VisionLeaf: Entropy-Guided Leaf-First Reasoning for Efficient and Accurate Think-with-Image

Supplementary Material

Contents

A. Training Data

As we adopt the same training dataset as DeepEyes, we conduct the following analysis to better illustrate the distribution and characteristics of the training data. Our analysis focuses on two key factors: task type and input resolution. The composition of the dataset is summarized below:

- **Visual Search [45]:** The training dataset consists of the Visual Search Dataset, which is derived from COCO2017 [21]. This dataset is essential for developing robust natural-image understanding, as accurate responses require recognizing subtle visual cues and making precise object-level distinctions.
- **ArxivQA [19]:** This dataset comprises samples with complex visual semantics, including scientific plots, diagrams, and schematic charts. Its inclusion enables the model to interpret abstract and symbolic visual representations that extend beyond conventional natural-image scenarios.

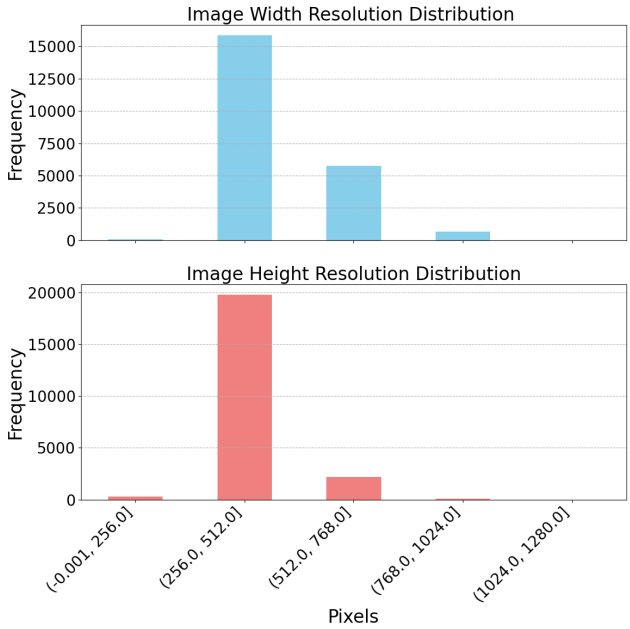

Figure 7. Distribution of image resolutions in the training dataset.

- **ThinkLite-VL [41]:** The training data also includes multimodal question–answering examples from ThinkLite-VL. This subset specifically targets tasks such as arithmetic reasoning, commonsense inference, and general problem-solving, which are critical for enhancing the model's overall reasoning robustness and mitigating modality-specific overfitting.

Furthermore, we conducted an analysis of the training data's resolution distribution, as illustrated in Fig. 7. The figure demonstrates that the majority of images are concentrated around resolutions such as 512×256 pixels. This presents a significant distribution shift when compared to high-resolution benchmarks like the VStar dataset and HR-Bench, which commonly feature resolutions exceeding $2K$. The model's strong performance across benchmarks with diverse input resolutions highlights the robust generalizability of our proposed methods.

B. Training Settings

The model training was conducted on a cluster of $\mathbf{8} \times \mathbf{H800}$ GPUs, each equipped with $\mathbf{80}$ GB of memory. We employed a train batch size of $\mathbf{128}$. To accommodate long-context tasks, the maximum prompt length was set to $\mathbf{8192}$ tokens, and the maximum response length was con-

figured at **16384** tokens.

For the optimization core, we utilized the vanilla GRPO algorithm to calculate the advantage during training. Consistent with prior work such as DAPO [50] and DeepEyes [60], we omitted the KL divergence term from the objective function. To further enhance model performance, we applied the standard data augmentation settings commonly used in segmental training approaches.

To significantly accelerate the sampling and inference procedure, we leveraged sglang as the inference engine. We adopted an asynchronous sampling method and restricted the maximum number of interactive turns to **5**. This limitation effectively mitigates potential overlong thinking cycles, thereby preventing unnecessary increases in overall training time.

C. Prompt

The following visualization illustrates the prompt instantiated by our tokenizer's chat template. It embeds the `image_zoom_in_tool` schema definition and places it within a multimodal interaction context, as shown in Fig. 8.

D. Evaluation Benchmarks

HR-Bench [40]. HR-Bench is a benchmark designed to evaluate the perceptual capabilities of multimodal large language models on high-resolution images. It addresses the limitations of existing datasets by providing images at 4K and 8K resolutions, offered in two variants: HR-Bench 8K and HR-Bench 4K.

VStar [45]. VStar is a dataset developed to assess the visual search and detailed perception capabilities of MLLMs on high-resolution images. It is constructed from 191 high-resolution images with an average resolution of approximately 2246×1582 pixels. The benchmark comprises two sub-tasks: an attribute recognition task with 115 samples and a spatial relationship reasoning task with 76 samples. Each sample is posed as a multiple-choice question designed to test a model's ability to process high-resolution inputs, focus on fine-grained visual details, and perform guided visual search.

MME-RealWorld [54]. MME-RealWorld is a large-scale benchmark tailored for real-world applications, featuring 13,366 high-resolution images with an average resolution of approximately 2000×1500 pixels. It contains 29,429 manually annotated question-answer pairs covering 43 subtasks across 5 real-world domains: OCR, diagrams and tables, remote sensing, autonomous driving, and monitoring. This dataset is designed to be highly challenging, requiring models to identify small or densely distributed objects and to perform robust high-resolution perception and reasoning in complex real-world scenarios. In our experiments, we adopt the official *MME-RealWorld-Lite* split, which preserves the original 5 domains and 43 subtasks while uniformly downsampling to at most 50 QA pairs per subtask (or all samples if fewer than 50), yielding a lightweight yet representative subset that significantly reduces evaluation cost for high-resolution benchmarks.

E. Discussion on Training Methods

With the growing emergence of visual agents that make decisions based on histories of visual observations, language, and actions, a variety of training paradigms—most notably supervised fine-tuning and reinforcement learning—have been adopted. Many prior works [16, 25] employ an SFT warm-up stage followed by RL training. To more clearly isolate the contribution of RL within the overall training pipeline, we conduct all experiments using only the RL stage. Recent studies [5, 14] further suggest that RL-trained models often exhibit stronger generalization capabilities than their SFT counterparts, particularly under out-of-distribution conditions.

F. More Comparision Results

F.1. Stability Analysis

To more clearly compare the performance characteristics of DeepEyes and VisionLeaf, we perform a statistical analysis of their total number of *crashed outputs* across all interaction turns. We define a crashed output as any case in which the model exhibits severe repetitive function calling or produces a substantially malformed output format. This metric serves as a direct indicator of model stability under diverse scenarios. As shown in Fig. 9, VisionLeaf exhibits a markedly lower crash rate than DeepEyes—almost reaching zero in our evaluation—indicating that VisionLeaf delivers more stable and reliable behavior.

Table 3. **Efficiency Analysis.** We report the training time and the inference time on the same settings.

Method	Training Time	Inference	
		Time	Tokens
DeepEyes	40.8h	22.5min	246
VisionLeaf	36.4h	12.2min	139

F.2. Efficiency Analysis

Furthermore, we assess the efficiency of VisionLeaf by comparing total inference time across diverse benchmarks. As depicted in Fig. 10, the proposed method achieves an approximate $3\times$ reduction in inference time under the full benchmark setting. To ensure a controlled and fair evaluation, all experimental configurations are held constant,

System Prompt

```
<|im_start|>system
You are a helpful assistant.

# Tools

You may call one or more functions to assist with the user query.

You are provided with function signatures within <tools></tools> XML tags:
<tools>
{
  "type": "function",
  "function": {
    "name": "image_zoom_in_tool",
    "description": "Zoom in on a specific region of an image by cropping it based on a
        bounding box (bbox) and an optional object label.",
    "parameters": {
      "type": "object",
      "properties": {
        "bbox_2d": {
          "type": "array",
          "items": { "type": "number" },
          "minItems": 4,
          "maxItems": 4,
          "description": "The bounding box of the region to zoom in, as [x1, y1, x2, y2
              ], where (x1, y1) is the top-left corner and (x2, y2) is the bottom-right
              corner."
        },
        "label": {
          "type": "string",
          "description": "The name or label of the object in the specified bounding box
              (optional)."
        }
      },
      "required": ["bbox_2d"]
    }
  }
}
</tools>

For each function call, return a json object with function name and arguments within <
    tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call><|im_end|>
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>
Can you verify the details of the object in the top right corner<|im_end|>
<|im_start|>assistant
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [500, 50, 800, 250], "label": "
    unknown_object"}}
</tool_call><|im_end|>
<|im_start|>tool
<tool_response>
Image cropped to [500, 50, 800, 250].
</tool_response><|im_end|>
<|im_start|>assistant
The zoomed-in region reveals a street sign. I can now read the text clearly.<|im_end|>
```

Figure 8. The prompt used during the training procedure.

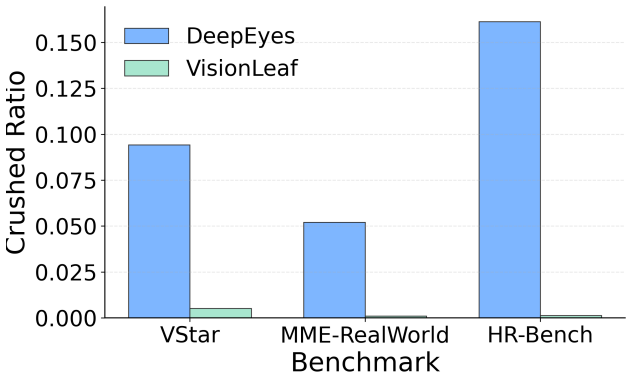

Figure 9. Comparision of the crushed ratio of VisionLeaf and DeepEyes.

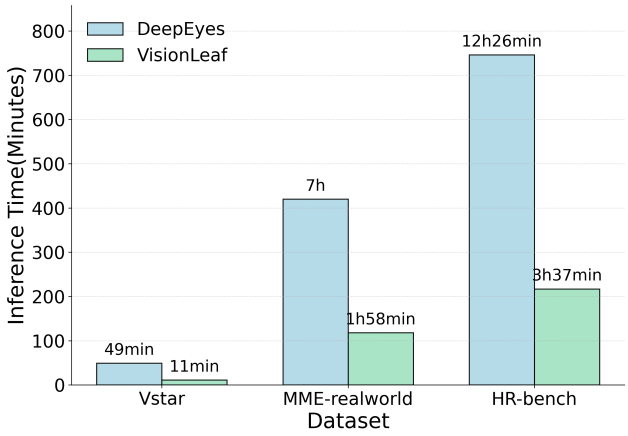

Figure 10. Comparison of the inference time of VisionLeaf and DeepEyes across multiple benchmarks.

and inference time is measured using a consistent evaluation framework. In addition, Table. 3 reports training and inference times on 8 GPUs, as well as the number of generated tokens. Collectively, these findings highlight that VisionLeaf not only improves task performance but also substantially reduces computational overhead in both training and inference phases.

F.3. Experiments on MiMO-VL

To further assess the generalization performance of our method, we conduct experiments on MiMO-VL, a representative model gaining increasing traction in the VLM community. As shown in Fig. 11, our approach achieves a 1.4% improvement over the baseline method, DeepEyes, providing clear evidence of its strong generalization capability across architectures.

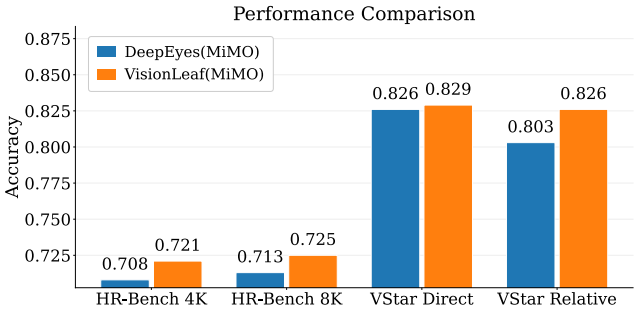

Figure 11. **Performance Comparison on MiMO-VL.** To further validate the generality of our method across different model architectures, we conduct additional experiments on MiMO-VL.

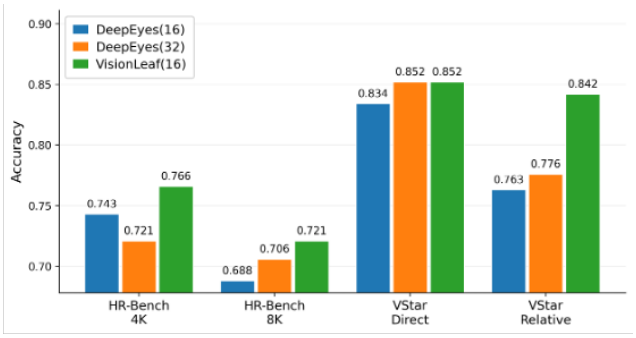

Figure 12. **Larger Rollout Comparison.** To provide a more thorough comparison, we evaluate VisionLeaf against the baseline using a larger number of rollouts.

F.4. Additional Ablation Study

To further investigate the efficiency of our approach, we conduct a comparative study against the baseline using an increased GRPO rollout budget. Specifically, we evaluate DeepEyes with 32 rollouts to understand the effect of scaling root-level rollouts. As shown in Fig. 12, the performance improvement is marginal, while the associated training cost rises considerably. These findings indicate that increasing the GRPO rollout number alone yields diminishing returns and is substantially less efficient than our method.

G. Background and Theorem Proof

G.1. Background on Conditional Expectation and Variance

Let X and Y be square-integrable random variables on a common probability space. The conditional expectation $\mathbb{E}[X \mid Y]$ is the (a.s. unique) random variable such that for every bounded measurable function ϕ,

$$\mathbb{E}\big[X\,\phi(Y)\big] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\,\phi(Y)\big]. \tag{12}$$

The conditional variance of X given Y is defined as

$$\mathrm{Var}[X \mid Y] = \mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2 \mid Y\big]. \tag{13}$$

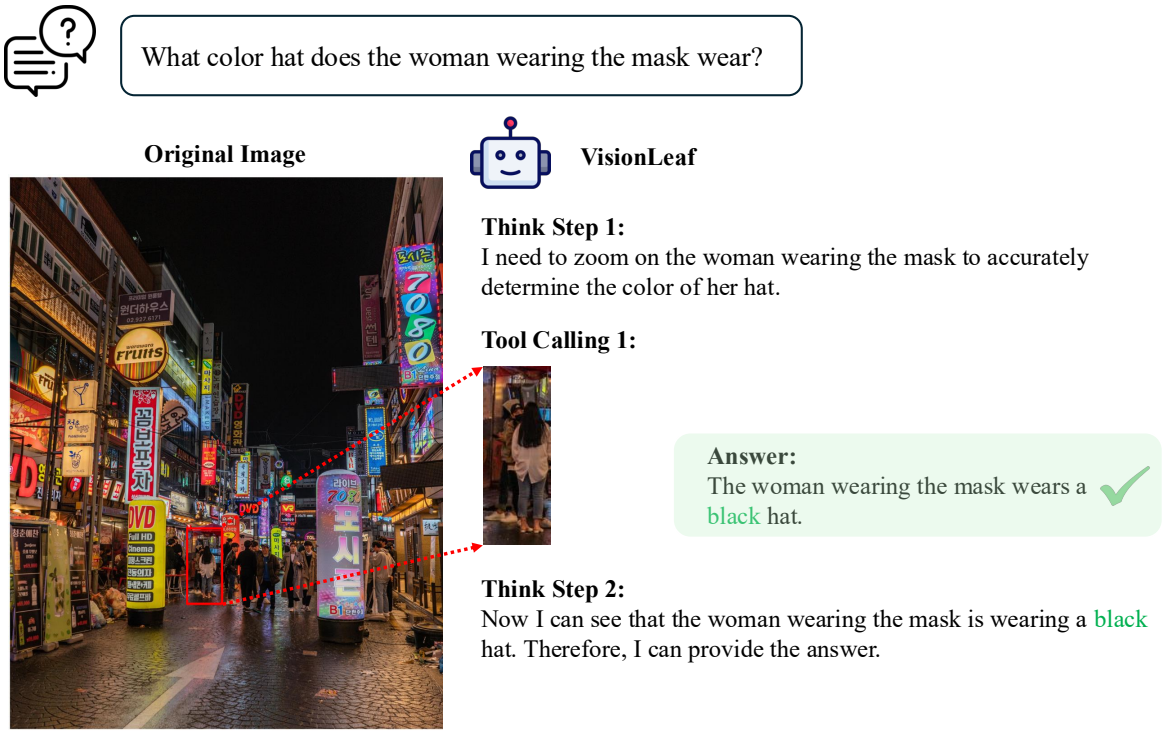

Figure 13. Fine-grained object perception case illustrating detailed visual reasoning. The correct match is highlighted in green, demonstrating the model's strong ability to perform fine-grained object perception.

We will repeatedly use the *law of total expectation*

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big], \tag{14}$$

and the *law of total variance*

$$\mathrm{Var}[X] = \mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] + \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big). \tag{15}$$

All such relationships have analogous versions when we additionally condition on another random variable Z, e.g.

$$\mathrm{Var}[X \mid Z] = \mathbb{E}\big[\mathrm{Var}(X \mid Y, Z) \mid Z\big] + \mathrm{Var}\big(\mathbb{E}[X \mid Y, Z] \mid Z\big). \tag{16}$$

G.2. Theorem Proof

Variance Analysis of Early Stage Response

We first recall a standard conditional variance identity that will be used throughout.

Lemma G.1 (Conditional law of total variance)**.** *Let X be a square-integrable random variable and let (Y, Z) be a pair of random variables. Then*

$$\mathrm{Var}[X \mid Z] = \mathbb{E}[\mathrm{Var}(X \mid Y, Z) \mid Z] + \mathrm{Var}(\mathbb{E}[X \mid Y, Z] \mid Z)\,. \tag{17}$$

This is the standard law of total variance applied conditionally on Z.

Lemma G.2 (Recursive conditional variance along a trajectory)**.** *Let $(s_t)_{t=1}^{T+1}$ be a sequence of context random variables in a rollout, and let R be a square-integrable terminal random variable. Assume that, for each t,*

$$\mathbb{P}(R \in \cdot \mid s_t, s_{t+1}) = \mathbb{P}(R \in \cdot \mid s_{t+1}), \tag{18}$$

i.e., s_{t+1} is sufficient for the future return. Then

$$\mathrm{Var}[R \mid s_t] = \mathbb{E}[\mathrm{Var}(R \mid s_{t+1}) \mid s_t] + \mathrm{Var}(\mathbb{E}[R \mid s_{t+1}] \mid s_t)\,. \tag{19}$$

Proof. Apply Lemma G.1 with $X = R$, $Y = s_{t+1}$ and $Z = s_t$:

$$\begin{aligned}\mathrm{Var}[R \mid s_t] = \mathbb{E}[\mathrm{Var}(R \mid s_t, s_{t+1}) \mid s_t] \\ + \mathrm{Var}(\mathbb{E}[R \mid s_t, s_{t+1}] \mid s_t)\,.\end{aligned} \tag{20}$$

The sufficiency assumption implies that, for each t, the conditional distribution of R given (s_t, s_{t+1}) coincides with that of R given s_{t+1}, hence

$$\begin{aligned}\mathrm{Var}(R \mid s_t, s_{t+1}) = \mathrm{Var}(R \mid s_{t+1}), \\ \mathbb{E}[R \mid s_t, s_{t+1}] = \mathbb{E}[R \mid s_{t+1}],\end{aligned} \tag{21}$$

and substituting these identities yields (19). □

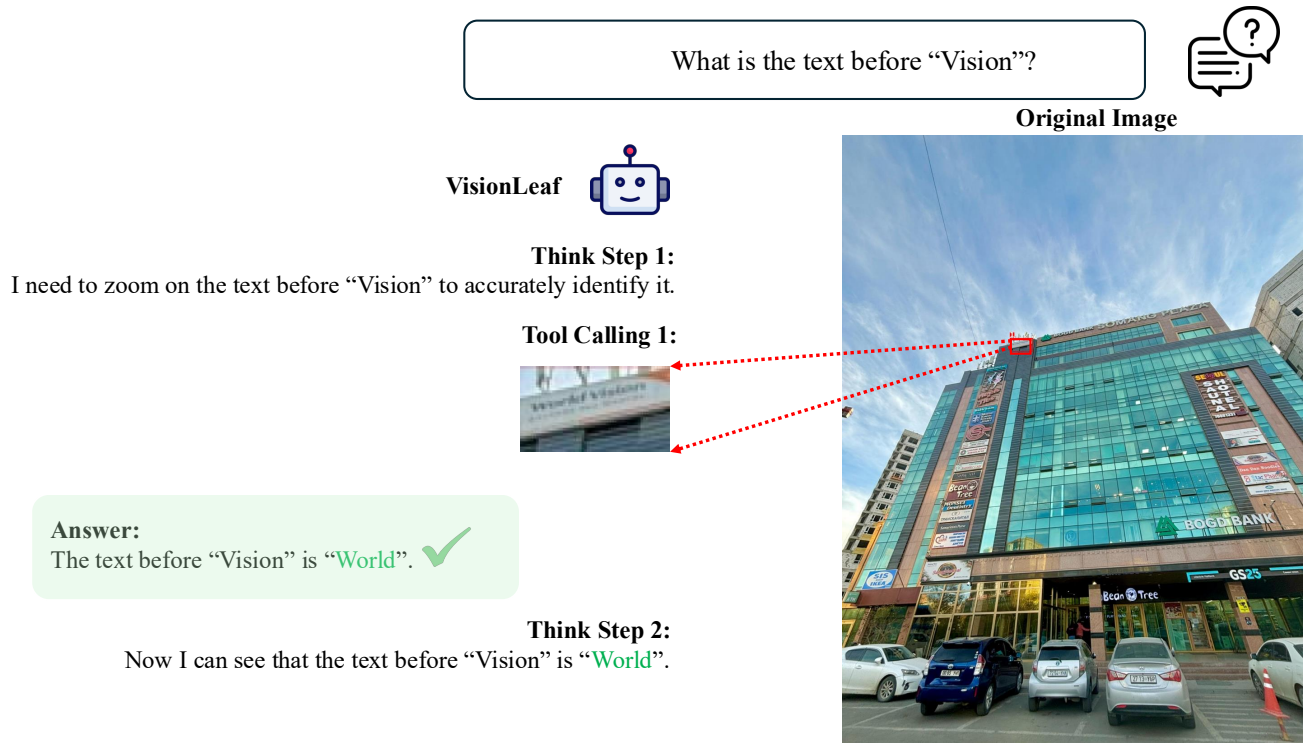

Figure 14. Fine-grained example of word-level OCR reasoning. The correct match is highlighted in green, demonstrating the model's strong ability to perform fine-grained word-level OCR reasoning.

Analysis on Prefix Sharing Sampling

We now derive the variance identities for group-centered returns and for tree sampling.

Lemma G.3 (Variance of group-centered returns)**.** *Let $R_1, \dots, R_G$ be conditionally independent and identically distributed copies of a square-integrable random variable R given s, with*

$$\mathbb{E}[R \mid s] = \mu_s, \qquad \mathrm{Var}[R \mid s] = \sigma_s^2 < \infty. \tag{22}$$

Define the group mean

$$\bar{R} := \frac{1}{G}\sum_{i=1}^{G} R_i, \tag{23}$$

and the centered quantities

$$\widehat{A}_i := R_i - \bar{R}, \qquad i = 1, \dots, G. \tag{24}$$

Then, for each i,

$$\mathrm{Var}[\widehat{A}_i \mid s] = \left(1 - \tfrac{1}{G}\right)\mathrm{Var}[R \mid s]. \tag{25}$$

Proof. Write $\sigma_s^2 = \mathrm{Var}[R \mid s]$. Since the R_i are conditionally i.i.d. given s,

$$\mathrm{Var}[R_i \mid s] = \sigma_s^2, \qquad \mathrm{Cov}[R_i, R_j \mid s] = 0 \quad (i \neq j). \tag{26}$$

We have

$$\begin{aligned} \widehat{A}_i &= R_i - \bar{R} \\ &= R_i - \frac{1}{G}\sum_{j=1}^{G} R_j \\ &= \Big(1 - \frac{1}{G}\Big) R_i - \frac{1}{G}\sum_{j \neq i} R_j. \end{aligned} \tag{27}$$

Thus

$$\begin{aligned} \mathrm{Var}[\widehat{A}_i \mid s] &= \mathrm{Var}\Big[\Big(1 - \frac{1}{G}\Big) R_i - \frac{1}{G}\sum_{j \neq i} R_j \;\Big|\; s\Big] \\ &= \Big(1 - \frac{1}{G}\Big)^2 \mathrm{Var}[R_i \mid s] + \mathrm{Var}\Big[\frac{1}{G}\sum_{j \neq i} R_j \;\Big|\; s\Big] \\ &\quad - 2\Big(1 - \frac{1}{G}\Big)\mathrm{Cov}\Big(R_i, \frac{1}{G}\sum_{j \neq i} R_j \;\Big|\; s\Big). \end{aligned} \tag{28}$$

The covariance term vanishes because R_i is conditionally independent of each R_j, $j \neq i$, given s. For the second variance term,

$$\begin{aligned} \mathrm{Var}\Big[\frac{1}{G}\sum_{j \neq i} R_j \;\Big|\; s\Big] &= \frac{1}{G^2}\sum_{j \neq i} \mathrm{Var}[R_j \mid s] \\ &= \frac{G-1}{G^2}\,\sigma_s^2. \end{aligned} \tag{29}$$

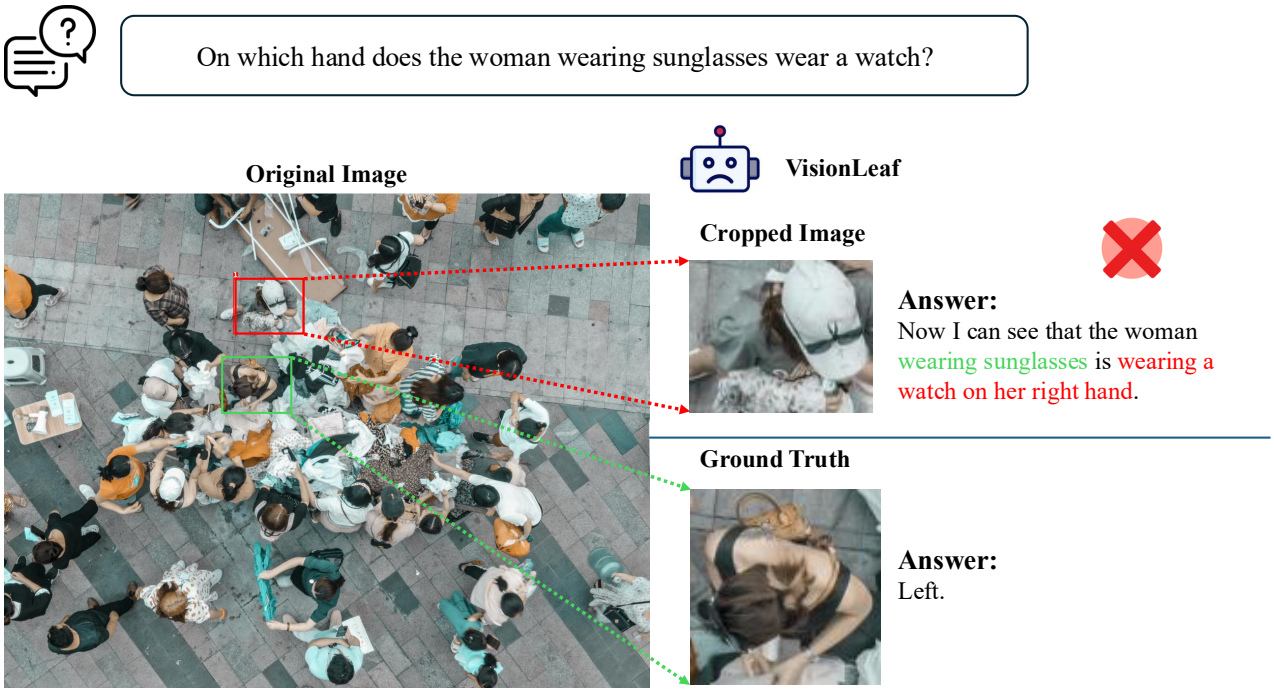

Figure 15. Failure case in which the reasoning procedure fails to capture all required details. We highlight the correct match in green and the mismatched prediction in red, illustrating that the model still encounters difficulties in complex reasoning scenarios.

Therefore

$$\begin{aligned}\mathrm{Var}[\widehat{A}_i \mid s] &= \Big(1-\frac{1}{G}\Big)^2\sigma_s^2 + \frac{G-1}{G^2}\,\sigma_s^2 \\ &= \Big(1-\frac{2}{G}+\frac{1}{G^2}+\frac{G-1}{G^2}\Big)\sigma_s^2 \\ &= \Big(1-\frac{2}{G}+\frac{1}{G}\Big)\sigma_s^2 \\ &= \Big(1-\frac{1}{G}\Big)\sigma_s^2 \\ &= \Big(1-\frac{1}{G}\Big)\mathrm{Var}[R \mid s].\end{aligned} \tag{30}$$

□

Lemma G.4 (Node-level group variance in a tree)**.** *Let n be a (random) node extending s. Suppose that, conditional on n,*

$$R_1^{(n)}, \ldots, R_G^{(n)} \tag{31}$$

are i.i.d. copies of a square-integrable random variable $R^{(n)}$ with finite variance. Define

$$\bar{R}^{(n)} := \frac{1}{G}\sum_{j=1}^{G} R_j^{(n)}, \qquad \widehat{A}_i^{(n)} := R_i^{(n)} - \bar{R}^{(n)}. \tag{32}$$

Then, for each i,

$$\mathrm{Var}[\widehat{A}_i^{(n)} \mid n] = \Big(1-\tfrac{1}{G}\Big)\,\mathrm{Var}[R^{(n)} \mid n]. \tag{33}$$

Proof. Fix n and apply Lemma G.3 with s replaced by n and R replaced by $R^{(n)}$. □

To relate $\mathrm{Var}[R^{(n)} \mid n]$ to $\mathrm{Var}[R \mid s]$, we use the conditional law of total variance once more.

Lemma G.5 (Average comparison between root and node variances)**.** *Let N be a random node extending s, and suppose that the conditional distribution of R given (s, N) coincides with that of $R^{(N)}$ given N. Then*

$$\mathbb{E}\big[\mathrm{Var}[R^{(N)} \mid N] \;\big|\; s\big] \le \mathrm{Var}[R \mid s]. \tag{34}$$

Proof. Apply Lemma G.1 with $X = R$, $Y = N$ and $Z = s$:

$$\mathrm{Var}[R \mid s] = \mathbb{E}\big[\mathrm{Var}(R \mid s, N) \;\big|\; s\big] + \mathrm{Var}\big(\mathbb{E}[R \mid s, N] \;\big|\; s\big). \tag{35}$$

By the assumption on the conditional distributions,

$$\mathrm{Var}(R \mid s, N) = \mathrm{Var}[R^{(N)} \mid N], \tag{36}$$

so

$$\begin{aligned}\mathrm{Var}[R \mid s] &= \mathbb{E}\big[\mathrm{Var}[R^{(N)} \mid N] \;\big|\; s\big] + \mathrm{Var}\big(\mathbb{E}[R^{(N)} \mid N] \;\big|\; s\big) \\ &\ge \mathbb{E}\big[\mathrm{Var}[R^{(N)} \mid N] \;\big|\; s\big],\end{aligned} \tag{37}$$

since the last term is a conditional variance and hence non-negative. □

Combining Lemma G.4 with Lemma G.5, we obtain that, on average over nodes N extending s,

$$\mathbb{E}\big[\mathrm{Var}[\widehat{A}_i^{(N)} \mid N] \;\big|\; s\big] = \Big(1 - \tfrac{1}{G}\Big)\,\mathbb{E}\big[\mathrm{Var}[R^{(N)} \mid N] \;\big|\; s\big] \leq \Big(1 - \tfrac{1}{G}\Big)\,\mathrm{Var}[R \mid s]. \tag{38}$$

This shows how the variance formulas for group-centered returns and their shared-prefix refinements follow from basic properties of conditional expectation and variance.

H. More Case Studies

H.1. Success Cases

To further illustrate the efficiency of VisionLeaf, we provide additional qualitative examples beyond those presented in the main paper. These cases are shown in Fig. 13 and Fig. 14. From these examples, we observe that the function-call mechanism plays a crucial role: the initial reasoning step determines which region of the image requires closer examination, and the subsequent zoom-in operation enables more focused, image-grounded analysis. This leads to a coherent "think-with-image" process that ultimately produces the correct answer. Moreover, this paradigm demonstrates strong performance in both OCR-based reasoning and general visual reasoning tasks.

H.2. Failed Cases

Despite the effectiveness of VisionLeaf, certain limitations remain when handling complex reasoning tasks. As illustrated in Fig. 15, the model focuses on the woman wearing sunglasses while overlooking an additional constraint—that the target individual must also be wearing a watch. This leads to an incorrect localization. For comparison, we show the correct bounding box corresponding to the question alongside the model's predicted crop region.

More specifically, this case is difficult even for humans: the black sunglasses blend closely with the subject's dark hair, resulting in low-contrast visual cues that are challenging to distinguish. Given the limited capacity of the 7B model used in our experiments, such fine-grained visual reasoning remains particularly difficult, highlighting the constraints imposed by model size.

I. Limitations and Future Work

Limitations. While the proposed reinforcement-learning pipeline effectively improves image-grounded reasoning, several limitations remain. All experiments in this work are conducted using a 7B model, whose relatively small capacity constrains both fine-grained visual perception and higher-level reasoning. Consequently, the model may underperform on tasks requiring accurate bounding-box construction or subtle visual discrimination, as discussed in Sec. H.2. We attribute these challenges to the restricted model size and its associated representational limitations.

Future Work. To mitigate these issues, future work will explore scaling the approach to larger backbone models with stronger visual and linguistic reasoning capabilities. Increasing model capacity is expected to significantly alleviate difficulties in detecting low-contrast regions, localizing subtle visual attributes, and maintaining stable reasoning under complex multimodal constraints. More broadly, integrating our method with more capable foundation models may unlock substantially improved performance in multi-turn, tool-augmented visual reasoning tasks.