SIEVES: Selective Prediction Generalizes through Visual Evidence Scoring

Hector G. Rodriguez
TU Darmstadt & hessian.AI, Germany

Marcus Rohrbach
TU Darmstadt & hessian.AI, Germany

Abstract

As the accuracy of Multimodal large language models (MLLMs) on traditional visual question answering benchmarks approaches saturation, reliable deployment requires satisfying low error tolerances in real-world out-of-distribution (OOD) scenarios. Selective prediction aims to improve coverage, i.e. the share of inputs the system answers, while adhering to a user-defined risk level. This is typically achieved by assigning a confidence score to each answer and abstaining on those that fall below a certain threshold. Existing selective prediction methods rely on text-only reasoning and learn to predict confidences simply based on whether the answer was correct or not. In contrast, we introduce SIEVES (Selective Prediction through Visual Evidence Scoring), which first requires the answering model to provide visual evidence supporting its answer, by returning crops of the image alongside their textual reasoning and answer. Second, a lightweight selector is trained to predict the quality of this evidence, assessing how well the provided crops localize the relevant visual elements and if the final answer is coherent with this evidence. We show that SIEVES improves coverage by up to six times on challenging OOD benchmarks (V* Bench, HR-Bench-8k, MME-RealWorld-Lite, VizWiz, and AdVQA), compared to non-grounding baselines. Beyond better generalization to OOD tasks, the design of the SIEVES selector enables transfer to proprietary reasoners without access to their weights or logits, such as o3 and Gemini-3-Pro, providing coverage boosts beyond those attributable to accuracy alone. We highlight that SIEVES generalizes across all tested OOD datasets and reasoner models, without benchmark- or reasoner-specific training or adaptation.

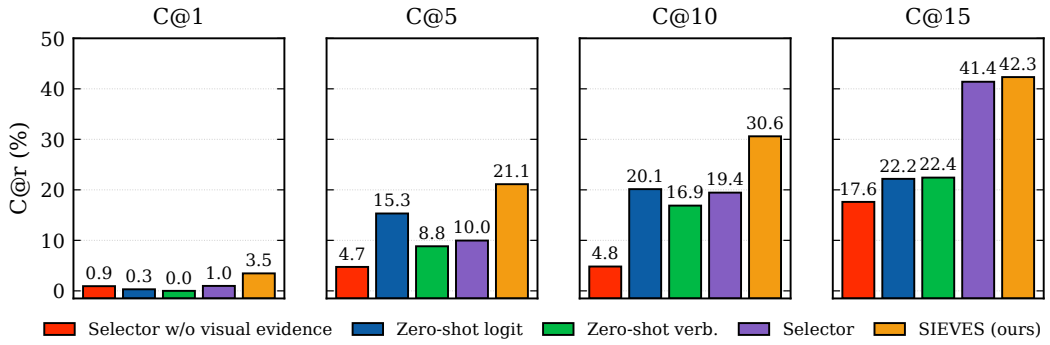


Figure 1: OOD coverage at varying risk levels for frontier proprietary reasoners not used for training the (SIEVES) selector. C@r averaged across OOD benchmarks and o3 and Gemini-3-Pro reasoners, excluding trivial cases with reasoner error $<$ risk. SIEVES achieves the highest coverage, especially at low risk.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Recent large language models (LLMs) [1, 18] and vision-language models (VLMs) [3, 36] can perform an increasing number of economically valuable tasks autonomously, from document understanding [31, 48] to assistive visual answering [11, 14]. Often in these tasks, mistakes can be costly, both in human and financial terms. Selective prediction enables AI systems to decide which tasks are beyond the reach of these models, and for which it should abstain or defer to an expert [50].

Visual Question Answering (VQA) has been a good testbed for VLMs. Nonetheless, even as traditional VQA benchmarks [2, 15] become saturated from an accuracy perspective [41], selective prediction remains a challenge [9]. There are several reasons why selective prediction continues to be difficult. Most importantly, machine learning (ML) systems are predominantly designed to answer as many questions correctly as possible, with little or no training to abstain when in doubt [25]. Since the user's risk tolerance is unknown in advance, a truly binary abstain/answer decision is too rigid. Selective prediction, therefore, assigns each answer a numerical confidence score, which makes the task harder: the selector must now rank the full distribution of answer-question pairs.

Existing selective prediction methods estimate confidence in a response given simply the answer [26], standard textual reasoning [23], or model internals [4, 34, 51]. Estimating confidence using these signals is hard, as confidence should depend not only on the final answer, but also on the evidence used to reach it.

Methods that rely on model internals, such as log-probabilities or hidden activations, are also model-specific [9, 34, 51] and cannot be applied to frontier models when those signals are unavailable through proprietary APIs[1]. Other approaches [27, 43] make selective prediction much more costly than the original answering task. Here, we aim to solve such limitations.

We propose Selective Prediction through Visual Evidence Scoring (SIEVES) for generalizable selective prediction. A strong reasoner that can zoom-in on the high-resolution image produces multimodal chain-of-thought reasoning with zoomed-in crops before providing a final answer. This exposes rich evidence that makes confidence prediction easier. We train the selector to take as input the full conversation, i.e. question, image, the multimodal chain-of-thought (MM-CoT) with crops, and the final answer, and output a scalar confidence that the answer is correct. Crucially, the SIEVES selector does not only predict a single confidence score. Since the MM-CoT exposes where the model looked at before answering, the selector also explicitly assesses the quality of the localization shown in the MM-CoT. Concretely, the selector predicts three signals: *correctness* (is the final answer accurate?), *localization* (did the model look at the right part of the image?), and *coherence* (does the visual evidence support the answer?). Training the selector to explicitly score this evidence at a fine-grained level helps distinguish well-supported answers from answers based on weak or irrelevant evidence, leading to more precise confidence estimates.

We evaluate on a wide range of VQA benchmarks, which are out-of-distribution for the reasoner and selector, requiring varying degrees of generalization. These benchmarks range from high-resolution natural image benchmarks (V*-bench, HR-Bench-8k), diverse high resolution tasks including diagrams, tables, and remote sensing (MME-RealWorld-Lite), to real-world questions from blind users (VizWiz) and Human-Adversarial problems (AdVQA). SIEVES improves the coverage at relevant risk tolerances by up to six times compared to non-grounding baselines, as shown in Table 1.

Our main contributions can be summarized as follows:

1. We introduce visual evidence for selective prediction. A reasoner uses a zoom-in tool to provide a crop which serves as visual evidence to support the final answer. The evidence is used by the selector to better assess how trustworthy the answer is. The selector uses precise observable outputs instead of model internals, so it can be applied to different reasoners.

2. We train the selector to score two fine-grained properties of the evidence in addition to answer correctness: whether the reasoner localized the relevant image region and whether the evidence supports the answer. This helps differentiate well-documented response from answers based on weak proof, hence allowing for improved confidence prediction.

[1] OpenAI: `https://community.openai.com/t/logprobs-deprecated-for-gpt-5-models/1355427`; Anthropic: `https://docs.anthropic.com/en/api/openai-sdk#detailed-openai-compatible-api-support`.

2 Related work

Reasoning with tools. Tool usage has become central to the recent explosion in reasoning capabilities of VLMs. For VQA tasks, providing VLMs with the ability to use zoom-in tools has significantly improved the ability to answer questions about high-resolution images, through a form of multimodal chain-of-thought reasoning [21, 44, 62]. Frontier VLMs can benefit from access to these tools without explicit training [21, 37], whereas smaller models are trained specifically to elicit such zooming in capability [44, 62]. However, this zooming-in ability has only been, to the best of our knowledge, studied as a means of improving accuracy. In this work, we propose leveraging the multimodal chain-of-thought that results from zooming in to assess how well grounded the final answer is. Therefore, we do not focus on improving the abilities of these models to zoom-in, but rather use these existing abilities to make answer correctness prediction easier. Unlike existing methods, which only show the benefits of zooming-in to improve accuracy in high-resolution VQA tasks, we employ the MM-CoT as a source of answer grounding evidence. Using this MM-CoT, the selector model has a richer context to output a confidence score for each answer. Intuitively, the ability to analyze where the model looked before answering eases confidence score prediction.

Test-time scaling via verification. Our work has connections to the use of external models to rank different model generations to the same question [7, 29], namely, using verifiers for test-time scaling. However, we focus on the selective prediction task, which requires ranking over the joint space of questions and answers, whereas verification effectively factors the problem by conditioning on the question. This makes the problem strictly harder, as scores must be comparable across different questions. Furthermore, verifiers are usually models of the same size and inference budget (i.e., chain-of-thought length) as the reasoners [42]. This implies that verification is as hard as answering the question itself. Here, we demonstrate that by having a reasoner provide visual evidence for the answer, and explicitly learning to estimate the quality of such evidence, confidence score prediction is made easier, and a smaller discriminative selector can be used.

Selective prediction. Selective prediction formalizes reliable deployment by maximizing the portion of answered queries (coverage) under a user-specified error rate (risk), typically summarized with risk-coverage curves or AURC. Early approaches rely on intrinsic model uncertainty, e.g., maximum softmax probability or logit margins, to decide when to abstain [6, 8, 12, 54]. Such approaches are not directly applicable to generative models and suffer from the gap between the accuracy-maximizing training objective and the selective prediction task. Whitehead et al. [51] demonstrate that it is effective to learn a dedicated lightweight selector to predict confidence in the final answer given both model's internal representations and question-image input. Dancette et al. [9] refine this for out-of-distribution settings, and Mushtaq et al. [34] abstain on answers from VLMs. However, hidden activations are entirely model dependent, tying the selector to a specific model architecture. Furthermore, such model internals are not available for frontier models queried through proprietary APIs, which limits the applicability of such selectors. In our work, we design a selector that relies only on observable outputs, and hence is model-agnostic.

Given the limitations of trained selectors that require model internals, recent works have developed alternative methods for abstaining with generative LLMs or VLMs. Khan and Fu [27] propose consistency as a proxy for certainty and use it to abstain. Srinivasan et al. [43] utilize an elaborate prompting scheme to use zero-shot selectors. Both of these methods make selective prediction much more expensive than answering, can only provide coarse confidence scores, and do not demonstrate significant coverage at demanding risk levels. In contrast, by using the multimodal reasoning process as a source of evidence for the selector, we avoid duplicating the compute devoted to solving the task, and can provide better confidence estimates with a smaller selector.

3 Selective Prediction through Visual Evidence Scoring

We approach selective prediction for visual question answering by combining a reasoner that can zoom-in on the image and a selector that outputs a confidence score for the answer.

3.1 Providing visual evidence via zooming-in tool

The reasoner model f is tasked with providing an answer A to the question Q about the image I. In the process, since it benefits answer accuracy, it produces multimodal chain-of-thought reasoning,

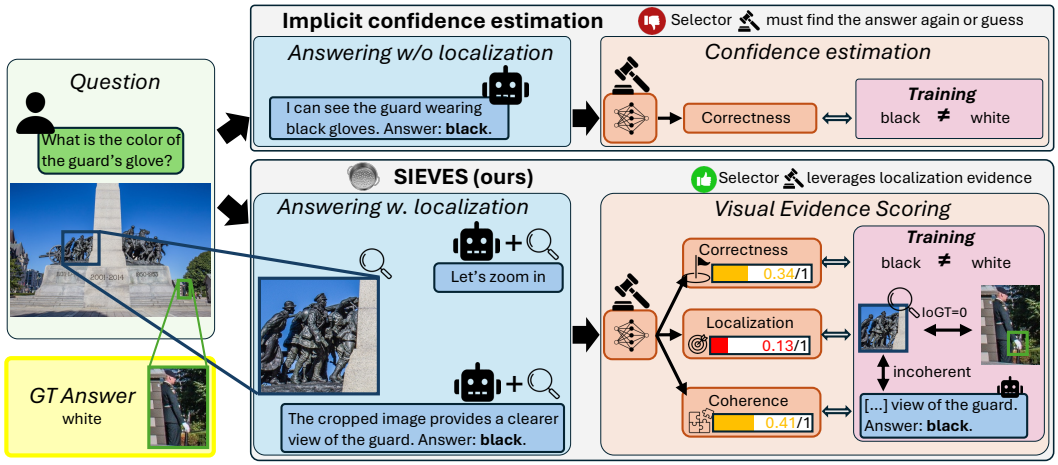


Figure 2: Confidence from visual evidence. Top: A correctness-only selector scores an answer from the image and text trace. **Bottom:** SIEVES also receives the reasoner's image crops and scores answer correctness, localization, and crop–answer coherence. Their weighted sum is the confidence used for abstention.

denoted as R. Therefore, after the reasoner answers

$$\{R, A\} = f(Q, I), \tag{1}$$

we have a conversation tuple $\{Q, I, R, A\}$. We use reasoners that can use a zoom-in tool. Recipes for teaching open-source weaker models to learn this ability have become commonplace [44], and several generations of frontier VLMs have been able to do this [37], so obtaining an answering system that is able to reason by pointing at the relevant sections of an image should not be a concern when tackling highest-stakes scenarios that selective prediction aims to solve.

3.2 Visual Evidence Scoring

Fine-grained confidence scoring using visual evidence. Standard selective prediction learns to predict confidence scores that infer whether a final answer matches the ground-truth label [9, 34, 51]. However, this signal alone does not capture *why* an answer is correct or incorrect, making it prone to overfitting to in-distribution patterns like question domain or reasoner patterns, and degrading under distribution shift. Intuitively, if the reasoner provides visual evidence for its answer, and the quality of that evidence can be evaluated, the confidence prediction task becomes easier and more robust. For VQA, in addition to the standard answer correctness confidence (c_{corr}), two natural ways of scoring visual evidence signals arise: whether the model localized the relevant region (c_{loc}), and whether its final answer is coherent with the visual evidence it provides (c_{coh}).

As shown in Figure 2, the selector receives question, image, MM-CoT and final answer tuple, and outputs three confidence values: c_{corr} that the answer is correct, c_{loc} that the relevant visual information was localized, and c_{coh} that the final answer is coherent with the evidence. Formally,

$$\{c_{corr}, c_{loc}, c_{coh}\} = s(Q, I, R, A) \in [0, 1]^3. \tag{2}$$

Finally, these individual confidence scores are aggregated to obtain a single confidence score c_{sel}. This confidence score, paired with a threshold, can enable abstention to maintain a desired error rate on the answered questions.

Learning localization confidence. The predicted localization confidence score c_{loc} aims to convey whether the model gathers the relevant visual information required to answer correctly. To learn this confidence score, the model's predicted crops are matched to the ground-truth bounding boxes. Concretely, the ground-truth localization signal g_{loc} is computed as the intersection-over-ground-truth (IoGT) of the predicted crop(s) and the ground truth bounding box(es). For the i-th ground-truth bounding box B^i_{gt} and the j-th predicted crop B^j_{crop},

$$\text{IoGT}^{i,j} = \frac{B^i_{gt} \cap B^j_{crop}}{B^i_{gt}}. \tag{3}$$

If more than one ground truth bounding box is provided, we average the IoGT over all N ground truth boxes, allowing for one predicted crop to match multiple ground truth boxes. Explicitly, we define the mean IoGT, $mIoGT$, as

$$\text{mIoGT} = \frac{1}{N}\sum_{i=1}^{N}\max_{j}\text{IoGT}^{i,j}. \tag{4}$$

Empirically, we find that discretizing this signal, namely using a binary spatial recall [33], performs better than regressing the continuous value. More concretely,

$$g_{loc} = \mathbb{1}[\text{mIoGT} \geq 0.75] \in \{0, 1\}. \tag{5}$$

Although IoGT is less common than IoU in object detection, it is better aligned with our setting: the crop need not tightly segment the object, but must contain enough of it for the reasoner to answer. Appendix D explicitly studies IoGT against IoU-based alternatives and threshold variants.

Learning crop-answer coherence. The coherence confidence score tries to model whether the answering model, after correctly localizing the relevant region, has an answer that is coherent with the relevant crop. Let R^j_{crop} denote the image crop corresponding to bounding box B^j_{crop}. The ground-truth coherence signal g_{coh} is computed as the maximum similarity between any crop and the text of the last message, which includes the final textual reasoning and answer. Formally,

$$g_{coh} = g_{loc} \cdot \max_{j} sim(R^j_{crop}, [R_{last}, A]) \in \{0, 1\}. \tag{6}$$

Empirically, we find it best to determine similarity by asking an external VLM (Qwen 2.5 VL 7B [3]) for a binary coherence judgment $g_{coh} \in \{0, 1\}$, given the zoomed-in crops and last message. Rather than requiring manual annotation, g_{coh} offers a practical and scalable training signal, which we show empirically improves OOD selective prediction. As a sanity check, we run a small human evaluation on 200 randomly chosen training samples and find that human annotations match the same g_{coh} label in 92% of cases. Notably, if the localization is not correct, the coherence score is 0. The inclusion of the coherence signal, g_{coh}, is motivated by the observation that, even after zooming in to the relevant region, the reasoner would in some scenarios provide an answer that was not correlated with the provided crop. Figure 2 shows an example of these labels. In Section 5, Table 2b analyzes the effect of model size on coherence labeling.

Training the selector. For training the selector, all objectives are learned using binary cross-entropy (BCE) loss. The final loss becomes a weighted sum of the losses per objective, with y indicating whether the answer is correct,

$$\mathcal{L} = \lambda_{corr} \cdot \text{BCE}(c_{corr}, y) + \lambda_{loc} \cdot \text{BCE}(c_{loc}, g_{loc}) + \lambda_{coh} \cdot \text{BCE}(c_{coh}, g_{coh}). \tag{7}$$

λ_{corr}, λ_{loc}, and λ_{coh} represent the scalar weights for the correctness, localization, and coherence objectives. Tables and figures abbreviate these weights as λ_c, λ_l, and λ_h, respectively. In Section 5, Table 2a demonstrates that including all objectives is better than excluding any, and shows that our method is robust to variations in the weight values.

Final confidence prediction. At inference, we use the predicted visual evidence scores and the answer correctness confidence prediction, to obtain a final confidence value. These three scores are combined using the same weights as for training, to obtain the final confidence score c_{sel}, i.e.

$$c_{\text{sel}} = \lambda_{corr} \cdot c_{corr} + \lambda_{loc} \cdot c_{loc} + \lambda_{coh} \cdot c_{coh}. \tag{8}$$

4 Experimental setup

4.1 Training setting

Here we overview the training setting for the selector, details in Appendix B.

Selector training data. A selector is tasked with mapping a conversation tuple {Q, I, R, A} into a confidence scalar value $c_{\text{sel}} \in [0, 1]$. Therefore, we first collect training data by generating answers using Pixel-Reasoner [44] as the reasoner model. Pixel-Reasoner [44] has been finetuned to use a zoom-in tool to improve high-resolution VQA accuracy. As the source of these questions, we use the Thyme [60] (high-resolution natural images) and TAT-DQA [63] (financial documents) datasets. All training labels can be generated automatically, as detailed in Appendix B.1.

Method	V* Bench				HR-Bench-8k				MME-RW-L				VizWiz				AdVQA			
	C@5	$\overline{C@r}$	AURC↓	Acc	C@1	$\overline{C@r}$	AURC↓	Acc	C@5	$\overline{C@r}$	AURC↓	Acc	C@20	$\overline{C@r}$	AURC↓	Acc	C@10	$\overline{C@r}$	AURC↓	Acc
Pixel-Reasoner w/o localization																				
Logprobs	0.2	52.7	14.7	79.6	0.0	6.2	30.3	65.6	0.1	8.3	35.3	51.6	0.0	0.1	48.8	38.1	0.1	0.3	42.5	54.5
Selector	2.3	54.6	15.7		0.9	30.8	20.3		0.0	17.8	**30.4**		1.6	3.5	45.6		0.0	7.7	33.7	
Pixel-Reasoner w/ localization																				
Logprobs	0.1	50.2	15.7		0.3	28.9	22.6		0.0	0.7	43.5		0.1	0.2	51.3		0.0	0.0	40.3	
Zero-shot verb.	0.0	42.9	17.2		0.0	13.3	28.9		0.0	0.0	43.5		0.0	0.0	50.8		0.0	0.0	40.7	
Zero-shot logit	0.0	47.1	15.6	80.5	0.0	32.8	21.3	68.4	0.0	4.7	36.4	51.3	0.0	0.0	53.1	38.0	0.0	0.0	38.6	58.1
Selector	1.4	53.1	14.1		0.0	**42.7**	**17.4**		0.0	16.9	31.9		0.1	1.5	44.2		0.2	6.3	33.0	
SIEVES (ours)	**9.7**	**59.5**	**12.8**		**2.3**	41.1	18.1		**4.0**	**18.7**	31.0		**7.6**	**6.6**	**42.3**		**2.8**	**9.3**	**32.2**	
o3 w/o localization																				
Selector	0.1	21.8	23.4	71.8	0.3	33.3	19.2	72.1	1.1	13.4	29.3	56.9	6.3	3.8	40.9	45.8	0.0	19.4	25.2	68.0
o3 w/ localization																				
Zero-shot verb.	0.0	64.7	9.1		0.0	54.2	13.5		0.0	0.0	36.1		0.0	0.0	41.6		0.0	27.5	24.9	
Zero-shot logit	1.7	65.1	9.9	85.9	0.0	56.3	13.1	82.0	0.0	4.8	35.3	58.7	0.0	0.0	39.0	46.9	0.0	30.6	21.9	72.4
Selector	0.0	61.0	9.7		1.0	61.7	10.5		0.1	16.6	28.6		0.0	3.7	35.2		3.8	33.3	20.1	
SIEVES (ours)	**23.1**	**70.8**	**8.4**		**8.0**	**67.7**	**9.3**		**1.7**	**17.5**	**28.4**		**11.7**	**12.4**	**33.3**		**8.2**	**36.5**	**19.6**	
Gemini-3-Pro w/o localization																				
Selector	8.8	68.4	8.8	86.4	5.8	74.5	5.4	85.6	5.1	**34.7**	21.9	60.9	3.8	4.2	41.9	44.1	15.0	55.0	12.8	80.3
Gemini-3-Pro w/ localization																				
Zero-shot verb.	88.3	84.0	3.1		0.0	71.4	6.4		0.0	9.3	28.2		0.0	0.0	46.2		0.0	42.9	17.0	
Zero-shot logit	93.5	84.8	**2.9**	94.3	1.4	79.9	4.9	90.1	0.0	19.5	26.4	63.2	0.0	0.0	47.7	43.6	0.0	42.9	17.3	81.4
Selector	40.7	77.6	3.5		4.5	80.1	**3.5**		1.2	28.2	22.8		0.0	2.7	39.5		32.5	58.9	11.9	
SIEVES (ours)	**95.0**	**85.6**	3.1		**14.4**	**80.8**	4.4		**12.2**	32.3	**21.5**		**8.1**	**8.7**	**37.2**		**33.3**	**59.8**	**11.5**	

Table 1: Selective prediction across reasoners and OOD benchmarks. C@r: coverage at risk; $\overline{C@r}$: average coverage across risk levels {1,5,10,15,20,25,30}%; AURC; Acc: answer accuracy (%), per reasoner, equal across abstention methods. "Selector" is the compute-matched trained correctness-only selector (one Gemma-3-4B call). SIEVES is trained on Pixel-Reasoner answers only; and generalizes to o3 and Gemini-3-Pro and these OOD benchmarks. **Bold** indicates best result per reasoner, underlined second best.

Selector model. We adopt a compact multimodal instruction-tuned model, specifically Gemma-3-4b-it [45], as the selector. Being smaller than the reasoner(s), this allows us to demonstrate that by using a reasoner that provides visual evidence, the selector can be made smaller and hence more efficient. We add three separate value heads, one for each of the confidence scores $c_{corr}, c_{loc}, c_{coh}$. These value heads project the final hidden state of the last sequence element to a scalar value, so predicting several confidence scores adds no overhead to the standard single confidence selector.

4.2 Evaluation

We evaluate on five complementary OOD benchmarks summarized in Table 33, spanning fine-grained high-resolution perception, diagrams and tables, remote sensing, blind-user mobile photos, and human-adversarial visual QA. This diversity stresses different failure modes and provides varied operating points for selective prediction, unlike traditionally used VQA benchmarks which tend to be in-distribution for recent models. Some benchmarks explicitly benefit from zooming, while others such as VizWiz and AdVQA do not strictly require it. We include both regimes because our goal is reliable selective prediction in diverse high-stakes OOD settings where abstention matters.

We report coverage at risk (C@r) at risk levels $r \in \{1, 5, 10, 15, 20, 25, 30\}\%$, together with the average $\overline{\text{C@r}}$ across those levels. C@r measures how much of the data can be answered while keeping the error rate below a target risk level. For small benchmarks, we generate up to five answers per question and compute metrics over the pooled question–answer pairs to reduce noise from answer generation. We also report area under the risk–coverage curve (AURC), a more global summary of ranking quality across the full operating range. For readability, Table 1 reports the lowest risk level at which at least one method reaches double-digit coverage; Appendix E contains results at all considered risk levels. Additional benchmark details, repeated-answer evaluation protocol, and metric definitions are provided in Appendices A and F.

4.3 Baselines

In this paper, we explore whether (i) having a reasoner which localizes the relevant visual evidence by using a zooming tool helps reliability, and (ii) having a selector that explicitly evaluates the grounding of the answer improves estimating answer correctness. Therefore, our baselines are chosen to test both of these hypotheses. Firstly, we run our reasoners (Pixel-Reasoner, o3, and Gemini-3-Pro) with

and without the zoom-in tool, and show whether the same abstention methods can answer more reliably when provided visual evidence. Secondly, we compare SIEVES with various abstention methods. Table 20 in the appendix summarizes the differences between these methods.

Intrinsic reasoner confidence. We extract the log-probability of the final message, i.e. the last reasoning step and the final answer, and use it as the confidence score [4, 57]. We experiment with different sections of the conversation to extract the log-probability from, and find this to perform best in general. Notably, log-probabilities are not available for frontier proprietary models, so we cannot use this baseline for our experiments with o3 or Gemini-3-Pro.

Zero-shot selector baselines. Given the strong zero-shot capabilities of recent VLMs across many domains, it is natural to consider whether they can estimate answer correctness without task-specific training. Therefore we present two zero-shot baselines which repurpose an existing VLM to produce confidence scores. In the zero-shot *logit* selector, the model is asked to think step by step and then output yes or no to express whether it thinks the answer is correct. We extract the log-probability of the yes token (normalized by the sum of the yes and no tokens) and use that as confidence score [26, 47]. Alternatively, we ask the zero-shot *verbalized* selector to output a numerical confidence from 0 to 100 using tokens. For conciseness, Table 1 only reports these baselines for the stronger setting where the answering model is required to provide visual evidence zero-shot. Additional results answering using plain language CoT are provided in Appendix E, where we show it achieves considerably lower coverage and AURC.

Selection without localization estimation. Lastly, we compare to a selector that is trained to predict only correctness, i.e. the scalar confidence that the answer is correct . This is our trained selector baseline: it uses the same Gemma-3-4B architecture, observable reasoner inputs, and single selector call as SIEVES, but is supervised only for answer correctness. Thus, the selector and SIEVES are directly compute matched at inference; the additional localization and coherence value heads are scalar projections of the same final hidden state and add negligible overhead . Notably, we run two versions of this baseline. One is trained on answers generated by a reasoner without access to zoom-in tools, where no localization evidence is provided. In this case the selector must rely on the reasoner's own textual reasoning to infer confidence, or solve the task again to estimate answer correctness. The other version is trained on answers generated by a reasoner with access to zoom-in tools, where localization evidence is provided.

5 Results

Visual evidence facilitates answer selection. Requiring localization for every answer simplifies selective prediction by producing explicit visual evidence. Across our benchmarks, we observe higher coverage and lower AURC in Table 1 when Pixel-Reasoner has access to the zoom-in tool and provides visual evidence. Notably, although zooming in has been predominantly used as means to improve answer accuracy [44, 62], we observe here that even when accuracy improvements are modest, the coverage gains obtained by the selector having access to visual evidence are much greater. This is most evident when comparing $C@10$ for Gemini-3-Pro on AdVQA. Even for the same standard selector, the coverage doubles from 15.0 to 32.5 with a mere 1.1% accuracy improvement. For SIEVES, this is even more pronounced, going from 1.6 C@20 for the correctness-only selector without localization in VizWiz, to 7.6 when using SIEVES, even though the reasoner accuracy actually decreases from 38.1 to 38.0. Importantly, to fully exploit the visual evidence produced by the reasoner, the selector must explicitly estimate the localization quality of each answer. This is evidenced by two correctness-only selectors with identical architectures but different training data: the selector trained on chains from the localized, tool-augmented answerer occasionally performs worse than the one trained on chains from the non-localized answerer, e.g. on MME-RealWorld-Lite and V* Bench (Table 1). Adding visual evidence scoring reverts this trend. At the risk levels reported in Table 1, SIEVES exceeds the compute-matched selector in all 15 reasoner–benchmark settings with localization, while using the same budget of one reasoner call and one Gemma-3-4B selector call. It also improves average coverage in 14/15 settings and AURC in 13/15 settings.

However, when explicitly looking at the generalization to a stronger reasoner, o3 and Gemini-3-Pro in Table 1, we observe that the selector which receives answers with localization transfers more effectively. This further confirms that visual evidence helps generalizing to OOD reasoners, as selectors can overfit textual expressions of confidence which are model-specific.

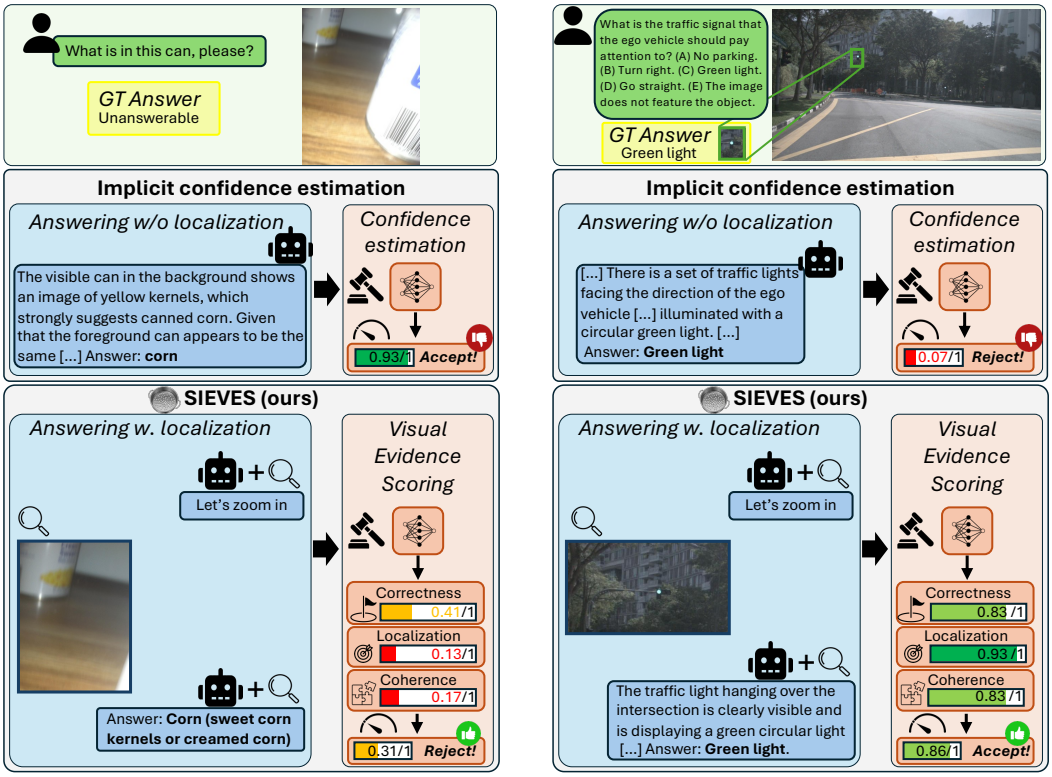


Figure 3: Examples where visual-evidence scores correct a selector error. Left: SIEVES abstains because o3's crop shows a background can and the VizWiz image does not support the requested foreground label. **Right:** SIEVES accepts Gemini-3-Pro's correct answer because its crop shows the green traffic light.

Localization estimation enables generalization. Estimating localization quality alongside answer correctness can maximally exploit the localization evidence provided by the reasoner, further improving the generalization beyond correctness-only selectors. By explicitly requiring that the selector evaluates the grounding of the answer, the selector can discount overconfident but weakly grounded answers and learn to prefer well-grounded ones at the same expressed reasoner confidence. Table 1 showcases SIEVES as the most consistent method, offering improved C@r and AURC.

Figure 3 shows high-stakes qualitative examples where, by explicitly scoring the quality of visual evidence, SIEVES correctly accepts or abstains when the implicit selector fails.

Notably, zero-shot selectors perform very poorly in general. We hypothesize that, although these models are strong zero-shot answerers, estimating comparable confidence scores across the entire distribution of question-answer pairs is a challenging task, for which they have not been trained.

SIEVES selector trained on Pixel-Reasoner generalizes out-of-the-box to o3 and Gemini-3-Pro . An advantage of our design, which takes only model outputs and not model internals like hidden activations or log-probabilities, is that it can be applied to any reasoner. This includes frontier models served through proprietary APIs, which generally do not provide access to such internal signals. Table 1 confirms that this selector design can generalize to stronger reasoners, such as o3 and Gemini-3-Pro. We perform no adaptation for the o3 or Gemini-3-Pro reasoners. They receive the same prompt as the Pixel-Reasoner, and we use the same selector checkpoints which are trained and tuned for abstaining on Pixel-Reasoner answers. Not only does the coverage at relevant risk levels not decrease under the change of reasoner, but it significantly increases in most cases. When selecting among these answers provided by o3 or Gemini-3-Pro with access to tools, the SIEVES selector shows an advantage over the correctness-only selector. Figure 1 highlights that at low risk tolerances, SIEVES consistently achieves highest coverage, often by large margins.

The accuracy results indicate that both zoom-in tool use and stronger reasoners tend to yield higher answer correctness. It is worth noting that the increase in coverage that our SIEVES selector provides is beyond the increase in accuracy provided by these stronger reasoner settings. This means that SIEVES is able to exploit the improved visual evidence provided by the stronger reasoner with

Table 2: Ablations. We study modifications to the default SIEVES configuration, reporting AURC and average coverage ($\overline{\text{C@r}}$). The OOD results are averaged over 5 benchmarks, Thyme MC holdout set is ID. This table reports Pixel-Reasoner; Table 19 reports o3 and Gemini-3-Pro. (a) The default gives the lowest OOD AURC and ties for the highest mean coverage. (b) We compare coherence labels from three VLMs. (c) Automatic boxes give results close to reference boxes. **Bold** indicates best result, underlined second best.

(a) Objective weight ablation.

			ID		OOD	
λ_c	λ_l	λ_h	$\overline{C@r}\uparrow$	AURC↓	$\overline{C@r}\uparrow$	AURC↓
1.0	0.0	0.0	**26.7**	25.8	24.1	28.1
0.0	1.0	0.0	3.0	38.4	17.3	33.0
0.0	0.0	1.0	9.1	34.9	21.4	29.9
0.6	0.4	0.0	24.3	27.0	**27.0**	27.4
0.6	0.0	0.4	21.4	28.0	25.2	27.6
0.6	0.3	0.1	26.6	**25.7**	**27.0**	**27.3**
0.6	0.2	0.2	23.5	27.0	25.6	27.5
0.3	0.3	0.3	19.4	29.6	24.2	28.9
0.2	0.4	0.4	14.1	33.1	21.9	30.2

(b) Crop-answer coherence annotator.

	ID		OOD	
Model	$\overline{C@r}\uparrow$	AURC↓	$\overline{C@r}\uparrow$	AURC↓
Qwen 2.5 VL 7B	26.6	25.7	**27.0**	**27.3**
Gemma 3 4B	24.4	27.1	25.5	28.1
Qwen 2.5 VL 72B	**26.8**	**25.5**	25.5	27.8

(c) Loc. bounding-box annotator.

	ID		OOD	
BBox annot.	$\overline{C@r}\uparrow$	AURC↓	$\overline{C@r}\uparrow$	AURC↓
Human	26.6	25.7	27.0	27.3
Gemini-3-Flash	24.2	27.5	26.3	27.6

Table 3: SIEVES improves reliability with fixed abstention thresholds transferred from a small held-out set. We choose thresholds on held-out Thyme-MC to maximize effective reliability $\Phi_2 = (1 \cdot N_{\text{correct}} - 2 \cdot N_{\text{wrong}} + 0 \cdot N_{\text{abstain}})/N$ and transfer them unchanged to five OOD benchmarks answered by o3. Each correct answer earns $+1$, each mistake costs 2, and each abstention contributes 0; N is the total number of questions.

Method	V*	HR-Bench	MME-RW-L	VizWiz	AdVQA	Mean Φ_2
Selector	6.49	10.42	5.54	1.94	8.40	6.56
SIEVES	**15.60**	**17.93**	**7.04**	**6.25**	**14.26**	**12.22**

access to zoom-in tools. When analyzing the behaviors of the reasoners with tools on V* Bench, we observe that o3 zooms in more (2.3 crops on average vs 1.1) and better (85% of answers have mIoGT ≥ 0.75 vs 37%). Appendix D analyzes this, including qualitative analysis of correct and incorrect localization and answering. This illustrates another generalization dimension, where o3 and Gemini-3-Pro generate significantly higher number of crops on average, since Pixel-Reasoner was explicitly trained to zoom-in just once, and yet our selector trained on Pixel-Reasoner can generalize to o3 and Gemini-3-Pro. Similarly, the correctness-only selector is able to generalize better to o3 and Gemini-3-Pro when observing localization evidence, which was not the case with the in-domain Pixel-Reasoner answers. This can be explained by the fact that, when presented only textual reasoning, the selector often relies on the verbally expressed confidence from the reasoner, which varies across reasoners. In contrast, the visual evidence can be directly assessed by the selector, and better visual evidence further helps this selector to achieve higher coverage. SIEVES also improves selective prediction when the abstention threshold is fixed at inference, as shown in Table 3.

SIEVES improves with each of the fine-grained evidence scoring losses . Table 2a ablates the weights $\lambda_{corr}, \lambda_{loc}, \lambda_{coh}$. The main paper reports the Pixel-Reasoner ablations, while the corresponding o3 and Gemini-3-Pro results are deferred to Table 19 in the appendix. Including all objectives is better than excluding any, and the framework is robust to changes in the weights. Table 2b varies the VLM used to annotate crop-image coherence during training. We find that a larger labeler offers little benefit, whereas a smaller model (in this case the same base model as the selector) can also be used with a modest decrease in performance. Table 2c shows that automatic bounding-box annotation with Gemini-3-Flash enables scalable SIEVES training.

SIEVES improves consistency-based test-time scaling. Sampling several answers and measuring their agreement makes use of the answerer's knowledge of the problem, but repeating long multimodal reasoning is expensive. A single-answer SIEVES selector provides competitive selective prediction at about one-fifth the cost of five-sample consistency (Table 4). Weighting each answer's contribution to agreement using SIEVES improves over both standard consistency and correctness-selector-weighted consistency, for only 0.3% extra cost.

Table 4: SIEVES improves consistency-based abstention at the same five-sample inference budget. Single-answer SIEVES provides competitive reliability at lower cost and latency; SIEVES-weighted consistency further improves results for only 0.3% extra cost. Results average over available reasoners from Pixel-Reasoner, o3, and Gemini-3-Pro; Table 23 gives the per-reasoner breakdown. R and S denote reasoner and trained selector calls, and Gen. S denotes a generative selector. Costs are USD per 100 questions, and latency is estimated assuming parallel calls.

	Cost		Latency		V* Bench			HR-Bench-8k			MME-RW-L			VizWiz		
Abstention method	Answer	Selector	Answer	Selector	C@5	$\overline{C@r}$	AURC↓	C@1	$\overline{C@r}$	AURC↓	C@5	$\overline{C@r}$	AURC↓	C@20	$\overline{C@r}$	AURC↓
Zero-shot logit (1R+1 Gen. S)	$1.845	$0.0117	16.4 s	4.7 s	31.7	65.7	9.5	0.5	56.3	13.1	0.0	9.7	32.7	0.0	0.0	46.6
SIEVES (1R+1S)	$1.845	$0.0070	16.4 s	0.5 s	42.6	72.0	8.1	5.2	54.4	13.7	2.9	18.1	29.7	7.6	6.6	42.3
Consistency (5R)	$9.826	NA	28.8 s	NA	59.4	77.2	5.2	0.0	55.9	12.5	0.0	4.0	33.0	0.0	1.6	45.7
Selector-weighted consistency (5R+5S)	$9.826	$0.0358	28.8 s	0.5 s	65.8	79.3	5.5	11.6	65.1	10.3	8.6	26.2	26.0	7.0	6.8	42.8
SIEVES-weighted consistency (5R+5S)	$9.826	$0.0358	28.8 s	0.5 s	**69.2**	**84.6**	**4.0**	**17.8**	**69.8**	**8.5**	**9.4**	**28.1**	**25.2**	**9.0**	**7.1**	**42.2**

6 Conclusion, limitations, and future work

We introduce SIEVES for generalizable selective prediction in challenging visual question answering tasks. SIEVES requires answering models to provide visual evidence that supports their answers, i.e. image crops. The SIEVES selector uses this evidence to assess how well localized the relevant visual elements are, and whether the crops are coherent with the reasoning and final answer. By explicitly scoring the quality of visual evidence, SIEVES filters out overconfident but weakly grounded answers. On a diverse set of challenging OOD VQA benchmarks, SIEVES achieves up to six times higher coverage at low risk when compared with non-grounding baselines. Additionally, because the SIEVES selector only consumes model outputs, not internals such as log-probabilities or hidden states, we showed that it can generalize from being trained on a weaker open-source model to evaluated using a proprietary frontier model such as o3 and Gemini-3-Pro. Selective prediction will be increasingly important for safely automating end-to-end tasks, and we hope our work can inspire the use of visual evidence to improve reliability in visual tasks.

Limitations. SIEVES improves over the correctness-only selector, but absolute coverage remains low at strict risk levels, especially for weaker reasoners and harder datasets. This limits full automation, as a large percentage of tasks remain unsolved by the automated system, and are instead deferred to an expert or unanswered. Additionally, our experiments cover VQA and use spatial crops as evidence. An ideal selector would generalize to any type of evidence, like code snippets or academic references.

Future work. SIEVES uses image crops as evidence, but the same principle could apply wherever an answer or action can be checked against observable evidence. Improving coverage while maintaining low risk is especially important in critical settings such as medicine and law. In medicine, recent work assesses whether medical sources or patient information support generated claims and diagnoses using zero- or few-shot prompted LLMs, without a task-trained selector [38, 52]. In our visual setting, zero-shot confidence selectors perform substantially worse than our selector trained with fine-grained evidence signals, motivating trained evidence selectors for medical confidence prediction. In law, systems already retrieve and cite legal evidence [30], yet commercial legal RAG systems still hallucinate on 17–33% of queries, including cases where a cited authority does not support the stated proposition [32]. A trained evidence selector could provide higher coverage at a fixed low risk by scoring cited legal evidence and abstaining when it is weak. In coding, current agent-safety evaluators and automated permission modules use prompted LLMs or separate model classifiers to assess whether actions are risky or authorized [22, 24, 39, 46, 58]. They do not assess whether task evidence, such as a failing test or error trace, the relevant code, and a passing test after a patch, supports the proposed edit or the need to execute a given command. A SIEVES-style selector could be trained with fine-grained signals for the relevance and coherence of this evidence, and use them to approve an edit, request more evidence, or defer to an expert before running a catastrophic command.

Acknowledgments and Disclosure of Funding

This research was partially funded by an Alexander von Humboldt Professorship in Multimodal Reliable AI, sponsored by the Federal Ministry of Research, Technology, and Space (BMFTR), by a LOEWE-Spitzen-Professur (LOEWE/4a//519/05.00.002(0010)/93), and has benefited from the Excellence Cluster "Reasonable AI" by the German Research Foundation (Deutsche Forschungsgemeinschaft – DFG) under Germany's Excellence Strategy – EXC-3057. We gratefully acknowledge support from the hessian.AI Service Center (funded by the Federal Ministry of Research, Technology and Space, BMFTR, grant no. 16IS22091) and the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. S-DIW04/0013/003). We thank Tobias Wieczorek and Mohamed Abdelsalam for their feedback on a previous version of this manuscript.

References

[1] Anthropic. System card: Claude opus 4 & claude sonnet 4. https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf, 2025.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. doi: 10.1109/ICCV.2015.279.

[3] 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.

[4] Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. Adaptation with self-evaluation to improve selective prediction in LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5190–5213, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.345. URL https://aclanthology.org/2023.findings-emnlp.345/.

[5] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016. URL https://arxiv.org/abs/1604.06174.

[6] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

[7] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[8] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *NeurIPS*, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/7634ea65a4e6d9041cfd3f7de18e334a-Paper.pdf.

[9] Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. In *CVPR*, pages 24049–24059, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Dancette_Improving_Selective_Visual_Question_Answering_by_Learning_From_Your_Peers_CVPR_2023_paper.html.

[10] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/98ed250b203d1ac6b24bbcf263e3d4a7-Abstract-Conference.html.

[11] Be My Eyes. Be my ai. https://www.bemyeyes.com/be-my-ai, 2026. Accessed on 03/05/2026.

[12] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NeurIPS*, 2017.

[13] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

[14] Google. Lookout: Assisted vision. https://support.google.com/accessibility/android/answer/9031274, 2026. Accessed on 03/05/2026.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. doi: 10.1109/CVPR.2017.670.

[16] Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models, 2024. URL https://arxiv.org/abs/2405.02917.

[17] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3512–3516, 2019.

[18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

[19] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018. doi: 10.1109/CVPR.2018.00380.

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[21] 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 *NeurIPS*, 2024. doi: 10.52202/079017-4423. URL https://doi.org/10.52202/079017-4423. NeurIPS 2024.

[22] John Hughes. How we built Claude Code auto mode. Anthropic, 2026. URL https://www.anthropic.com/engineering/claude-code-auto-mode.

[23] Chaeyun Jang, Moonseok Choi, Yegon Kim, Hyungi Lee, and Juho Lee. Verbalized confidence triggers self-verification : Emergent behavior without explicit reasoning supervision. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=Ub3eXwQ0uA.

[24] Zimo Ji, Zongjie Li, Pingchuan Ma, Wenyuan Jiang, Yudong Gao, and Shuai Wang. Measuring the permission gate: A stress-test evaluation of Claude Code's auto mode, 2026. URL https://arxiv.org/abs/2604.04978.

[25] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

[26] Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In *NeurIPS*, 2024. doi: 10.52202/079017-2729. URL https://doi.org/10.52202/079017-2729.

[27] Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *CVPR*, pages 10854–10863, 2024. doi: 10.1109/CVPR52733.2024.01032.

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[29] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

[30] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275, 2024. doi: 10.1609/aaai.v38i20.30232. URL https://ojs.aaai.org/index.php/AAAI/article/view/30232.

[31] Dan Lucarini and Alan Pelz-Sharpe. Intelligent document processing market analysis 2025-2028: Idp at the crossroads. Technical report, Deep Analysis, 2025. URL https://www.deep-analysis.net/intelligent-document-processing-market-analysis-2025-2028/.

[32] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-free? assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025. doi: 10.1111/jels.12413. URL https://doi.org/10.1111/jels.12413.

[33] Patrick Müller, Mattis Brummel, and Alexander Braun. Spatial recall index for machine learning algorithms. In Graham D. Finlayson and Sophie Triantaphillidou, editors, *London Imaging Meeting 2021: Imaging for Deep Learning, LIM 2021, online, September 20-22, 2021*, pages 58–62. Society for Imaging Science and Technology, 2021. doi: 10.2352/ISSN.2694-118X.2021.LIM-58. URL https://doi.org/10.2352/issn.2694-118X.2021.LIM-58.

[34] Erum Mushtaq, Zalan Fabian, Yavuz Faruk Bakman, Anil Ramakrishna, Mahdi Soltanolkotabi, and Salman Avestimehr. Harmony: Hidden activation representations and model output-aware uncertainty estimation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/TMM-OpenWorld/papers/Mushtaq_HARMONY_Hidden_Activation_Representations_and_Model_Output-Aware_Uncertainty_Estimation_for_CVPRW_2025_paper.pdf.

[35] OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025.

[36] OpenAI. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf, 2025.

[37] OpenAI. Thinking with images. https://openai.com/index/thinking-with-images/, 2025.

[38] Zhiyao Ren, Yibing Zhan, Siyuan Liang, Guozheng Ma, Baosheng Yu, and Dacheng Tao. Towards reliable medical LLMs: Benchmarking and enhancing confidence estimation of large language models in medical consultation, 2026. URL https://arxiv.org/abs/2601.15645.

[39] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.15817.

[40] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. In *NeurIPS*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/aa97d584861474f4097cf13ccb5325da-Abstract.html.

[41] Ayush Shrivastava, Yash Goyal, Dhruv Batra, Devi Parikh, and Aishwarya Agrawal. Real open-ended track leaderboard results, vqa challenge 2021. https://visualqa.org/roe.html, 2021.

[42] Nishad Singhi, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, and Anna Rohrbach. When to solve, when to verify: Compute-optimal problem solving and generative verification for llm reasoning. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.01005.

[43] Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12935–12948, 2024. doi: 10.18653/v1/2024.findings-acl.767. URL https://aclanthology.org/2024.findings-acl.767.

[44] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *NeurIPS*, 2025. doi: 10.52202/085713-0280. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0c38f54740062529aa4117a04b583f3c-Abstract-Conference.html.

[45] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

[46] Maja Trębacz, Sam Arnesen, Ollie Matthews, Dylan Hurd, Won Park, Owen Lin, and Joe Gershenson. Auto-review of agent actions without synchronous human oversight. OpenAI Alignment Research Blog, 2026. URL https://alignment.openai.com/auto-review/.

[47] Neeraj Varshney and Chitta Baral. Post-abstention: Towards reliably re-attempting the abstained instances in QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 967–982, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.55. URL https://aclanthology.org/2023.acl-long.55/.

[48] Megon Venter. 40 intelligent document processing statistics. https://www.pdfreaderpro.com/blog/document-processing-statistics, 2025.

[49] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915, 2025. doi: 10.1609/aaai.v39i8.32852. URL https://ojs.aaai.org/index.php/AAAI/article/view/32852.

[50] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00754.

[51] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *Computer Vision – ECCV 2022*, pages 148–166. Springer, Cham, 2022. doi: 10.1007/978-3-031-20059-5_9.

[52] Kevin Wu, Eric Wu, Kevin Wei, Angela Zhang, Allison Casasola, Teresa Nguyen, Sith Riantawan, Patricia Shi, Daniel Ho, and James Zou. An automated framework for assessing how well LLMs cite relevant medical references. *Nature Communications*, 16:3615, 2025. doi: 10.1038/s41467-025-58551-6. URL https://www.nature.com/articles/s41467-025-58551-6.

[53] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[54] Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. doi: 10.18653/v1/2021.acl-long.84. URL https://aclanthology.org/2021.acl-long.84.

[55] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[56] Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for LLMs. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025. URL https://openreview.net/forum?id=CVRdNQvFPE.

[57] Hiyori Yoshikawa and Naoaki Okazaki. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2017–2028, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.150. URL https://aclanthology.org/2023.findings-eacl.150/.

[58] Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. R-Judge: Benchmarking safety risk awareness for LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1467–1490. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.79. URL https://aclanthology.org/2024.findings-emnlp.79/.

[59] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/df29d63af05cb91d705cf06ba5945b9d-Abstract-Conference.html.

[60] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8080912b919ad0344c0d34e68983dfcb-Abstract-Conference.html.

[61] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

[62] 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. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cdb347d7516d52a9280ea9d6708911f6-Abstract-Conference.html.

[63] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *ACM MM*, 2022. doi: 10.1145/3503161.3548422. URL https://doi.org/10.1145/3503161.3548422.

Supplementary Material

We briefly describe how the appendix is organized. Section A defines the selective-prediction metrics and repeated-answer evaluation protocol. Section B provides training details, automatic-label generation, and train–test overlap checks. Section C studies optional zooming and broad crops. Section D studies localization targets, supervision choices, and localization-related failure modes. Section E provides baseline comparisons, cost and latency, threshold transfer, coverage at all risk levels, and additional qualitative examples. Section F describes the OOD benchmarks. Section G presents the annotation and judging prompts.

A Selective prediction metrics and repeated-answer evaluation

Coverage at risk (C@r). Selective prediction seeks to maximize coverage (fraction of answered, i.e., non-abstained, questions) at a user-chosen risk level (error rate on answered examples) [12, 13]. For sample i, let A^i be the predicted answer and y_i the ground-truth answer, both represented in language space. Let $c^i_{\text{sel}} \in [0, 1]$ denote the final scalar confidence prediction for answer A^i, obtained from the individual confidence scores $\{c_{corr}, c_{loc}, c_{coh}\}$ as described in Section 3, and let $\tau \in [0, 1]$ be the abstention threshold. The coverage at risk r is

$$\text{C@}r \triangleq \text{C@R}(r) = \max_{\tau} \ \text{C}(\tau) \quad \text{s.t.} \quad \text{R}(\tau) \leq r,$$

where

$$C(\tau) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[c^i_{\text{sel}} \geq \tau], \tag{9}$$

$$R(\tau) = \frac{\sum_{i=1}^{N} \mathbb{1}[A^i \neq y_i] \, \mathbb{1}[c^i_{\text{sel}} \geq \tau]}{\sum_{i=1}^{N} \mathbb{1}[c^i_{\text{sel}} \geq \tau]}. \tag{10}$$

Area under the risk–coverage curve (AURC). Reporting coverage at a single risk level provides a snapshot of the performance of a confidence prediction method. Alternatively, one can aggregate the error rate (i.e. risk) across all coverage levels, sorting samples by confidence. The continuous version of AURC is given by:

$$\text{AURC} = \int_0^1 R(C) \, dC = \int_{\tau_{\text{max}}}^{\tau_{\text{min}}} R(\tau) \left| \frac{dC(\tau)}{d\tau} \right| d\tau$$

In practice, we compute the AURC by sorting all samples by confidence, $c_{[1]} \geq c_{[2]} \geq \cdots \geq c_{[N]}$, and then computing risk at each sample (which is equivalent to each coverage level), summing over all samples, starting from the highest confidence ones. We define the risk at the k-th ordered sample as:

$$R_{[k]} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}[A^{[i]} \neq y_{[i]}]$$

from which the discrete AURC follows:

$$\text{AURC} = \frac{1}{N} \sum_{k=1}^{N} R_{[k]}.$$

While AURC is a convenient single-number summary of ranking quality over the entire coverage range, if a desired risk level is known, coverage at risk can be more relevant for deployment.

Repeated answers and aggregation. As described in Section 4, we often sample five answers per question to measure answer-sampling variance. We generate five answers for all three reasoners on V* Bench, which has only 191 questions. Pixel-Reasoner also has five runs on HR-Bench-8k, MME-RealWorld-Lite, and VizWiz; o3 has five runs on HR-Bench-8k and MME-RealWorld-Lite. Other settings have one matched run due to inference cost. Low-risk coverage on a small dataset can change because of only a few answers. The main evaluation pools the independently sampled answer traces across runs to reduce answer-sampling noise. As a consistency check, we also evaluate each

Table 5: Single-answer results on V* Bench (191 questions). Values are mean ± sample SD across repetitions; $n = 1$ rows have no SD. Trained selectors use the fixed ID-selected checkpoint for each metric.

Method	n	C@1	C@5	C@10	C@15	C@20	C@25	C@30	$\overline{C@r}$	AURC	Acc.
Pixel-Reasoner											
Selector, no crops	5	2.5 ± 0.9	2.5 ± 0.9	26.1 ± 17.8	67.2 ± 19.2	95.4 ± 6.5	100.0 ± 0.0	100.0 ± 0.0	56.2 ± 5.6	15.3 ± 3.6	79.6 ± 1.8
Selector	5	2.9 ± 3.1	8.8 ± 11.5	20.6 ± 8.6	66.6 ± 14.8	97.2 ± 6.0	100.0 ± 0.0	100.0 ± 0.0	56.6 ± 2.3	14.1 ± 0.8	80.5 ± 1.6
SIEVES	5	5.5 ± 3.3	16.3 ± 9.1	35.2 ± 18.3	77.1 ± 12.6	99.4 ± 1.1	100.0 ± 0.0	100.0 ± 0.0	61.9 ± 3.5	12.7 ± 1.8	80.5 ± 1.6
ZS logit	5	0.9 ± 2.1	2.3 ± 5.2	9.1 ± 10.1	54.0 ± 33.3	98.7 ± 2.5	100.0 ± 0.0	100.0 ± 0.0	52.2 ± 7.0	15.5 ± 2.7	80.5 ± 1.6
ZS verbalized	5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	17.6 ± 39.3	98.4 ± 2.2	100.0 ± 0.0	100.0 ± 0.0	45.1 ± 5.8	17.1 ± 2.0	80.5 ± 1.6
Logprobs	5	2.3 ± 1.8	2.3 ± 1.8	13.1 ± 11.8	52.1 ± 36.3	98.8 ± 1.8	100.0 ± 0.0	100.0 ± 0.0	52.7 ± 7.0	15.6 ± 2.9	80.5 ± 1.6
o3											
Selector, no crops	5	0.5 ± 0.5	0.5 ± 0.5	0.5 ± 0.5	2.1 ± 2.0	10.5 ± 15.3	46.1 ± 13.1	91.0 ± 20.1	21.6 ± 4.9	23.6 ± 3.5	71.8 ± 1.5
Selector	5	0.3 ± 0.5	1.9 ± 2.0	50.7 ± 24.0	98.4 ± 2.2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	64.5 ± 3.7	9.8 ± 1.4	85.9 ± 1.5
SIEVES	5	12.9 ± 8.7	22.5 ± 11.6	75.7 ± 15.9	98.4 ± 2.2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	72.8 ± 4.8	8.4 ± 1.4	85.9 ± 1.5
ZS logit	5	5.8 ± 5.2	12.1 ± 10.7	80.5 ± 3.4	93.3 ± 9.2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	70.2 ± 1.8	9.2 ± 2.7	85.9 ± 1.5
ZS verbalized	5	0.0 ± 0.0	15.3 ± 34.2	65.9 ± 36.9	93.6 ± 8.8	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	67.8 ± 7.3	8.3 ± 2.0	85.9 ± 1.5
Gemini-3-Pro											
Selector, no crops	5	7.0 ± 6.6	22.3 ± 16.7	62.8 ± 29.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	70.3 ± 6.4	8.8 ± 3.1	86.5 ± 1.1
Selector	5	8.1 ± 3.6	77.0 ± 29.5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	83.6 ± 3.9	3.6 ± 1.2	94.3 ± 1.4
SIEVES	5	14.7 ± 9.6	84.4 ± 20.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	85.6 ± 2.9	3.1 ± 1.0	94.3 ± 1.4
ZS logit	5	22.9 ± 25.2	91.3 ± 7.2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	87.7 ± 3.9	2.9 ± 1.1	94.3 ± 1.4
ZS verbalized	5	0.0 ± 0.0	90.5 ± 5.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	84.4 ± 0.8	3.4 ± 0.9	94.3 ± 1.4

run separately, giving one answer per question in each run. Tables 5 to 9 report the mean and sample standard deviation across runs. These standard deviations measure answer-sampling variability on the fixed benchmark, rather than uncertainty from sampling another test set. The pooled main results are in Table 1. The per-run tables use uniform answer extraction from the stored model outputs. Both trained selectors are evaluated on the full benchmarks.

To formalize the difference between aggregating and averaging over repetitions, let $q \in \{1, \ldots, M\}$ index questions and $t \in \{1, \ldots, T\}$ index sampled answers per question, with $T = 5$, confidence $c_{\text{sel}}^{q,t}$, prediction $A^{q,t}$, and ground-truth answer y_q. The pooled evaluation applies the coverage definition to the full set of MT question–answer pairs:

$$C_{\text{pool}}(\tau) = \frac{1}{MT} \sum_{q=1}^{M} \sum_{t=1}^{T} \mathbb{1}[c_{\text{sel}}^{q,t} \geq \tau]. \tag{11}$$

The pooled operating threshold is therefore

$$\tau_{\text{pool}}^{\star}(r) = \arg\max_{\tau} C_{\text{pool}}(\tau) \quad \text{s.t.} \quad R_{\text{pool}}(\tau) \leq r,$$

and pooled C@r is $C_{\text{pool}}(\tau_{\text{pool}}^{\star}(r))$.

For comparison, an *averaged per-repetition* alternative evaluates each repetition t separately:

$$C_t(\tau_t) = \frac{1}{M} \sum_{q=1}^{M} \mathbb{1}[c_{\text{sel}}^{q,t} \geq \tau_t]. \tag{12}$$

For each repetition t, we choose its own operating threshold

$$\tau_t^{\star}(r) = \arg\max_{\tau} C_t(\tau) \quad \text{s.t.} \quad R_t(\tau) \leq r,$$

and the averaged per-repetition summary is

$$\frac{1}{T} \sum_{t=1}^{T} C_t(\tau_t^{\star}(r)),$$

with sample standard deviations reported across the T repetitions.

B Experimental details

To improve the share of answers with localization, we modify the reasoner's prompt to force it to zoom-in at least once. Since the questions in the training data are open-ended, when there is no string

Table 6: Single-answer results on HR-Bench-8k (800 questions). Values are mean ± sample SD across repetitions; $n = 1$ rows have no SD. Trained selectors use the fixed ID-selected checkpoint for each metric.

Method	n	C@1	C@5	C@10	C@15	C@20	C@25	C@30	$\overline{C@r}$	AURC	Acc.
Pixel-Reasoner											
Selector, no crops	5	1.8 ± 1.0	3.0 ± 3.2	7.0 ± 8.2	20.0 ± 11.9	44.7 ± 6.5	64.5 ± 4.1	81.0 ± 2.9	31.7 ± 3.4	20.4 ± 1.3	65.5 ± 1.0
Selector	5	0.9 ± 0.8	4.4 ± 3.9	21.6 ± 11.3	42.9 ± 6.9	60.8 ± 5.4	78.9 ± 5.2	95.0 ± 2.7	43.5 ± 2.2	17.3 ± 1.5	68.3 ± 1.0
SIEVES	5	2.6 ± 0.8	7.8 ± 2.1	19.3 ± 4.7	35.8 ± 6.9	54.3 ± 6.4	73.5 ± 5.0	94.8 ± 4.5	41.1 ± 2.8	18.1 ± 1.1	68.3 ± 1.0
ZS logit	5	0.7 ± 0.9	0.7 ± 0.9	2.2 ± 2.4	14.2 ± 14.0	43.0 ± 15.6	73.9 ± 3.1	94.2 ± 3.0	32.7 ± 4.7	21.2 ± 1.6	68.3 ± 1.0
ZS verbalized	5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	75.6 ± 42.3	10.8 ± 6.0	29.0 ± 1.0	68.3 ± 1.0
Logprobs	5	0.3 ± 0.3	0.3 ± 0.3	0.3 ± 0.3	1.3 ± 2.2	2.9 ± 5.8	32.5 ± 20.4	84.5 ± 11.3	17.4 ± 3.6	28.1 ± 1.2	68.3 ± 1.0
o3											
Selector, no crops	5	1.3 ± 0.9	1.4 ± 1.1	5.3 ± 3.8	13.7 ± 4.9	32.4 ± 10.8	83.0 ± 9.1	100.0 ± 0.0	33.9 ± 2.9	19.2 ± 0.6	72.1 ± 1.0
Selector	5	3.0 ± 3.0	5.8 ± 7.2	41.1 ± 9.8	88.6 ± 9.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	62.6 ± 3.4	10.4 ± 1.2	82.1 ± 1.0
SIEVES	5	8.2 ± 1.5	22.2 ± 1.2	51.5 ± 8.5	91.2 ± 5.4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	67.6 ± 1.6	9.3 ± 0.8	82.1 ± 1.0
ZS logit	5	0.7 ± 0.7	2.9 ± 5.5	17.1 ± 16.9	86.7 ± 2.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	58.2 ± 2.8	12.7 ± 1.4	82.1 ± 1.0
ZS verbalized	5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	87.8 ± 2.4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	55.4 ± 0.3	13.2 ± 1.0	82.1 ± 1.0
Gemini-3-Pro											
Selector, no crops	1	5.8	31.4	85.0	100.0	100.0	100.0	100.0	74.6	5.5	85.9
Selector	1	4.5	56.5	100.0	100.0	100.0	100.0	100.0	80.1	3.5	90.1
SIEVES	1	14.4	50.9	100.0	100.0	100.0	100.0	100.0	80.8	4.4	90.1
ZS logit	1	1.4	58.0	100.0	100.0	100.0	100.0	100.0	79.9	4.9	90.1
ZS verbalized	1	0.0	0.0	100.0	100.0	100.0	100.0	100.0	71.4	6.4	90.1

Table 7: Single-answer results on MME-RealWorld-Lite (1,782 questions). Values are mean ± sample SD across repetitions; $n = 1$ rows have no SD. Trained selectors use the fixed ID-selected checkpoint for each metric.

Method	n	C@1	C@5	C@10	C@15	C@20	C@25	C@30	$\overline{C@r}$	AURC	Acc.
Pixel-Reasoner											
Selector, no crops	5	0.1 ± 0.1	0.1 ± 0.1	4.9 ± 5.7	17.7 ± 2.3	26.6 ± 1.5	34.5 ± 1.4	44.7 ± 1.6	18.4 ± 1.3	30.4 ± 0.4	51.6 ± 0.3
Selector	5	0.2 ± 0.1	0.3 ± 0.5	4.9 ± 1.4	13.2 ± 2.7	26.0 ± 1.1	33.9 ± 1.2	41.1 ± 1.9	17.1 ± 0.8	31.9 ± 0.4	51.3 ± 0.8
SIEVES	5	1.0 ± 1.1	5.3 ± 1.9	10.3 ± 2.1	17.9 ± 0.9	26.0 ± 1.5	32.5 ± 2.0	41.6 ± 2.0	19.2 ± 0.9	31.0 ± 0.5	51.3 ± 0.8
ZS logit	5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.3 ± 1.9	5.4 ± 4.1	28.2 ± 5.0	5.0 ± 0.8	36.3 ± 0.3	51.3 ± 0.8
ZS verbalized	5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	43.5 ± 1.0	51.3 ± 0.8
Logprobs	5	0.2 ± 0.4	0.2 ± 0.4	0.4 ± 0.6	1.0 ± 1.5	2.3 ± 3.0	7.2 ± 4.2	14.3 ± 4.1	3.6 ± 1.7	39.5 ± 1.7	51.3 ± 0.8
o3											
Selector, no crops	5	0.7 ± 0.5	1.6 ± 0.2	2.4 ± 0.5	5.2 ± 1.2	9.3 ± 1.8	32.2 ± 2.1	45.4 ± 2.7	13.8 ± 0.6	29.3 ± 0.8	56.9 ± 1.1
Selector	5	1.6 ± 0.6	0.4 ± 0.2	10.2 ± 1.4	14.8 ± 1.5	18.5 ± 3.2	26.6 ± 3.5	48.9 ± 6.6	17.3 ± 1.9	28.6 ± 1.3	58.9 ± 0.9
SIEVES	5	1.0 ± 0.2	4.0 ± 1.7	8.9 ± 2.4	15.5 ± 2.2	22.2 ± 1.8	31.0 ± 2.9	45.1 ± 2.8	18.2 ± 1.5	28.3 ± 0.8	58.9 ± 0.9
ZS logit	5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.3 ± 1.8	12.9 ± 3.6	22.8 ± 3.1	5.3 ± 1.0	35.2 ± 0.8	58.9 ± 0.9
ZS verbalized	5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	36.1 ± 0.9	58.9 ± 0.9
Gemini-3-Pro											
Selector, no crops	1	2.4	5.1	24.2	37.4	47.1	57.0	70.2	34.8	21.9	61.3
Selector	1	0.7	1.2	18.2	26.8	37.3	48.5	64.4	28.2	22.9	63.2
SIEVES	1	3.4	12.2	17.2	31.6	39.6	53.5	68.9	32.3	21.5	63.2
ZS logit	1	0.0	0.0	0.0	0.0	28.6	44.9	63.0	19.5	26.4	63.2
ZS verbalized	1	0.0	0.0	0.0	0.0	0.0	0.0	65.0	9.3	28.2	63.2

match between the ground truth and predicted answer, we prompt Qwen3-8B [55] using the ground truth and predicted answer to evaluate whether the predicted answer is correct.

We finetune all linear layers of the backbone using LoRA [20] with rank 512. We train with 10% label smoothing. We select the best checkpoint for each risk level using a multiple-choice held-out set from Thyme [60].

All selectors are trained for 74880 steps with batch size 8, using LLaMA-Factory [61]. Each selector is trained on one node with eight H100 GPUs (80 GB per GPU) and finishes in under 15 hours, or under 120 GPU-hours per run. This is a one-time training cost; the same checkpoint is used across OOD datasets and reasoners. To reduce the memory footprint, we use gradient checkpointing [5] and flash attention 2.0 [10]. We use cosine learning rate scheduling with a warmup period of 211 steps and a maximum learning rate of 1e-6.

Thyme is a high-resolution natural image QA, where objects are often small and localized, comprising 8083 question-image pairs. Since no validation or test splits are provided for Thyme, we construct a holdout set by randomly choosing 750 of the questions (approximately 10%). TAT-DQA is a document QA dataset based on financial reports, annotated by experts, which contains 16,558 question-answer pairs, out of which 13,251 are included in the training split. We generate 5 responses from the reasoner model to each training question.

Table 8: Single-answer results on VizWiz (4,319 questions). Values are mean ± sample SD across repetitions; $n = 1$ rows have no SD. Trained selectors use the fixed ID-selected checkpoint for each metric.

Method	n	C@1	C@5	C@10	C@15	C@20	C@25	C@30	$\overline{C@r}$	AURC	Acc.
Pixel-Reasoner											
Selector, no crops	5	0.1 ± 0.1	0.1 ± 0.1	0.4 ± 0.3	1.1 ± 0.4	2.3 ± 1.6	5.0 ± 5.0	16.7 ± 2.0	3.7 ± 0.9	45.6 ± 0.5	38.1 ± 0.4
Selector	5	0.1 ± 0.1	0.1 ± 0.1	0.3 ± 0.2	1.2 ± 0.8	0.2 ± 0.3	1.2 ± 1.6	9.6 ± 1.0	1.8 ± 0.3	44.2 ± 0.3	38.0 ± 0.1
SIEVES	5	0.1 ± 0.1	0.2 ± 0.4	0.4 ± 0.8	1.9 ± 2.1	6.6 ± 3.3	13.8 ± 2.4	22.1 ± 1.8	6.5 ± 1.3	42.3 ± 0.5	38.0 ± 0.1
ZS logit	5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	53.1 ± 0.8	38.0 ± 0.1
ZS verbalized	5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	50.8 ± 0.5	38.0 ± 0.1
Logprobs	5	0.1 ± 0.1	0.1 ± 0.1	0.1 ± 0.1	0.1 ± 0.1	0.1 ± 0.1	0.2 ± 0.2	0.7 ± 0.8	0.2 ± 0.2	50.6 ± 1.1	38.0 ± 0.1
o3											
Selector, no crops	1	0.0	0.0	0.0	0.0	6.3	5.4	14.6	3.8	40.9	45.8
Selector	1	0.1	0.0	0.2	5.5	0.0	1.3	19.1	3.8	35.2	46.9
SIEVES	1	0.2	0.0	0.0	7.2	11.7	27.2	40.5	12.4	33.3	46.9
ZS logit	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	39.0	46.9
ZS verbalized	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.6	46.9
Gemini-3-Pro											
Selector, no crops	1	0.0	0.0	0.0	0.7	3.8	6.5	18.4	4.2	41.9	44.1
Selector	1	0.0	0.1	1.0	4.2	0.0	4.5	9.4	2.7	39.5	43.6
SIEVES	1	0.0	0.0	0.0	2.0	8.1	21.3	29.2	8.7	37.2	43.6
ZS logit	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	47.7	43.6
ZS verbalized	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.2	43.6

Table 9: Single-answer results on AdVQA (10,000 questions). Values are mean ± sample SD across repetitions; $n = 1$ rows have no SD. Trained selectors use the fixed ID-selected checkpoint for each metric.

Method	n	C@1	C@5	C@10	C@15	C@20	C@25	C@30	$\overline{C@r}$	AURC	Acc.
Pixel-Reasoner											
Selector, no crops	1	0.0	0.0	0.0	0.0	1.0	19.0	33.6	7.7	33.7	54.5
Selector	1	0.0	0.0	0.2	0.9	3.0	11.2	28.8	6.3	33.0	58.1
SIEVES	1	0.0	1.3	2.8	3.5	7.5	15.7	34.1	9.3	32.2	58.1
ZS logit	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	38.6	58.1
ZS verbalized	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	40.7	58.1
Logprobs	1	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	40.3	58.1
o3											
Selector, no crops	1	0.0	0.0	0.0	2.5	7.6	40.8	85.2	19.4	25.2	68.0
Selector	1	0.0	0.1	3.8	15.8	30.1	83.1	100.0	33.3	20.1	72.5
SIEVES	1	0.1	0.1	8.2	18.5	43.6	84.9	100.0	36.5	19.6	72.5
ZS logit	1	0.0	0.0	0.0	0.0	38.0	75.9	100.0	30.6	21.9	72.5
ZS verbalized	1	0.0	0.0	0.0	0.0	0.0	92.4	100.0	27.5	24.9	72.5
Gemini-3-Pro											
Selector, no crops	1	0.1	0.1	15.0	69.6	100.0	100.0	100.0	55.0	12.8	80.3
Selector	1	0.0	0.1	32.5	79.9	100.0	100.0	100.0	58.9	11.9	81.4
SIEVES	1	0.1	6.1	33.3	78.9	100.0	100.0	100.0	59.8	11.5	81.4
ZS logit	1	0.0	0.0	0.0	0.0	100.0	100.0	100.0	42.9	17.3	81.4
ZS verbalized	1	0.0	0.0	0.0	0.0	100.0	100.0	100.0	42.9	17.0	81.4

To increase the frequency of answers with localization, we modify the reasoner's prompt to force it to zoom-in at least once, and only accept the final answer if a zoom-in crop is generated or the maximum number of turns reached, whichever comes first. Appendix C.1 also evaluates optional cropping. This guarantees that for 94% of the questions in Thyme and 77% in TAT-DQA, a zoom-in crop is generated, with 51% and 31% respectively having a mIoGT ≥ 0.75.

For the model selection we employ Thyme's holdout set, cast into a multiple-choice (MC) format for more similarity with MC benchmarks. To construct this MC version, we pad the ground-truth answer with an incorrect response from Pixel-Reasoner (if a wrong one was generated), and GPT-5 [35] generated additional distractor options, until 4 answers per question are obtained.

Table 10: Automatic selector supervision. The selector has no evidence targets. The reference-box row uses a mix of document-derived and reference boxes. The final row uses automatic boxes for every training example. Coherence labels are automatic in both SIEVES rows.

Training setup	ID $\overline{C@r}$	OOD $\overline{C@r}$	OOD AURC↓
Selector	26.7	24.1	28.1
SIEVES, reference boxes	26.6	**27.0**	**27.3**
SIEVES, automatic boxes	24.2	26.3	27.6

To generate multiple choices for the Thyme hold out set, we draw incorrect answers (hard negatives) from Pixel-Reasoner. If these are not enough to reach 4 choices (including the ground-truth), we generate distractor options with GPT-5. The exact prompt templates used for distractor options generation, answering with localization, correctness judging, coherence labeling, and localization annotation are shown in Appendix G.

B.1 Automatic labels and train–test overlap

All coherence labels are generated with Qwen2.5-VL-7B. The supervision and loss-weight ablations use the same pooled protocol as the main evaluation. In the fully automatic localization setting, 63.74% of the training traces use boxes extracted from TAT-DQA document structure, and Gemini-3-Flash generates boxes for the remaining natural images. This natural-image box generation cost about $21 in total. Neither labeler is called at test time. Table 10 shows that this automatic setting remains close to the setting that uses reference boxes.

We check 20,277 unique training QA pairs against 17,092 evaluation QA pairs for exact train–test overlap. One evaluation example shares an image with training data (0.006%). There are no shared questions, QA pairs, or question–image pairs. One image appears in both Thyme and MME-RealWorld-Lite, but with different questions and answers: Thyme asks for the word on the second line ("PANAMA"), while MME asks for the full ship inscription ("MSC NEW YORK PANAMA").

C Further analysis and implementation details

C.1 Optional zooming and broad crops

The main setup requests at least one crop before the final answer to expose visual evidence. This is a data-collection choice, not a requirement that every VQA question needs zooming. We replay the stored Pixel-Reasoner traces while allowing an answer before the first crop. The prompt and sampled model outputs are unchanged; only the stopping rule differs. Of 36 , 547 traces, 6,230 (17.05%) end earlier under the optional policy. Among the 6,183 early traces with comparable parsed answers, 997 (16.12%) change their answer. Table 11 reports selective-prediction results.

We also test VizWiz and AdVQA examples for which the smallest returned crop covers more than 25% or 50% of the image. These questions do not depend on a tiny target region. Table 12 shows that SIEVES still improves mean coverage on both benchmarks and improves AURC in three of four comparisons.

D Localization target and threshold analyses

D.1 Additional supervision choices

Table 13 compares binary localization with continuous mIoGT regression. Binary localization is better on the held-out ID set and the OOD mean.

For coherence label creation, we also considered judging the complete reasoning trace against all returned crops. Compared with judging the final reasoning step and answer against the best-supporting crop, this changes 8.0% of the coherence labels across the training and validation examples, after

Table 11: Optional versus requested zooming with Pixel-Reasoner. The paired replay keeps the prompt and stored samples fixed and changes only whether an answer may be returned before a crop. Higher $\overline{C@r}$ and lower AURC are better.

	V* Bench		HR-Bench-8k		VizWiz		AdVQA	
Zoom policy	$\overline{C@r}$	AURC	$\overline{C@r}$	AURC	$\overline{C@r}$	AURC	$\overline{C@r}$	AURC
Optional	54.7	14.07	40.7	18.53	1.6	46.80	8.6	32.68
Requested	**60.3**	**11.68**	**42.1**	**17.94**	**3.8**	**44.55**	**9.3**	**32.37**

Table 12: Tasks that do not require a tight crop. Results are averaged over Pixel-Reasoner, o3, and Gemini-3-Pro on the same questions. The subset is defined by the smallest returned crop as a share of image area.

		VizWiz			AdVQA		
Min. crop area	Method	C@20	$\overline{C@r}$	AURC	C@10	$\overline{C@r}$	AURC
$> 25\%$	Selector, no tool	10.9	11.0	34.0	**13.6**	34.7	21.5
	SIEVES, tool	**21.9**	**17.0**	**30.1**	11.7	**38.1**	**19.9**
$> 50\%$	Selector, no tool	19.2	16.6	30.1	**19.3**	35.3	**20.2**
	SIEVES, tool	**28.5**	**22.1**	**26.8**	15.4	**39.8**	20.3

applying the localization requirement. The final-step variant performed better on the held-out ID set used for model choice, so we retained it.

D.2 Why IoGT instead of IoU

The localization target should measure whether the relevant object is visible in the crop used by the reasoner. This differs from object detection, where a predicted box is expected to tightly match the object. In our setting, a useful crop may intentionally include context around a small object, so its IoU with the object box can be very small even when the crop contains the full object. This makes IoU a poor target for deciding whether the model looked at the relevant visual evidence.

By contrast, IoGT directly measures object containment: IoGT $= 1$ when the crop contains the entire ground-truth object, regardless of how much surrounding context the crop includes. This is the behavior we want from a zoom-in tool.

D.3 Ablating threshold for binarizing localization

Here, we also ablate the mIoGT threshold used to binarize the localization target, which then propagates to the coherence target:

$$g_{loc} = \mathbb{1}[\text{mIoGT} \geq t], \tag{13}$$

$$g_{coh} = g_{loc} \cdot \max_j sim(R^j_{crop}, [R_{last}, A]), \tag{14}$$

Table 13: Localization label ablation with Pixel-Reasoner. Binary localization matches the decision of whether the needed region is visible.

Label design	ID $\overline{C@r}$	ID AURC	OOD $\overline{C@r}$	OOD AURC
Continuous mIoGT + MSE	26.5	25.8	26.2	27.9
Binary mIoGT ≥ 0.75 + BCE	**26.6**	**25.7**	**27.0**	**27.3**

Table 14: IoU is highly skewed for zoom-in crops. We recover IoU from the stored inverse-log IoU signal on the SIEVES training data. Even permissive IoU thresholds label only a small fraction of crops as localized, because crops are designed to make small objects visible rather than tightly segment them.

Split	IoU ≥ 0.25	IoU ≥ 0.50	IoU ≥ 0.75
Train	3.36%	0.61%	0.04%
Holdout	1.37%	0.72%	0.00%

Table 15: Spatial recall threshold ablation. We vary the mIoGT threshold t in $g_{loc} = \mathbb{1}[\text{mIoGT} \geq t]$, used to define the ground truth label indicating whether localization is correct or not for a reasoning trace. Reported are average coverage ($\overline{\text{C@r}}$; ↑) and AURC (↓), on the ID tuning set (Thyme) and averaged across five OOD benchmarks, for Pixel-Reasoner with localization. The row highlighted in gray denotes the final SIEVES setting ($t = 0.75$), as in the main text.

	ID		OOD	
t	$\overline{\text{C@r}}$ ↑	AURC↓	$\overline{\text{C@r}}$ ↑	AURC↓
0.25	<u>26.5</u>	<u>25.8</u>	25.6	27.8
0.50	24.0	27.1	<u>26.9</u>	<u>27.7</u>
0.75	**26.6**	**25.7**	**27.0**	**27.3**

where $t \in \{0.25, 0.50, 0.75\}$. For Pixel-Reasoner with localization, Table 15 compares these three choices on the Thyme tuning split and across the five OOD benchmarks. The selected value $t = 0.75$ is best in-distribution, achieving the highest average coverage and the lowest AURC. This transfers to OOD, where it also gives the highest average coverage and the lowest AURC.

D.4 Visual evidence quality on V* Bench

One possible concern with IoGT is that a reasoner could satisfy the target with a very broad crop. We therefore analyze the actual V* Bench crops from Pixel-Reasoner and o3 with localization, using the same five repeated answers as in the main evaluation. Table 16 shows that the crops are not broad: the median final crop covers only 4.74% of the image for o3 and 2.67% for Pixel-Reasoner, and no final crop covers more than 25% of the image.

The key difference is instead whether the crop actually contains the relevant object. For o3, the median final crop contains the entire ground-truth object, whereas Pixel-Reasoner often produces a crop that is too tight or mislocalized.

D.5 Failure modes and stratified localization behavior

Figure 4 stratifies V* Bench answers by answer correctness and localization correctness. This highlights two qualitatively different regimes. Pixel-Reasoner often answers correctly even when the crop does not meet the localization target, indicating that answer correctness alone does not guarantee that the provided visual evidence is reliable. In contrast, o3 's improvements are concentrated in the upper-left cell: it is much more often correct while also localizing the relevant evidence. This supports

Table 16: V* Bench visual-evidence quality. We report statistics for the final valid crop on V* Bench, pooled over the five repeated answers used in the main evaluation. A valid crop trace has at least one nonzero-area predicted crop; its percentage is computed over all answer traces. "Crop/image" is the crop area divided by the image area. "Object/crop" is the union area of the ground-truth objects divided by the crop area. "GT recall" is the fraction of the ground-truth object covered by the crop.

Reasoner	Valid crop traces	Median crop/image	Median object/crop	Median GT recall	Crop >25% image
Pixel-Reasoner	90.7%	2.67%	1.38%	54.74%	0.00%
o3	98.0%	4.74%	1.15%	100.00%	0.00%

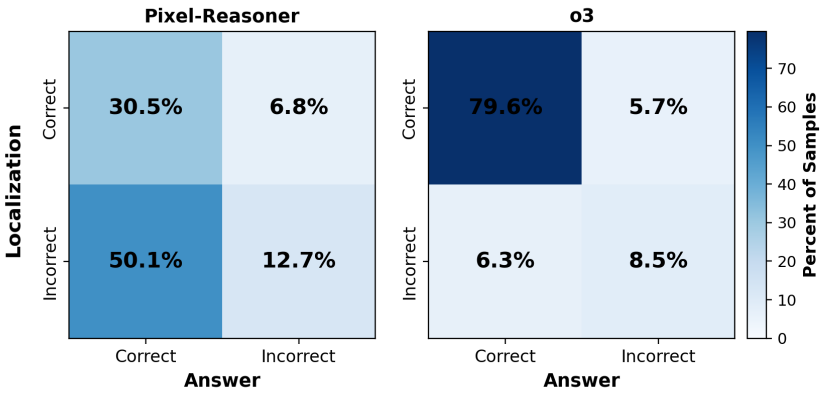


Figure 4: Answer correctness vs. localization correctness on V* Bench. We pool the five repeated answers used in the main evaluation for Pixel-Reasoner and o3 with localization. Localization is correct when mIoGT is at least 0.75. The answer-correct column sums to the corresponding V* Bench answer accuracy in Table 1. o3's accuracy gains coincide with substantially better localization, while Pixel-Reasoner frequently answers correctly despite weak localization evidence.

Table 17: SIEVES preferentially accepts correct answers with correct localization on V* Bench. Columns report mean confidence and the percentage accepted within each group at 5% risk.

Reasoner	Cell	Selector conf.	SIEVES conf.	Selector selected	SIEVES selected
Pixel-Reasoner	Loc. corr. / Ans. corr.	0.60	0.54	3.1%	16.8%
	Loc. corr. / Ans. incorr.	0.47	0.41	0.0%	1.5%
	Loc. incorr. / Ans. corr.	0.54	0.45	1.0%	8.4%
	Loc. incorr. / Ans. incorr.	0.34	0.32	0.0%	2.5%
o3	Loc. corr. / Ans. corr.	0.51	0.48	0.0%	26.7%
	Loc. corr. / Ans. incorr.	0.53	0.38	0.0%	14.8%
	Loc. incorr. / Ans. corr.	0.58	0.41	0.0%	11.7%
	Loc. incorr. / Ans. incorr.	0.41	0.30	0.0%	3.7%

the need to score visual evidence explicitly rather than relying only on final-answer correctness. Additional qualitative examples of incorrect localization and weak crop-answer grounding are shown in Figure 7.

The answer-correct/localization-incorrect cell is not a target success case for SIEVES: even when the final answer is correct, the trace does not provide reliable visual evidence for trusting that answer. We nevertheless analyze this regime because it diagnoses whether a visual-evidence selector simply discards all such answers or can still retain some when the remaining evidence is strong. Figure 5 and table 17 compare the correctness-only selector with SIEVES using the method-specific V* C@5 checkpoints. SIEVES assigns the highest selection rate when both the answer and localization are correct, but it can still accept some traces when only one of these signals is correct. In contrast, traces with both incorrect answers and incorrect localization are selected least often. This is consistent with our goal: the selector should prefer answers supported by visible evidence, without requiring answer correctness and localization correctness to be perfectly correlated.

Table 18 repeats the same analysis after removing traces with no localization attempt or no parseable answer. The trend remains the same: SIEVES selects the answer-correct/localization-correct slice most often, but still accepts some traces when either the answer or the localization signal is correct, and selects the both-incorrect slice least often. In this filtered subset, the o3 correctness-only selector even assigns higher average confidence to the answer-incorrect/localization-incorrect slice than to the answer-correct/localization-correct slice.

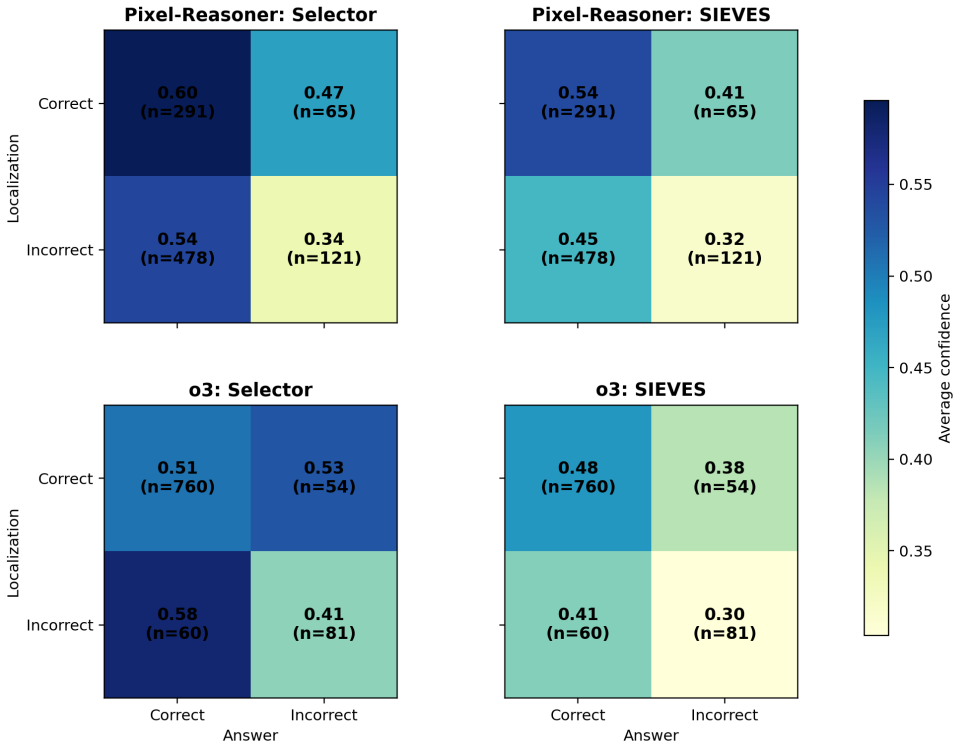


Figure 5: SIEVES confidence better reflects answer and evidence quality. Each cell shows mean confidence and the number of examples in parentheses. For o3, the correctness-only selector is more confident in incorrect answers with correct localization than in answers where both are correct. SIEVES instead assigns its highest confidence when both the answer and localization are correct.

Table 18: SIEVES preferentially accepts correct answers with correct localization on V* Bench. We include examples with a crop and a parseable answer. Columns report mean confidence and the percentage accepted within each group at 5% risk.

Reasoner	Cell	Selector conf.	SIEVES conf.	Selector selected	SIEVES selected
Pixel-Reasoner	Loc. corr. / Ans. corr.	0.60	0.54	3.1%	16.8%
	Loc. corr. / Ans. incorr.	0.47	0.41	0.0%	1.5%
	Loc. incorr. / Ans. corr.	0.54	0.46	1.2%	9.4%
	Loc. incorr. / Ans. incorr.	0.36	0.34	0.0%	2.9%
o3	Loc. corr. / Ans. corr.	0.51	0.48	0.0%	26.7%
	Loc. corr. / Ans. incorr.	0.57	0.40	0.0%	16.0%
	Loc. incorr. / Ans. corr.	0.58	0.41	0.0%	12.3%
	Loc. incorr. / Ans. incorr.	0.53	0.39	0.0%	5.1%

For completeness, Table 19 reports the same selector-design ablations on the two stronger OOD-only reasoners, o3 and Gemini-3-Pro.

E Additional results

E.1 Qualitative examples

Figure 7 shows additional examples where SIEVES correctly accepts or abstains when the implicit selector fails.

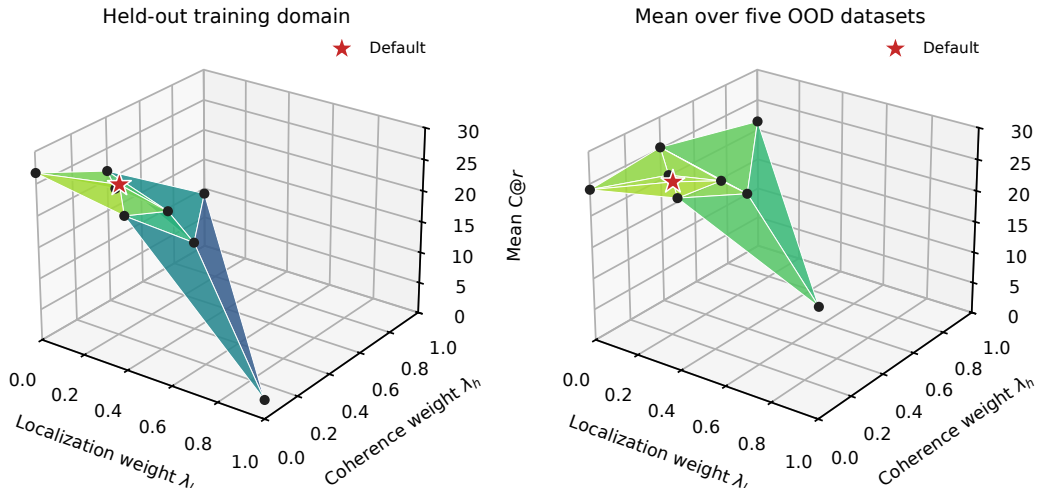


Figure 6: Coverage across the tested objective weights. Each point is one row of Table 2a; the surface only helps show the trend between tested points. Point positions use weights normalized to sum to one. Several nearby mixtures perform similarly. The default $(\lambda_c, \lambda_l, \lambda_h) = (0.6, 0.3, 0.1)$ gives the best OOD AURC in the table.

(a) Objective weight ablation.

				OOD	
f	λ_c	λ_l	λ_h	$\overline{C@r}\uparrow$	AURC$\downarrow$
o3	1.0	0.0	0.0	35.3	20.8
	0.0	1.0	0.0	37.5	20.7
	0.0	0.0	1.0	39.2	19.3
	0.6	0.4	0.0	38.4	20.4
	0.6	0.0	0.4	39.8	20.4
	0.6	0.3	0.1	41.0	19.8
	0.6	0.2	0.2	37.7	21.3
	0.3	0.3	0.3	**42.6**	**19.0**
	0.2	0.4	0.4	41.3	19.6
G3P	1.0	0.0	0.0	49.5	16.2
	0.0	1.0	0.0	36.4	22.9
	0.0	0.0	1.0	40.6	19.8
	0.6	0.4	0.0	52.1	16.5
	0.6	0.0	0.4	53.6	16.6
	0.6	0.3	0.1	53.4	15.5
	0.6	0.2	0.2	**54.7**	15.7
	0.3	0.3	0.3	54.1	**15.2**
	0.2	0.4	0.4	51.6	16.3

(b) Crop-answer coherence annotator.

f	Model	$\overline{C@r}\uparrow$	AURC$\downarrow$
o3	Qwen 2.5 VL 7B	41.0	**19.8**
	Gemma 3 4B	**41.2**	20.3
	Qwen 2.5 VL 72B	38.5	21.0
G3P	Qwen 2.5 VL 7B	53.4	**15.5**
	Gemma 3 4B	53.6	16.2
	Qwen 2.5 VL 72B	**54.7**	15.8

(c) Loc. bounding-box annotator.

f	BBox annot.	$\overline{C@r}\uparrow$	AURC$\downarrow$
o3	Human	41.0	19.8
	Gemini-3-Flash	40.0	21.0
G3P	Human	53.4	15.5
	Gemini-3-Flash	52.1	16.2

Table 19: Additional ablations on stronger OOD reasoners. We report the same selector-design ablations as in Table 2, but for o3 and Gemini-3-Pro only. G3P denotes Gemini-3-Pro. Since these are evaluated purely out-of-distribution, we report OOD metrics only. **Bold** indicates best result per reasoner and subtable, underlined second best, and the gray row marks the default SIEVES configuration.

E.2 Baseline comparisons, cost, and latency

Table 4 separates answering cost from confidence-scoring cost. Costs are estimated from visible token counts and public prices. Pixel-Reasoner uses the hosted Qwen2.5-7B rate as a size-based cost proxy ($0.04/$0.10 per million input/output tokens); this text-model rate may underestimate vision-model hosting cost. The o3 rates are $2/$8, Gemini-3-Pro rates are $2/$12, and Gemma-3-4B rates are $0.05/$0.10 per million input/output tokens. For text, we use the exact local tokenizers for open models, `o200k_base` for o3, and about four characters per token for Gemini. Each selector

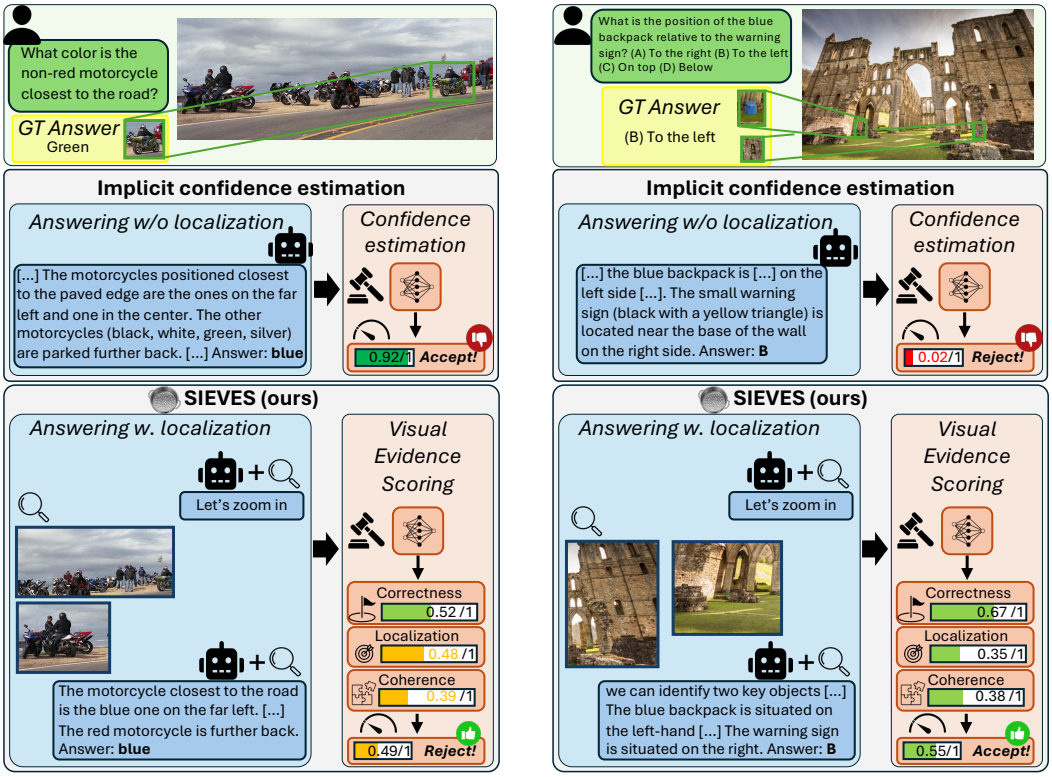


Figure 7: More selector errors corrected by visual evidence. Left: SIEVES abstains because the answer refers to the wrong motorcycle. **Right:** SIEVES accepts the correct relative-position answer because the crop supports it.

Table 20: Comparison with selective-prediction methods. "Internal states" means that a method needs hidden states, logits, or model-specific features. Calls count full reasoner calls per question; ReCoVERR may ask up to 100 auxiliary questions.

Method	Internal states	Reasoner calls	Learned selector	Scores visual evidence
Reliable VQA [51]	yes	1	yes	no
Learning from Peers [9]	yes	1	yes	no
HARMONY [34]	yes	1	yes	no
Consistency [27]	no	> 1	no	no
ReCoVERR [43]	no	1+ auxiliary	no	indirect
SIEVES	no	1	yes	yes

crop adds 256 input tokens; o3 uses 75 base tokens plus 150 per 512-pixel tile, and Gemini uses 1,120 tokens per image. Multi-turn reasoners are charged for the accumulated context at every turn. For closed models, hidden reasoning tokens are unavailable and are not counted. Latency uses provider time-to-first-token and throughput values and assumes independent calls run in parallel. The Gemma pooling selectors generate no output tokens. Cost is averaged within each of the eight reasoner–benchmark settings, then across settings.

The selector has the same Gemma-3-4B backbone, visible inputs, and one-pass budget as SIEVES. We add two black-box comparisons. First, consistency scores five stochastic answers by their agreement, so it uses five full reasoner calls. Second, ReCoVERR asks auxiliary questions until it finds enough evidence or reaches its budget. Table 21 compares ReCoVERR and SIEVES on the same Pixel-Reasoner answers from V* Bench. SIEVES has better C@5, mean coverage, and AURC at much lower estimated cost and latency. Table 22 adds completed searches across all five benchmarks using Gemma for evidence confidence. These searches allow up to 100 auxiliary questions, keep the original answer fixed, and measure whether evidence search improves its confidence ranking. We evaluate

Table 21: Black-box comparison on V* Bench using the same Pixel-Reasoner answers. We evaluate one answer per question from a single repetition on all 191 V* Bench questions. ReCoVERR may ask up to 100 auxiliary questions.

Method	Models used for confidence	Cost $\times 100$	Latency	C@5	$\overline{C@r}$	AURC↓
ReCoVERR	InstructBLIP + GPT-3.5/Flan-T5	\$0.379	51 s	1.6	44.1	24.04
ReCoVERR	Pixel-Reasoner + Gemma-3-4B	\$0.795	$\sim$2.5 min	13.6	54.2	14.05
SIEVES	Pixel-Reasoner + Gemma-3-4B	**\$0.0070**	**0.5 s**	**21.5**	**62.9**	**11.71**

Table 22: ReCoVERR search across five benchmarks. We use a single repetition of Pixel-Reasoner answers. Gemma supplies direct confidence and evidence checks; Pixel-Reasoner answers auxiliary questions. Search allows up to ten rounds with ten questions per round. Scores come from a search run at 5% risk on V* and 1% elsewhere; we evaluate their rankings at the reported target risk. VizWiz uses 4,317 traces.

		Direct Gemma score			ReCoVERR search		
Benchmark	Questions	C@r	$\overline{C@r}$	AURC	C@r	$\overline{C@r}$	AURC
V* Bench	191	0.0	42.9	15.7	0.0	54.5	15.4
HR-Bench-8k	800	2.9	34.0	20.9	0.6	28.6	24.5
MME-RealWorld-Lite	1782	3.5	11.8	35.3	0.2	0.3	41.6
VizWiz	4317	0.7	1.2	47.9	0.3	0.3	54.8
AdVQA	10000	0.0	3.4	36.1	0.0	0.4	36.4

two ReCoVERR configurations: InstructBLIP with GPT-3.5 and Flan-T5, and Pixel-Reasoner with Gemma-3-4B. Both gather evidence through auxiliary visual questions using the benchmark image, without external object detections or image captions. We use the available GPT-3.5 endpoint for the first configuration. Answers that do not meet the recovery criterion retain their direct confidence score.

E.3 Additional baselines and more coverage levels on OOD benchmarks

When evaluating on OOD benchmarks, the SIEVES selector consistently shows an advantage over the correctness-only selector and other baselines. The overall trends are generally consistent across risk levels. We additionally evaluate five-sample consistency-based abstention in Table 23. This baseline assigns each stochastic answer the frequency with which its parsed answer occurs among five samples. The aggregate consistency comparison averages all three reasoners on V* Bench, Pixel-Reasoner and o3 on HR-Bench-8k and MME-RW-L, and Pixel-Reasoner only on VizWiz, where it is the only reasoner with five sampled answers per question. For Pixel-Reasoner it obtains zero coverage at every demanding risk level shown in the main table, although its global V* Bench AURC is competitive. For the more accurate o3 and Gemini-3-Pro reasoners it is highly competitive on V* Bench, but it requires five full reasoner calls and remains weak at the the risk levels reported in Table 1 on HR-Bench and MME-RW-L. Table 24 holds the five sampled answers fixed and compares agreement-based abstention with SIEVES-weighted voting. Weighting each answer by its SIEVES scores improves target-risk coverage, mean coverage, and AURC in 6/7 settings. The exception is o3 on V* Bench, where the high base accuracy and agreement make plain consistency particularly effective. We also compare to a *verbalized* zero-shot selector, which is asked to output a confidence score between 0 and 100 in natural language [16, 23, 56]. Tables 25 to 29 show coverage at all risk levels for the OOD benchmarks. As for the ablation results, low risk levels can be less informative, e.g. in VizWiz, if the task is challenging for the answerer. Similarly, coverage at high risk levels are saturated for V* Bench.

E.4 Frozen-threshold transfer

C@r and AURC test how well a score ranks answers. They do not test whether one raw threshold stays calibrated after a distribution shift. For this stricter test, we choose one threshold for each

	V* Bench			HR-Bench-8k			MME-RW-L			VizWiz		
Reasoner and method	C@5	$\overline{C@r}$	AURC↓	C@1	$\overline{C@r}$	AURC↓	C@5	$\overline{C@r}$	AURC↓	C@20	$\overline{C@r}$	AURC↓
Pixel-Reasoner, consistency	0.0	54.1	12.2	0.0	33.7	18.9	0.0	0.0	36.8	0.0	1.6	45.7
Pixel-Reasoner, SIEVES	9.7	59.5	12.8	2.3	41.1	18.1	4.0	18.7	31.0	7.6	6.6	42.3
o3, consistency	81.3	92.2	2.2	0.0	78.1	6.0	0.0	8.1	29.1	–	–	–
o3, SIEVES	23.1	70.8	8.4	8.0	67.7	9.3	1.7	17.5	28.4	11.7	12.4	33.3
Gemini-3-Pro, consistency	96.9	85.3	1.4	–	–	–	–	–	–	–	–	–
Gemini-3-Pro, SIEVES	95.0	85.6	3.1	14.4	80.8	4.4	12.2	32.3	21.5	8.1	8.7	37.2

Table 23: Five-sample consistency-based abstention. The confidence of each stochastic answer is the fraction of the five samples producing the same parsed answer. Consistency uses five full reasoner calls; SIEVES uses one reasoner call and one smaller 4B selector call. Due to inference cost, o3 has one VizWiz repetition and Gemini-3-Pro has one repetition outside V* Bench; these consistency entries are marked with dashes. AdVQA has one repetition per reasoner and is omitted.

		Agreement / self-consistency			SIEVES-weighted self-consistency		
Reasoner	Benchmark (r)	C@r	$\overline{C@r}$	AURC↓	C@r	$\overline{C@r}$	AURC↓
Pixel-Reasoner	V* Bench (5%)	0.0	54.1	12.1	**28.4**	**70.2**	**8.6**
Pixel-Reasoner	HR-Bench-8k (1%)	0.0	34.2	18.9	**7.9**	**59.5**	**12.1**
Pixel-Reasoner	MME-RW-L (5%)	0.0	0.0	36.3	**11.5**	**27.2**	**26.3**
Pixel-Reasoner	VizWiz (20%)	0.0	1.7	43.2	**12.2**	**9.2**	**38.7**
o3	V* Bench (5%)	**91.1**	**96.0**	**1.4**	85.8	89.3	2.1
o3	HR-Bench-8k (1%)	0.0	80.4	5.2	**30.5**	**84.2**	**3.6**
o3	MME-RW-L (5%)	0.0	8.6	28.5	**7.6**	**30.2**	**22.7**

Table 24: SIEVES complements five-sample test-time scaling. Both methods use the same five stochastic reasoner outputs. Agreement selects the majority answer and uses its frequency as confidence (five reasoner calls, no selector calls). SIEVES-weighted self-consistency sums the SIEVES scores of samples producing each answer, selects the largest sum, and uses that sum divided by five as confidence (five reasoner and five 4B selector calls). SIEVES weighting improves all three selective metrics in 6/7 settings; plain agreement is stronger for o3 on V* Bench.

method on 3,695 held-out Thyme-MC Pixel-Reasoner traces and apply it unchanged to five datasets answered by o3. No target labels are used to choose the threshold.

We report effective reliability $\Phi_c = (N_{\text{correct}} - cN_{\text{wrong}})/N$, where an accepted correct answer is worth $+1$, an accepted error costs c, and an abstention is worth zero. Table 3 reports $c = 2$, for which SIEVES improves all five targets. Across costs $c \in \{1, 2, 4, 5, 10\}$, SIEVES has higher mean transferred utility at four costs; the selector is slightly better at $c = 1$. At $c = 100$, both methods can have negative mean utility and the selector is better. Thus, the fixed-threshold results support score transfer in many settings but do not establish OOD risk control.

Table 31 also reports the aggregate over Pixel-Reasoner, o3, and Gemini-3-Pro. At costs $c = 2, 4, 5$, SIEVES has higher mean effective reliability than the correctness-only selector with or without localization.

E.5 In-distribution results

We also report the coverage at different risk levels for the hold-out sets of the training dataset in Table 32. The SIEVES selector achieves the best coverage at lower risk, being a close second to correctness-only selector for lower-stakes scenarios (i.e. higher risk). However, the differences in performance are not as pronounced as in some OOD cases, again highlighting the generalization capabilities of explicitly estimating localization quality.

Table 25: V* Bench baselines. 🔍 indicates whether the reasoner f explicitly localizes the answer or not. $\overline{C@r}$: average coverage across risk levels 1–30%.

f	🔍	Abstention Method	Acc↑	Coverage at Risk (C@r) ↑							$\overline{C@r}$ ↑	AURC↓
				@1	@5	@10	@15	@20	@25	@30		
Pixel-Reasoner	✗	Logprobs	79.6	0.2	0.2	9.7	61.5	97.5	100.0	100.0	52.7	14.7
		Zero-shot verb.		0.0	0.0	0.0	0.0	93.8	100.0	100.0	42.0	17.3
		Zero-shot logit		0.0	0.0	0.0	0.0	98.0	100.0	100.0	42.6	16.9
		Selector		1.9	2.3	8.4	73.0	96.3	100.0	100.0	54.6	15.7
	✓	Logprobs	80.5	0.1	0.1	9.8	41.3	100.0	100.0	100.0	50.2	15.7
		Zero-shot verb.		0.0	0.0	0.0	0.0	100.0	100.0	100.0	42.9	17.2
		Zero-shot logit		0.0	0.0	5.5	24.2	100.0	100.0	100.0	47.1	15.6
		Selector		1.5	1.4	8.4	60.4	100.0	100.0	100.0	53.1	14.1
		SIEVES (ours)		1.9	9.7	31.7	73.4	100.0	100.0	100.0	59.5	12.8
o3	✗	Zero-shot verb.	71.8	0.0	0.0	0.0	0.0	0.0	0.0	100.0	14.3	26.2
		Zero-shot logit		0.0	0.0	0.0	0.0	23.6	77.3	100.0	28.7	22.7
		Selector		0.1	0.1	0.1	0.1	0.6	51.5	100.0	21.8	23.4
	✓	Zero-shot verb.	85.9	0.0	0.0	84.4	84.4	84.4	100.0	100.0	64.7	9.1
		Zero-shot logit		1.7	1.7	84.2	84.2	84.2	100.0	100.0	65.1	9.9
		Selector		0.0	0.0	26.7	100.0	100.0	100.0	100.0	61.0	9.7
		SIEVES (ours)		3.4	23.1	69.3	100.0	100.0	100.0	100.0	70.8	8.4
Gemini-3-Pro	✗	Zero-shot verb.	86.4	0.0	0.0	0.0	100.0	100.0	100.0	100.0	57.1	12.0
		Zero-shot logit		0.0	0.0	0.0	100.0	100.0	100.0	100.0	57.1	13.8
		Selector		0.6	8.8	69.1	100.0	100.0	100.0	100.0	68.4	8.8
	✓	Zero-shot verb.	94.3	0.0	88.3	100.0	100.0	100.0	100.0	100.0	84.0	3.1
		Zero-shot logit		0.0	93.5	100.0	100.0	100.0	100.0	100.0	84.8	2.9
		Selector		2.3	40.7	100.0	100.0	100.0	100.0	100.0	77.6	3.5
		SIEVES (ours)		4.3	95.0	100.0	100.0	100.0	100.0	100.0	85.6	3.1

F Details on the OOD benchmarks

For VizWiz and AdVQA, we use the validation splits, since the test splits are not public and the evaluation server for the test set does not support selective prediction metrics. These two benchmarks are open-ended, so we use a hard accuracy metric with the same LLM judge used during training.

V* Bench. Wu and Xie [53] target fine-grained perception in high-resolution images, focusing on attribute recognition ("What is the material of the glove?") and spatial relationships ("Is the red suitcase on the left or right side of the red stool?"). Images are 2246 $\times$ 1582 on average, sourced from SA-1B [28], and relevant objects represent a small percentage of image area.

HR-Bench-8k. Wang et al. [49] extend the V* Bench by increasing both the size and resolution by 4. It contains 800 questions on "Fine-grained Single-instance Perception" (aligns with V* Bench's attribute recognition), and "Fine-grained Cross-instance Perception" (aligns with V* Bench's spatial relationship reasoning). We use the 8K resolution version. The domain of the images is broader. Gathered from DIV8K [17], questions pertain not only natural images but various domains like map and chart analysis. This benchmark is more challenging and less sensitive to noise (due to the size) than V* Bench, while still explicitly stresses tiny-object recognition and reading fine details that only appear when zooming.

MME-RealWorld-Lite. Zhang et al. [59] focus on real-world tasks that require perceiving small or far-away elements in high-resolution imagery. The Lite split contains 1,782 unique multiple-choice questions sampled from the full set for faster evaluation, while preserving the benchmark's breadth. The benchmark organizes tasks into five top-level domains, Remote Sensing (RS), Autonomous Driving (AD), Monitoring (MO), Diagram and Table (DT), and Optical Character Recognition (OCR), and up to 43 sub-tasks beneath them. Images are 2,000 $\times$ 1,500 pixels on average, with relevant objects frequently being small. However, the difficulty of this benchmark (top model in the public leaderboard barely reaches 60% accuracy) can be attributed mostly to the nature of the questions, which are much more out of distribution for models compared to the other common VQA benchmarks.

VizWiz. Gurari et al. [19] aim to evaluate models on a real-world setting they are often not trained for: answering questions pertaining images taken by blind users, relating to their daily life. These images are often blurry or even don't contain the object in question (32% of questions in the validation set are marked as non-answerable), and the textual questions are more conversational, as they are

Table 26: HR-Bench-8k baselines. 🔍 indicates whether the reasoner f explicitly localizes the answer or not. $\overline{C@r}$: average coverage across risk levels 1–30%.

f	🔍	Abstention Method	Acc↑	Coverage at Risk (C@r) ↑							$\overline{C@r}$ ↑	AURC↓
				@1	@5	@10	@15	@20	@25	@30		
Pixel-Reasoner	✗	Logprobs	65.6	0.0	0.0	0.3	0.4	3.9	9.1	29.7	6.2	30.3
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	30.6
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	47.1	75.8	17.6	25.9
		Selector		0.9	1.8	2.9	20.2	44.2	64.9	80.7	30.8	20.3
	✓	Logprobs	68.4	0.3	0.5	0.7	7.0	38.0	63.3	92.8	28.9	22.6
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	93.4	13.3	28.9
		Zero-shot logit		0.0	0.0	2.4	12.1	47.3	73.9	93.9	32.8	21.3
		Selector		1.0	3.5	19.4	44.5	58.3	78.4	94.5	42.7	17.4
		SIEVES (ours)		2.3	7.1	19.0	35.0	55.5	73.0	95.6	41.1	18.1
o3	✗	Zero-shot verb.	72.1	0.0	0.0	0.0	0.0	0.0	84.3	100.0	26.3	24.6
		Zero-shot logit		0.0	0.0	0.0	0.0	12.6	82.0	100.0	27.8	22.4
		Selector		0.3	0.8	3.3	12.7	30.3	86.0	100.0	33.3	19.2
	✓	Zero-shot verb.	82.0	0.0	0.0	0.0	89.7	89.7	100.0	100.0	54.2	13.5
		Zero-shot logit		0.0	0.0	16.5	88.7	88.7	100.0	100.0	56.3	13.1
		Selector		1.0	0.8	37.9	92.2	100.0	100.0	100.0	61.7	10.5
		SIEVES (ours)		8.4	22.1	50.5	93.0	100.0	100.0	100.0	67.7	9.3
Gemini-3-Pro	✗	Zero-shot verb.	85.6	0.0	0.0	0.0	100.0	100.0	100.0	100.0	57.1	10.9
		Zero-shot logit		0.0	0.0	80.0	100.0	100.0	100.0	100.0	68.6	10.1
		Selector		5.8	31.4	84.6	100.0	100.0	100.0	100.0	74.5	5.4
	✓	Zero-shot verb.	90.1	0.0	0.0	100.0	100.0	100.0	100.0	100.0	71.4	6.4
		Zero-shot logit		1.4	58.0	100.0	100.0	100.0	100.0	100.0	79.9	4.9
		Selector		4.5	56.5	100.0	100.0	100.0	100.0	100.0	80.1	3.5
		SIEVES (ours)		14.4	50.9	100.0	100.0	100.0	100.0	100.0	80.8	4.4

transcriptions of enunciated questions by the blind users. Unlike the prior three benchmarks we presented, this benchmark is open-ended. Each visual question has ten crowd answers, and items are labeled answerable vs non-answerable. The validation split contains 4,319 examples, of which 2,934 (68%) are answerable. We evaluate on validation rather than test, since test labels are not public and the official server does not support selective-prediction metrics. We do not use train or validation splits for model development. For evaluation to be consistent with that of our other benchmarks, we follow the stricter "hard" accuracy: we select the most frequent human answer per item and count a prediction correct only if it matches such ground truth answer, either by exact match or with the same LLM judge we use during training. For questions which are annotated as unanswerable, any response is considered incorrect, as the optimal strategy is to abstain. Although nominally high resolution (1224 $\times$ 1224 pixels on average), unlike in the other benchmarks, zooming in is generally not required or expected to be beneficial. Nonetheless, the OOD nature of this benchmark, together with the high-stakes nature of the tasks, makes it a valuable testbed for our grounding-aware selective prediction framework.

AdVQA. Sheng et al. [40] introduce an adversarially collected VQA benchmark where crowd workers interact with a strong model and craft valid questions that fool it. This "stress test" distribution is substantially harder than prior benchmarks (e.g. VQAv2 [15]), and highlights brittle reasoning and spurious priors. The validation split (v1.0[2]) contains 10,000 questions. Following our setup for VizWiz described above, we use the validation split using a hard accuracy metric. AdVQA is the largest of the benchmarks we use by total question count and is particularly useful for selective prediction because it is intentionally designed to expose failure cases where abstention should be preferred.

[2] `https://adversarialvqa.org/download.html`

Table 27: MME-RealWorld-Lite baselines. 🔍 indicates whether the reasoner f explicitly localizes the answer or not. $\overline{C@r}$: average coverage across risk levels 1–30%.

f	🔍	Abstention Method	Acc↑	Coverage at Risk (C@r) ↑ @1	@5	@10	@15	@20	@25	@30	$\overline{C@r}$ ↑	AURC↓
Pixel-Reasoner	✗	Logprobs	51.6	0.1	0.1	1.1	4.5	7.0	16.7	28.4	8.3	35.3
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.0
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	27.7	4.0	37.7
		Selector		0.0	0.0	0.0	18.1	27.1	34.8	44.8	17.8	30.4
	✓	Logprobs	51.3	0.0	0.0	0.2	0.2	0.2	1.4	3.0	0.7	43.5
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	43.5
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	4.0	28.7	4.7	36.4
		Selector		0.0	0.0	4.0	13.5	25.9	33.8	41.0	16.9	31.9
		SIEVES (ours)		0.0	4.0	9.9	17.9	25.9	32.4	41.1	18.7	31.0
o3	✗	Zero-shot verb.	56.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	36.8
		Zero-shot logit		0.0	0.0	0.0	0.0	4.5	16.6	32.6	7.7	34.1
		Selector		0.0	1.1	2.2	5.3	8.1	31.1	45.9	13.4	29.3
	✓	Zero-shot verb.	58.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	36.1
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	10.8	23.1	4.8	35.3
		Selector		1.2	0.1	9.6	14.6	16.7	26.4	47.9	16.6	28.6
		SIEVES (ours)		0.7	1.7	7.7	14.4	21.7	30.9	45.2	17.5	28.4
Gemini-3-Pro	✗	Zero-shot verb.	60.9	0.0	0.0	0.0	0.0	0.0	0.0	51.2	7.3	29.5
		Zero-shot logit		0.0	0.0	0.0	12.7	25.8	43.5	57.9	20.0	26.9
		Selector		2.4	5.1	24.2	37.4	47.2	56.8	70.1	34.7	21.9
	✓	Zero-shot verb.	63.2	0.0	0.0	0.0	0.0	0.0	0.0	65.0	9.3	28.2
		Zero-shot logit		0.0	0.0	0.0	0.0	28.6	44.9	63.0	19.5	26.4
		Selector		0.7	1.2	18.2	26.8	37.3	48.5	64.4	28.2	22.8
		SIEVES (ours)		3.4	12.2	17.2	31.6	39.6	53.5	68.9	32.3	21.5

Table 28: VizWiz baselines. 🔍 indicates whether the reasoner f explicitly localizes the answer or not. $\overline{C@r}$: average coverage across risk levels 1–30%.

f	🔍	Abstention Method	Acc↑	Coverage at Risk (C@r) ↑ @1	@5	@10	@15	@20	@25	@30	$\overline{C@r}$ ↑	AURC↓
Pixel-Reasoner	✗	Logprobs	38.1	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.1	48.8
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.5
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	48.1
		Selector		0.0	0.0	0.1	0.7	1.6	5.4	16.5	3.5	45.6
	✓	Logprobs	38.0	0.0	0.0	0.0	0.0	0.1	0.3	1.1	0.2	51.3
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	50.8
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	53.1
		Selector		0.0	0.0	0.1	0.6	0.1	0.1	9.3	1.5	44.2
		SIEVES (ours)		0.0	0.0	0.0	1.1	7.6	14.7	22.5	6.6	42.3
o3	✗	Zero-shot verb.	45.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	43.1
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	47.1
		Selector		0.0	0.0	0.0	0.0	6.3	5.4	14.6	3.8	40.9
	✓	Zero-shot verb.	46.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.6
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	39.0
		Selector		0.1	0.0	0.2	5.5	0.0	1.3	19.1	3.7	35.2
		SIEVES (ours)		0.2	0.0	0.0	7.2	11.7	27.2	40.5	12.4	33.3
Gemini-3-Pro	✗	Zero-shot verb.	44.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	43.8
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	45.0
		Selector		0.0	0.0	0.0	0.7	3.8	6.5	18.4	4.2	41.9
	✓	Zero-shot verb.	43.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.2
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	47.7
		Selector		0.0	0.1	1.0	4.2	0.0	4.5	9.4	2.7	39.5
		SIEVES (ours)		0.0	0.0	0.0	2.0	8.1	21.3	29.2	8.7	37.2

Table 29: AdVQA baselines. 🔍 indicates whether the reasoner f explicitly localizes the answer or not. $\overline{C@r}$: average coverage across risk levels 1–30%.

f	🔍	Abstention Method	Acc↑	Coverage at Risk (C@r) ↑ @1	@5	@10	@15	@20	@25	@30	$\overline{C@r}$ ↑	AURC↓
Pixel-Reasoner	✗	Logprobs	54.5	0.0	0.0	0.1	0.1	0.1	0.5	1.5	0.3	42.5
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	42.7
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	42.2
		Selector		0.0	0.0	0.0	0.0	1.0	19.0	33.6	7.7	33.7
	✓	Logprobs	58.1	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	40.3
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	40.7
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	38.6
		Selector		0.0	0.0	0.2	0.9	3.0	11.2	28.8	6.3	33.0
		SIEVES (ours)		0.0	1.3	2.8	3.5	7.5	15.7	34.1	9.3	32.2
o3	✗	Zero-shot verb.	68.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.8
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	33.6	4.8	30.2
		Selector		0.0	0.0	0.0	2.5	7.6	40.8	85.2	19.4	25.2
	✓	Zero-shot verb.	72.4	0.0	0.0	0.0	0.0	0.0	92.4	100.0	27.5	24.9
		Zero-shot logit		0.0	0.0	0.0	0.0	38.0	75.9	100.0	30.6	21.9
		Selector		0.0	0.1	3.8	15.8	30.1	83.1	100.0	33.3	20.1
		SIEVES (ours)		0.1	0.1	8.2	18.5	43.6	84.9	100.0	36.5	19.6
Gemini-3-Pro	✗	Zero-shot verb.	80.3	0.0	0.0	0.0	0.0	100.0	100.0	100.0	42.9	16.6
		Zero-shot logit		0.0	0.0	0.0	0.0	100.0	100.0	100.0	42.9	17.4
		Selector		0.1	0.1	15.0	69.6	100.0	100.0	100.0	55.0	12.8
	✓	Zero-shot verb.	81.4	0.0	0.0	0.0	0.0	100.0	100.0	100.0	42.9	17.0
		Zero-shot logit		0.0	0.0	0.0	0.0	100.0	100.0	100.0	42.9	17.3
		Selector		0.0	0.1	32.5	79.9	100.0	100.0	100.0	58.9	11.9
		SIEVES (ours)		0.1	6.1	33.3	78.9	100.0	100.0	100.0	59.8	11.5

Table 30: Transferred thresholds at several error costs. One threshold per method and cost is chosen on held-out Thyme-MC and applied unchanged to o3 answers. Coverage, realized error rate, and effective reliability Φ_c are reported in percent. Error rates after transfer may exceed the source risk level.

		V* Bench			HR-Bench-8k			MME-RW-L			VizWiz			AdVQA			
Method	c	Cov.	Err.↓	Φ_c ↑	Cov.	Err.↓	Φ_c ↑	Cov.	Err.↓	Φ_c ↑	Cov.	Err.↓	Φ_c ↑	Cov.	Err.↓	Φ_c ↑	Mean Φ_c ↑
Selector	1	43.0	14.4	30.7	34.7	12.8	25.8	27.7	25.5	13.6	34.9	35.5	10.2	39.6	21.3	22.8	20.60
SIEVES	1	27.6	7.6	23.5	31.1	7.4	26.5	18.5	17.5	12.0	31.0	26.8	14.4	40.1	19.5	24.5	20.17
Selector	2	16.2	20.0	6.5	15.0	10.2	10.4	10.5	15.8	5.5	14.9	29.0	1.9	20.3	19.6	8.4	6.56
SIEVES	2	17.8	4.1	15.6	20.9	4.7	17.9	11.4	12.7	7.0	21.0	23.4	6.3	29.6	17.3	14.3	12.22
Selector	4	2.0	0.0	2.0	5.7	1.8	5.2	3.3	4.4	2.6	6.2	15.4	1.4	10.8	13.8	3.4	2.91
SIEVES	4	6.4	3.3	5.3	8.8	2.0	7.9	3.7	4.5	2.9	8.5	16.9	1.3	14.9	13.2	5.1	4.51
Selector	5	2.0	0.0	2.0	5.7	1.8	5.1	3.3	4.4	2.5	6.2	15.4	0.5	10.8	13.8	1.8	2.37
SIEVES	5	3.9	2.7	3.2	6.5	0.4	6.4	2.7	3.8	2.1	5.7	14.5	0.7	11.3	11.8	3.3	3.15
Selector	10	0.1	100.0	-1.0	0.4	0.0	0.4	0.2	0.0	0.2	2.5	19.4	-2.8	1.6	12.7	-0.7	-0.78
SIEVES	10	0.1	0.0	0.1	0.9	0.0	0.9	0.3	0.0	0.3	1.1	12.5	-0.4	2.5	7.3	0.5	0.28
Selector	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	50.0	-1.0	-0.19
SIEVES	100	0.0	0.0	0.0	1.3	0.0	1.3	0.4	0.0	0.4	0.5	13.0	-6.5	1.9	9.2	-15.3	-4.02

Table 31: Fixed-threshold transfer averaged over three reasoners. Effective reliability Φ_c (%) averaged over Pixel-Reasoner, o3, and Gemini-3-Pro. Thresholds are chosen on held-out training-domain data and applied unchanged to the OOD benchmarks.

Reasoning + abstention method	c	V*	HR-Bench	MME-RW-L	VizWiz	AdVQA	Mean Φ_c
No localization + Selector	2	10.89	19.12	8.54	2.16	9.63	10.07
Localization + Selector	2	12.70	16.82	8.98	0.32	9.11	9.59
Localization + SIEVES	2	**18.15**	**25.64**	**10.21**	**4.15**	**14.40**	**14.51**
No localization + Selector	4	2.23	2.76	2.59	-0.23	0.49	1.57
Localization + Selector	4	3.84	8.20	2.94	**-0.11**	1.90	3.35
Localization + SIEVES	4	**7.19**	**12.65**	**4.97**	-0.79	**4.54**	**5.71**
No localization + Selector	5	2.16	2.31	2.22	**-0.94**	-0.26	1.10
Localization + Selector	5	3.67	7.96	2.44	-1.71	-0.22	2.43
Localization + SIEVES	5	**5.31**	**9.69**	**3.32**	-1.37	**2.66**	**3.92**

Table 32: Thyme-750 (MC) baselines. 🔍 indicates whether the reasoner f explicitly localizes the answer or not. $\overline{C@r}$: average coverage across risk levels 1–30%.

f	🔍	Abstention Method	Acc↑	Coverage at Risk (C@r) ↑							$\overline{C@r}$ ↑	AURC↓
				@1	@5	@10	@15	@20	@25	@30		
Pixel-Reasoner	✗	Logprobs	53.7	0.3	0.3	0.8	1.1	2.1	4.8	19.3	4.1	36.3
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	36.8
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	39.2
		Selector		0.2	0.2	1.4	10.0	17.1	35.0	53.4	16.8	30.1
	✓	Logprobs	54.0	0.1	0.1	0.1	0.1	0.4	1.1	1.2	0.4	42.0
		Zero-shot verb.		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	44.6
		Zero-shot logit		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	42.5
		Selector		0.8	1.1	8.4	22.3	39.6	51.9	63.0	26.7	25.8
		SIEVES (ours)		0.7	2.1	9.9	24.9	38.7	50.3	59.6	26.6	25.7

Benchmark	#Q	Format	Avg. res.	Domain
V* Bench [53]	191	MC	2246×1582	Natural images: Attribute recognition, relative position
HR-Bench-8k [49]	800	MC	7680×4320	Attribute recognition, OCR, map & chart analysis, spatial reasoning
MME-RW-L [59]	1,782	MC	2000×1500	Remote sensing, autonomous driving, monitoring, diagrams & tables, OCR
VizWiz [19]	4,319	OE	1224×1224	Blind-user mobile photos: yes/no, counting, other (32% unanswerable)
AdVQA [40]	10,000	OE	640×480	Adversarially crafted: counting, OCR, rare concepts, reasoning

Table 33: Out-Of-Distribution Benchmarks. We evaluate SIEVES on diverse benchmarks of varying sizes (#Q), using Multiple-Choice (MC) and Open-Ended (OE) formats, some of which don't require to zoom-in.

Thyme multiple-choice distractor generation prompt

```
System message:
You are an expert at creating plausible but incorrect distractor options for visual
     question answering tasks. Given a question, image, ground truth answer, and
     optionally a wrong answer from a model, generate additional plausible distractor
     options that are wrong but could seem reasonable.

User message (same-category example):
Category: {category} Question: {question} Ground truth answer: {ground_truth}

Here is the cropped region of interest: [cropped ground-truth box image]

Generate 2-3 plausible but incorrect distractor options.

Assistant message (example distractors):
[example distractor 1]
[example distractor 2]
[example distractor 3]

User message (current Thyme example):
Category: {category} Question: {question} Ground truth answer: {ground_truth}
Wrong answer from a model: {wrong_answer} [included only when available]

Here is / Here are the cropped region(s) of interest from the image: [one or more
     cropped ground-truth box images]

Generate exactly {num_distractors_needed} plausible but incorrect distractor options.
     Make them specific and realistic based on the image. Return only the distractors,
     one per line, without numbering.
```

Figure 8: Thyme multiple-choice distractor generation prompt. We use this prompt to turn Thyme examples into multiple-choice questions by generating plausible wrong options that stay close to the image and question. The template is shown as a four-message chat: a system instruction, one fixed same-category example, its assistant response, and the current Thyme example to complete. The Thyme categories are `ocr_recognition`, `quantity_recognition`, `object_recognition`, `attribute_recognition`, `position_recognition`, and `chart_understanding`; the few-shot example is chosen from the same category as the current question. If an incorrect Pixel-Reasoner answer is available, it is included so the generated distractors can target realistic model mistakes.

G Prompt templates

The exact prompts referenced in the experimental details are collected here for completeness.

Thyme multiple-choice distractor generation prompt. We use this prompt to convert Thyme into a multiple-choice setting by generating hard distractors conditioned on the question, the relevant crop, and optionally a wrong model answer. In the main experiments, this chat template is queried with GPT-5. The full message sequence is defined in Figure 8.

Reasoner with localization prompt. To encourage explicit visual evidence, we augment the reasoner prompt with the instruction block in Figure 9. This tells the model to use the crop tool before giving its first final answer, so the resulting trace contains localization evidence . The reasoners emit plain-text `crop_image_normalized` calls with `bbox_2d` corner coordinates in $[0, 1]$ and a one-based `target_image` index. The executor converts the box to pixels, clips it to the image bounds, and returns the crop as another image in the conversation. The spatial-evaluation files express boxes in $[0, 1000]$; these are converted coordinates for measuring overlap, rather than the coordinates requested in the reasoner's tool schema .

Correctness judge prompt. For open-ended questions, when normalized string match fails, we run the prompt in Figure 10 to decide whether the prediction should still count as correct. This makes correctness labels more robust to valid rewordings and semantically equivalent answers.

Grounding-coherence labeling prompt. We use the prompt in Figure 11 to create the crop-answer coherence labels used for selector training. The judge receives the question, the reasoner's final

Prompt for reasoner with localization

```
Guidelines: Understand the given visual information and the user query. Employ the given
    visual operations (tools) to observe better the visual elements necessary to answer
    the question. For an image, we can look closer by ‘crop_image_normalized‘. Please
    always crop into the relevant region before answering the question. You must use
    the cropping tool first and finish your turn. After the user returns the zoomed-in
    image, you can reason about what you observe and give your final answer. You must
    always use the cropping tool in the first turn, end your turn, wait for the cropped
    image. If you see the relevant object or objects needed to answer the question
    clearly, you can give the final answer. Otherwise, try to use the
    ‘crop_image_normalized‘ tool again, until you find the relevant object of objects.
    You have up to 10 turns. Reason with the visual information step by step, and put
    your final answer within \boxed{}. It is very important to follow these
    instructions.
```

Figure 9: Prompt for reasoner with localization. We prompt the reasoner with the tool definition and the guidelines presented here, which encourage it to use the zoom-in tool to provide visual evidence for its answer.

answer, and one predicted crop, and decides whether the crop contains enough evidence and whether the answer is supported by that crop.

Localization labeling prompt. For automatic localization annotation on Thyme, we prompt Gemini-3-Flash with the two-stage pipeline defined in Figure 12. The first stage extracts the target object phrase from the question, and the second localizes it in the full image.

Open-ended correctness-judge prompt

```
Compare the predicted answer to the ground truth answer and determine if they convey the
    same meaning, and if the model was likely referring to the same object or situation
    when answering visual questions.

You can think step-by-step about whether the predicted answer conveys the same meaning
    as the ground truth answer, but after that, output only ANSWER: and yes or no after
    that.
Sometimes, the predicted answer will also contain the model's thinking process and
    justificatioon. You can take it into account, but focus on the validity of the
    final answer, which will generally appear at the end.
If the ground truth answer is only 't', this means the question is not answerable. In
    that case, you should always mark the question as wrong, because no answer can be
    correct, not even saying that there is no answer.

For example:

Question: What is the person holding?
Predicted answer: Blue pullover
Ground truth answer: Sweater

A pullover is a type of sweater. The colour is not mentioned in the ground truth answer,
    so we can assume it is correct. ANSWER: yes

## Another example ##
Question: Which object was put down by the person?
Predicted answer: jacket
Ground truth answer: The shoe.
A jacket is a type of clothing, similar to a shoe, which is also a type of
    clothing/accessory.
However, a jacket does not look similar to a shoe, so it is unlikely that the model
    confused them, and instead simply answered incorrectly. ANSWER: no
- The other options (food, blanket, sandwich) are not related to clothing or accessories.

## Another example ##
Question: What verification is this paper for?
Predicted answer: freeboard verification
Ground truth answer: Freeboard

Clearly the model shows in its response it refers to the verification being a freeboard
    verification, which conveys the same meaning. ANSWER: yes

## Real user request ##
Question: {question}
Predicted answer: {pred_answer}
Ground truth answer: {gt_answer}
```

Figure 10: Open-ended judge prompt. For open-ended questions, we use this prompt when normalized exact match fails to decide whether the predicted answer should still count as correct. This makes correctness labels more robust to valid rewordings and semantically equivalent answers that exact match would reject. We prompt Qwen3-8B with a single user message at temperature 0. If multiple ground-truth answers are available, the judge is queried sequentially and stops at the first positive match.

Grounding-coherence labeling prompt for selector training

```
You are an expert evaluator assessing whether a model's response to a visual question is
    grounded in the provided image crop.

Given:
- Question: {question}
- Model's Response (with final answer in \boxed{}): {last_message_with_boxed_answer}
- Image: [Provided image crop]

Please evaluate two aspects:

1. **Crop Sufficiency**: Is the provided image crop sufficient to support the model's
    response? Does it contain all the necessary visual information referenced in the
    response? If the model explicitly states they use the global view to answer this
    question, you should consider this as not grounded in the prompt. Note you are not
    provided this final image, and only the crop, which the model should only use to
    give the final answer.

2. **Answer Coherence**: Is the model's response coherent with what is actually visible
    in the image? Or is the model hallucinating information or obtaining it from
    elsewhere (not from the image)?

Think step by step about both aspects, then provide your final assessment.

Output your final decision as \boxed{Yes} if the answer is well-grounded in the image
    crop (both crop is sufficient AND answer is coherent), or \boxed{No} if there are
    issues with either aspect.

Examples:
- If the crop shows a clear view of a red car and the model answers "red car" ->
    \boxed{Yes}
- If the crop shows a partial view that doesn't contain enough information to answer ->
    \boxed{No}
- If the crop shows a dog but the model answers "cat" -> \boxed{No}
- If the crop shows a room but the model mentions specific details not visible in the
    crop -> \boxed{No}

Your response:
```

Figure 11: Crop-answer grounding coherence labeling prompt for selector training. We use this prompt to create the crop-answer coherence labels used to train the selector. The judge receives the question, the reasoner's final answer, and the visual evidence (crop), and decides whether that crop contains enough evidence and whether the answer is supported by what is visible. The parser reads `\boxed{Yes}` or `\boxed{No}` as a binary coherence label.

Subprompt 1: target object extraction

```
System:
You extract the target object from a visual question. Return only a short noun phrase
    describing the object to locate in the image. If the question uses relational or
    positional language, keep that context (e.g., 'object on the woman's left ring
    finger').
User:
{question}
```

Subprompt 2: full-image grounding

```
System:
You are a precise visual grounding assistant. Return only JSON.
User:
Please return the bounding box coordinates of "{target_object}". Use normalized 0-1000
    coordinates in [top, left, bottom, right] order. Return a JSON list like:
    [{"box_2d": [top, left, bottom, right], "label": "..."}].
Image: [Provided full image]
```

Figure 12: Localization annotation prompt for SIEVES training. We use this two-stage pipeline to automatically annotate Thyme with bounding boxes indicating the object referred to in the question. The two subprompts are shown separately. The first call extracts the target object phrase from the question, and the second call takes the full image plus that extracted phrase and returns one or more boxes.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main empirical claims and scope, namely grounding-aware selective prediction, transfer to stronger proprietary reasoners, and evaluation on five OOD VQA benchmarks. See the abstract and Sections 1 and 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss that SIEVES uses bounding-box annotations and a separate VLM for crop-answer coherence labeling during training, which adds annotation/modeling cost. We argue this is a one-time training cost for improved reliability, and ablate scalable alternatives: bounding boxes can be generated automatically with little loss, and the same base model can be used for coherence labeling with competitive coverage; see Section 3 and tables 2b, 2c and 19.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical and does not present theorems or formal proofs requiring this kind of treatment.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the method, selector objectives, evaluation protocol, benchmark setup, training hyperparameters, and prompt templates used for judging and pseudo-annotation. See Sections 3 and 4 and appendices A, B, F and G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The submission does not currently provide an anonymized code release or reproduction package. We use public benchmarks and document the setup in the paper and appendix, but the code and exact run scripts are not part of this submission, although the code will be made public at a later date.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The training data sources, selector architecture, loss, benchmark choices, splits, repeated-answer protocol, and key hyperparameters are described in the main text and appendix. See Sections 3 and 4 and appendices B and F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We improve statistical robustness by sampling multiple inference responses per question and aggregating selective-prediction metrics over the pooled set of answers, as detailed in Appendix A. We also study pooled evaluation versus per-repetition evaluation and averaging in Table 5, finding that pooling is more robust while both estimates are empirically similar; this is complemented by evaluation across five OOD benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss compute resources in the checklist: selector training uses H100 GPUs with 80GB memory, and training completes in under 15 hours on one 8-GPU node.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and the work is consistent with it to the best of our knowledge. The paper studies reliability and abstention on public benchmarks and aims to reduce harmful overconfident behavior in high-stakes settings.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss that improving selective reliability can make VLMs more useful in high-stakes settings where abstention is preferable to unreliable answers. At the same time, increasing reliability also raises the feasibility of more autonomous systems, so deployment should still consider the consequences of incorrect accepted answers and deferred decisions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model or a scraped dataset. Its primary contribution is a selective-prediction method intended to make deployment safer by enabling abstention.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper credits the datasets, models, and software we build on via citations, but it does not exhaustively document licenses and terms of use for every external asset in the submission itself.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The submission does not release a new dataset, model, or code asset as part of the paper package.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not report a crowdsourcing study or a study with recruited human participants. It uses public datasets and includes only a small internal manual sanity check for label agreement in Section 3.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not present a human-subjects study requiring IRB review.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.