Entropy-based Visual Re-perception Inference for Multimodal Models

Supplementary Material

1. More Details of EVRI

1.1. Why Entropy?

EVRI relies on token-level and answer-level entropy as a proxy for model uncertainty. This choice is motivated by both prior theoretical and empirical evidence, as well as our own observations in preliminary experiments with other MLLM reasoning frameworks.

First, from a probabilistic perspective, entropy directly characterizes the dispersion of the predictive distribution. For multiple-choice answers, a peaky distribution with one dominant option corresponds to low answer entropy H_{ans} and a large logit margin C, which are commonly used as confidence indicators in calibration and selective prediction. Conversely, a flatter distribution leads to higher H_{ans} and smaller margins, reflecting ambiguity among candidate answers. Second, recent work has shown that *high-entropy tokens* in the generation process are particularly informative for reasoning. "Beyond the 80/20 Rule" demonstrates that minority, high-entropy tokens play a disproportionate role in driving effective reinforcement learning for LLM reasoning, and that intervening on these tokens can significantly influence final performance [3]. This suggests that not all tokens contribute equally to the reasoning trajectory: segments with elevated token-level entropy tend to correspond to hard decision points, implicit branching, or unstable hypotheses. Empirically, a number of studies have observed a positive correlation between low entropy (or high top-1 logit) and answer correctness in chain-of-thought style decoding, and have used entropy or confidence-based criteria to guide self-consistency, verification, or early stopping in text-only LLMs. In our own preliminary experiments on related test-time reasoning setups, we also observed that tokens around incorrect or unstable reasoning steps typically exhibit higher entropy than tokens in clearly correct segments.

Based on these considerations, EVRI combines answer-level entropy H_{ans} and thinking-level entropy $\widehat{H}_{\text{cot}}$ to form a unified uncertainty measure

$$U = \lambda\, H_{\text{ans}} + (1-\lambda)\, \widehat{H}_{\text{cot}}, \tag{1}$$

which is used both to trigger visual re-perception and to select high-entropy windows along the CoT for multimodal intervention.

1.2. Uncertainty Estimation and Window Extraction

In the main paper, we define answer-level quantities such as the option entropy H_{ans} and the logit margin C, as well as token-level entropies $\{h_t\}_{t=1}^{T}$ and the aggregated thinking uncertainty $\widehat{H}_{\text{cot}}$. In EVRI, these signals play complementary roles: H_{ans} and C serve as coarse-grained confidence indicators for early exit and rollback, while token-level entropy is used to *localize* the most uncertain span in the Chain-of-Thought (CoT), which then guides where multimodal intervention should happen.

Given a generated CoT sequence $y_{1:T}$ and its token-wise entropies $\{h_t\}$, we slide a fixed-length window of size L over the sequence with stride s (we use a fixed stride of $s = 10$ in all experiments). For each window $w = [t, t+L-1]$, we compute its average entropy

$$\bar{h}(w) = \frac{1}{L} \sum_{i=t}^{t+L-1} h_i, \tag{2}$$

and rank all windows according to $\bar{h}(w)$. The indices of the Top-k windows

$$\mathcal{S} = \text{TopK}_{\text{win}}\big(\{\bar{h}(w)\}, k\big) \tag{3}$$

are treated as high-entropy windows, from which we extract entity-centric phrases and focus cues for region selection. Unless otherwise stated, we set $k = 1$ so that EVRI concentrates on the single most uncertain segment in the CoT. In preliminary trials, using $k > 1$ brought limited additional gains while increasing the number of visual calls and implementation complexity, so we adopt $k = 1$ as a simple and effective default. Likewise, we did not observe consistent benefits from varying the stride, and therefore keep $s = 10$ as a stable choice that provides sufficient coverage with moderate computational cost.

We empirically explored several window lengths and found that a moderate scale works best in practice. When the window is too short (e.g., $L = 20$), the selected window often cuts through meaningful phrases or clauses, producing fragmented tokens that are hard to map to concrete visual referents. When the window is too long (e.g., $L \geq 80$), the window tends to mix multiple reasoning steps and object mentions, making the extracted targets overly diffuse and less informative for zooming. Based on these observations, we adopt a fixed window length of $L = 40$ in all experiments. This setting provides a good trade-off between locality and coherence: the high-entropy window is long enough to capture a semantically coherent reasoning fragment, yet focused enough to highlight the specific part of the CoT where EVRI should request additional visual evidence.

Algorithm 1 Uncertainty Estimation and High-Entropy Window Extraction

Require: image I, question q, options $\mathcal{O}$, MLLM f; window length L, stride s (default L=40, s=10).

// Step 1: Initial CoT reasoning and answer
$(y_{1:T}, \ell^*) \leftarrow \textsc{GenerateCoT}(f, I, q, \mathcal{O})$
obtain answer distribution $p_\theta(\ell \mid I, q)$ over $\mathcal{O}$
compute logits $\{z_\ell\}_{\ell \in \mathcal{O}}$
$C \leftarrow z_{\max} - z_{2\text{nd}}$ ▷ logit margin
$H_{\text{ans}} \leftarrow -\sum_{\ell \in \mathcal{O}} p_\theta(\ell \mid I, q) \log p_\theta(\ell \mid I, q)$

// Step 2: Token-level entropies along the CoT
for $t = 1$ **to** T **do**
  get next-token distribution $p_t(v) = p(y_t{=}v \mid y_{<t}, I, q)$
  $h_t \leftarrow -\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v)$
end for

// Step 3: Sliding-window aggregation
$\mathcal{W} \leftarrow \emptyset$
for $t = 1$ **to** $T{-}L{+}1$ **step** s **do**
  $w \leftarrow [t, t{+}L{-}1]$
  $\bar{h}(w) \leftarrow \frac{1}{L} \sum_{i=t}^{t+L-1} h_i$
  $\mathcal{W} \leftarrow \mathcal{W} \cup \{(w, \bar{h}(w))\}$
end for
select Top-k windows by $\bar{h}(w)$ (default k=1): $\mathcal{S} \leftarrow \text{TopK}_{\text{win}}(\mathcal{W}, k)$

// Step 4: Thinking-level and combined uncertainty
$\widehat{H}_{\text{cot}} \leftarrow \frac{1}{\sum_{s \in \mathcal{S}} |s|} \sum_{s \in \mathcal{S}} \sum_{t \in s} h_t$
$U \leftarrow \lambda H_{\text{ans}} + (1{-}\lambda) \widehat{H}_{\text{cot}}$

return $(\ell^*, C, H_{\text{ans}}, \widehat{H}_{\text{cot}}, U, \mathcal{S})$

Ablation on high-entropy window length. We further study the effect of the high-entropy window length L on EVRI with Qwen2.5-VL-7B on HR-Bench-4K. As shown in Tab. 1, using a short window ($L = 20$) yields 77/55/66 on FSP/FCP/Overall, while a moderate window ($L = 40$) improves the scores to 79/61/70. Increasing the window length to $L = 80$ leads to a drop to 74/57/65.5. These results are consistent with the qualitative examples in Fig. 1: very short windows tend to produce fragmented windows that are hard to ground visually, and very long windows mix multiple reasoning steps and relations, making the zoom targets diffuse. A moderate window length of $L = 40$ therefore offers the best trade-off between locality and coherence, and is adopted as the default configuration in EVRI.

1.3. Region Selection and Zoom Search

In the main paper, we describe how EVRI selects informative visual regions based on uncertainty. Here we briefly summarize the zoom search procedure and its connection to high-entropy windowing.

Table 1. Ablation on high-entropy window length L for EVRI with Qwen2.5-VL-7B on HR-Bench-4K.

	HR-Bench 4K		
Window length L	FSP	FCP	Overall
$L = 20$	77.0	55.0	66.0
$L = 40$	**79.0**	**61.0**	**70.0**
$L = 80$	74.0	57.0	65.5

Candidate regions from a zoom tree. For high-resolution images, we build a hierarchical "zoom tree" over the image.The root node corresponds to the full image, and each node is recursively split into smaller tiles until the patch size is small enough to be processed by the vision encoder. Each node in this tree represents a candidate region v with spatial coordinates and scale. At each reflection round, EVRI operates on a frontier of nodes that can either be zoomed in (expanded to children) or pruned, enabling adaptive exploration between global and local views.

Text-guided focus phrase builder. The text queries $\mathcal{A} = \{a_i\}$ used for region selection are produced by a dedicated *focus phrase builder* rather than simple keyword splitting. Given the question q and the high-entropy CoT snippet (Sec. 1.2), we first apply a *JSON-formatted list prompt* to the same MLLM, asking it to output a short JSON array of concrete visual targets that, if zoomed in, would help resolve the uncertainty. The prompt explicitly instructs the model to avoid generic terms ("object", "thing", "text", "number", etc.) and pronouns. The raw JSON array is then cleaned by rule-based filters that remove boilerplate, overly long/short phrases, and non-visual or generic nouns, yielding CoT-driven uncertain targets `cot_terms`. When this step fails or returns too few valid targets, we fall back to two simple heuristics. First, we use an IC-style one-sentence prompt that lists the objects needed to answer the question and parse additional noun phrases from it. Second, we apply a lightweight heuristic extractor that scans for uncertainty markers ("maybe", "uncertain", "appears", etc.) and collects nearby noun phrases.In parallel, we derive question-driven targets `base_targets` by prompting the MLLM with a few in-context examples to answer which objects' information are required to solve q, and parsing the resulting list. The final focus phrase set is obtained by merging and deduplicating these two sources,

$$\mathcal{A} = \{a_i\} = \textsc{MergeFocus}(\texttt{base\_targets}, \texttt{cot\_terms}), \tag{4}$$

with a global cap of 1–3 phrases. The main prompt templates used in this process are summarized in Tab. 2.

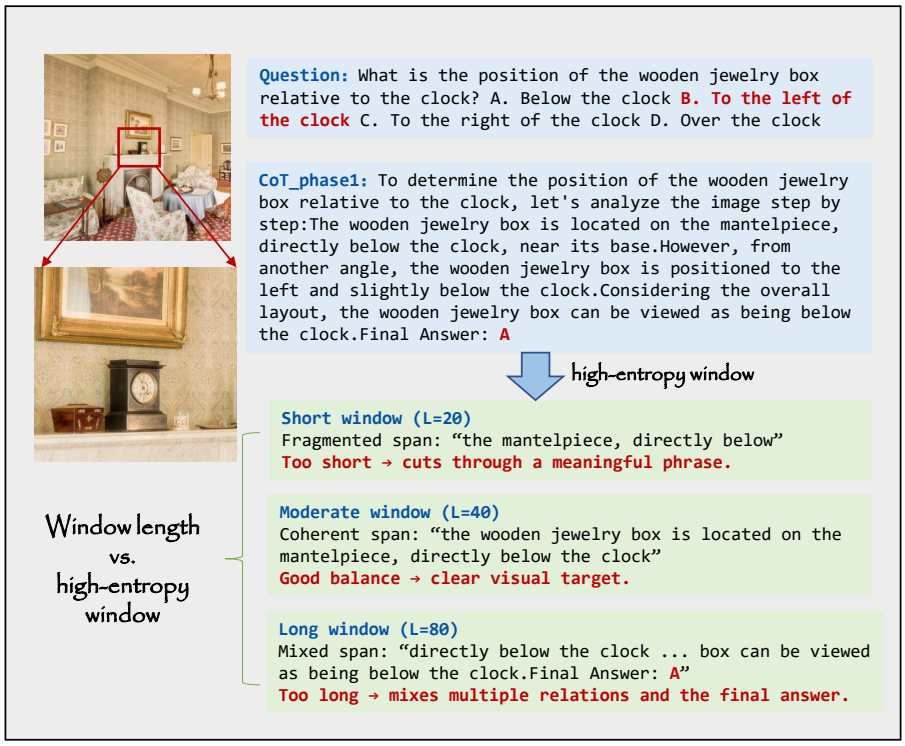

Figure 1. High-entropy windows selected under different window lengths. On the same CoT segment, a short window ($L = 20$) produces a fragmented window, a moderate window ($L = 40$) captures a coherent description with a clear visual target, and a long window ($L = 80$) mixes multiple relations and the final answer, leading to diffuse zoom targets. This illustrates why EVRI adopts a moderate window length ($L = 40$) for uncertainty localization.

As discussed in Sec. 1.2, the choice of high-entropy window length L strongly affects the quality of $\mathcal{A}$. Very short windows (e.g., L=20) tend to cut through clauses and yield fragmented, overly generic targets, so the zoom tree is often guided to fuzzy or irrelevant regions. Very long windows (e.g., L=80) mix multiple reasoning steps and objects, making the focus phrases diffuse and harder to ground. A moderate window length (L=40) more often corresponds to a semantically coherent reasoning fragment, from which the extracted targets are both specific and concentrated, as illustrated in Fig. 1 and Tab. 1. This, in turn, steers the zoom search toward truly critical local regions on HR-Bench, which is why we adopt L=40 as the default in all experiments.

Uncertainty-based scoring and patch selection. The focus phrases $\mathcal{A}$ are then used to pre-filter candidate regions in the zoom tree: for each region v, a short relevance prompt conditioned on q and $\mathcal{A}$ is used to discard patches that are unrelated to all a_i. The remaining regions are ranked, and only the top-r proceed to uncertainty-based scoring. For each such region v, we perform a lightweight forward pass by appending v as an additional view while keeping q and the CoT prefix fixed, obtaining updated confidence and uncertainty (C', U') and computing

$$\Delta C = C' - C, \qquad \Delta U = U - U'. \tag{5}$$

A region is considered beneficial if

$$\Delta C \geq \tau_{\text{gain}} \quad \vee \quad \Delta U \geq \tau_{\text{drop}}, \tag{6}$$

otherwise it is discarded; if no region is beneficial at the current scale, the search falls back to coarser nodes. Among all beneficial regions in a reflection round, we compute a composite score

$$g(v) = \alpha\, \Delta C(v) + (1 - \alpha)\, \Delta U(v), \tag{7}$$

and select the top-r patches under a global cap on the number of visual inputs (Sec. 1.5). These selected regions are then interleaved with the CoT prefix preceding the high-entropy window using the before/after prompts from the main paper, forming the augmented multimodal context for the next-stage reasoning.

1.4. Multi-round Reflection Strategy

As described in the main paper, EVRI first performs a standard CoT-style reasoning pass on the full image to obtain an

Table 2. Prompts used in EVRI for focus phrase construction and region selection.

Prompt type	Brief template description
JSON visual-target extraction	Input: question + high-entropy CoT snippet. Ask the MLLM to return *only* a JSON array of 1–4 short, concrete visual target phrases (1–3 words), avoiding generic terms ("object", "thing", "text", "number", etc.) and pronouns. Used to build `cot_terms`.
IC-style uncertain-object sentence	Input: question + uncertain CoT snippet. Ask for exactly one sentence of the form "To answer the question, I need to know the locations/details of <OBJECTS> ... So I need the information about the following objects: <obj-1> and <obj-2> ..." and parse the object list as additional targets.
Question-based target extraction	Input: question only. Prompt: "If you want to answer this question, which objects' information do you need?" and parse the listed objects as `base_targets`.
Patch relevance check	Input: question, focus phrases $\mathcal{A}$, and one image patch. Ask the MLLM whether this region contains or is closely related to any target phrase in the context of answering the question, and use the result as a relevance score for region pre-filtering.

initial answer and rationale, together with its margin $C^{(0)}$ and answer entropy $H^{(0)}_{\text{ans}}$ (and combined uncertainty $U^{(0)}$). For completeness, we here spell out the reflection loop and the associated gain criteria in more detail.If the initial run is already confident ($C^{(0)} \geq \tau_{\text{conf}}$ or $U^{(0)} \leq \tau_{\text{ent}}$), no reflection is triggered and the initial answer is returned.

Otherwise, EVRI may perform up to K entropy-guided reflection rounds. At reflection round $t \geq 1$, we start from the current reasoning state of round $t-1$, with scores $(C^{(t-1)}, H^{(t-1)}_{\text{ans}})$. We then run the region selection module to obtain a set of candidate patches. For each candidate patch v, we perform a short forward evaluation to obtain its local scores (C', H'_{ans}) and define the patch-level gain relative to the current round:

$$\Delta C(v) = C' - C^{(t-1)}, \qquad \Delta U(v) = H^{(t-1)}_{\text{ans}} - H'_{\text{ans}}. \quad (8)$$

Only patches that satisfy

$$\Delta C(v) \geq \tau_{\text{gain}} \ \text{ or } \ \Delta U(v) \geq \tau_{\text{drop}} \quad (9)$$

are considered beneficial and added to the candidate set $\mathcal{C}^{(t)}$. If $\mathcal{C}^{(t)}$ is empty, the reflection loop stops early and EVRI keeps the answer from round $t-1$.

When $\mathcal{C}^{(t)} \neq \emptyset$, we rank beneficial patches by a combined score

$$g(v) = \alpha\, \Delta C(v) + (1-\alpha)\, \Delta U(v), \quad (10)$$

select at most r patches, and inject them into the CoT at the high-entropy window identified in Sec. 1.2. The MLLM then continues reasoning with this augmented multimodal context, producing updated scores $(C^{(t)}, H^{(t)}_{\text{ans}})$. We measure the *round-level* gain relative to the previous round:

$$\Delta C^{(t)} = C^{(t)} - C^{(t-1)}, \qquad \Delta H^{(t)} = H^{(t-1)}_{\text{ans}} - H^{(t)}_{\text{ans}}. \quad (11)$$

A new reflection round $t+1$ is attempted only if the current round achieves a non-trivial gain,

$$\Delta C^{(t)} \geq \tau_{\text{gain}} \ \text{ or } \ \Delta H^{(t)} \geq \tau_{\text{drop}}, \quad (12)$$

and $t < K$. Otherwise, the loop terminates and the answer from the last executed round is taken as the *reflected* candidate.

Finally, we apply a simple do-no-harm safeguard when choosing between the reflected candidate (from round $t^\star \geq 1$) and the initial answer (round 0). We compute the global improvement of the best round with respect to the initial run:

$$\Delta C^\star = C^{(t^\star)} - C^{(0)}, \qquad \Delta H^\star = H^{(0)}_{\text{ans}} - H^{(t^\star)}_{\text{ans}}. \quad (13)$$

The reflected answer is accepted only if

$$\Delta C^\star \geq \tau_{\text{gain\_e}} \ \text{ or } \ \Delta H^\star \geq \tau_{\text{drop\_e}}, \quad (14)$$

otherwise EVRI rolls back to the initial CoT answer. This ensures that entropy-guided visual re-perception can overwrite the base MLLM only when it brings a clear, measurable improvement, and never harms a strong initial prediction.

1.5. Hyperparameters and Sensitivity

EVRI introduces a small set of hyperparameters that control (i) when to trigger visual re-perception, (ii) how uncertainty is aggregated along the CoT, and (iii) how much visual budget is allocated per reflection round.

Sensitivity to early-stop thresholds. EVRI uses two scalar thresholds $(\tau_{\text{conf}}, \tau_{\text{ent}})$ to decide whether the first CoT pass is already reliable enough to skip visual reflection. We study the sensitivity of this gating on HR-Bench-4K by varying $\tau_{\text{conf}} \in \{0.5, 0.8, 0.85, 0.9, 0.95\}$ and $\tau_{\text{ent}} \in \{0.04, 0.05, 0.06, 0.5, 0.7\}$, while keeping all other hyperparameters fixed.

The results are summarized in Tab. 3. We observe that EVRI is quite robust in a reasonable neighbourhood of our default setting $(\tau_{\text{conf}}, \tau_{\text{ent}}) = (0.9, 0.05)$: for $0.85 \leq \tau_{\text{conf}} \leq 0.95$ and $0.04 \leq \tau_{\text{ent}} \leq 0.06$, the FSP/FCP/Overall scores on HR-Bench-4K remain essentially unchanged (around

Algorithm 2 Region Selection and Zoom-based Re-perception (single round)

Require: image I, question q, options $\mathcal{O}$, MLLM f; high-entropy windows $\mathcal{S}$, focus phrases $\mathcal{A} = \{a_i\}$; initial confidence and entropy $(C^{(0)}, H^{(0)}_{\text{ans}})$; thresholds $\tau_{\text{gain}}, \tau_{\text{drop}}$; max kept patches r.

// Step 1: Candidate patches from zoom tree
build a coarse-to-fine partition of I
$\mathcal{V} \leftarrow$ PROPOSECANDIDATES(I) ▷ e.g., from a zoom tree

// Step 2: Text-guided relevance pre-filtering
for all $v \in \mathcal{V}$ **do**
  $r(v) \leftarrow$ PATCHRELEVANCE$(f, q, \mathcal{A}, v)$ ▷ Sec. 1.3
end for
$\widetilde{\mathcal{V}} \leftarrow$ top-r patches by $r(v)$

// Step 3: Quick scoring by confidence/entropy gain
$\mathcal{C} \leftarrow \emptyset$
for all $v \in \widetilde{\mathcal{V}}$ **do**
  build a short prompt with q and $\mathcal{A}$, attach patch v
  $(C', H'_{\text{ans}}) \leftarrow$ QUICKSCORE$(f, I, q, \mathcal{O}, v)$
  $\Delta C(v) \leftarrow C' - C^{(0)}$
  $\Delta U(v) \leftarrow H^{(0)}_{\text{ans}} - H'_{\text{ans}}$
  if $\Delta C(v) \geq \tau_{\text{gain}}$ **or** $\Delta U(v) \geq \tau_{\text{drop}}$ **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{v\}$
  end if
end for
if $\mathcal{C} = \emptyset$ **then**
  return `no_beneficial_patch`
end if

// Step 4: Final patch set for CoT injection
for all $v \in \mathcal{C}$ **do**
  $g(v) \leftarrow \alpha\, \Delta C(v) + (1-\alpha)\, \Delta U(v)$
end for
sort $\mathcal{C}$ by $g(v)$ in descending order
$\mathcal{R} \leftarrow$ top-r patches from $\mathcal{C}$

return $\mathcal{R}$ ▷ patches to be injected into CoT

$79/61/70$), with fluctuations within ± 0.5 points. In contrast, when the gate is made too loose (e.g., $\tau_{\text{conf}} = 0.5$ or $\tau_{\text{ent}} \geq 0.5$), many low-confidence, high-entropy answers bypass reflection, leading to a noticeable performance drop of 3–7 points in Overall accuracy. These trends indicate that our default choice $(\tau_{\text{conf}}, \tau_{\text{ent}}) = (0.9, 0.05)$ lies in a stable plateau: it is high enough to trigger reflection on genuinely uncertain cases, yet not overly strict, so the behaviour of EVRI is insensitive to small perturbations of these thresholds.

Patch selection gain thresholds. Recall that a candidate region v from the zoom tree is kept only when its local gain satisfies $\Delta C \geq \tau_{\text{gain}}$ or $\Delta U \geq \tau_{\text{drop}}$ in the current reflection round. We study the sensitivity of these thresholds on HR-Bench-4K with Qwen2.5-VL-7B by varying $(\tau_{\text{gain}}, \tau_{\text{drop}})$ around the default setting $(0.12, 0.05)$. As summarized in Tab. 4, EVRI is very stable within a small neighborhood: using $(0.10, 0.05)$, $(0.12, 0.05)$, $(0.15, 0.05)$, or pairing τ_{gain}=0.12 with $\tau_{\text{drop}} \in \{0.04, 0.06\}$ all yield the same 79/61/70 FSP/FCP/Overall accuracy on HR-Bench-4K. This suggests that as long as the thresholds require a moderate improvement in logit margin or answer entropy, the patch selection is robust.

When we set very strict thresholds, such as $(\tau_{\text{gain}}, \tau_{\text{drop}}) = (0.30, 0.30)$ or $(0.50, 0.50)$, performance drops noticeably since only a few patches satisfy the gain criterion and EVRI effectively degenerates to almost text-only CoT. Based on these observations, we adopt $\tau_{\text{gain}} = 0.12$ and $\tau_{\text{drop}} = 0.05$ as a balanced default: they are stringent enough to filter out noisy patches while still allowing a sufficient number of helpful regions to be injected into the CoT.

Table 3. Sensitivity of EVRI to early-stop thresholds $(\tau_{\text{conf}}, \tau_{\text{ent}})$ on HR-Bench-4K with Qwen2.5-VL-7B.

$(\tau_{\text{conf}}, \tau_{\text{ent}})$	FSP	FCP	Overall
(0.90, 0.04)	79.0	61.0	70.0
(0.90, 0.05)	**79.0**	**61.0**	**70.0**
(0.90, 0.06)	79.0	61.0	70.0
(0.90, 0.50)	75.0	56.0	65.5
(0.90, 0.70)	73.0	52.0	62.5
(0.85, 0.05)	79.0	61.0	70.0
(0.80, 0.05)	79.0	61.0	69.5
(0.95, 0.05)	79.0	61.0	70.0
(0.50, 0.05)	71.0	52.0	61.5

Table 4. Sensitivity of patch selection thresholds $(\tau_{\text{gain}}, \tau_{\text{drop}})$ on HR-Bench-4K with Qwen2.5-VL-7B.

$(\tau_{\text{gain}}, \tau_{\text{drop}})$	FSP	FCP	Overall
(0.10, 0.05)	79.0	61.0	70.0
(0.12, 0.05)	**79.0**	**61.0**	**70.0**
(0.15, 0.05)	79.0	61.0	70.0
(0.12, 0.04)	79.0	61.0	70.0
(0.12, 0.06)	79.0	61.0	70.0
(0.30, 0.30)	75.0	52.0	63.5
(0.50, 0.50)	69.0	50.0	59.5

Combined uncertainty U. For token-level uncertainty, we define a combined score

$$U = \lambda\, H_{\text{ans}} + (1 - \lambda)\, \widehat{H}_{\text{cot}},$$

where H_{ans} is the answer entropy and $\widehat{H}_{\text{cot}}$ is the aggregated thinking uncertainty over high-entropy windows. In prac-

Table 5. Sensitivity of global rollback thresholds $(\tau_{\text{gain\_e}}, \tau_{\text{drop\_e}})$ on HR-Bench-4K.

$(\tau_{\text{gain\_e}}, \tau_{\text{drop\_e}})$	FSP	FCP	Overall
(0.15, 0.04)	79.0	61.0	70.0
(0.15, 0.05)	**79.0**	**61.0**	**70.0**
(0.15, 0.06)	79.0	61.0	70.0
(0.12, 0.06)	79.0	61.0	70.0
(0.17, 0.06)	79.0	61.0	70.0
(0.70, 0.70)	63.0	46.0	53.5
(0.02, 0.01)	72.0	51.0	61.5

tice U is used as a soft trigger for visual re-perception. In all experiments we simply set $\lambda = 1$, so that U reduces to H_{ans} when deciding whether to enter the reflection stage. The CoT-level uncertainty $\widehat{H}_{\text{cot}}$ is still exploited for localizing high-entropy windows and constructing focus phrases, as detailed in the main paper.

Global rollback thresholds. EVRI adopts a simple do-no-harm rule when deciding whether to keep the reflected answer or revert to the initial CoT prediction. This rule is controlled by two global thresholds $(\tau_{\text{gain\_e}}, \tau_{\text{drop\_e}})$ on the logit-margin gain and answer-entropy drop between the final reflection round and the initial run. Tab. 5 reports a sensitivity study on HR-Bench-4K.

When the thresholds are too strict, EVRI rarely accepts reflected answers and degenerates towards the initial CoT baseline, leading to much lower accuracy. When they are too permissive, many low-quality reflections are accepted and performance again degrades. In contrast, a broad band of moderate values around $(\tau_{\text{gain\_e}}, \tau_{\text{drop\_e}}) \approx (0.15, 0.05)$ is very stable: several nearby settings all yield the same best score of 79/61/70 on FSP/FCP/Overall. We therefore adopt $(\tau_{\text{gain\_e}}, \tau_{\text{drop\_e}}) = (0.15, 0.05)$ as our default configuration, which strikes a good balance between allowing useful corrections and preventing harmful revisions.

Windowing and entropy aggregation. The choice of high-entropy window length L follows Sec. 1.2. Token entropies $\{h_t\}$ are aggregated over a sliding window of length $L = 40$ with stride $s = 10$, and we select the Top-k windows with the highest average entropy, with $k = 1$ by default. This configuration is used consistently for uncertainty estimation and for text-guided focus phrase construction. Ablations in Tab. 1 and Fig. 1 show that $L = 40$ provides a good trade-off between locality and coherence, so we do not further tune L in this section.

Visual budget and reflection depth. In EVRI, the visual budget is mainly controlled by two integers: the maximum number of selected regions per injection r, and the maximum number of reflection rounds K. Following the implementation in the main paper, we use

$$r = 4, \qquad K = 3$$

in all main experiments. These choices are consistent with the ablations in the main paper, where the first column is denoted as (m, K) and m corresponds exactly to the number of visual cues per injection (i.e., $m \equiv r$).

From Tab.5 of the main paper, increasing the number of regions per injection from $r = 1$ to $r = 3$–4 clearly improves FSP/FCP and Overall accuracy, while using more than one effective reflection round (larger K) brings only marginal gains but increases computation. This is consistent with our multi-round reflection strategy in Sec. 1.4, where a new round is entered only if the current round passes the gain thresholds; in practice, most examples stop after the first successful reflection, and the actual number of rounds rarely reaches the cap K. We therefore adopt a configuration with (at most) one strong interleaving step, in which a small but carefully selected set of zoomed regions is injected, while allowing up to $K = 3$ rounds as a safety margin for harder cases.

Other settings and robustness. CoT decoding is performed greedily with a total budget of 320 tokens, shared across all methods in our framework. Overall, we observe that EVRI is relatively robust to moderate perturbations of the uncertainty and gain thresholds: varying $(\tau_{\text{conf}}, \tau_{\text{ent}})$ and $(\tau_{\text{gain}}, \tau_{\text{drop}})$ within a small range around the default values does not change the qualitative behavior of the system. For simplicity and reproducibility, we thus report results with this single shared hyperparameter setting across all experiments.

2. Additional Results on HR-Bench Option Combinations

As described in the main paper, each question in HR-Bench is associated with four official multiple-choice option sets, which leads to four distinct question–option configurations for every question. In the main results, we follow the common evaluation protocol and randomly sample one option set for each question, then report accuracy on the resulting 200 question–option instances at each resolution.

For completeness, we further evaluate all methods on all four option sets. This produces in total 800 question–option combinations at 4K resolution and another 800 combinations at 8K resolution. The evaluation protocol is identical to that in the main paper: for each question–option configuration, the model must output exactly one option label, and we compute FSP, FCP, and Overall accuracy in the same

Table 6. Full results on HR-Bench-4K with all four option combinations (800 question–option pairs) using Qwen2.5VL-7B as the base model. Accuracy (%).

Method	HR-Bench 4K		
	FSP	FCP	Overall
Qwen2.5VL-7B [1] (direct)	73.8	58.2	66.0
Qwen2.5VL-7B + ZoomEye [2]	78.5	59.8	69.1
Qwen2.5VL-7B + CoT	65.0	50.2	57.6
Qwen2.5VL-7B + ZoomEye + CoT	74.2	52.8	63.5
Qwen2.5VL-7B + **EVRI**	**79.8**	**60.8**	**70.2**

Table 7. Full results on HR-Bench-8K with all four option combinations (800 question–option pairs) using Qwen2.5VL-7B as the base model. Accuracy (%).

Method	HR-Bench 8K		
	FSP	FCP	Overall
Qwen2.5VL-7B [1] (direct)	66.2	50.8	58.5
Qwen2.5VL-7B + ZoomEye [2]	71.0	51.2	61.1
Qwen2.5VL-7B + CoT	51.8	46.5	49.1
Qwen2.5VL-7B + ZoomEye + CoT	67.2	50.8	59.0
Qwen2.5VL-7B + **EVRI**	**72.2**	**51.8**	**62.0**

way as in the main results. The full results on HR-Bench-4K are summarized in Tab. 6, and the results on HR-Bench-8K are shown in Tab. 7.

On HR-Bench-4K, EVRI improves over direct Qwen2.5VL-7B from 73.8% to 79.8% on FSP, from 58.2% to 60.8% on FCP, and from 66.0% to 70.2% on Overall accuracy. EVRI also outperforms ZoomEye and ZoomEye+CoT on all three metrics. On HR-Bench-8K, EVRI raises FSP accuracy from 66.2% to 72.2%, FCP from 50.8% to 51.8%, and Overall from 58.5% to 62.0%, again achieving the best performance among all test-time strategies considered.

The relative behavior of different methods is consistent with the trends in the main paper. Pure CoT reasoning without visual re-perception can degrade performance, especially on 8K images. ZoomEye brings moderate gains by purely visual zooming. EVRI provides the strongest and most stable improvements on both FSP and FCP when averaged over all 800 question–option combinations. These results confirm that the benefits of entropy-guided visual re-perception are not tied to a particular random sampling of option sets, but remain robust across all official option configurations in HR-Bench.

3. Summary of Supplementary Analyses

For clarity, we briefly summarize here the main findings from the supplementary experiments and analyses:

- **Uncertainty estimation.** EVRI uses answer-level entropy and token-level entropy as complementary uncertainty signals. Answer entropy provides a reliable coarse measure for early exit and rollback, while high-entropy windows along the CoT allow us to localize the most unstable reasoning windows and target them with visual re-perception.
- **High-entropy window length.** A moderate window length is crucial for extracting coherent uncertainty windows. Very short windows lead to fragmented phrases that are hard to ground visually, and very long windows mix multiple reasoning steps. On HR-Bench-4K, a fixed length of $L = 40$ achieves the best trade-off and is therefore used as the default setting.
- **Early-stop and gain thresholds.** The early-stop thresholds $(\tau_{\text{conf}}, \tau_{\text{ent}})$ and the patch-level gain thresholds $(\tau_{\text{gain}}, \tau_{\text{drop}})$ are stable in a neighbourhood of the default configuration. When these thresholds are set too loose, many low-confidence answers bypass reflection or noisy patches are accepted, which degrades performance. When they are too strict, EVRI degenerates towards text-only CoT. The default values reported in the main paper lie in a broad, robust plateau.
- **Do-no-harm rollback.** The global rollback thresholds $(\tau_{\text{gain\_e}}, \tau_{\text{drop\_e}})$ implement a simple do-no-harm principle. A wide range of moderate values around $(0.15, 0.05)$ consistently yields the best accuracy on HR-Bench-4K, while overly aggressive or overly permissive settings both harm performance. This confirms that the rollback mechanism is effective and not overly sensitive to fine tuning.
- **Visual budget and reflection depth.** Ablations on the number of selected regions per injection and the maximum reflection depth show that allowing a small set of informative zoomed regions (around $r = 3$–4) is important for performance, whereas adding many reflection rounds brings little benefit but increases latency. In practice, most examples terminate after one successful reflection round, and the cap $K = 3$ mainly serves as a safety margin for harder cases.
- **HR-Bench option configurations.** Evaluating on all four official option sets on HR-Bench-4K/8K confirms that the improvements brought by EVRI are consistent across 800 question–option combinations at each resolution. The relative ordering between direct Qwen2.5VL-7B, ZoomEye, ZoomEye+CoT, and EVRI matches the trends reported in the main paper, indicating that the gains of entropy-guided visual re-perception are robust to different option configurations.

Overall, these analyses indicate that EVRI is effective and robust across benchmarks and settings, supporting the central idea of entropy-guided visual re-perception.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7

[2] Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6602–6618. Association for Computational Linguistics, 2025. 7

[3] Shenzhi Wang, Le Yu, Chang Gao, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *CoRR*, abs/2506.01939, 2025. 1