# Selective "Selective Prediction": Reducing Unnecessary Abstention in Vision-Language Reasoning

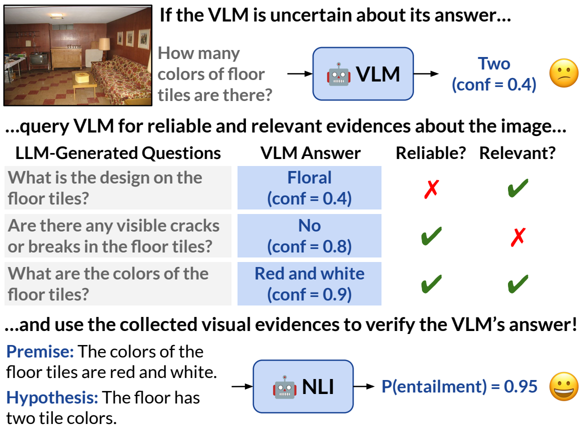
**Tejas Srinivasan**[1*] **Jack Hessel**[2] **Tanmay Gupta**[3] **Bill Yuchen Lin**[3]
**Yejin Choi**[3,4] **Jesse Thomason**[1] **Khyathi Raghavi Chandu**[3]
[1]University of Southern California [2]Samaya AI
[3]Allen Institute for Artificial Intelligence [4]University of Washington
tejas.srinivasan@usc.edu

## Abstract

Selective prediction minimizes incorrect predictions from vision-language models (VLMs) by allowing them to abstain from answering when uncertain. However, when deploying a vision-language system with low tolerance for inaccurate predictions, selective prediction may be over-cautious and abstain too frequently, even on many correct predictions. We introduce ReCoVERR, an inference-time algorithm to reduce the over-abstention of a selective vision-language system without increasing the error rate of the system's predictions. When the VLM makes a low-confidence prediction, instead of abstaining ReCoVERR tries to find relevant clues in the image that provide additional evidence for the prediction. ReCoVERR uses an LLM to pose related questions to the VLM, collects high-confidence evidences, and if enough evidence confirms the prediction the system makes a prediction instead of abstaining. ReCoVERR enables three VLMs (BLIP2, InstructBLIP and LLaVA-1.5) to answer up to 20% more questions on the VQAv2 and A-OKVQA tasks without decreasing system accuracy, thus improving overall system reliability. Our code is available at https://github.com/tejas1995/ReCoVERR.

## 1 Introduction

Instruction-tuned vision-and-language models (VLMs) (Dai et al., 2023; Liu et al., 2023; Laurençon et al., 2023; Bai et al., 2023) have achieved strong accuracy on reasoning benchmarks, which typically require VLMs to produce an answer for each instance. For downstream use, however, these systems should abstain from answering when uncertain (e.g., by saying "I don't know") (Rajpurkar et al., 2018). Selective prediction systems (De Stefano et al., 2000; El-Yaniv et al., 2010) aim to balance the number of predictions made (*coverage*) and the error rate on predicted instances (*risk*).


If the VLM is uncertain about its answer...
How many colors of floor tiles are there?
VLM
Two (conf = 0.4)
...query VLM for reliable and relevant evidences about the image...

| LLM-Generated Questions | VLM Answer | Reliable? | Relevant? |
|---|---|---|---|
| What is the design on the floor tiles? | Floral (conf = 0.4) | ✗ | ✓ |
| Are there any visible cracks or breaks in the floor tiles? | No (conf = 0.8) | ✓ | ✗ |
| What are the colors of the floor tiles? | Red and white (conf = 0.9) | ✓ | ✓ |

...and use the collected visual evidences to verify the VLM's answer!
Premise: The colors of the floor tiles are red and white.
Hypothesis: The floor has two tile colors.
NLI
P(entailment) = 0.95

Figure 1: Illustration of ReCoVERR. The VLM predicts that the floor has two tile colors with low confidence. Instead of abstaining, ReCoVERR collects reliable and relevant visual evidences related to the question. ReCoVERR makes salient the evidence that the floor tiles are red and white, helping to verify the VLM's original answer.

However, a vanilla selective prediction system with low tolerance for incorrect predictions will abstain too frequently to be practical, even when the model answer may be correct (Whitehead et al., 2022a). For example, if a user specifies that the BLIP2 (Li et al., 2023) predictions should be right at least 90% of the time, vanilla selective prediction will make a prediction for **just 4%** of A-OKVQA (Schwenk et al., 2022) questions, with 94% of the correct predictions being abstained on.

We introduce ReCoVERR (**Re**ason by **Co**llecting **V**isual **E**vidences that are **R**eliable and **R**elevant), an algorithm that increases the number of questions that a selective VLM system can answer confidently while adhering to a specified risk tolerance, *without any additional training*. When the VLM is uncertain about its prediction for a given question, instead of abstaining outright, ReCoVERR tries to verify the prediction by recovering reliable supporting (or contradicting) evidence. This ability is predicated on two characteristics of VLMs. First,

[0] Work done by Tejas as part of an internship at AI2.

*Findings of the Association for Computational Linguistics: ACL 2024*, pages 12935–12948
August 11-16, 2024 ©2024 Association for Computational Linguistics

VLMs can produce well-calibrated confidence estimates (§ 4.2.1). Second, VLMs can often correctly and confidently recover information in the image that entails a low-confidence initial prediction. For instance, in Figure 1, when asked about the number of floor tile colors, a VLM (here, BLIP2) correctly answers "two" but with low confidence: simple threshold-based abstention would abstain on this instance. But, when asked to identify the colors of the floor tiles, the model confidently answers "red and white". While it is obvious that "red and white" entails "two colors", the VLM was unable to confidently make that inference. Based on these two insights, ReCoVERR searches for additional visual evidence by iteratively posing relevant questions to the VLM, and collecting answers as visual evidence if they are: a) *reliable*, i.e, the VLM is highly confident in its answer, and b) *relevant* to the question the VLM is trying to answer in the first place.

We experiment on the VQAv2 and A-OKVQA visual reasoning tasks using three VLMs: BLIP2, InstructBLIP and LLaVA-1.5. For all three VLMs, ReCoVERR substantially increases the number of questions answered by the selective VLM system, while keeping system risk under the specified risk tolerance (Section 5). ReCoVERR is particularly helpful for BLIP2, which has not been trained on the target task, by improving coverage by 20% and recall by 25-30%. Our analysis reveals that the ability to give accurate confidence estimates, and high estimates for correct predictions, is crucial to ReCoVERR's performance. Further experiments demonstrate the importance of ensuring evidences are both reliable and relevant (Section 5.1), and that ReCoVERR tuned for a single task can be directly applied to new tasks without further tuning (Section 5.3). Our findings suggest that ReCoVERR is a promising solution towards building more reliable multimodal reasoning systems.

## 2 Multimodal Selective Prediction

We consider the task of answering a textual question about an image by drawing inferences from what is visually observed. A vision-language model (VLM) is given an input $x = (I, Q) \in \mathcal{X}$ consisting of an image $I$ and a question $Q$, and aims to produce an answer $a \in A$ from a closed or open set of possible answers.

Different from popular benchmarks of this form which require models to make a prediction for each instance, we evaluate models on the selective prediction setting, where abstention on individual instances is allowed (De Stefano et al., 2000). A decision function $g$, which has access to the image, question, VLM prediction, and associated confidence[1] determines whether the system produces an answer or abstains (denoted by $\varnothing$). The selective VLM system $\mathcal{S}_{\mathsf{VLM}} : \mathcal{X} \rightarrow \mathcal{A} \cup \{\varnothing\}$ is defined as:

$$a = \mathcal{M}_{\mathsf{VLM}}(x)$$

$$\mathcal{S}_{\mathsf{VLM}}(x) = \begin{cases} a, & \text{if } g(a) = 1 \\ \varnothing, & \text{if not } g(a) = 0 \end{cases}$$

Prior work (Whitehead et al., 2022a) employs confidence-based selection, where instead of assuming access to a VLM which has been trained to abstain with explicit supervision, we assume a VLM that can produce both an answer, as well as a confidence score for that answer $\pi_{\mathsf{VLM}} : \mathcal{A} \rightarrow [0, 1]$ that estimates $\pi_{\mathsf{VLM}}(a) = P(a|Q, I; \mathcal{M}_{\mathsf{VLM}})$. Confidence-based selection relies on the confidence $\pi_{\mathsf{VLM}}(a)$ being higher for correct answers than incorrect ones, on average. The decision function $g$ is a simple thresholding function, with a threshold $\gamma$: $g(a; \gamma) = \mathbb{1}\{\pi_{\mathsf{VLM}}(a) \geq \gamma\}$. We refer to this method as *vanilla selective prediction*.

### 2.1 Evaluating Selective Predictors

We evaluate selective prediction systems via coverage and risk (El-Yaniv et al., 2010). Given a labeled evaluation dataset $\mathcal{D} = \{(I_i, Q_i, a_i)\}_{i=1}^{N}$. *Coverage* ($\mathcal{C}$) is the percentage of questions in $\mathcal{D}$ where the system chooses to make a prediction. *Risk* ($\mathcal{R}$) is the error rate on the questions where a prediction is made. For a $\gamma$-threshold selective prediction system, these metrics are computed as:

$$\mathcal{R}(\gamma) = \frac{\sum_{x_i \in \mathcal{D}} (1 - \mathrm{Acc}(a_i)) \cdot g(a_i; \gamma)}{\sum_{x_i \in \mathcal{D}} g(a_i; \gamma)} \quad (1)$$

$$\mathcal{C}(\gamma) = \frac{\sum_{x_i \in \mathcal{D}} g(a_i; \gamma)}{|\ \mathcal{D}\ |} \quad (2)$$

where a lower $\gamma$ trades off increased coverage for increased risk. In practice, when deploying a selective VLM system with an application-specific risk tolerance $r \in [0, 1]$, the system designer would determine an appropriate setting for the hyperparameters of $g$ using a calibration set. For example, if $g$ is the simple $\gamma$-thresholding function, one would choose $\gamma_{@r}$ to maximize coverage while being below the risk threshold on the calibration set:

$$\gamma_{@r} = \underset{\gamma \in [0,1]}{\arg\min}\, R(\gamma) \leq r$$

[1] For brevity, we note arguments for $g$ when relevant.

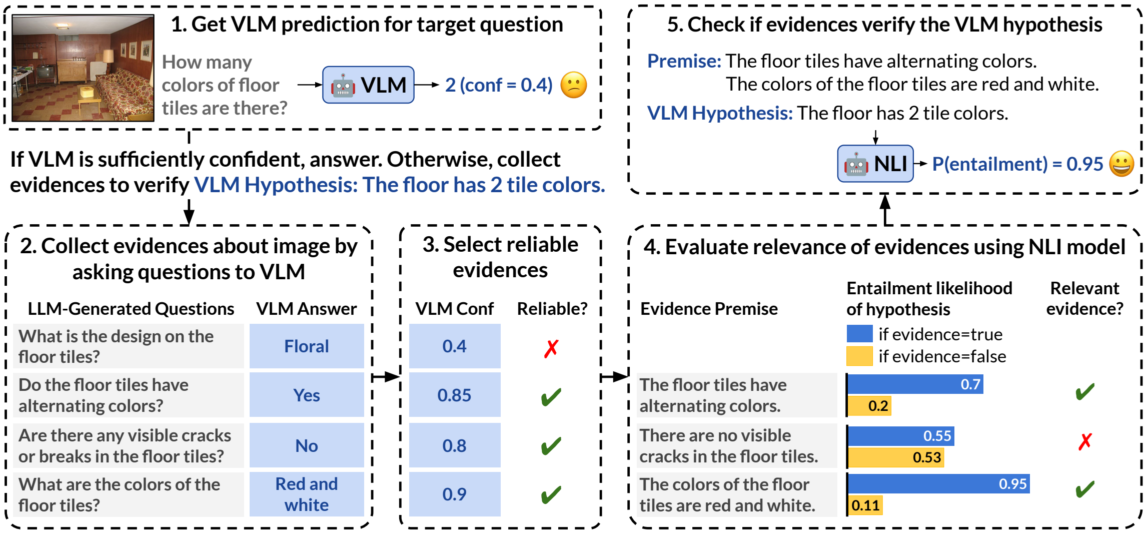

1. Get VLM prediction for target question
How many colors of floor tiles are there?
VLM
2 (conf = 0.4)
If VLM is sufficiently confident, answer. Otherwise, collect evidences to verify VLM Hypothesis: The floor has 2 tile colors.
2. Collect evidences about image by asking questions to VLM
LLM-Generated Questions
VLM Answer
What is the design on the floor tiles?
Floral
Do the floor tiles have alternating colors?
Yes
Are there any visible cracks or breaks in the floor tiles?
No
What are the colors of the floor tiles?
Red and white
3. Select reliable evidences
VLM Conf
Reliable?
0.4
✗
0.85
✓
0.8
✓
0.9
✓
4. Evaluate relevance of evidences using NLI model
Evidence Premise
Entailment likelihood of hypothesis
Relevant evidence?
if evidence=true
if evidence=false
The floor tiles have alternating colors.
0.7
0.2
✓
There are no visible cracks in the floor tiles.
0.55
0.53
✗
The colors of the floor tiles are red and white.
0.95
0.11
✓
5. Check if evidences verify the VLM hypothesis
Premise: The floor tiles have alternating colors.
The colors of the floor tiles are red and white.
VLM Hypothesis: The floor has 2 tile colors.
NLI
P(entailment) = 0.95

Figure 2: The ReCoVERR algorithm. If the VLM is uncertain in its prediction (1), ReCoVERR tries to verify the VLM hypothesis by collecting evidences. ReCoVERR undertakes multiple turns of evidence collection, which involves generating visual evidences by using an LLM to ask questions to the VLM (2), retaining the reliable (3) and relevant (4) evidences, and checking whether the collected evidence entails the hypothesis (5).

## 3 ReCoVERR

In practice, reasonable risk tolerance thresholds often lead to untenable coverage, with many accurate predictions being discarded. A threshold-selective BLIP2 (Li et al., 2023) system, for example, answers fewer than 4% of questions in A-OKVQA at 10% risk, but these represent only 6% of questions for which the model prediction was correct.

We introduce ReCoVERR (**Re**ason by **Co**llecting **V**isual **E**vidences that are **R**eliable and **R**elevant), an algorithm that increases a selective VLM system's coverage while remaining under the given risk tolerance (Figure 2). For each question $Q$ for which the VLM predicts an answer $a$ with $\pi_{\texttt{VLM}}(a) < \gamma_{@r}$, instead of abstaining, ReCoVERR uses large language models to generate follow-up questions about the image. These questions are answered by the VLM, and the resulting QA-pair is added as *evidence* if it is both *reliable*—the VLM confidence in the answer is high—and *relevant*—the introduction of the new evidence affects the downstream confidence in the hypothesized answer $a$ to $Q$. If sufficient evidences are collected to confidently entail the hypothesis, ReCoVERR elects to make a prediction instead of abstaining.

Let $\mathcal{D}_\varnothing$ be the subset of test set $\mathcal{D}$ where the selective prediction system would have abstained, based on the confidence threshold $\gamma_{@r}$, and $\mathcal{D}_S$ be the set of questions that were answered.

$$\mathcal{D}_\varnothing = \{x_i \in \mathcal{D}; \pi_{\texttt{VLM}}(a_i) < \gamma_{@r}\}; \mathcal{D}_S = \mathcal{D} \backslash \mathcal{D}_\varnothing$$

ReCoVERR aims to answer additional instances $\mathcal{D}_{\texttt{R}} \subseteq \mathcal{D}_\varnothing$, while keeping the combined risk of $\mathcal{D}_S \cup \mathcal{D}_{\texttt{R}}$ under $r$. Since risk of $\mathcal{D}_S$ is already estimated to be $r$, the risk of $\mathcal{D}_{\texttt{R}}$ must also be $\leq r$. For each instance in $\mathcal{D}_\varnothing$, ReCoVERR makes an online decision of whether to answer the question, independent of other instances in $\mathcal{D}_\varnothing$. Therefore, the expected risk of any instance $x_i \in \mathcal{D}_{\texttt{R}}$ should be, at most, $r$; in other words, the likelihood of each instance in $\mathcal{D}_{\texttt{R}}$ being correct should be $\geq 1 - r$. For example, if our system has a specified risk tolerance of 20%, ReCoVERR aims to answer additional questions of which 80% are correct.

Algorithm 1 describes ReCoVERR in detail, including all hyperparameters. Given an image $I$ and a question $Q$, ReCoVERR begins by generating an answer $a$ using the VLM $\mathcal{M}_{\texttt{VLM}}$. The VLM also returns a confidence score $\pi_{\texttt{VLM}}(a) \in [0, 1]$. If the confidence is higher than our confidence threshold for selective prediction $\gamma_{@r}$, we can simply return the answer. If not, we use a model $\mathcal{M}_{\texttt{QA}\rightarrow\texttt{S}}$ that converts the question-answer pair $(Q, a)$ into a hypothesis statement $\mathcal{H}$. ReCoVERR will now try to verify this hypothesis by collecting reliable and relevant visual evidences.

**Initialize Evidences from Vision Tools:** We first gather some general information about the image by invoking a set of vision models, $\mathbb{M}_{\texttt{Vis}}$, that capture visual information in language form (the specific tools we use in our instantiation of

**Algorithm 1:** ReCoVERR Pseudocode

```
Selective prediction hyperparameters:
$r \in [0,1]$ : Risk tolerance of system
$\gamma_{@r}$ : VLM conf threshold corresponding to risk $r$
ReCoVERR Model-based Tools:
$\mathbb{M}_{\texttt{Vis}}$: Set of additional vision models that output information about the image in text form
$\mathcal{M}_{\texttt{QGen}}$: Question generation model
$\mathcal{M}_{\texttt{QA}\rightarrow\texttt{S}}$: LM to paraphrase QA pair into sentence
$\mathcal{M}_{\texttt{NLI}}$: NLI model
ReCoVERR hyperparameters:
$\delta_{\texttt{min}}.$ : minimum relevance of "relevant" evidences
$\pi_{\texttt{NLImin}}$: Minimum entailment confidence to answer
$N$: Maximum number of turns of evidence collection
$K$: Questions generated at each turn
Inputs: Question $Q$, image $I$, VLM $\mathcal{M}_{\texttt{VLM}}$

$a, \pi_{\texttt{VLM}}(a) \leftarrow \mathcal{M}_{\texttt{VLM}}(I, Q)$
if $\pi_{\texttt{VLM}}(a) \geq \gamma_{@r}$ then
    return $a$
$\mathcal{H} \leftarrow \mathcal{M}_{\texttt{QA}\rightarrow\texttt{S}}(Q, a)$
$\mathcal{E}_R, \mathcal{E}_{RR} \leftarrow$ InitImageEvidences($I; \mathbb{M}_{\texttt{Vis}}$)
for $i = 1$ to $N$ do
    $e_{1 \ldots K} \leftarrow$ CollectKEvidences($I, Q, \mathcal{E}_R;$
                                  $\mathcal{M}_{\texttt{QGen}}, \mathcal{M}_{\texttt{VLM}}$)
    for $j = 1$ to $K$ do
        if VLMConfidence($e_j; \mathcal{M}_{\texttt{VLM}}$) $\geq 1 - r$
          then
            $\mathcal{E}_R \leftarrow \mathcal{E}_R + [e_j]$
            if Relevance($e_j; \mathcal{M}_{\texttt{NLI}}$) $\geq \delta_{\texttt{min}}.$ then
                $\mathcal{E}_{RR} \leftarrow \mathcal{E}_{RR} + [e_j]$
    if $P(\mathcal{E}_{RR}$ entails $\mathcal{H}; \mathcal{M}_{\texttt{NLI}}) \geq \pi_{\texttt{NLIMin}}$ then
        return $a$
return $\varnothing$
```

ReCoVERR are highlighted in Section 4.3). Information from these tools instantiate two evidence sets: $\mathcal{E}_R$, a set of reliable evidences about the image, and $\mathcal{E}_{RR}$, a set of reliable *and* relevant evidences.

ReCoVERR performs up to $N$ turns of evidence collection. In each turn, $K$ new visual evidences are generated from the VLM, the reliable and relevant ones are retained, and we check whether the collected evidences sufficiently entail the hypothesis. If we are unable to verify the hypothesis at the end of $N$ turns, we abstain from answering.

**Generating Visual Evidences:** We prompt a question generation model $\mathcal{M}_{\texttt{QGen}}$ to generate a set of $K$ sub-questions $q_{1 \ldots K}$, conditioned on the target question $Q$ and the already-collected reliable evidences $\mathcal{E}_R$. For each sub-question $q_j, j \in \{1 \ldots K\}$, the VLM produces an answer $a_j$ with confidence $\pi_{\texttt{VLM}}(a_j)$. The paraphraser model $\mathcal{M}_{\texttt{QA}\rightarrow\texttt{S}}$ paraphrases the pair $(q_j, a_j)$ to a declarative sentence $S_j$. Each 4-tuple $\big(q_i, a_j, \pi_{\texttt{VLM}}(a_j), S_j\big)$ represents an evidence $e_j$.

**Check Evidence Reliability:** Since ReCoVERR decides to make a prediction based on the collected evidences, the correctness likelihood of a given evidence is an upper bound for the correctness likelihood of the prediction. Since ReCoVERR aims to make predictions with correctness likelihood $\geq 1 - r$, we only consider evidences $e_j$ that satisfy $\pi_{\texttt{VLM}}(a_j) \geq 1 - r$ as *reliable* for $\mathcal{E}_R$. **ReCoVERR requires that the confidence estimate $\pi_{\texttt{VLM}}(a)$ is calibrated**. A confidence estimate being calibrated means that for all predictions whose confidence is $\alpha \in [0,1]$, $\alpha\%$ of the predictions are actually correct. In our experiments, we first evaluate calibration of our VLMs' confidence estimates (§ 4.2.1).

**Check Evidence Relevance:** To decide whether a reliable evidence $e_j$ is also relevant, we adopt a defeasible reasoning approach (Rudinger et al., 2020). An evidence $e_j$ is considered relevant if its truth value affects the entailment probability of the hypothesis. We measure the absolute difference between the entailment probabilities of the hypothesis $\mathcal{H}$ conditioned on the evidence premise $S_j$ and its negated counterfactual i.e. $\bar{S}_j$.

$$\delta(e_j) = \mid \pi_{\texttt{NLI}}(\mathcal{H} | S_j) - \pi_{\texttt{NLI}}(\mathcal{H} | \bar{S}_j) \mid$$

Evidence $e_j$ is added to the relevant evidence set $\mathcal{E}_{RR}$ if the relevance $\delta(e_j) \geq \delta_{\texttt{min}}$.

**Check Sufficiency of Collected Evidences:** After each round of evidence collection, we concatenate the $S_j$ for all evidences $e_j \in \mathcal{E}_{RR}$ into a premise sentence $S_{RR}$ and calculate the hypothesis's entailment probability $\pi_{\texttt{NLI}}(\mathcal{H} | S_{RR})$. If that probability meets $\pi_{\texttt{NLImin}}$, we return prediction $a$, increasing coverage with sufficient estimated confidence that risk will not increase above $r$ as a result. Else, after $N$ rounds, we abstain and return $\varnothing$.

## 4 Experiments

We instantiate a selective prediction task using the A-OKVQA benchmark and compare ReCoVERR to simple threshold-based selective prediction with two different backbone VLMs with calibrated confidence estimates, using metrics that capture aspects and tradeoffs of risk and coverage.

### 4.1 Vision-Language Reasoning Tasks: A-OKVQA and VQAv2

A-OKVQA (Schwenk et al., 2022) is a VQA task designed to require reasoning over external knowledge and commonsense alongside image-based information. VQAv2 (Goyal et al., 2017) is a VQA

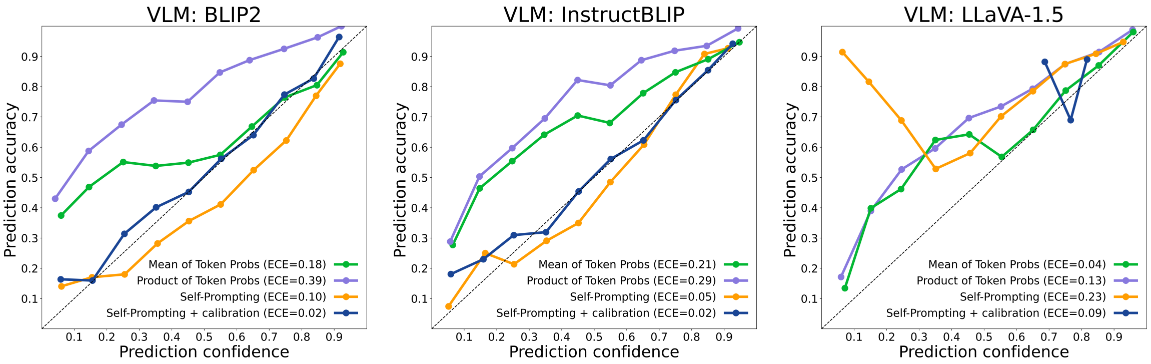


Figure 3: Calibration curves for BLIP-2, InstructBLIP and LLaVA-1.5 on A-OKVQA questions.

task that requires reasoning about images, sometimes involving commonsense reasoning. We evaluate on the A-OKVQA validation set ($n = 1,075$) and 1,000 examples from the VQAv2 validation set. Both tasks involve open-ended answer generation, without any choices provided to the model.

For evaluating answer correctness, the standard VQA accuracy metric (Antol et al., 2015) has been shown to penalize correct VLM answers if they do not exactly match reference answers (Agrawal et al., 2023). Therefore, we use LAVE$_{\text{GPT-3.5}}$ (Mañas et al., 2023) to measure the accuracy of predicted answers. LAVE uses a large language model[2] to estimate the semantic similarity of each predicted answer to the 10 crowdsourced answers in the benchmark.

### 4.2 Vision-Language Models

We experiment with three VLMs: BLIP2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), and LLaVA-1.5 (Liu et al., 2023). Both BLIP models use FlanT5-XL as the LLM backbone.[3] InstructBLIP and LLaVA-1.5 are instruction tuned on both VQAv2 and A-OKVQA; BLIP2 has not been trained on either task.

#### 4.2.1 Calibrated VLM Confidence Estimates

One key requirement for ReCoVERR is that, for a prediction $a$, the VLM can also return a *calibrated* confidence score $\pi_{\texttt{VLM}}(a) \in [0, 1]$. Thus, a first step when applying ReCoVERR to a new VLM is identifying a confidence function that produces calibrated confidence estimates.

Confidence estimates for generative VLMs are not well defined. Straightforward estimates such as the product of answer token likelihoods can severely underestimate model confidence due to factors like *surface form competition* (Holtzman et al., 2021). Inspired by Tian et al. (2023), we devise a Self-Prompting technique for estimating $\pi_{\texttt{VLM}}(a)$ by prompting the VLM to verify "yes" or "no" correctness of its predictions and examine the resulting probability distribution (full details in Appendix A). The Self-Prompting confidence can be further calibrated using Platt scaling (Platt, 1999).

We compare three methods for extracting VLM confidence scores: product of token probabilities, mean of token probabilities, and Self-Prompting. For each VLM, we evaluate the calibration error of these confidence functions on a set of 5,000 examples from the A-OKVQA training data. The remaining 12,000 examples are used to calibrate Self-Prompting with Platt scaling. We find that Self-Prompting (both *off-the-shelf* and *calibrated*) has the lowest calibration error for the BLIP models, whereas the mean of token probabilities has lowest error for LLaVA-1.5 (Figure 3). We therefore use these respective methods for estimating VLM confidence in our experiments with ReCoVERR.

### 4.3 ReCoVERR Implementation Details

ReCoVERR can be instantiated with different choices of models and system hyperparameters.

#### 4.3.1 ReCoVERR Model-based Tools

ReCoVERR leverages a set of vision tools $\mathbb{M}_{\texttt{Vis}}$ for extracting general visual information. In our experiments, the vision tools $\mathbb{M}_{\texttt{Vis}}$ consists of LVIS (Gupta et al., 2019) for object detection, and Qwen-VL (Bai et al., 2023) for region captioning. We use FlanT5-XL as a zero-shot sentence paraphraser $\mathcal{M}_{\texttt{QA}\rightarrow\texttt{S}}$ and entailment model $\mathcal{M}_{\texttt{NLI}}$. Since our VLMs use FlanT5-XL as the LLM backbone, our ReCoVERR instantiation does not require additional models. We use GPT-3.5 with temperature 1.0 for $\mathcal{M}_{\texttt{QGen}}$. See Appendix B for prompts.

[2] We utilize the `gpt-3.5-turbo-16k-0613` checkpoint.

[3] `https://github.com/salesforce/LAVIS`

| VLM | Method | Risk tolerance $r$ = 10% | | | | Risk tolerance $r$ = 20% | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\mathcal{R}(\downarrow)$ | $\Phi_1(\uparrow)$ | $\mathcal{C}(\uparrow)$ | $\mathbb{R}_{\mathsf{SP}}(\uparrow)$ | $\mathcal{R}(\downarrow)$ | $\Phi_1(\uparrow)$ | $\mathcal{C}(\uparrow)$ | $\mathbb{R}_{\mathsf{SP}}(\uparrow)$ |
| BLIP2 (Off-the-shelf) | Vanilla SelPred | 6.1 | 3.4 | 3.8 | 6.0 | 14.9 | 17.1 | 24.1 | 34.2 |
| | Vision Tools | 13.8 | 17.4 | 23.5 | 33.5 | 17.0 | 23.5 | 35.1 | 48.3 |
| | ReCoVERR | 14.3 | **18.8** | **26.0** | **37.1** | 21.7 | **27.1** | **47.3** | **61.5** |
| BLIP2 (Calibrated) | Vanilla SelPred | 4.1 | 3.2 | 3.4 | 5.5 | 11.9 | 16.8 | 21.9 | 32.0 |
| | Vision Tools | 13.2 | 17.5 | 23.4 | 33.7 | 15.1 | 23.9 | 33.6 | 47.2 |
| | ReCoVERR | 14.0 | 17.3 | 23.6 | 33.6 | 16.1 | **25.3** | **36.7** | **51.0** |
| InstructBLIP (Off-the-shelf) | Vanilla SelPred | 9.3 | 20.5 | 25.1 | 34.8 | 17.2 | 38.3 | 57.7 | 72.2 |
| | Vision Tools | 10.5 | 26.1 | 32.9 | 44.8 | 17.6 | 39.6 | 60.5 | 75.3 |
| | ReCoVERR | 10.9 | 26.3 | 33.5 | 45.2 | 17.9 | **41.3** | **63.7** | **78.9** |
| InstructBLIP (Calibrated) | Vanilla SelPred | 8.5 | 22.0 | 26.4 | 36.9 | 17.5 | 37.2 | 56.6 | 70.5 |
| | Vision Tools | 10.4 | 27.2 | 34.1 | 46.5 | 17.8 | 38.8 | 59.7 | 74.2 |
| | ReCoVERR | 11.8 | **29.8** | **38.7** | **51.8** | 18.6 | **41.4** | **65.0** | **79.8** |
| LLaVA-1.5 | Vanilla SelPred | 8.2 | 21.5 | 25.1 | 33.0 | 16.5 | 40.1 | 59.3 | 69.9 |
| | Vision Tools | 9.6 | 24.4 | 30.0 | 38.3 | 17.3 | 40.8 | 61.8 | 72.1 |
| | ReCoVERR | 11.3 | **34.7** | **45.4** | **55.9** | 19.5 | **44.6** | **72.7** | **81.9** |

Table 1: Metric results as percentages on the A-OKVQA task at two risk tolerance levels. We evaluate selective prediction methods on the overall system risk ($\mathcal{R}$), effective reliability ($\Phi_1$), coverage ($\mathcal{C}$) and recall ($\mathbb{R}_{\mathsf{SP}}$). System risks in red exceeded tolerance. Measurements in **blue** indicate when ReCoVERR outperformed both baselines.

#### 4.3.2 ReCoVERR Hyperparameters

We set $N = 10$ rounds of evidence collection with $K = 10$ evidences generated per turn. To count an evidence $e$ as reliable, we set relevance threshold $\delta_{\texttt{min}} = 0.2$. The minimum final hypothesis entailment confidence is $\pi_{\texttt{NLImin}} = 0.9$.

### 4.4 Selective Prediction Baselines

We compare ReCoVERR to a **Vanilla Selective Prediction** baseline (Whitehead et al., 2022a), where the model abstains if the VLM prediction confidence $g(a) = \{1 \text{ if } \pi_{\mathsf{VLM}}(a) \leq \gamma_{@r} \text{ else } 0\}$, and to a **Vision Tools** baseline. For **Vision Tools**, image caption from the VLM, objects detected by LVIS, and region captions from Qwen-VL are provided as evidence to the NLI model $\mathcal{M}_{\texttt{NLI}}$ directly, with no rounds of evidence collection, equivalent to lines 1-5 of Algorithm 1 followed by lines 14-15.

### 4.5 Evaluation Metrics

We measure **risk** ($\mathcal{R}$) and **coverage** ($\mathcal{C}$) (Eqs 1, 2) across system predictions and abstentions. Additionally, we calculate **Effective Reliability** $\Phi_c$ (Whitehead et al., 2022a), a metric that rewards correct predictions, assigns a penalty $c$ to incorrect predictions, and assigns zero reward to abstentions. We assign a penalty of $c = 1$ for incorrect answers ($\Phi_1$). Finally, we define **Selective Prediction Recall** ($\mathbb{R}_{\mathsf{SP}}$), the percentage of correct answers ($\text{Acc}(a) = 1$) in the dataset that were answered.

$$\mathbb{R}_{\mathsf{SP}} = \frac{\sum_{x_i \in \mathcal{D}} g(a_i) \cdot \mathbb{1}\{\text{Acc}(a_i) = 1\}}{\sum_{x_i \in \mathcal{D}} \mathbb{1}\{\text{Acc}(a_i) = 1\}}$$

We note that the only stochastic component in our instantiation of ReCoVERR is the question generation model. We run three seeds for each ReCoVERR experiment and report average metric results. Table 8 shows full results with standard deviation.

## 5 Results and Analysis

We evaluate selective prediction methods on A-OKVQA at two specified risk tolerances: 10% and 20% risk. Table 1 shows the results of ReCoVERR when applied to InstructBLIP and BLIP2, both with and without Platt scaling calibration.

At 20% risk tolerance, we find that ReCoVERR improves coverage, effective reliability, and recall over both baselines while staying at the specified risk tolerance. At 10% risk tolerance, ReCoVERR and the Vision Tools baseline tend to slightly overshoot the risk tolerance (by 1-2% for InstructBLIP and LLaVA-1.5, and 3-4% for BLIP2). We further see that ReCoVERR shows large improvements

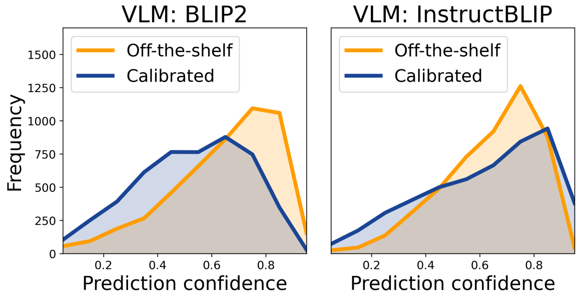


Figure 4: Distribution of VLM prediction confidences on A-OKVQA calibration set.

| VLM | Method | $\mathcal{R}(\downarrow)$ | $\Phi_1(\uparrow)$ | $\mathcal{C}(\uparrow)$ | $\mathbb{R}_{\texttt{SP}}(\uparrow)$ |
|---|---|---|---|---|---|
| BLIP2 (Off-the-shelf) | Vanilla SelPred | 7.9 | 13.5 | 15.9 | 21.3 |
| | Vision Tools | 11.8 | 20.5 | 26.6 | 34.0 |
| | ReCoVERR | 11.9 | **21.3** | **27.8** | **35.4** |
| BLIP2 (Calibrated) | Vanilla SelPred | 7.7 | 13.1 | 15.5 | 20.9 |
| | Vision Tools | 11.3 | 21.8 | 27.9 | 35.9 |
| | ReCoVERR | 11.8 | **22.9** | **29.7** | **38.1** |
| InstructBLIP (Off-the-shelf) | Vanilla SelPred | 10.0 | 31.8 | 39.5 | 45.8 |
| | Vision Tools | 10.8 | 34.8 | 43.1 | 50.2 |
| | ReCoVERR | 10.7 | 35.5 | **45.0** | **51.9** |
| InstructBLIP (Calibrated) | Vanilla SelPred | 10.3 | 32.3 | 41.1 | 47.5 |
| | Vision Tools | 11.2 | 35.4 | 44.3 | 51.1 |
| | ReCoVERR | 11.0 | 35.9 | **45.8** | **52.6** |
| LLaVA-1.5 | Vanilla SelPred | 10.4 | 49.4 | 62.3 | 74.5 |
| | Vision Tools | 11.0 | 49.7 | 63.1 | 75.5 |
| | ReCoVERR | 11.8 | **51.5** | **67.9** | **79.2** |

Table 2: Metric results as percentages on the VQAv2 task at a risk tolerance of 10%. We evaluate selective prediction methods on overall system risk ($\mathcal{R}$), effective reliability ($\Phi_1$), coverage ($\mathcal{C}$) and recall ($\mathbb{R}_{\texttt{SP}}$). System risks in red exceeded tolerance. Measurements in **blue** indicate when ReCoVERR outperformed both baselines.

of up to 20% on coverage, 15% on effective reliability and 33% on recall for off-the-shelf BLIP2, calibrated InstructBLIP and LLaVA-1.5.

We highlight several takeaways (**T***). **T1**: For the same risk tolerance, vanilla selective prediction has lower coverage and recall for BLIP2 than InstructBLIP and LLaVA-1.5. BLIP2 has not been trained on A-OKVQA, unlike the other two VLMs, and therefore produces more uncertain answers. We further see that ReCoVERR results in larger improvements for BLIP2. **T2**: Off-the-shelf BLIP2 sees largest risk increase with ReCoVERR. The calibration curves in Figure 3 indicate that off-the-shelf BLIP2 makes overconfident estimates, leading to more unreliable evidences being collected and resulting in a large risk increase for ReCoVERR. **T3**: At 10% risk, most of the additional risk and recall is caused by Vision Tools, rather than ReCoVERR's evidence collection. **T4**: At 20% risk tolerance, ReCoVERR's evidence collection leads to greatest recall improvement for off-the-shelf BLIP2, calibrated InstructBLIP and LLaVA-1.5.

**T3&4** can be understood by examining the confidence distribution of the VLMs' predictions (Figure 4). We see that off-the-shelf BLIP2 has more predictions with greater than 80% confidence compared to calibrated BLIP2, with the trend reversed for InstructBLIP. This means that ReCoVERR finds more reliable evidences ($\pi_{\texttt{VLM}}(a) \geq 1 - r$) with off-the-shelf BLIP2 and calibrated InstructBLIP, resulting in higher recall for ReCoVERR with those two VLMs, thus explaining **T3&4**. These findings indicate that ReCoVERR works best when the VLM is *strongly confident* on correct answers.

We also experiment on VQAv2, at 10% risk tolerance.[4] In Table 2, we see that ReCoVERR improves system reliability and recovers more correct answers than the baselines, while crossing the risk threshold by only 1–2%. However, the improvements are smaller compared to the more complex A-OKVQA task, indicating that ReCoVERR may be more useful for complex reasoning tasks. We also observe that, similar to A-OKVQA, the largest improvements are for BLIP2, which wasn't finetuned on the VQAv2 task.

## 5.1 Reliability and Relevance Ablations

We perform an ablation of the reliability and relevance components of ReCoVERR for calibrated BLIP2 and InstructBLIP, at a risk tolerance of 20%. In the ReCoVERR formulation, for an evidence $e_j$ to be considered reliable, the VLM confidence $\pi_{\texttt{VLM}}(a_j|I, Q_j)$ must be at least $1 - r$, and for it to be considered relevant, the defeasible relevance $\delta(e_j)$ must be at least $\delta_{\texttt{min}} = 0.2$. For the reliability ablation, we lower the minimum evidence confidence to be 0.5 instead of $1 - r = 0.8$. For the relevance ablation, we set $\delta_{\texttt{min}} = 0$.

Table 3 shows the results of the reliability and relevance ablation. We see that loosening the evidence confidence bound to be 0.5 instead of $1 - r$ causes a marked increase in risk for both VLMs, with system risk crossing the risk tolerance significantly. Ablating the relevance component results in a decrease in recall and effective reliability for InstructBLIP, but virtually no effect for BLIP2.

[4] Since the VLMs have strong performance on VQAv2, we do not perform experiment at 20% risk tolerance.

| Method | $\mathcal{R}(\downarrow)$ | $\Phi_1(\uparrow)$ | $\mathbb{R}_{\texttt{SP}}(\uparrow)$ |
|---|---|---|---|
| VLM : Calibrated BLIP2 | | | |
| ReCoVERR | 16.1 | 25.3 | 51.0 |
| - Reliability | 28.4 | 31.6 | 85.9 |
| - Relevance | 16.1 | 25.3 | 51.1 |
| VLM : Calibrated InstructBLIP | | | |
| ReCoVERR | 18.6 | 41.4 | 79.8 |
| - Reliability | 25.9 | 40.5 | 91.9 |
| - Relevance | 18.6 | 39.7 | 76.9 |

Table 3: Ablation of the reliability and relevance components of ReCoVERR, at risk tolerance of 20%.

| Calibrated VLM | Method | $\mathcal{R}(\downarrow)$ | $\mathbb{R}_{\texttt{SP}}(\uparrow)$ |
|---|---|---|---|
| Task: OK-VQA (20% risk tolerance) | | | |
| BLIP2 | Vanilla | 15.4 | 25.4 |
| | ReCoVERR | 25.7 | 45.0 |
| InstructBLIP | Vanilla | 18.5 | 56.3 |
| | ReCoVERR | 22.8 | 70.8 |
| Task: Sherlock (10% risk tolerance) | | | |
| BLIP2 | Vanilla | 11.1 | 24.9 |
| | ReCoVERR | 15.7 | 25.8 |
| InstructBLIP | Vanilla | 9.7 | 33.4 |
| | ReCoVERR | 10.9 | 35.5 |

Table 4: ReCoVERR using a VLM calibrated for A-OKVQA can be directly applied to new tasks.

### 5.2 Question Generation Model Ablations

The specific instantiation of ReCoVERR involves a choice of LLM for the question generation model, $\mathcal{M}_{\texttt{QGen}}$. While our experiments so far have used GPT-3.5, we demonstrate that ReCoVERR works similarly well with open-source LLMs as well. We compare ReCoVERR performance with GPT-3.5 against two open-source LLMs: Mistral-7B-Instruct (Jiang et al., 2023) and Tulu-2-7B (Ivison et al., 2023). Our results in Table 5 demonstrate that the specific instantiation of $\mathcal{M}_{\texttt{QGen}}$ did not make a significant difference to ReCoVERR's performance, for both calibrated BLIP2 and InstructBLIP.

| $\mathcal{M}_{\texttt{QGen}}$ | $\mathcal{R}(\downarrow)$ | $\Phi_1(\uparrow)$ | $\mathcal{C}(\uparrow)$ | $\mathbb{R}_{\texttt{SP}}(\uparrow)$ |
|---|---|---|---|---|
| VLM : Calibrated BLIP2 | | | | |
| ChatGPT | 16.1 | 25.3 | 36.7 | 51.0 |
| Mistral-7B | 16.9 | 25.4 | 37.9 | 52.2 |
| Tulu-2-7B | 15.8 | 24.8 | 35.8 | 50.0 |
| VLM : Calibrated InstructBLIP | | | | |
| ChatGPT | 18.6 | 41.4 | 65.0 | 79.8 |
| Mistral-7B | 19.1 | 39.9 | 63.8 | 77.8 |
| Tulu-2-7B | 18.0 | 40.9 | 63.3 | 78.2 |

Table 5: Effect of different question generation models on ReCoVERR performance, at 20% risk tolerance.

### 5.3 Generalizing to New Tasks

We examine whether our instantiation of ReCoVERR calibrated for the A-OKVQA task can be directly applied to new tasks without additional tuning (Table 4). When we apply ReCoVERR with A-OKVQA-calibrated VLMs to OK-VQA (Marino et al., 2019) at 20% risk tolerance, we observe strong improvements in recall (15-20%), but also increased risk (2-5% above the risk tolerance). Applying ReCoVERR to Sherlock (*Hessel et al., 2022), an abductive reasoning task, at 10% risk we observe smaller improvements in recall. These results indicate that ReCoVERR requires some degree of task-specific tuning. Appendix C contains full task details.

### 5.4 Qualitative ReCoVERR Examples

We present some examples of evidences collected by ReCoVERR in Table 6. The collected evidences $\mathcal{E}_{RR}$ can be used to either corroborate correct VLM predictions, or contradict incorrect ones. In example 1, when the VLM predicts that a vegetarian is likely to eat this meal, ReCoVERR collects supporting evidences indicating that the food does not contain any meat products. In example 2, when the VLM predicts that the colors of the bus match those of the U.K. flag, ReCoVERR finds that the color of the bus is highly likely to be yellow and blue, thus contradicting the VLM's prediction (since the U.K. flag does not contain yellow).

## 6 Conclusion and Related Works

We introduce ReCoVERR, an algorithm to improve the coverage of a selective VLM system while respecting a user-specified risk tolerance level. ReCoVERR verifies low-confidence VLM predictions by recovering high-confidence evidences in the image that support the prediction. We instantiate a selective prediction task on the A-OKVQA reasoning benchmark and demonstrate the quantitative advantages of ReCoVERR for inference-time selective prediction that holds across different back-

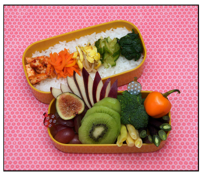
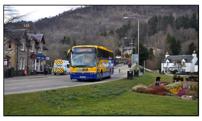
| Image | Initial prediction | Evidences $\mathcal{E}_{RR}$ collected by ReCoVERR<br>{LLM Question $q_j$} **{VLM Answer $a_j$}** [$\pi_{\text{VLM}}(a_j)$] |
|---|---|---|
|  | **Question:** What kind of a person usually eats food like this?<br>**G.T. answers:** vegetarian, vegan, healthy<br>**VLM:** vegetarian (conf=0.57) | Does the image depict a variety of plant-based ingredients? **yes** [0.95]<br>Does the food predominantly consist of fruits and vegetables? **yes** [0.80]<br>Are there any meat items in the image? **no** [0.86]<br>**Result: Correctly predicted** |
|  | **Question:** The colors on the bus match the colors on what flag?<br>**G.T. answers:** Sweden, Ukraine<br>**VLM:** U.K. (conf=0.53) | Is there any blue color on the bus? **yes** [0.95]<br>What is the color of the bus? **yellow and blue** [0.93]<br>**Result: Correctly abstained** |

Table 6: Examples from the A-OKVQA task where, given a low-confidence VLM prediction, ReCoVERR collects a set of highly-confident evidences $\mathcal{E}_{RR}$ that corroborate (example 1) or contradict (example 2) the prediction.

bone VLMs, choice of underlying LLM generators, and even gains some ground when transferred to different benchmarks without recalibration.

A large breadth of prior work has studied the ability of models to abstain from answering (Chow, 1957; De Stefano et al., 2000; El-Yaniv et al., 2010). This ability is especially important when the model's prediction cannot be trusted (Jiang et al., 2018), particularly for OOD (Kamath et al., 2020) and adversarial (Varshney et al., 2022) inputs. Similar to our work, low coverage caused by low risk tolerance has also been observed and mitigated in text-only selective prediction systems (Varshney and Baral, 2023).

In the multimodal domain, previous works have explored abstention for error recovery (Wu et al., 2023; Khan and Fu, 2023). Whitehead et al. (2022b) establish the ReliableVQA framework, wherein VQA models have the option to abstain from answering. Shukor et al. (2023) evaluates VQA models on their ability to abstain when a question does not apply to the image. Dancette et al. (2023) trains VQA models to abstain for out-of-distribution inputs. In contrast to *learning to abstain*, we *reduce abstention* of multimodal selective prediction while maintaining reliability.

Using LLMs to query a pool of vision experts, either through program generation (Gupta and Kembhavi, 2023; Surís et al., 2023; Subramanian et al., 2023), or iterative querying (Zeng et al., 2022; Shen et al., 2023; You et al., 2023; Yang et al., 2023), can decompose vision-language reasoning. We use LLMs to query an image for visual evidence, similar to RepARe (Prasad et al., 2023) which extracts additional visual context to rephrase underspecified VQA questions.

## Limitations

The plug-and-play capability of ReCoVERR, while facilitating adaptability, does introduce a degree of engineering complexity due to the sequential dependencies in predictions. While it opens up flexible choices in modules, it also introduces an overhead on response time. Our primary method black-box APIs, and we also tested it with open-source models, as detailed in the Table 5. We leave conducting comprehensive testing with various open-source models for future work.

The curated benchmarks tailored for VLM evaluation, while being beneficial, may necessitate further exploration into scenarios that better align with ecological validity (de Vries et al., 2020) and real-world applications for abstaining. Notably, our focus on image understanding excludes considerations of underspecified, ambiguous, or safety-related abstaining, opening avenues for expanding the scope of our approach in the future.

Our experiments are restricted to English, a consequence of leveraging pretrained VLMs and LLMs, offers opportunities for future advancements in multilingual contexts.

Finally, we assume that the base VLMs are calibrated for confidence estimation. Although this is effective to some extent, it leaves room for future exploration into alternative methodologies for making the models well-calibrated.

## 7 Acknowledgements

Jack Hessel partially completed this work while at AI2.

## References

Aishwarya Agrawal, Ivana Kajic, Emanuele Bugliarello, Elnaz Davoodi, Anita Gergely, Phil Blunsom, and Aida Nematzadeh. 2023. Reassessing evaluation practices in visual question answering: A case study on out-of-distribution generalization. In *Findings of the European Association for Computational Linguistics (EACL)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

C.K. Chow. 1957. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. 2023. Improving selective visual question answering by learning from your peers. In *Computer Vision and Pattern Recognition(CVPR)*.

Claudio De Stefano, Carlo Sansone, and Mario Vento. 2000. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(1):84–94.

Harm de Vries, Dzmitry Bahdanau, and Christopher D. Manning. 2020. Towards ecologically valid research on language user interfaces. *CoRR*, abs/2007.14435.

Ran El-Yaniv et al. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Computer Vision and Pattern Recognition (CVPR)*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Computer Vision and Pattern Recognition (CVPR)*.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Computer Vision and Pattern Recognition (CVPR)*.

Jack *Hessel, Jena D *Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning. In *European Conference on Computer Vision (ECCV)*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. 2018. To trust or not to trust a classifier. *Neural Information Processing Systems (NeurIPS)*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5684–5696. Association for Computational Linguistics.

Zaid Khan and Yun Fu. 2023. Selective prediction for open-ended question answering in black-box vision-language models. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2023. Improving automatic vqa evaluation using large language models. *arXiv preprint arXiv:2310.02567*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Computer Vision and Pattern Recognition (CVPR)*.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2023. Rephrase, augment, reason: Visual grounding of questions for vision-language models. *arXiv preprint arXiv:2310.05861*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of Empirical Methods in Natural Language Processing (EMNLP)*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Mustafa Shukor, Alexandre Ramé, Corentin Dancette, and Matthieu Cord. 2023. Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. *CoRR*, abs/2310.00647.

Sanjay Subramanian, Medhini Narasimhan, Kushal Khangaonkar, Kevin Yang, Arsha Nagrani, Cordelia Schmid, Andy Zeng, Trevor Darrell, and Dan Klein. 2023. Modular visual question answering via code generation. *arXiv preprint arXiv:2306.05392*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Neeraj Varshney and Chitta Baral. 2023. Post-abstention: Towards reliably re-attempting the abstained instances in qa. In *Annual Meeting of the Association for Computational Linguistics*.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. In *Findings of the Association for Computational Linguistics (ACL)*.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022a. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision (ECCV)*.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022b. Reliable visual question answering: Abstain rather than answer incorrectly. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 148–166. Springer.

Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E. Gonzalez, and Trevor Darrell. 2023. See, say, and segment: Teaching lmms to overcome false premises. *CoRR*, abs/2312.08366.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mmreact: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. *arXiv preprint arXiv:2305.14985*.

Andy Zeng, Maria Attarian, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations (ICLR)*.

# A Self-Prompting for Producing Calibrated VLM Confidence Estimates

In order to do selective prediction, it is important that we use a calibrated confidence estimate *i.e.* the model confidence $g(x)$ reflects the true probability of its output being correct (Platt, 1999). The calibration of a model's confidence measure is typically evaluated using the Expected Calibration Error (Guo et al., 2017) over a calibration set.

Calibration has typically been studied in discriminative models, using the probability of the predicted class as the confidence estimate. However, it is unclear what makes for a good confidence estimate for a generative VLM that produces an answer $a$ with multiple tokens. One solution is to sum or average all the log probabilities in the answer token sequence string; however, this will be an underestimate of the model's confidence due to surface form competition (Holtzman et al., 2021).

Instead, we experiment with a *self-prompting*, approach where the VLM is prompted to verify the correctness of its answer. Specifically, after predicting an answer $a$ from the VLM for the question $Q$, we present the following prompt $\texttt{prompt}(Q, a)$ to the VLM:

> Question: $\{Q\}$
> Answer: $\{a\}$
> Is the given answer correct for the question? Answer yes or no:

We look at the VLM's next token probability distribution $P(\cdot|\texttt{prompt}, I)$, and compute confidence estimate by looking at the next token probabilities for the yes and no tokens:

$$P_{\texttt{VLM}}(a|Q) = \frac{P(\texttt{yes}|\texttt{prompt}, I)}{P(\texttt{yes}|\texttt{prompt}, I) + P(\texttt{no}|\texttt{prompt}, I)}$$

## A.1 Platt Scaling

Platt scaling (Platt, 1999) is a technique for calibrating classifiers by training a logistic regression model over the logits for the output classes. To train a Platt scaling calibrator, for 12,000 examples in the A-OKVQA training data, we use Self-Prompting as described above for the off-the-shelf VLMs to extract logits for the yes and no tokens. We train a logistic regression model over these logits to predict whether the VLM prediction was correct or not. Figure 3 shows the effect of Platt scaling calibration on the VLM confidence estimates.

# B Prompts for Text-Only Language Models

Table 7 contains the prompts we used for the various language-only functions of ReCoVERR.

# C OK-VQA and Sherlock Task Details

We apply ReCoVERR using A-OKVQA-calibrated VLMs for two new tasks, OK-VQA and Sherlock.

OK-VQA (Marino et al., 2019) is a VQA task that, similar to A-OKVQA, requires external knowledge and commonsense reasoning. We evaluate on the OK-VQA validation set, consisting of 5,046 instances.

Sherlock (*Hessel et al., 2022) is an abductive reasoning task, where images and image regions are paired with inferences. The original task formulations in Sherlock (retrieval, localization, human rating comparison) do not directly apply to generative VLMs that take image and text inputs and produce text outputs. We re-formulate the comparison task into a binary judgment task, where the VLM must output whether a given inference is true for an image or not. An image-region-inference triple is paired with two human ratings that rate the inference as true, false, or neutral. We treat instances where both humans agree that the inference is true as positive instances, and both humans agree that it is false as negative instances. We end up with a dataset of 561 image-region-inference triples. Each triple is presented to the VLM by drawing a region box onto the image, and forming a templated question in the following format:

> Is the given inference true for the given image or not? { Inference } Options: yes, no

| Function | Model | Inputs | Prompt |
|---|---|---|---|
| Question generation | $\mathcal{M}_{\texttt{QGen}}$ | $Q$: The question the VLM is trying to confidently answer<br>$a$:The VLM prediction for above question<br>$\mathcal{E}_R = [e_1, e_2, ...]$: A list of reliable evidences that have been collected about the image so far<br>$K$: The number of questions to generate in this turn. | You are an AI assistant who has rich visual commonsense knowledge and strong reasoning abilities. You will be provided with:<br>1. A target question about an image that you are trying to answer.<br>2. Although you won’t be able to directly view the image, you will receive a general caption that might not be entirely precise but will provide an overall description.<br>3. You may receive some additional evidences about the image.<br>Your goal is: To effectively analyze the image and select the correct answer for the question, you should break down the main question into several sub-questions that address the key aspects of the image.<br><br>What you already know about the image:<br>$e_1$<br>$e_2$<br>...<br><br>Target question: $\{Q\}$. Generate $K$ sub-questions that might help you confirm whether the answer to the target question is ’$\{a\}$’.<br>Here are the rules you should follow when listing the sub-questions:<br>1. Ensure that each sub-question is independent. It means the latter sub-questions shouldn’t mention previous sub-questions.<br>2. The sub-questions should be separated by a newline character.<br>3. Each sub-question should start with "What" or "Is".<br>4. Each sub-question should be short (less than 10 words) and easy to understand.<br>5. The sub-question are necessary to distinguish the correct answer. |
| Sentence paraphrasing | $\mathcal{M}_{\texttt{QA}\rightarrow\texttt{S}}$ | $Q$, $a$: The question-answer pair that needs to be paraphrased into a sentence | Rephrase the question and answer into a single statement.<br>The re-phrased statement should summarize the question and answer.<br>The re-phrased statement should not be a question.<br><br>Question: Is the dog herding or guiding the cows?<br>Answer: guiding<br>Statement: The dog is guiding the cows.<br><br>Question: Are there any other written numbers visible in the image?<br>Answer: no<br>Statement: There are no other written numbers visible in the image.<br><br>Question: Which color of clothing is unique to just one of the two people here?<br>Answer: black<br>Statement: The color of clothing that is unique to just one of the two people here is black.<br><br>Question: Does the picture on the screen involve any human subjects or animals?<br>Answer: human<br>Statement: The picture on the screen involves human subjects.<br><br>Question: $\{Q\}$<br>Answer: $\{a\}$<br>Statement: |
| Checking evidence sufficiency | $\mathcal{M}_{\texttt{NLI}}$ | $\mathcal{H}$: Hypothesis statement<br>$\mathcal{E}_{RR}$: A list of reliable and relevant evidences $[e_1, e_2, ...]$ | Premise: {evidence statements in $\mathcal{E}_{RR}$ concatenated together}<br><br>Hypothesis: $\{\mathcal{H}\}$<br>Can we infer the hypothesis from the premise? Options: yes, no.<br>Answer: |

Table 7: Prompts for text-only language models

|  |  | Risk tolerance $r$ = 10% |  |  |  | Risk tolerance $r$ = 20% |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| VLM | Method | $\mathcal{R}(\downarrow)$ | $\Phi_1(\uparrow)$ | $\mathcal{C}(\uparrow)$ | $\mathbb{R}_{\text{SP}}(\uparrow)$ | $\mathcal{R}(\downarrow)$ | $\Phi_1(\uparrow)$ | $\mathcal{C}(\uparrow)$ | $\mathbb{R}_{\text{SP}}(\uparrow)$ |
| BLIP2 (Off-the-shelf) | Vanilla SelPred | 6.1 | 3.4 | 3.8 | 6.0 | 14.9 | 17.1 | 24.1 | 34.2 |
|  | Vision Tools | 13.8 | 17.4 | 23.5 | 33.5 | 17.0 | 23.5 | 35.1 | 48.3 |
|  | ReCoVERR | 14.3 | **18.8**$_{\pm 0.2}$ | **26.0**$_{\pm 0.5}$ | **37.1**$_{\pm 0.6}$ | 21.7 | **27.1**$_{\pm 0.8}$ | **47.3**$_{\pm 0.3}$ | **61.5**$_{\pm 0.8}$ |
| BLIP2 (Calibrated) | Vanilla SelPred | 4.1 | 3.2 | 3.4 | 5.5 | 11.9 | 16.8 | 21.9 | 32.0 |
|  | Vision Tools | 13.2 | 17.5 | 23.4 | 33.7 | 15.1 | 23.9 | 33.6 | 47.2 |
|  | ReCoVERR | 14.0 | 17.3 | 23.6 | 33.6 | 16.1 | **25.3**$_{\pm 0.5}$ | **36.7**$_{\pm 0.2}$ | **51.0**$_{\pm 0.6}$ |
| InstructBLIP (Off-the-shelf) | Vanilla SelPred | 9.3 | 20.5 | 25.1 | 34.8 | 17.2 | 38.3 | 57.7 | 72.2 |
|  | Vision Tools | 10.5 | 26.1 | 32.9 | 44.8 | 17.6 | 39.6 | 60.5 | 75.3 |
|  | ReCoVERR | 10.9 | 26.3 | 33.5 | 45.2 | 17.9 | **41.3**$_{\pm 0.3}$ | **63.7**$_{\pm 0.3}$ | **78.9**$_{\pm 0.4}$ |
| InstructBLIP (Calibrated) | Vanilla SelPred | 8.5 | 22.0 | 26.4 | 36.9 | 17.5 | 37.2 | 56.6 | 70.5 |
|  | Vision Tools | 10.4 | 27.2 | 34.1 | 46.5 | 17.8 | 38.8 | 59.7 | 74.2 |
|  | ReCoVERR | 11.8 | **29.8**$_{\pm 0.8}$ | **38.7**$_{\pm 1.3}$ | **51.8**$_{\pm 1.4}$ | 18.6 | **41.4**$_{\pm 0.3}$ | **65.0**$_{\pm 0.7}$ | **79.8**$_{\pm 0.8}$ |

Table 8: Metric results as percentages on the A-OKVQA task at two risk tolerance levels. We evaluate selective prediction methods on the overall system risk ($\mathcal{R}$), effective reliability ($\Phi_1$), coverage ($\mathcal{C}$) and recall ($\mathbb{R}_{\text{SP}}$). System risks in red exceeded tolerance. Measurements in **blue** indicate when ReCoVERR outperformed both baselines.