Enhancing Visual Question Answering for Smart Glasses Using Vision-Language Models

Xinxi Chen
Department of Computer Science
Stanford
xchn@stanford.edu

Tianyang Chen
Department of Computer Science
Stanford
tiche@stanford.edu

Abstract

We propose a method to improve Visual Question Answering (VQA) with Retrieval-Augmented Generation (RAG) by introducing text-grounded object localization. Rather than retrieving information based on the entire image, our approach enables the model to generate a bounding box around the object most relevant to the question, allowing for targeted image cropping and focused retrieval. This reduces background noise, improves alignment between visual and textual cues, and helps mitigate hallucinations. Our RAG method enhances context-aware VQA responses increased the accuracy from 22.19% to 25.64%, with an absolute increase of 3.45 percentage points, compared to the baseline Llama-3.2-Vision-11B agent. We also proposed a de-hallucination method based on question type which can effectively reduce the hallucination rate from 65.79% to 19.14% and improves the truthfulness score.

1. Introduction

Visual Question Answering (VQA) [1] sits at the intersection of computer vision and natural language processing, requiring systems to reason over both images and text to produce meaningful answers. Recent advances in Vision-Language Models (VLMs) [6] have significantly enhanced the ability of machines to jointly understand visual and linguistic content, enabling more accurate and context-aware interpretations of complex visual scenes. However, these models are inherently limited by the knowledge encoded in their training data. To address this, Retrieval-Augmented Generation (RAG) [11] introduces an external knowledge retrieval step that grounds model outputs in up-to-date or domain-specific information, bridging the gap between perception and world knowledge. The combination of VLMs with RAG is particularly important for VQA, as it allows systems not only to interpret what they see, but also to reason with additional contextual or factual information—ultimately leading to more robust, informed, and trustworthy responses. In this paper we investigate methods that can improve VQA performance.

A core challenge in combining VQA with RAG lies in identifying and retrieving external knowledge that is simultaneously relevant to both the textual query and the visual content. Unlike pure text-based RAG, where the query alone guides document selection, multimodal RAG must interpret features from an image—objects, scenes, spatial relationships—and align them with the user's question to construct a precise retrieval request. For example, when give the question "How much does this cost?" and an image of users holding one object in hand, while the background contains multiple other objects, the VLM must be able to understand which object is the question referring to in order to perform effective informational retrieval. If the retrieval system focuses too narrowly on one modality (e.g., only on keywords in the text), it risks ignoring critical visual cues; conversely, overemphasizing visual attributes may surface facts that have little bearing on the question's intent. Moreover, the retrieved facts must be filtered and integrated in a way that respects the visual context—misaligned or tangential information can lead to confident but incorrect answers. Balancing these two streams of information to surface grounded, image-aware knowledge is therefore a delicate orchestration that remains an open research frontier in multimodal language understanding.

Furthermore, dehallucinating VLMs is critical to ensuring the reliability and safety of their outputs, especially in applications like VQA where users may rely on responses for decision-making. VLMs, like their language-only counterparts, are prone to hallucination—generating plausible but factually incorrect or visually inconsistent answers—when they lack sufficient understanding or context. This issue is exacerbated when VLMs must reason about complex scenes or incorporate external knowledge, as they may confidently assert false claims not supported by the image or retrieved content. For the example provided by the previous section, a better answer is "I don't know" rather

than responding with wrong price for the wrong object.

2. Related Work

Improving VQA has been a dynamic area of research, with recent efforts focusing on enhancing the gorunding of visual and textual information. GLIP [13] (Grounded Language-Image Pre-training) is a unified vision-language architecture designed to bridge object detection and phrase grounding by reformulating object detection as a vision-language matching task. The core idea is to align image regions (bounding boxes) with phrases from a natural language prompt, enabling the model to detect and ground objects based on open-vocabulary textual queries rather than a fixed set of class labels. GLIP handles both object detection and phrase grounding with a single architecture allowing the model to localize objects of interest that can best answer a given question prompt. Similarly, Grounding DINO [16] tightly integrates language and vision to enable detection and localization of arbitrary objects specified by natural language prompts, rather than being limited to a fixed set of classes. Its architecture features dual backbones for image and text, a feature enhancer module that deeply fuses visual and linguistic features via cross-attention, a language-guided query selection mechanism that dynamically chooses relevant image regions based on the prompt, and a cross-modality decoder that refines predictions by alternating attention between image and text features. Grounding DINO is designed to handle referring expressions in text prompts—including pronouns like ”it”—as part of its referring expression comprehension (REC) capability. The model can localize and identify specific objects or regions within an image based on a given textual description, which may include coreferences such as ”it” if the context in the prompt is clear enough for the model to resolve what ”it” refers to, which is crucial for targeted retrieval augmentation for VQA. The ”Chain-of-Spot” [7] introduces a novel and efficient approach to enhancing the visual reasoning capabilities of large vision-language models (LVLMs) through an interactive reasoning process. What sets Chain-of-Spot apart is its focus on dynamically identifying and attending to key regions of interest (ROI) within an image that are most relevant to the posed question or instruction, rather than processing the entire image at a fixed (often low) resolution. This is achieved by prompting the model to first localize the critical region in response to a query, cropping or zooming in on that region, and then generating the answer based on both the original and the focused image. This interactive, multi-step procedure allows the model to access more detailed and multi-granularity visual features without increasing the computational cost associated with higher-resolution processing.

Visual grounding aims to localize the image region referred to by a given language expression. Methodologies in this field have evolved from multi-modal fusion over a fixed set of detected objects to direct bounding-box prediction with open-vocabulary capabilities. Early approaches integrated object-level visual features into textual representations to enhance generic VQA performance (e.g., via object-text fusion strategies [4]). Later work introduced more structured two-stage pipelines: for instance, a ”Locate Then Generate” framework first predicts the relevant scene-text region and then generates the answer from the cropped area [24]. Recent efforts extend this paradigm to the video domain, where grounding scene-text temporally across frames proves beneficial for text-based video QA [21].

3. Data

We analyzed a VQA dataset collected from Meta Ray-Ban smart glasses [3], which contains both single-turn and multi-turn image-question-answer pairs across 14 diverse domains, including shopping, food, and science. Used for evaluation of our work, this dataset presents a significant challenge due to its varying image quality and ambiguous questions, requiring models to extract the most relevant information from noisy inputs. For instance, consider an image showing several cars near buildings. Performing image retrieval directly on the full image would likely yield results focused on street scenes or buildings, since large background elements tend to dominate the image. Therefore, if the question is "How many passengers can the red car seat?", a retrieval system unaware of the object of interest will fail to provide accurate information (see figure 1).

The dataset features a variety of question types, such as color, counting, location, object recognition, reasoning, and yes/no queries. According to the distribution of question types in the v2 dataset (see Figure 10b), object recognition questions constitute the largest portion. This skew towards object recognition highlights the need for models to possess strong, text-grounded visual understanding, and sometimes be able to localize the object of interest among multiple other objects in the background based on the input text.

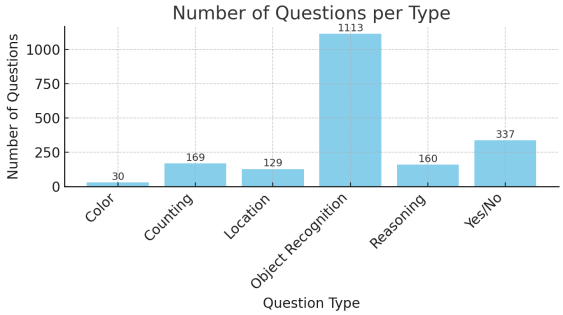
(a) Input image and region of interest

(b) Retrieved image doesn't match the object in question

Figure 1: Example image retrivel without text grounding

3.1. Image Search API

We are utilizing a prebuilt image search API provided by Meta [10], which contains a database of images and associated metadata for 900K items. We noticed that knowing region of interest is crucial to RAG to find useful information as the image search would return very different results for the aforementioned example. The unrelated information retrieved does not help the model answer the question, and



(a) Question-type count

may even promote hallucination, as demonstrated later in the Results section 5. Listing 2 shows using the image from the previous section would result in completely useless information retrieval when user question actually cares about the car parked on the street rather than the street itself.

```
'entities': [{'entity_name': 'Toyota Prius
    v', 'entity_attributes': {'
    alternative_names': ['Prius    ', 'Prius+
    '], 'production_start': 'May 2011', '
    production_end': 'March 2021', '
    body_style': 'compact MPV'...
```

Listing 1: Example Image Search Result Using Cropped Region Of Interest

```
'entities': [{'entity_name': 'Ocean Grove,
    New Jersey', 'entity_attributes': {'
    official_name': 'Ocean Grove, New Jersey
    ', 'settlement_type': '[[Census-
    designated place]]', 'image_skyline': '
    Ocean_Grove_Welcome_Sign.jpg', '
    imagesize': '250x200px', 'image_caption'
    : 'Ocean Grove welcome sign'...
```

Listing 2: Example Image Search Without Knowing Region Of Interest

3.2. Training Data for Bounding Box Detection

We use the Toloka Visual Question Answering Benchmark (WSDMCup2023) dataset by [18] for training to detect the location of bounding box, which can be used for cropping image to improve the image search results used in the RAG system. The dataset provides a corpus of *45,199* image–question pairs, each annotated with a ground-truth bounding box that localizes the visual answer to the corresponding natural language question.

Below is a representative examples from the training set, shown with the question in Table 1. The bounding box is rendered here, which corresponds to the coordinates given.

Question: What do we use to support the immune system and get vitamin C?

Table 1: Example question with corresponding bounding box rendered in the image

4. Methods

To address the problem discussed previously, we explored extending the capabilities of the existing VLM by introducing two key functionalities: (1) localizing the region of interest (ROI), and (2) de-hallucinating uncertain answers. These enhancements are aimed at improving the robustness and accuracy of our RAG agent.

Our main system design follows a Retrieval-Augmented Generation (RAG) framework composed of the following components:

- **Vision-Language Model (VLM):** We experimented with several models including BLIP, QWen, and LLaMA 3.2, ultimately selecting LLaMA 3.2 for its performance and compatibility.
- **Region of Interest Proposer:** A module responsible for identifying the object or region most relevant to answering the question.
- **Image-Based Information Retriever:** Performs web-based or local database search using cropped image regions as queries.

In addition, we fine-tune the VLM to reduce hallucination in uncertain scenarios. As discussed in the Results section, hallucination remains a significant bottleneck to overall performance.

While each module—the VLM, ROI proposer, and retriever—can be individually improved, our current focus is on enhancing the Region of Interest Proposer. Preliminary results show that models often generate irrelevant answers based on distracting background elements. Improving object localization is therefore a promising direction. Future work may explore optimization of the VLM and retriever components.

4.1. Localizing Region of Interest

We reformulate visual understanding as a text-guided object localization problem [14]. Given a question, the model is tasked with identifying the most relevant object by outputting a bounding box around it. This localized ROI is then used to crop the image, reducing background noise and enabling a more focused retrieval process. The retrieved content is passed back to the model to generate the final answer (see Figure 3). We explore different approaches to make our RAG agent region-aware:

4.1.1 Fine-tuning the Language Head

Inspired by recent trends where downstream tasks are framed as natural language generation, we formulate bounding box prediction as a text output task. The VLM is trained to generate bounding box coordinates as part of its text response. This approach is flexible and minimally invasive, requiring no changes to the model architecture. However, it may suffer from imprecise localization, as the model is not explicitly optimized for numerical accuracy. To improve geometric precision, training incorporates both standard language modeling loss and a CIOU loss [20] on the generated coordinates. 4 shows how training dataset and prompt are formatted.

4.1.2 Fine-tuning a Bounding Box Head

An alternative approach involves attaching a dedicated bounding box regression head to the frozen VLM, similar to methods proposed by [22, 23]. This head leverages the final cross-attention features (last hidden layer) between vision and text modalities to produce accurate, grounded box predictions. This design improves spatial precision by directly optimizing for localization, though it adds architectural complexity and reduces generality across open-ended queries. Training is supervised using only CIOU loss on the predicted boxes. In this case, the ground truth label are floating point numbers instead of strings.

4.1.3 Using a Pretrained Localizer

Although integrating a trained localization component keeps the system self-contained, it incurs training costs and complexity. As an alternative, we also experimented with incorporating a pretrained localizer, such as Grounding DINO [15], to identify regions of interest prior to retrieval. The cropped region is then used to perform visual search, and the resulting information is fed back into the VLM to answer the original question.

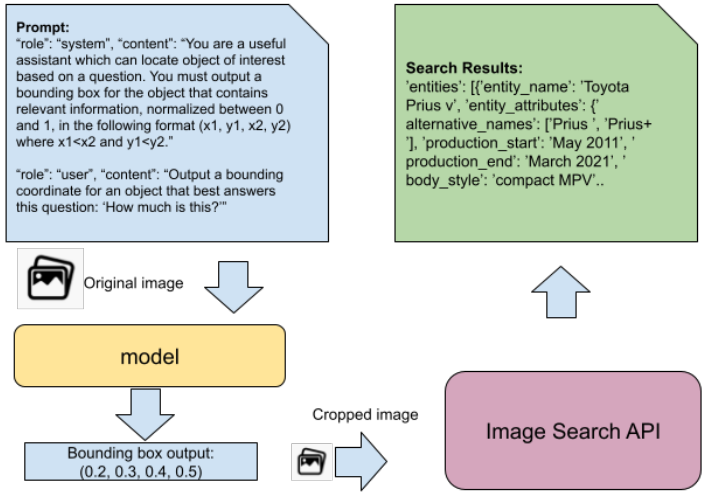


Figure 3: Overview of the Retrieval-Augmented Generation (RAG) Pipeline.

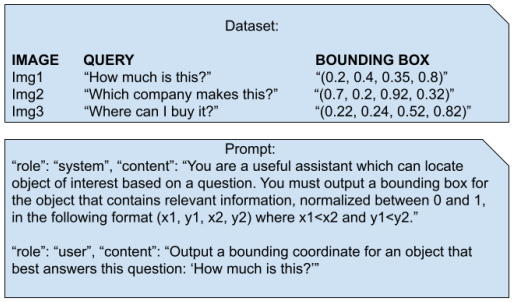


Figure 4: Training the VLM to Output Text-Grounded Bounding Boxes.

4.1.4 Multi-task Bounding Box Head

Overall Architecture Design Different from the simple pooling of cross attention states described in 4.1.2, we also investigated more sophisticated bounding box regression head design. The core idea is to extend the LLaMA 3.2 vision-language model so that, after it produces pooled visual features and text embeddings, a small fusion MLP concatenates these vectors and learns to regress normalized bounding boxes directly, as shown in Figure 5. By "late-fusing" the text vector (LLaMA 3.2 outputs) and the vision vector, we avoid adding separate detection networks or region proposals. During training, the fusion head minimizes a combination of generalized IoU loss, a size L1 loss, and an IoU-guided classification loss, letting the model to "point" at the correct object without changing LLaMA's core Transformer stacks. This keeps the final RAG pipeline simple (only one VLM backbone) while enabling precise spatial grounding for question-driven box prediction.

- **1. Vision Encoding:** We use a lightweight vision encoder to extract a compact representation $\mathbf{v} \in \mathbb{R}^D$ from each image. The current setup leverages pretrained transformer-based features (e.g., from Dinov2 [17]) and serves as a skeleton design for rapid integration, with planned extensions toward richer ViT-style embeddings in future iterations.

- **2. Text Encoding:** The natural-language question is tokenized and passed through a transformer-based text encoder (i.e. a LLaMA 3.2 language model). Once we obtain the final-layer hidden states for all tokens, we apply linear projection layer then maps $\mathbf{t} \in \mathbb{R}^H$ into a lower-dimensional vector $\mathbf{t}' \in \mathbb{R}^d$.

- **3. Feature Fusion:** We fuse the visual vector $\mathbf{v}$ and the text vector $\mathbf{t}'$ via a simple MLP to obtain the joint embedding $\mathbf{h} \in \mathbb{R}^d$. This baseline fusion strategy is lightweight and extensible, with room for future upgrades such as cross-modal attention.

- **4. Prediction Heads:** From $\mathbf{h}$, two parallel heads are applied:

  - **Bounding-Box Head:** A two-layer MLP followed by a sigmoid activation outputs four values in $[0, 1]^4$, representing normalized box center and size $(\hat{c}_x, \hat{c}_y, \hat{w}, \hat{h})$.
  - **Classification Head:** A single linear layer produces one logit $s \in \mathbb{R}$, intended to estimate how well the predicted box overlaps the ground truth.

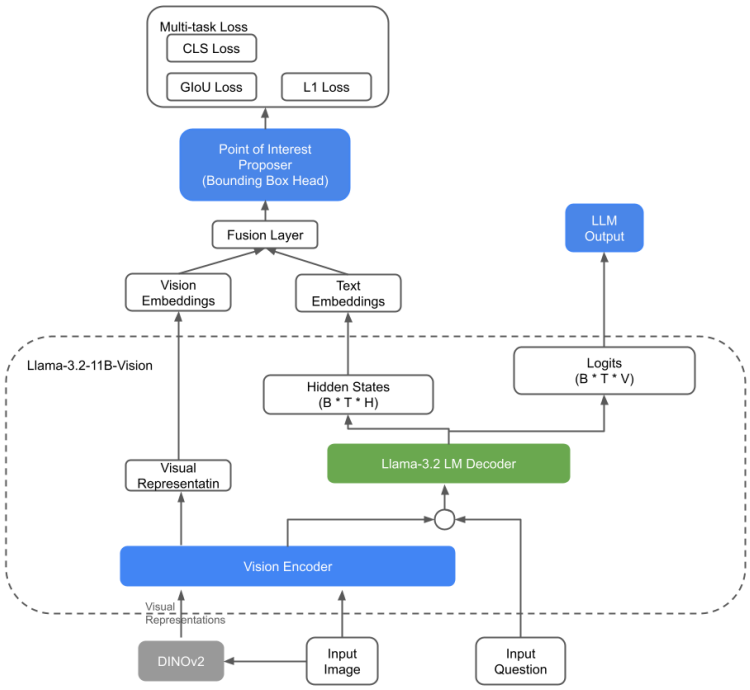


Figure 5: Multi-task Model Architecture

Weighted Multi-Task Loss Enhances Granularity and Mitigates Overgeneralization When trained with only a GIoU loss, the model can converge to predicting a single, large bounding box that overlaps the target by roughly 40%—thereby including excessive background. To counteract this, we add two complementary terms:

- **Size L1 Term:** Penalizes predictions whose width/height are too large. Denote the predicted box as $\hat{\mathbf{b}} = (\hat{c}_x, \hat{c}_y, \hat{w}, \hat{h})$ and ground truth as $\mathbf{b} = (c_x, c_y, w, h)$. We apply

$$L_{\ell_1} = |\hat{w} - w| + |\hat{h} - h|.$$

  This term grows whenever the model attempts to enlarge $\hat{w}$ or $\hat{h}$ beyond the true size, discouraging overly large boxes that merely satisfy a modest IoU.

- **IoU-Guided Classification Term:** Encourages the classification head's logit s to reflect actual overlap quality. Let the predicted and ground-truth boxes (in corner format) be p and g, and define the clamped GIoU target

$$\tau = \max\big(0, \mathrm{GIoU}(p, g)\big) \in [0, 1].$$

  We then train

$$L_{\mathrm{cls}} = -\big[\tau \log\big(\sigma(s)\big) + (1 - \tau) \log\big(1 - \sigma(s)\big)\big].$$

  If a predicted box overlaps poorly (low GIoU), the target τ is small, and $\sigma(s)$ is driven down. Hence, boxes that merely "cover" 40% of the object but include too much background incur a low IoU target and higher classification loss.

- **GIoU Term:** Retains the original overlap-based objective:

$$L_{\mathrm{GIoU}} = 1 - \mathrm{GIoU}(p, g).$$

  Alone, this term can be satisfied by expanding the box until a minimal overlap threshold is met; combined with the other terms, it ensures precise alignment around the object.

The total loss is

$$L_{\mathrm{total}} = L_{\mathrm{GIoU}} + \lambda_{\mathrm{cls}} L_{\mathrm{cls}} + \lambda_{\mathrm{size}} L_{\ell_1},$$

where λ_{cls} and λ_{size} are constant parameters.

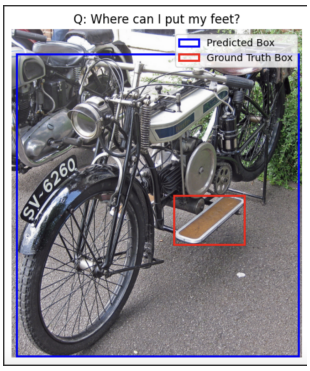
Effect of the Multi-task Loss Function:

- *GIoU alone* encourages any box that overlaps enough—often "lazy" large boxes covering ≈40% of the target.

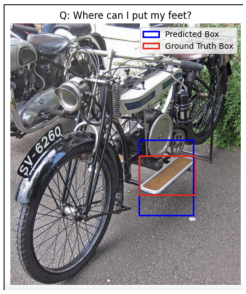
- *Adding* L_{ℓ_1} penalizes excessive width/height, so the model must shrink the box rather than inflate it.

- *Adding* L_{cls} further penalizes low overlap: a large box with only 40% IoU yields a small τ, forcing $\sigma(s)$ to drop and incurring classification loss.

Together, these terms prevent overly general boxes and drive the model toward tighter, more precise localization. As shown in Figure 6, in an example about a bicycle, the left prediction shows that only using GIoU loss generates a relatively large bounding box, while the right predicted bounding boxes are more precise, which is better for cropping and downstream image search task.



(a) GIoU Loss Only



(b) Weighted Multi-task Loss

Figure 6: Comparison of Results with Different Loss Function

4.2. Finetuning to Reduce Hallucination

We also fine tuned the model to say "I don't know" in some cases to reduce hallucination. For example, for the following question in Table 2, Llama-3.2-Vison-11B outputs an incorrect answer. We can train the model to answer "I don't know" for "who" type question, since this typically requires external knowledge. We investigate the questions based on the type shown in Figure 10b, and then fine tune the model to only answer the questions with high confidence.

5. Results

We have collected baseline performance for three popular open source VLMs: BLIP (Bootstrapping Language-Image Pre-training) [12] , Llama 3.2 (11B) [8] and Qwen 2.5 (3B) [5], on this dataset without any information retrieval implemented.

5.1. Results: Training a Simple Bounding Box Regression Head

We trained the bounding box regression head described in Section 4.1.2 for 4 epochs on Llama 3.2 (11B). However, the loss curve remained flat, indicating minimal learning progress. Despite experimenting with various hyperparameters, we observed signs of instability—likely due to the fact that the regression head is initialized from scratch. 7 Successful training may require careful weight initialization and constraints to ensure valid bounding box predictions early on (e.g., enforcing $x_1 < x_2$, $y_1 < y_2$ to calculate IOU loss). While we attempted clamping to maintain validity, this led to vanishing gradients, preventing effective learning and leaving the training process in a suboptimal state.

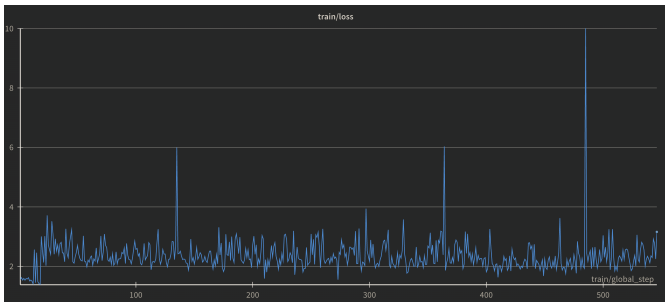


Figure 7: Bounding Box Regression Head Training loss

5.2. Results: Fine-Tuning the Language Head

For fine-tuning the language head, we used the Unsloth framework [2] to perform QLoRA [9] training. The LlaMA 3.2 11B model was loaded in 4-bit quantized mode, and LoRA adapters were applied to the language head, attention layers, and MLP layers using $\alpha = 16$ and $r = 16$. This resulted in approximately 52k trainable parameters, accounting for only 0.48% of the full 11B model. The training loss showed a clear downward trend, indicating effective learning 8. However, due to computational constraints, we had to stop training early. Despite this limitation, the approach appears promising and scalable with sufficient training data and compute resources.

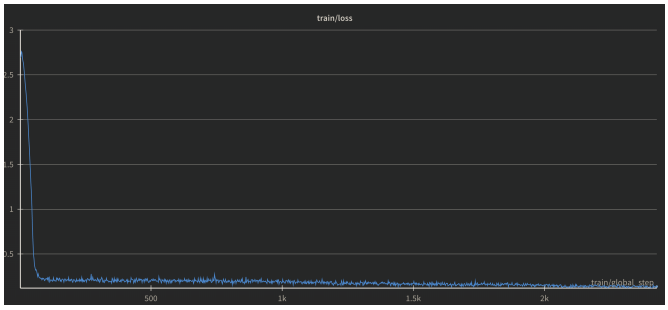


Figure 8: Fine-tuning Language Head Training Loss

Table 2: Evaluation record for Interaction ID `00a48`

ID	Query	Agent Response	Ground Truth	Result
00a48	Who invented this kind of tape?	The tape measure was invented by the French tailor *Pierre-Frédéric Guillaume*, who patented the first practical retractable tape measure in 1829 (prototype in 1821).	James Chesterm…	INCORRECT

5.3. Results: Multi-task Bounding Box Head

With a cosine learning rate scheduler, the multi-task loss decreased fast in the first hundreds of steps, and then decreased slowly in the following steps as shown in Figure 9.

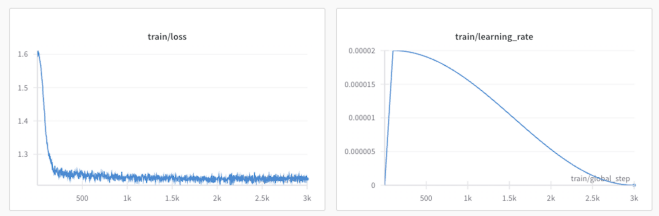
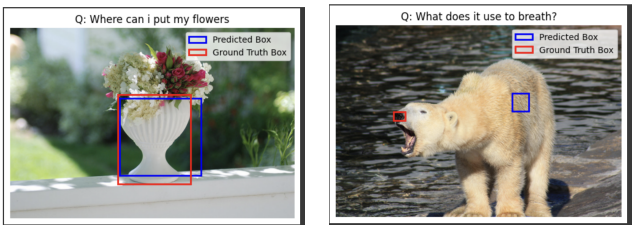


Figure 9: Training loss over steps for Multi-task Loss

The following example in Figure 10 shows a positive example from the predictions that it can detect the bounding box of the vase. There is also a negative example shown in the Figure that the location of the bounding box is away from the ground truth place. Due to time limitation of the project, we did not integrate this into the RAG flow yet, and we will experiment with RAG in the future if we have time.



(a) Positive Example (b) Negative Example

Figure 10: Example of Predictions of Multi-task Bounding Box Detector

5.4. End To End Evaluation Method

We use `GPT-4o-mini` as a judge, guided by three rules: (i) a prediction is correct when it contains all key information in the ground truth, (ii) paraphrasing is acceptable if the meaning is unchanged, and (iii) a prediction is incorrect if it introduces errors or omits essentials. For each question we assign a scalar score—**Perfect** (1.0), **Acceptable** (0.5, minor non-harmful flaws), **Missing** (0.0, refusal/"I don't know"), or **Incorrect** (–1.0, wrong or irrelevant). A system's *Truthfulness Score* is the mean of these values across the evaluation set, yielding a range of –1 (all wrong) to 1 (all perfect). Listing 3 shows the full prompt for the judge.

5.5. End-to-end Performance Comparison

5.5.1 Baseline on V1 Dataset

The CRAG MM dataset has two versions, where the first version v1 includes 1548 single turn questions, and the second version v2 includes 1938 single turn questions. We evaluated our baseline on v1 dataset with three different pre-trained models: BLIP, Qwen-VL-2.5-3B and Llama-3.2-Vision-11B. Since Llama-3.2-Vision-11B performs relatively great, we use Llama-3.2-Vision-11B for our customized solutions and evaluate only Llama-3.2-Vision-11B on v2 dataset as baseline for our customized solutions.

Based on the results of three pretrained unmodified models on v1 1548 questions from Table 4, the blip-vqa-base model performs poorly on the real world data set with only 3.49% accuracy. The Qwen-VL-2.5-3B and Llama-3.2-Vision-11B perform relatively better with 18.09% and 26.23% accuracy. The Hallucination rate is relatively higher in Llama-3.2-Vision-11B compared to Qwen-VL-2.5-3B. As a result, Llama-3.2-Vision-11B has the highest truthfulness score, and Qwen-VL-2.5-3B ranks the second.

5.5.2 De-Hallucination Results on V2 Dataset

With the fine tuning strategy on question type, we can effectively enable the model to answer "I don't know" in the complicated question types like reasoning to avoid hallucination. As a result, we can reduce the hallucination rate from 65.79% to 19.14%, and effectively improve the truthfulness score from -0.4360 to -0.0738, as shown in Table 5 for v2 dataset.

5.5.3 End To End Results on V2 Dataset

We evaluated our RAG agent on the new version of the dataset using two different prompting strategies. Both experiments use Grounding DINO (grounding-dino-tiny, 172M) for bounding box extraction instead of our internally trained bounding box heads due to limited compute budget.

Table 6 compares the performance of the baseline model (without RAG) against the two prompting approaches. In the first strategy, we crop the input image using the detected bounding box, perform a retrieval based on the cropped region, and use the retrieved information to answer the question. In the second strategy, we follow a Chain-of-Spot-style approach [7]: the model first summarizes the region of interest, then we crop the image using Grounding DINO, and finally feed both the summary and search results back into the model to answer the original question.

Surprisingly, using Grounding DINO alone in the first experiment led to a nearly 5% drop in accuracy, with a hallucination rate comparable to the baseline. In contrast, the Chain-of-Spot-style prompting improved accuracy beyond the baseline, but introduced a higher hallucination rate (5%), which ultimately lowered the overall score.

Upon analyzing outputs from different stages of our RAG agent, we observed the following:

- When image-based information retrieval returns completely irrelevant content, it can mislead the model into producing incorrect answers (Table 7). This limits the effectiveness of RAG and results in performance comparable to the baseline. In contrast, Chain-of-Spot prompting mitigates this issue by first asking the model to describe what it sees before incorporating retrieved information. We hypothesize that this approach encourages the model to rely more confidently on its own visual understanding than on potentially misleading external sources.

- However, Chain-of-Spot prompting can also increase hallucination. That is, once the model identifies the region of interest, it may become overly confident in its predictions. As shown in Table 8, this can sometimes override correct prior knowledge, leading to confidently incorrect answers. This behavior contributes to a higher hallucination rate and lowers the overall accuracy. This is evident given the low missing rate and the model is less likely to output "I don't know".

- Computation wise, the gounding DINO is extremely efficient to generate bounding boxes, while the Chain-of-Spot prompting requires the model to look at the image twice, leading to almost double inference time. On A100 with 80 GB VRAM using a batch size of 36, one full evaluation on v2 dataset takes 1 hour, and on average each batch takes 1.8 seconds. This is consistent with other test-time scaling technique like Chain-of-Thought [19].

6. Conclusion

In summary, this work demonstrates that incorporating text-grounded object localization into retrieval-augmented VQA systems enables models to produce more accurate and context-aware answers by focusing on the most relevant image regions for each question. By leveraging Chain-of-Spot-style prompting, our RAG agent is able to effectively combine retrieved web content with the model's own visual understanding. Experiments on challenging real-world datasets show that this localization-based strategy improves accuracy over baseline methods, though at the cost of increased hallucination.

Interestingly, our de-hallucinated model, which more frequently responds with "I don't know," achieves the highest overall score—highlighting a valuable real-world insight: providing an incorrect answer can be more detrimental than admitting uncertainty. This underscores a practical challenge of using RAG agents in safety-critical applications.

Future work could explore combining Chain-of-Spot prompting with a fine-tuned, de-hallucinated VLM, enabling the agent to retain low hallucination rates while still leveraging external information to enhance accuracy.

7. Acknowledgment

We sincerely thank Sabri Eyuboglu for the valuable insights on the topic of vision-language models.

A. Additional Results

```
"You are an expert evaluator for question answering systems. "
"Your task is to determine if a prediction correctly answers a question based on the ground truth.\n\n"
"Rules:\n"
"1. The prediction is correct if it captures all the key information from the ground truth.\n"
"2. The prediction is correct even if phrased differently as long as the meaning is the same.\n"
"3. The prediction is incorrect if it contains incorrect information or is missing essential details.\n"
"Output a JSON object with a single field 'accuracy' whose value is true or false."
```

Listing 3: LLM judge prompt

Table 3: Overall evaluation metrics for Llama-3.2-Vision-11B on V1 Dataset (Baseline vs. De-hallucination)

Model	**Total conversations**	**Total turns**	**Exact accuracy** (%)	**Accuracy** (%)	**Missing rate** (%)	**Hallucination rate** (%)	**Truthfulness score**
Llama-3.2-Vision-11B (Baseline)	1548	1548	0.84	26.23	13.24	60.53	−0.3430
Llama-3.2-Vision-11B (De-hallucination)	1548	1548	0.90	12.60	69.64	17.76	−0.0517

Table 4: Overall evaluation metrics for three vision–language models without RAG

Model	**Total conversations**	**Total turns**	**Exact accuracy** (%)	**Accuracy** (%)	**Missing rate** (%)	**Hallucination rate** (%)	**Truthfulness score**
blip-vqa-base	1548	1548	0.00	3.49	0.00	96.51	−0.9302
Qwen-VL-2.5-3B	1548	1548	0.78	18.09	33.72	48.19	−0.3010
Llama-3.2-Vision-11B	1548	1548	0.84	26.23	13.24	60.53	−0.3430

Table 5: Overall evaluation metrics for Llama-3.2-Vision-11B on V2 Dataset (Baseline vs. De-hallucination)

Model	**Total conversations**	**Total turns**	**Exact accuracy** (%)	**Accuracy** (%)	**Missing rate** (%)	**Hallucination rate** (%)	**Truthfulness score**
Llama-3.2-Vision-11B (Baseline)	1938	1938	0.46	22.19	12.02	65.79	−0.4360
Llama-3.2-Vision-11B (De-hallucination)	1938	1938	0.67	11.76	69.09	19.14	−0.0738

Table 6: Overall evaluation metrics for RAG implementation

Model	**Total conversations**	**Total turns**	**Exact accuracy** (%)	**Accuracy** (%)	**Missing rate** (%)	**Hallucination rate** (%)	**Truthfulness score**
Llama-3.2-Vision-11B (Baseline no RAG)	1938	1938	0.46	22.19	12.02	65.79	−0.4360
GDINO	1938	1938	0.15	17.75	16.87	65.38	−0.4763
GDINO + chain of spot	1938	1938	0.83	25.64	3.77	70.59	−0.4494

Table 7: Model Output When Search Result Is Not Useful

Question	**Image**	**Image Crop + Search**	**Chain-of-Spot**	**Baseline**
What is the typical filling of this Chinese steamed bun? **Groundtruth:** The typical filling is pork.	**Original:**  **Cropped:** 	**Search result:** Blood soup. **Response(incorrect):** The typical filling of this Chinese steamed bun is not blood soup, as the image shows a steamed bun with a brown filling, not a soup.	**Search result:** Blood soup. **ROI summary:** The typical filling of this Chinese steamed bun is pork. **Response(correct):** The typical filling of this Chinese steamed bun is pork.	**Response(correct):** The typical filling of this Chinese steamed bun is pork.

Table 8: Model Output When Chain-of-spot makes over confident outputs

Question	**Image**	**Image Crop + Search**	**Chain-of-Spot**	**Baseline**
How old was this artist when he started hosting his own show on NBC? **Groundtruth:** Nat King Cole was 37 years old.	**Original:**  **Cropped:** 	**Search result:** Levi Strauss & Co. is an American clothing company **Final answer:** I don't know.	**Search result:** Levi Strauss & Co. is an American clothing company **ROI summary:** The object of interest is a mural of Nat King Cole, an American singer and musician. **Response(incorrect):** Nat King Cole was 31 years old when he started hosting his own show on NBC, "The Nat King Cole Show," in 1956.	**Response(correct):** Nat King Cole was born on March 17, 1919, and he started hosting his own show on NBC in 1956. Therefore, he was 37 years old when he started hosting his own show on NBC.

References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh. Vqa: Visual question answering, 2015. arXiv:1505.00468.

[2] U. AI. Unsloth: Efficient fine-tuning of large language models, 2025. Accessed: 2025-06-04.

[3] AIcrowd. Meta crag-mm challenge 2025. `https://www.aicrowd.com/challenges/meta-crag-mm-challenge-2025`, 2025. Accessed: 2025-04-25.

[4] C. Alberti, J. Ling, M. Collins, and D. Reitter. Fusion of detected objects in text for visual question answering. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2131–2140, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report, 2025.

[6] F. Bordes et al. An introduction to vision-language modeling. 2024. Provides a comprehensive introduction to VLMs, their architectures, training methods, and evaluation strategies.

[7] Y. Dong et al. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024.

[8] A. Grattafiori, A. Dubey, and A. J. et al. The llama 3 herd of models, 2024.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[10] Jyotish. Search api — meta comprehensive rag benchmark starter kit, May 2025. Accessed 16 May 2025.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[12] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[13] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, 2022.

[14] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11032–11042, 2022.

[15] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[16] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[17] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[18] D. Ustalov, N. Pavlichenko, S. Koshelev, D. Likhobaba, and A. Smirnova. Toloka visual question answering benchmark, 2023.

[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[20] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12993–13000, 2020.

[21] S. Zhou, J. Xiao, X. Yang, P. Song, D. Guo, A. Yao, M. Wang, and T.-S. Chua. Scene-text grounding for text-based video question answering, 2025.

[22] Y. Zhou, X. Wang, X. Li, Y. Li, F. Yu, T. Darrell, L. Zhang, Z. Yu, Z. Liu, Y. Li, et al. F-vlm: Open-vocabulary object detection with frozen vision-language models. *arXiv preprint arXiv:2306.17107*, 2023.

[23] Y. Zhou, X. Wang, X. Li, Y. Li, F. Yu, T. Darrell, L. Zhang, Z. Yu, Z. Liu, Y. Li, et al. Teaching vision-language models to detect novel objects. *arXiv preprint arXiv:2411.18207*, 2024.

[24] Y. Zhu, Z. Liu, Y. Liang, X. Li, H. Liu, C. Bao, and L. Xu. Locate then generate: Bridging vision and language with bounding box for scene-text vqa, 2023.