Q_{t,ji}, P_{t,ji}) = D_{t,j}. \end{aligned} \tag{42}$$

Thus, the overlap gap, positive residual mass, and total-variation distance are equivalent under the normalization in Equations 6 and 8.

E Proof of Proposition 1

For an active demand j, define the nonnegative token weight

$$w_{t,ji} = Q_{t,ji} \, \mathbf{1}[E_{t,ji} > 0] \tag{43}$$

and the capped coverage function

$$g_{t,j}(S) = \min \left\{ D_{t,j}, \sum_{i \in S} w_{t,ji} \right\}. \tag{44}$$

Because $w_{t,ji} \geq 0$, adding a token cannot decrease the inner sum. Therefore, $g_{t,j}$ is monotone.

For any $A \subseteq B \subseteq \{1, \ldots, N\}$ and $i \notin B$, the marginal gain of adding token i is

$$g_{t,j}(A \cup \{i\}) - g_{t,j}(A) = \min \left\{ w_{t,ji}, \left[D_{t,j} - \sum_{k \in A} w_{t,jk} \right]_+ \right\}. \tag{45}$$

The residual debt in brackets is no smaller for A than for B. Therefore, the marginal gain of i under A is at least its marginal gain under B. This establishes diminishing returns and hence the submodularity of $g_{t,j}$. Because a nonnegative weighted sum of monotone submodular functions remains monotone submodular, $F_t(S) = |\mathcal{K}_t|^{-1} \sum_j g_{t,j}(S)$ is also monotone and submodular. The classical cardinality-constrained greedy result (Nemhauser et al., 1978) therefore gives

$$F_t\left(S_t^{\text{greedy}}\right) \geq \left(1 - \frac{1}{e}\right) \max_{|S| \leq B} F_t(S) \tag{46}$$

when greedy selection is run to the full budget B. If selection stops early because the largest remaining marginal repayment is below λ_{tok}, the resulting adaptive policy is not claimed to inherit this full-budget guarantee.