ACTVIS: ACTIVE MAINTENANCE OF STRUCTURED VISUAL EVIDENCE FOR EFFICIENT VISUAL CHAIN-OF-THOUGHT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal chain-of-thought reasoning requires models to acquire and reuse visual evidence across multiple inference steps. Existing training-free Visual CoT methods mainly optimize when and where new visual evidence should be acquired, while previously acquired observations are typically appended to a linear reasoning trajectory without explicit maintenance. As reasoning proceeds, useful evidence may become obscured, whereas redundant, outdated, or conflicting observations continue to occupy the multimodal context. We introduce **ActVis**, a training-free framework for active visual evidence maintenance. ActVis organizes task-critical observations in a structured and editable *Evidence Canvas*, where each evidence unit remains grounded in its visual source and can be added, retained, revised, merged, or deleted as reasoning evolves. ActVis further employs an entropy-guided evidence gate to estimate the marginal utility of candidate observations based on predictive uncertainty, retaining informative evidence, filtering redundancy, and revising conflicts. Extensive experiments across multiple multimodal large language models and diverse reasoning tasks demonstrate that ActVis consistently improves reasoning accuracy over linear text–image reasoning and training-free Visual CoT baselines.

1 INTRODUCTION

Chain-of-thought (CoT) reasoning has substantially improved the problem-solving capabilities of large language models by decomposing complex problems into intermediate steps Wei et al. (2022); Kojima et al. (2022). Extending this paradigm to multimodal large language models (MLLMs) requires the evolving reasoning process to remain grounded in relevant visual evidence Su et al. (2025); Hu et al. (2024). However, visual inputs are typically encoded at the beginning of inference, whereas subsequent reasoning is increasingly dominated by generated text. As reasoning proceeds, critical observations may therefore be overlooked, while incorrect perceptual assumptions can propagate across multiple steps.

Existing methods address visual-grounding failures through either *post-hoc look-back* or *online visual interaction*. Look-back methods revisit the image after uncertainty or errors emerge, but incorrect visual premises may already have influenced subsequent deductions Yang et al. (2025); Zhou et al. (2026). Online approaches instead acquire or refine visual information during reasoning. They can be training-based, requiring specialized data and parameter updates Zheng et al. (2025); Hong et al. (2025); Wei et al. (2026), or training-free, controlling inference through prompting, tools, or test-time interaction. Among training-free methods, latent approaches manipulate hidden representations or continuous visual thoughts with relatively low decoding overhead Li et al. (2025); Liu et al. (2025); Zhang et al. (2025); Wang et al. (2026b), whereas explicit approaches expose intermediate evidence through selected regions, annotations, sketches, or interleaved image–text states Shao et al. (2024); He et al. (2026); Hu et al. (2024). We focus on explicit reasoning because its intermediate evidence can be inspected, visually grounded, and directly manipulated.

Importantly, recent explicit visual-reasoning methods already allow intermediate visual artifacts or observations to be revisited and reused across reasoning steps Hu et al. (2024); Gao et al. (2024). However, *persistence alone does not determine how previously acquired evidence should evolve as*

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.

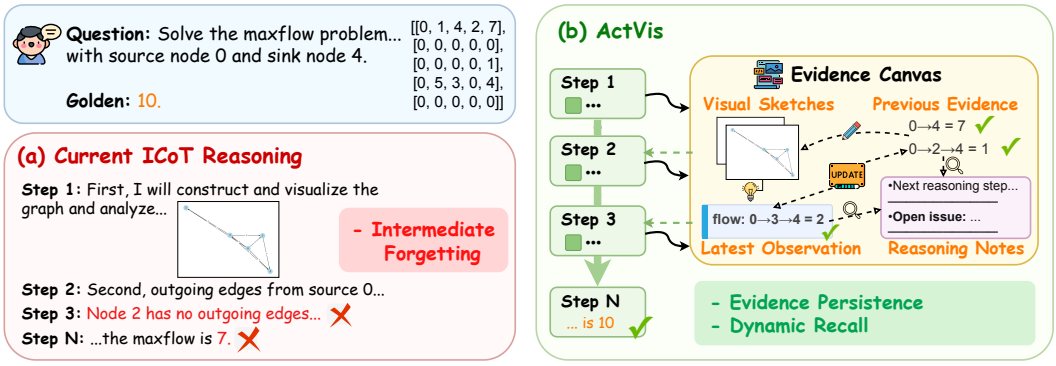


Figure 1: Comparison between current interleaved CoT and ActVis. While conventional interleaved reasoning accumulates intermediate observations along the reasoning trajectory, ActVis maintains a compact and editable evidence state that can be reused and updated across reasoning steps.

reasoning proceeds. Accumulated observations may become redundant, incomplete, outdated, or mutually conflicting, yet remain present in the reasoning context and continue to influence later deductions. What is missing is an explicit mechanism for managing the *current valid evidence state*: deciding which evidence should remain active, which should be revised or consolidated in light of new observations, and which should be removed once invalidated. We refer to this problem as *active visual evidence maintenance*.

To address this problem, we propose **ActVis**, a training-free framework that maintains a compact and editable visual-evidence state throughout multimodal reasoning. ActVis represents source-grounded observations in an *Evidence Canvas* and explicitly supports five evidence-lifecycle operations: ADD, RETAIN, REVISE, MERGE, and DELETE. At each reasoning step, newly acquired observations are evaluated relative to the current canvas rather than preserved indiscriminately. An entropy-guided gate estimates their marginal predictive utility, while a Step Audit verifies whether the selected evidence is visually supported before updating the canvas. The resulting evidence state is reinjected into subsequent reasoning, allowing frozen MLLMs to reuse established observations, correct outdated premises, and maintain a bounded set of task-relevant evidence without modifying model parameters.

We evaluate ActVis across multiple MLLMs and diverse multimodal reasoning tasks. ActVis consistently improves reasoning accuracy over conventional CoT and training-free visual-reasoning baselines. Ablation and evidence-level analyses further show that editable evidence maintenance and adaptive evidence selection play complementary roles in maintaining a reliable reasoning state.

Our contributions are summarized as follows:

- We identify an underexplored problem in training-free multimodal reasoning: persistent visual evidence should not only be accumulated or revisited, but actively maintained as a mutable state that can be revised, consolidated, or removed as reasoning evolves.
- We introduce **ActVis**, a training-free framework that organizes source-grounded observations in a bounded and editable Evidence Canvas and actively updates the maintained evidence state through utility-guided selection and visual verification.
- Extensive experiments across multiple MLLMs and diverse reasoning tasks demonstrate consistent improvements over visual-reasoning baselines, while ablations reveal the complementary effects of editable evidence maintenance and adaptive evidence selection.

2 MULTIMODAL CHAIN-OF-THOUGHT REASONING

The paradigm of Chain-of-Thought (CoT) reasoning Wei et al. (2022); Kojima et al. (2022) has fundamentally enhanced the reasoning capabilities of large language models (LLMs). Early extensions to multimodal contexts primarily relied on text-centric reasoning conditioned on static image inputs Zhang et al. (2024). However, as reasoning chains lengthen, these models often exhibit a

progressive degradation of visual groundingWu et al. (2025); Yang et al. (2025). For instance, temporal visual forgetting has been identified as a key bottleneck, prompting efforts to counteract it via visually-guided policy optimization Wang et al. (2026c). To encourage models to actively engage with visual information, reinforcement learning (RL) has been adopted to improve multimodal reasoning, including self-evolving frameworks that enable VLMs to autonomously improve from unlabeled images He et al. (2026). Moving beyond text-only reasoning, recent research advocates for externalized multimodal reasoning Su et al. (2025), where intermediate visual artifacts are explicitly generated. For instance, models are prompted to draw on visual sketchpads Hu et al. (2024), dynamically interleave cropped visual tokens into reasoning steps Chen et al. (2025), or learn multimodal tool via RL to perform explicit visual editing Wu et al. (2026). Nevertheless, explicit image generation and editing are plagued by generative fragility and high computational costs. To circumvent pixel-level synthesis, another trajectory represents intermediate reasoning states in continuous latent spacesZhang et al. (2025). Recent works explore reasoning with continuous thought tokens in a latent space Hao et al. (2025), while Monet Wang et al. (2026a) and Render-of-Thought Wang et al. (2026b) extend this idea to the visual domain by encoding visual rationales via continuous embeddings or rendering textual CoT as latent visual representations. While computationally efficient, these latent states sacrifice human interpretability and external verifiability, making structured intervention and explicit reuse considerably more difficult.

3 ACTVIS: ACTIVE VISUAL EVIDENCE MAINTENANCE

3.1 OVERVIEW

ActVis is a training-free framework that augments a frozen multimodal large language model (MLLM) with an editable *Evidence Canvas*. Rather than allowing previously acquired observations to accumulate indiscriminately, ActVis maintains a bounded set of task-relevant visual and semantic evidence whose contents can evolve as reasoning proceeds.

ActVis follows an iterative REASON–ACQUIRE–SELECT–MAINTAIN loop. At reasoning step t, the model conditions on the task input and the current canvas to determine what information is still required. When proposing a visual tool action, the model also generates a *self-assessed tool-use confidence score* that reflects its estimated necessity for acquiring additional visual information. Only actions whose confidence exceeds a threshold are executed, preventing unnecessary visual operations when the current evidence state is already sufficient. The resulting observations are then evaluated for predictive utility and validated before modifying the maintained evidence state:

$$\begin{aligned}
(h_t, \mathcal{A}_t, \mathbf{s}_t) &= \mathrm{Reason}_\theta\big(x, \mathrm{Inject}(W_{t-1})\big), \\
\widetilde{\mathcal{A}}_t &= \{a_{t,j} \in \mathcal{A}_t \mid s_{t,j} \geq \tau_A\}, \\
\mathcal{C}_t &= \mathrm{Execute}\big(\widetilde{\mathcal{A}}_t; I, W_{t-1}\big), \\
c_t^* &= \mathrm{Gate}_H\big(x, W_{t-1}, \mathcal{C}_t\big), \\
\rho_t &= \mathrm{Audit}\big(x, W_{t-1}, h_t, c_t^*\big), \\
W_t &= \mathrm{Maintain}\big(W_{t-1}, c_t^*, \rho_t\big).
\end{aligned} \tag{1}$$

Here, h_t denotes the current reasoning state, $\mathcal{A}_t$ the proposed tool actions, $\mathbf{s}_t = \{s_{t,j}\}$ their associated tool-use confidence scores, and $\widetilde{\mathcal{A}}_t$ the actions that pass the invocation threshold τ_A. $\mathcal{C}_t$ denotes the resulting candidate evidence, c_t^* the candidate selected by the entropy-guided gate, and ρ_t the Step Audit.

ActVis therefore applies two complementary controls over visual interaction. A self-assessed tool-use confidence score acts as a *pre-acquisition invocation filter*, suppressing low-necessity tool calls before execution. This score is generated directly by the reasoning model and is used only as an invocation criterion rather than as a calibrated probability. In contrast, the entropy-guided gate operates *after acquisition*, determining whether newly obtained evidence provides sufficient predictive utility to enter the maintained state. ActVis uses vision tools for evidence acquisition and refinement, including localization, cropping, zooming, highlighting, and annotation, together with canvas operations for retrieving and updating maintained evidence.

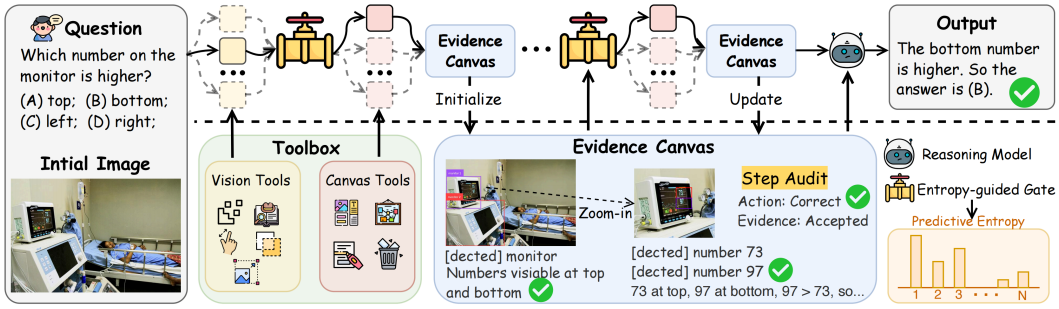


Figure 2: Overview of ActVis. The reasoning model acquires candidate visual evidence, selects informative observations through an entropy-guided gate, validates them with a Step Audit, and updates the Evidence Canvas. The maintained evidence state is reinjected into subsequent reasoning steps.

An iteration may leave the canvas unchanged when no proposed tool action passes the invocation threshold, when no acquired candidate provides sufficient predictive utility, or when the selected evidence is rejected by the Step Audit. Inference terminates when the model judges the maintained evidence sufficient, no further informative action can be proposed, or the interaction budget is exhausted. All operations are performed through prompting the same frozen MLLM, requiring neither parameter updates nor architectural modification.

3.2 Structured Evidence Canvas

Persistent visual evidence alone does not specify which observations remain valid or useful as reasoning evolves. ActVis therefore represents previously acquired evidence as a mutable *current state*, rather than an accumulated record of the interaction history:

$$W_t = (E_t, R_t, A_t), \tag{2}$$

where E_t contains active evidence units, R_t records their relations, and A_t stores recent audit feedback. The canvas retains only evidence that is currently considered relevant and supported for subsequent reasoning; earlier observations may be revised, consolidated, superseded, or removed as new evidence becomes available.

Source-Grounded Evidence. Each evidence unit contains a persistent identifier, evidence type, visual provenance and spatial reference, a semantic interpretation, and a reliability status. An evidence unit may retain both a visual artifact and its semantic description: the visual component preserves fine-grained perceptual information for later verification, while the semantic component provides a compact form for subsequent reasoning. The canvas further records lightweight relations such as Derived-From, Supports, Contradicts, and Same-Entity, preserving dependencies between observations and their visual sources. The complete evidence schema is provided in the Appendix.

Editable Evidence State. ActVis maintains the canvas through five operations:

$$\mathcal{O} = \{\textsc{Add}, \textsc{Retain}, \textsc{Revise}, \textsc{Merge}, \textsc{Delete}\}. \tag{3}$$

Add introduces new evidence, Retain preserves a valid unit, Revise corrects incomplete or inaccurate evidence, Merge consolidates redundant or complementary observations, and Delete removes invalidated evidence. The operation is determined relative to the current canvas rather than from the candidate alone. Consequently, incorporating a new observation does not necessarily enlarge the evidence state: it may instead revise an existing unit, merge with related evidence, or invalidate a stale premise. This distinguishes active maintenance from simply appending selected observations to persistent memory.

To prevent unbounded accumulation, the active canvas contains at most B evidence units. Invalidated, superseded, and duplicated evidence is removed first, followed by low-utility evidence when further pruning is required. The canvas is independent of its concrete serialization; we use compact

JSON by default and additionally evaluate an HTML variant with the same underlying evidence state and maintenance policy. More detials about tools can be find in A.2.

3.3 ENTROPY-GUIDED CANVAS UPDATE

The self-assessed tool-use score introduced in Section 3.1 controls whether additional visual information should be acquired, but it does not determine whether the resulting observation should persist in the evidence state. After this pre-acquisition filtering, ActVis applies an entropy-guided gate to evaluate the marginal predictive utility of each acquired candidate before it is admitted to the canvas.

Predictive Utility. Each executed visual action produces a candidate observation together with its visual artifact and preliminary semantic interpretation. Comparing candidate $c_{t,i}$ with the current canvas induces a tentative maintenance operation $\phi_{t,i}$ and a corresponding tentative state $W_{t-1} \oplus \phi_{t,i}$.

Given a fixed answer probe, the frozen MLLM defines the predictive distribution

$$p_\theta(y \mid I, q, W). \tag{4}$$

We measure its uncertainty using

$$H(W) = -\sum_{y \in \mathcal{Y}} p_\theta(y \mid I, q, W) \log p_\theta(y \mid I, q, W), \tag{5}$$

where $\mathcal{Y}$ denotes the candidate answer space. For closed-set tasks, the distribution is normalized over valid answers; for open-ended tasks, it is approximated using the top-K candidates produced by the same answer probe.

The utility of candidate i is then measured by its normalized reduction in predictive entropy:

$$\Delta H_{t,i} = \frac{H(W_{t-1}) - H(W_{t-1} \oplus \phi_{t,i})}{\max\left(H(W_{t-1}), \epsilon\right)}, \tag{6}$$

where ϵ ensures numerical stability. A positive value indicates reduced predictive uncertainty, a near-zero value suggests redundancy, and a negative value indicates that the tentative update introduces additional ambiguity.

Importantly, utility is evaluated relative to the current evidence state rather than from the candidate in isolation. The same observation may be informative when first acquired, redundant when equivalent evidence is already active, or corrective when it revises an inaccurate premise.

Selection and Validation. Among candidates whose entropy reduction exceeds a threshold τ_H, ActVis selects

$$i^* = \arg \max_{i:\Delta H_{t,i} \geq \tau_H} \Delta H_{t,i}, \qquad c_t^* = c_{t,i^*}. \tag{7}$$

If no candidate passes the threshold, the active evidence state remains unchanged. Candidates that explicitly contradict existing evidence are nevertheless forwarded to the Step Audit, since correcting a confidently incorrect premise may temporarily increase predictive uncertainty.

The entropy gate and Step Audit serve complementary roles. The gate estimates whether a candidate is *useful* for the current prediction, whereas the audit determines whether it is *visually supported and reliable*. The audit returns an ACCEPT, PARTIAL, or REJECT verdict. Accepted evidence triggers the corresponding maintenance operation, partially supported evidence retains only its verified content, and rejected evidence does not modify the active canvas.

After maintenance, the evidence budget is enforced and the resulting canvas is reinjected into the next reasoning step together with the original task input and recent audit feedback. Only the current active evidence state, rather than the complete history of previous observations and tool calls, is retained for subsequent reasoning, allowing established evidence to be directly reused, revised, or replaced.

4 EXPERIMENTS

4.1 DATASETS AND METRICS

Datasets. We evaluate our method across four reasoning domains. Following VisualSketchpad Hu et al. (2024), we use Geometry for diagram-based geometric reasoning; Graph Maxflow and Connectivity for graph-algorithm reasoning; and Winner ID for predicting the outcome of a chess position from its FEN representation. To further evaluate general multimodal reasoning, we include MathVista Lu et al. (2024), which covers mathematical reasoning over diagrams, plots, scientific figures, and real-world images; Mantis Jiang et al. (2024), which requires comparative and relational reasoning over multiple images; and MMVP Tong et al. (2024), which focuses on fine-grained visual differences that are difficult for CLIP-based visual encoders.

Metrics. *Task accuracy* is the percentage of correctly answered instances after task-specific answer normalization and matching. *Token consumption* is the average total number of input and output tokens consumed across all model calls for each instance. *Reasoning turns* is the average number of Planner iterations required to produce the final answer for each instance.

Implementation Details. We evaluate ActVis on three open-source MLLMs from two model families: Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and InternVL-3.5-8B. For training-free comparisons on the Qwen backbones, we include Visual Sketchpad and DaP-ICoT under the same task prompts, visual tools, decoding configurations, and interaction budgets. We additionally report VTool-R1-7B as a training-based visual tool-use baseline.

To isolate the effect of evidence maintenance, we construct *Append-only* variants in both JSON and HTML formats. These variants preserve accumulated intermediate evidence throughout reasoning but do not revise, merge, or delete previously stored evidence, whereas ActVis maintains an editable evidence state. The JSON and HTML variants of ActVis share the same canonical evidence state, tool-use policy, entropy-guided evidence selection, Step Audit, and maintenance procedure, differing only in their serialization format.

For local models, we use deterministic decoding with temperature 0, set `model_max_length` to 8192, and allow at most 10 agent turns. The Evidence Canvas retains at most 3 active evidence units at each reasoning step, and predictive entropy is computed from the final-layer next-token logits.

4.2 MAIN RESULTS

Table 1: Comparison with representative multimodal reasoning methods across different MLLM backbones. ActVis is applied to frozen MLLMs without additional training. Best and second-best results under each backbone are shown in **bold** and underline, respectively. Accuracy (%).

Backbone	Method	Geometry	Graph		Game	General		
		Geometry	Maxflow	Connectivity	Winner ID	MathVista	Mantis	MMVP
Qwen2.5-VL-7B	Base ModelBai et al. (2025b)	0	15.62	53.91	29.57	27.60	31.34	34.33
	Visual SketchpadHu et al. (2024)	16.67	22.66	69.53	49.42	23.90	32.72	47.00
	DaP-ICoTLiu et al. (2026)	2.08	3.13	33.59	26.46	37.28	34.53	48.67
	VTool-R1Wu et al. (2026)	45.83	**95.28**	**98.43**	**65.62**	43.60	**64.81**	**75.33**
	ActVis (JSON)	45.83	92.19	97.66	41.63	45.00	53.46	58.33
	ActVis (HTML)	**47.92**	91.34	**98.43**	39.69	**48.00**	57.14	57.00
Qwen3-VL-8B	Base ModelBai et al. (2025a)	0	46.09	95.31	34.24	67.8	42.40	50.33
	Visual SketchpadHu et al. (2024)	31.25	65.62	92.97	36.19	37.60	66.82	68.00
	DaP-ICoTLiu et al. (2026)	37.50	98.44	94.53	21.01	**73.20**	56.84	**75.33**
	ActVis (JSON)	60.42	**100.00**	**100.00**	52.92	70.10	74.54	68.67
	ActVis (HTML)	**62.50**	**100.00**	**100.00**	**56.42**	69.80	**77.78**	66.33
InternVL3.5-8B	Base ModelWang et al. (2025)	54.17	21.09	80.47	24.90	**78.40**	70.05	71.67
	ActVis (JSON)	35.98	93.75	99.22	52.31	71.50	69.50	70.67
	ActVis (HTML)	**56.25**	**99.22**	**100.00**	**56.25**	76.00	**78.80**	**74.67**

Overall Performance. Table 1 compares ActVis with direct inference and representative multimodal reasoning methods across three MLLM backbones. In addition to Qwen2.5-VL-7B and

Table 2: Controlled comparison between append-only evidence accumulation and active evidence maintenance. Both settings preserve previously acquired evidence, while only ActVis explicitly revises, merges, or removes existing evidence as reasoning evolves. Δ denotes the absolute accuracy improvement of ActVis over the corresponding append-only variant in percentage points (pp).

Backbone	Evidence Strategy	Geometry	Graph		Game	General		
		Geometry	Maxflow	Connectivity	Winner ID	MathVista	Mantis	MMVP
Qwen2.5-VL-7B	Append-only (JSON)	31.25	6.25	55.47	22.57	10.80	14.29	13.00
	ActVis (JSON)	45.83	92.19	97.66	41.63	45.00	53.46	58.33
	Δ Gain (JSON)	**+14.58**	**+85.94**	**+42.19**	**+19.06**	**+34.20**	**+39.17**	**+45.33**
	Append-only (HTML)	18.75	16.41	53.91	24.90	40.70	46.54	17.67
	ActVis (HTML)	47.92	82.03	98.44	39.69	48.00	57.14	57.00
	Δ Gain (HTML)	**+29.17**	**+65.62**	**+44.53**	**+14.79**	**+7.30**	**+10.60**	**+39.33**
Qwen3-VL-8B	Append-only (JSON)	37.50	84.38	90.62	29.96	20.40	67.74	43.00
	ActVis (JSON)	60.42	100.00	100.00	52.92	70.10	74.54	68.67
	Δ Gain (JSON)	**+22.92**	**+15.62**	**+9.38**	**+22.96**	**+49.70**	**+6.80**	**+25.67**
	Append-only (HTML)	37.50	80.47	82.03	26.46	21.00	66.82	59.69
	ActVis (HTML)	62.50	100.00	100.00	56.42	69.80	77.78	66.33
	Δ Gain (HTML)	**+25.00**	**+19.53**	**+17.97**	**+29.96**	**+48.80**	**+10.96**	**+6.64**

Qwen3-VL-8B, we evaluate ActVis on InternVL3.5-8B to examine whether its gains transfer across model families. Overall, ActVis achieves strong and consistent improvements on most tasks, with particularly large gains on geometry, graph reasoning, and game-state analysis. On Qwen3-VL-8B, ActVis reaches 100% accuracy on both graph tasks and substantially improves Geometry, Winner ID, and Mantis. The improvements also transfer to InternVL3.5-8B: the HTML variant improves over the base model on six of the seven tasks, including large gains on Maxflow ($21.09 \rightarrow 99.22$), Connectivity ($80.47 \rightarrow 100.00$), and Winner ID ($24.90 \rightarrow 56.25$). These results indicate that active evidence maintenance is not specific to the Qwen architecture.

We further compare against VTool-R1-7B, a training-based visual tool-use model, on Qwen2.5-VL-7B. Despite requiring no additional model training, ActVis remains competitive with the fine-tuned model, matching or exceeding it on Geometry, Connectivity, and MathVista, although VTool-R1 retains an advantage on several other tasks. This comparison suggests that explicit inference-time evidence maintenance can recover part of the benefit otherwise obtained through specialized tool-use training. DaP-ICoT, in contrast, requires memory-intensive visual-feature computation during interleaved reasoning and is more prone to memory constraints on multi-image inputs in our implementation. This limitation is particularly apparent on Mantis, where ActVis substantially outperforms DaP-ICoT on both Qwen backbones while maintaining a bounded evidence state.

An apparent exception occurs on Winner Identification with Qwen2.5-VL-7B, where Visual Sketchpad obtains the second-highest accuracy. However, trajectory inspection reveals that it predicts a White win for almost all examples, while the dataset contains nearly the same proportion of White win instances. This result therefore mainly reflects label distribution bias rather than effective game-state reasoning.

Effect of Active Evidence Maintenance. Table 2 directly compares ActVis with an append-only setting that preserves acquired evidence throughout reasoning without actively revising, merging, or removing existing observations. Across both Qwen backbones and both serialization formats, active maintenance produces substantial accuracy improvements over append-only accumulation, with particularly large gains on multi-step reasoning tasks such as Graph and MathVista.

We attribute these improvements to two effects. First, indiscriminately accumulating intermediate observations causes the reasoning context to grow rapidly. Under our maximum context length of 8192 tokens, append-only trajectories are more likely to exhaust the available context budget, preventing some reasoning processes from being completed normally. Second, even when the context limit is not reached, retaining long sequences of redundant, outdated, or conflicting observations increases contextual interference and can dilute the influence of task-critical visual evidence. ActVis instead maintains a compact current evidence state by selectively retaining useful observations and revising, merging, or removing evidence as reasoning evolves. The results therefore suggest that

effective visual reasoning benefits not only from persistent access to previously acquired evidence, but also from actively controlling which evidence remains in the reasoning context.

Effect of Representation Format. JSON and HTML exhibit complementary characteristics as evidence representations. JSON provides a compact and regular structure with relatively low contextual overhead, which can be advantageous for symbolic or graph-structured reasoning. HTML, in contrast, provides a more flexible representation for organizing multimodal evidence, allowing visual cues such as highlighting, auxiliary lines, spatial grouping, and hierarchical structure to be expressed more naturally.

The results in Table 1 suggest that this additional expressiveness becomes increasingly useful when the backbone has sufficient multimodal reasoning capacity. On Qwen3-VL-8B, HTML improves over JSON on Geometry, Winner ID, and Mantis while matching it on both graph tasks, and on InternVL3.5-8B the HTML variant consistently outperforms JSON across all evaluated tasks. In contrast, on the weaker Qwen2.5-VL-7B backbone, the two formats show more mixed behavior: the compact JSON representation remains competitive and outperforms HTML on several tasks, including Maxflow, Winner ID, and MMVP. These observations suggest a trade-off between representational richness and contextual simplicity: HTML can better preserve and organize complex visual evidence when the backbone is capable of exploiting such structure, whereas the more concise JSON format may be easier to utilize when model capacity is more limited.

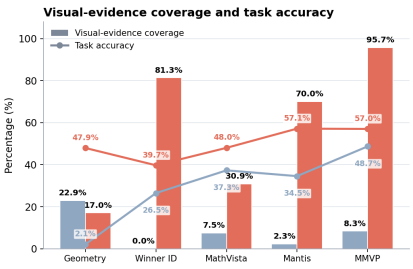


Figure 3: Visual-evidence coverage and task accuracy of DaP-ICoT and ActVis on Qwen2.5-VL-7B-Instruct.

Table 3: Component ablation of structured evidence and entropy-guided selection on Qwen3-VL-8B-Instruct. Accuracy (%).

Components		Task Category			
Structured	Entropy Gate	Geometry	Graph	Game	General
✗	✗	31.25	79.45	36.19	47.79
✓	✗	37.50	87.50	29.96	31.64
✗	✓	35.42	74.22	28.02	53.85
✓	✓	**62.50**	**100.00**	**56.42**	**70.25**

Visual Evidence Maintaince. DaP-ICoT performs only sparse visual insertion on most tasks, whereas ActVis maintains substantially more effective visual evidence on Winner Identification, MathVista, Mantis, and MMVP. This difference reflects their distinct mechanisms: DaP-ICoT introduces visual information only when an additional visual operation is triggered, while ActVis preserves useful observations in the Evidence Canvas for reuse throughout subsequent reasoning.

Figure 3shows that higher visual-evidence coverage is generally associated with stronger task accuracy, particularly on Winner Identification, Mantis, and MMVP. Without explicit maintenance, previously acquired visual information may lose influence as the reasoning context grows, leading to incomplete grounding or repeated visual acquisition. ActVis instead keeps task-relevant observations accessible across steps, reducing evidence loss and supporting more stable reasoning. Geometry further indicates that evidence quality and task relevance are more important than quantity alone, since a small amount of targeted evidence is sufficient to produce a substantial performance gain.

4.3 Analysis and Ablation Studies

Token Consumption. Table 4 evaluates the effect of the entropy-guided gate on inference efficiency. Compared with ungated variants, the entropy gate substantially reduces both token consumption and reasoning turns on most task categories. For example, it achieves considerable reductions on Graph and General tasks for both JSON and HTML representations, showing that active evidence maintenance effectively filters redundant observations and avoids repeated visual inspection, unnecessary reflection, and post-hoc look-back. The consistent reductions across both serialization formats further indicate that the efficiency gains arise from the maintenance mechanism rather than a specific representation choice.

Geometry and Game show a different trend, where the gate introduces slightly higher token usage and reasoning turns. This is because these tasks require richer visual reasoning and more fine-grained evidence. Instead of aggressively compressing the context, ActVis preserves additional task-relevant visual observations in the Evidence Canvas, leading to stronger visual grounding and improved reasoning accuracy. The additional computation therefore reflects deliberate evidence acquisition rather than uncontrolled context expansion. Overall, these results demonstrate that the entropy gate adaptively balances efficiency and evidence sufficiency rather than simply minimizing context length.

Table 4: Ablation study of the entropy-guided evidence gate using Qwen3-VL-8B as the base model. ActVis w/o Gate performs evidence maintenance without adaptive evidence selection, while ActVis incorporates entropy-guided filtering. The gate reduces redundant evidence updates and improves inference efficiency, especially on Graph and General reasoning tasks. All averages are weighted by the number of valid examples.

Method	Geometry		Graph		Game		General	
	Tok.↓	Turns↓	Tok.↓	Turns↓	Tok.↓	Turns↓	Tok.↓	Turns↓
Base Model	30,124	4.58	23,539	3.46	8,750	2.09	16,952	2.69
ActVis w/o Gate (JSON)	37,080	3.05	46,406	3.34	71,997	4.34	64,216	4.03
ActVis (JSON)	42,653 (↑ 15.0%)	3.17 (↑ 3.9%)	20,020 (↓ 56.9%)	2.01 (↓ 39.8%)	77,231 (↑ 9.4%)	4.55 (↑ 5.5%)	35,785 (↓ 44.3%)	2.52 (↓ 37.5%)
ActVis w/o Gate (HTML)	24,839	2.26	38,312	3.15	67,560	4.13	39,783	2.93
ActVis (HTML)	25,670 (↑ 3.3%)	2.31 (↑ 2.2%)	19,526 (↓ 49.0%)	2.00 (↓ 36.5%)	93,940 (↑ 39.0%)	4.78 (↑ 19.7%)	35,370 (↓ 11.1%)	2.50 (↓ 14.7%)

Percentages indicate relative changes compared with the corresponding ActVis w/o Gate variant. Green and red indicate reduced and increased inference cost, respectively.

Effect of Structured Evidence Maintenance and Entropy-Guided Selection. Table 3 studies the contribution of structured evidence representation and entropy-guided selection using Qwen3-VL-8B. The setting without either component corresponds to the original VisualSketchpad-style interleaved image–text reasoning process, while the complete setting with both components represents ActVis.

Adding only the structured representation provides limited and inconsistent improvements. Although organizing the reasoning trajectory into a structured format can benefit some tasks, it may also introduce degradation on others, indicating that simply preserving intermediate observations is insufficient. Without evidence selection, redundant and low-utility information remains in the reasoning context, increasing interference and reducing attention to task-critical evidence.

In contrast, combining the structured Evidence Canvas with the entropy-guided gate consistently achieves the best performance across all task categories. These results demonstrate that the two components are complementary: the Evidence Canvas provides an explicit state for organizing and revising visual evidence, while the entropy-guided gate controls which observations should be retained. Together, they enable proactive evidence maintenance instead of merely recording the reasoning trajectory.

5 Conclusion

In this work, we propose **ActVis**, a training-free framework for active visual evidence maintenance in multimodal reasoning, which extends visual interaction from evidence acquisition to evidence selection, revision, and preservation through a structured Evidence Canvas and an entropy-guided gate. Extensive evaluations on multiple multimodal reasoning benchmarks and advanced MLLMs demonstrate that ActVis consistently improves reasoning accuracy while reducing redundant visual acquisition and context growth. Further analyses show that structured evidence organization and adaptive selection are complementary: maintaining an explicit evidence state enables reliable reuse of intermediate observations, while entropy-guided filtering prevents ineffective accumulation. We believe ActVis provides a general inference-time paradigm for building more efficient, interpretable, and robust multimodal reasoning systems.

6 AI USE STATEMENT

Generative Al tools were used solely to assist with language editing, clarity, and formatting duringmanuscript preparation, They were not used to conduct cxperiments, derive scientific results, or makrescarch decisions, "The authors revicwed all Al-assisted content and remain fully responsible for thesicntife content,analysis, and conclusions

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. MINT-CoT: Enabling interleaved visual tokens in mathematical Chain-of-Thought reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. *arXiv preprint arXiv:2411.19488*, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Conference on Language Modeling (COLM)*, 2025.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. VisPlay: Self-evolving vision-language models from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual Sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:22199–22213, 2022.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought, 2025. URL https://arxiv.org/abs/2501.07542.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyu Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space, 2025. URL https://arxiv.org/abs/2512.12623.

Xu Liu, Yongheng Zhang, Qiguang Chen, Yao Li, Sheng Wang, and Libo Qin. Let's think with images efficiently! an interleaved-modal chain-of-thought reasoning framework with dynamic and precise visual thoughts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32213–32221, 2026.

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 (ICLR)*, 2024.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Zhaochen Su, Peng Xia, Hangyu Guo, et al. Thinking with images for multimodal reasoning: Foundations, Methods, and Future Frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, June 2024.

Qixun Wang, Yang Shi, Yuanxing Zhang, Pengfei Wan, Kun Gai, Yifei Wang, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yifan Wang, Shiyu Li, Peiming Li, Xiaochen Yang, Zheng Wei, and Yang Tang. Render-of-Thought: Rendering textual Chain-of-Thought as images for visual latent reasoning. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2026b.

Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xiangxiang Chu. Visually-Guided policy optimization for multimodal reasoning. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2026c.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-Thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 24824–24837, 2022.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. VTool-R1: VLMs learn to think with images via reinforcement learning on multimodal tool use. In *International Conference on Learning Representations (ICLR)*, 2026.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models, 2025. URL https://arxiv.org/abs/2503.12799.

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

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, Tieniu Tan, and Furu Wei. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms, 2025. URL https://arxiv.org/abs/2510.24514.

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 (TMLR)*, 2024.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Jiazhou Zhou, Yucheng Chen, Hongyang Li, Qing Jiang, Hu Zhou, Ying-Cong Chen, and Lei Zhang. V-reflection: Transforming mllms from passive observers to active interrogators, 2026. URL https://arxiv.org/abs/2604.03307.