Active Visual Reasoning as Sequential Bayesian Optimal Experimental Design

Overview

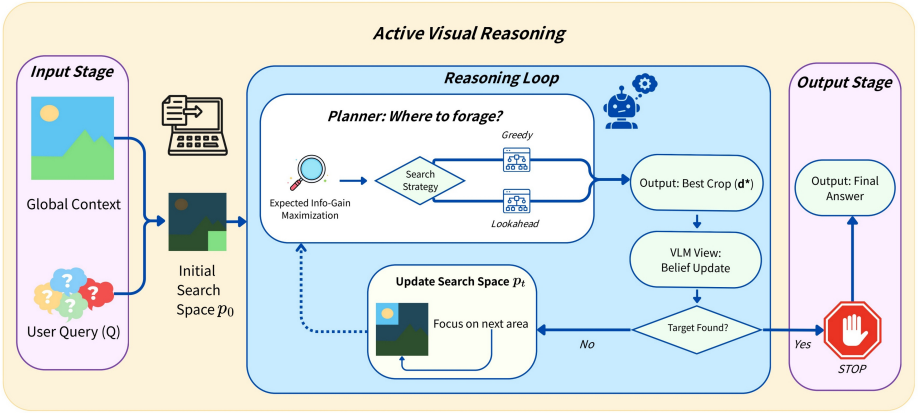

Figure: The S-BOED Framework for active visual reasoning. Given a global context and query, the agent iteratively optimises its foveation design $\mathbf{d}^*$ by maximising expected information gain. The loop continues by implicitly updating the search space p_t using interaction history until the semantic target is resolved.

Motivation: why do we need active visual reasoning?

Core issue

A VLM has a fixed visual token budget. On a very large image, a global view gives broad coverage but loses the fine details needed for OCR, small objects, and precise localization.

- A wide crop helps answer *where to look*, but often cannot resolve the target.
- A tight crop helps answer *what it is*, but may miss the target entirely.
- So the problem is naturally sequential: we should decide **where to zoom next** in order to reduce uncertainty as efficiently as possible.

Main idea

We formulate this as **sequential Bayesian optimal experimental design (S-BOED)**: each crop is an experiment, and the next crop should maximize expected information gain.

Formulation: from EIG to a simple objective

Let d be a crop, ℓ the latent target location, and y the semantic answer.

Ideal objective

$$\mathrm{EIG}(d) = I(z; \ell, y \mid d)$$

where z is the observation after executing crop d.

Exact EIG is hard to compute, so we use a tractable proxy:

$$J(d) = \underbrace{\int_{x \in d} p_t(x)\, dx}_{\text{coverage}} \times \underbrace{\phi(d)}_{\text{resolution}}$$

- **Coverage**: how likely the target is inside this crop.
- **Resolution**: how likely this crop is detailed enough to resolve the target.
- Interpretation: Good crops are not just large or just sharp. They sit at the right trade-off between **finding** the target and **resolving** it.

How we instantiate it in practice

1. Start from the global image and query.
2. Ask the VLM for an initial region proposal.
3. Sample a few nearby candidate crops.
4. Probe each candidate with a simple resolvability question, e.g. whether this crop is likely to answer the query.
5. Pick the best crop greedily, or do one-step look-ahead when the task is more sequential.

Why the loop works

After seeing an uninformative crop, we treat it as **negative evidence**: probability mass should move away from the explored region and toward unexplored regions.

$$p_{t+1}(\ell) \propto p(z_t \mid \ell, d_t)\, p_t(\ell)$$

- In the paper this belief is represented implicitly through the interaction history in context.
- A practical active search loop without maintaining an explicit dense posterior grid.

Remote sensing experiment: the clearest stress test

Why this setting matters

Remote sensing images are huge and targets are sparse, so the perceptual bandwidth bottleneck is especially severe.

Method	Accuracy (%)
Base	38.4
ReAct	45.1
Naive Sampling	47.6
MCMC	49.1
Look-ahead	**54.7**
Oracle	68.0

- Even simple BOED-style search beats the heuristic ReAct baseline.
- Look-ahead helps most in the gigapixel regime, where information is sparse and myopic zooming is not enough.