Visual Grounding for Object Questions

Martin N. Everaert[1]*, Hiroyuki Takeda[2], Xiruo Liu[2], Vivek Yadav[2], Vidya Narayanan[2], Raja Bala[2]
[1]EPFL, Switzerland [2]Amazon Inc.
[1]martin.everaert@epfl.ch [1]{hrtakeda, xiruoliu, ydvivek, vidyanrn, rajabl}@amazon.com

Abstract

Current visual grounding research remains limited for practical applications, because existing tasks primarily focus on direct visual queries (e.g., "find the red car") or reading visible text (e.g., "what is the title of this book?"), rather than supporting general questions about objects (e.g., "how comfortable are these earbuds?"). We introduce the novel problem of Visual Grounding for Object Questions (VGOQ). Unlike previous tasks that ground only what is directly visible in images, VGOQ handles open-ended general questions about objects, including concepts such as ease and comfort of use, and aims to identify visual evidence or context that would support an answer. This unexplored problem has immediate practical value, particularly in designing and optimizing product imagery in e-commerce stores. As initial steps toward this task, we develop two automated data generation techniques, which serve to train a lightweight visual grounding model, and to evaluate visual grounding approaches on the resulting synthetic benchmarks, ABO-VGOQ and VizWiz-VGOQ. Our results provide initial evidence that VGOQ represents a meaningful research direction: current SoTA visual grounding performance decreases from 52% gIoU to 37% gIoU when questions are rephrased from visual questions (segmentation of the answer) to general object questions (segmentation of visual evidence). On our new benchmarks, our lightweight model outperforms prior models while being much smaller. Project page: https://martin-ev.github.io/vgoq.

1. Introduction

We introduce the problem of *Visual Grounding* for *Object Questions* (VGOQ), where we aim to find visual evidence or context in images that is useful for answering general questions about objects. Traditional visual grounding research-focuses primarily on directly visible elements in images. In contrast, VGOQ tackles general object questions whose answers may not be immediately apparent in the images.

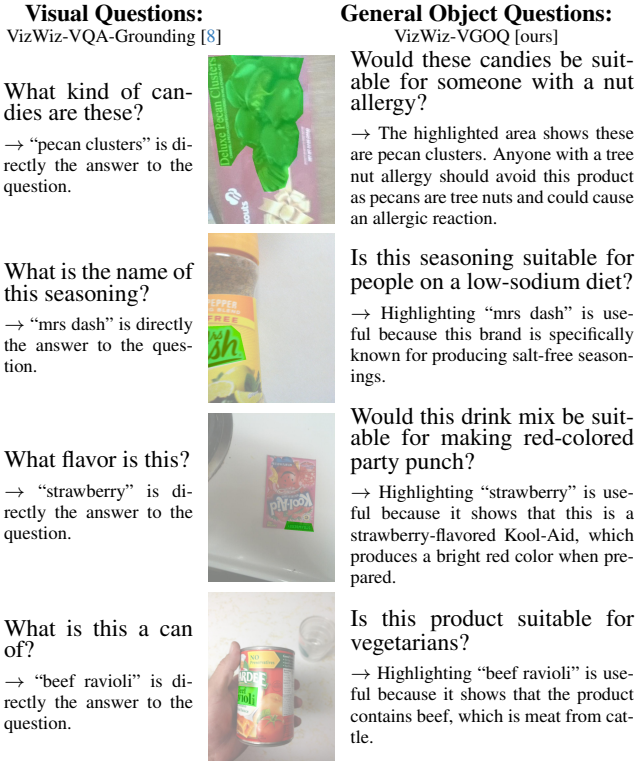


Figure 1. **From Visual Grounding of Visual Questions to Visual Grounding of Object Questions:** Existing visual grounding datasets, such as VizWiz-VQA-Grounding [8] (left), focus on locating directly visible answers to visual questions or referring expressions. In this study, we focus on the more challenging task of finding visual evidence/context that is helpful for general, open-ended, object questions (right). **Creating the synthetic VizWiz-VGOQ dataset:** One of the ways to address the absence of data for our task is to rewrite existing Visual Questions (left) into general Object Questions (right). The same segmentation masks (middle column) now highlight visual evidence or context useful for answering the questions, rather than being the direct answer.

Among many applications (Section 7), the ability to segment image regions that show relevant visual evidence or context to answer the question can enhance interpretability and trust in question answering systems, particularly for e-commerce applications where customers ask abstract ques-

*This work was conducted during Martin's internship at Amazon.

tions about product comfort, durability, or suitability.

In traditional visual grounding datasets, *e.g.*, *Visual Question Answering* (VQA) grounding [8, 27] and *Referring Expression Segmentation* (RES) [15], most samples focus on what is visible in the images, such as locating specific objects (*e.g.*, find the "white bowl with vertical stripes"), directly reading visible text (*e.g.*, "what brand of smartphone is this?"), or describing spatial relationships (*e.g.*, "is the red sphere left of the blue cube?"). These tasks involve direct matching between language descriptions and visible image elements. Often, the visual annotations (*e.g.*, segmentation masks) for these tasks are the answer to the question (*e.g.*, segmentation of the red car, of the title of the book) or a segmentation of the objects mentioned in the question (*e.g.*, segmentation of the red sphere and the blue cube). VGOQ is more challenging, starting with open-ended general questions about objects, *e.g.*, "how comfortable are these earbuds?", and locating visual evidence or context (*e.g.*, the silicon eartips of the earbuds) that would support answers to the question.

One key challenge of VGOQ is the lack of datasets containing images, questions about objects, and visual grounding of evidence rather than visual grounding of the answer directly. As a first step toward addressing this problem, we create two automated techniques: (1) transforming existing visual questions from VQA grounding datasets into general object questions (Figure 1), and (2) a zero-shot pipeline using Claude [1] and traditional grounding models [10, 29, 36] to create visual grounding in a zero-shot manner for generating customer questions about products (Figure 2).

To enable practical applications in e-commerce stores that need to process millions of product images and queries, we train a lightweight CLIP-based [26] grounding model inspired by CLIPSeg [22]. The model is trained jointly on traditional visual grounding tasks and our VGOQ task.

We evaluate existing SoTA visual grounding models on existing VQA grounding benchmarks (VizWiz-VQA-Grounding [8], TextVQA-X [27], and Toloka [33]) and our new VGOQ benchmarks. Results show most existing visual grounding models struggle with finding visual evidence for general object questions. Performance drops significantly from 52.2% mean intersection-over-union (gIoU) to 37.2% gIoU when shifting from visual questions (segmentation of the answer) to object questions (segmentation of visual evidence/context) on VizWiz-VQA-Grounding images [12]. On our new benchmarks, for the tasks of locating specific visual evidence or context related to the question, our lightweight model has higher gIoU than previous SoTA approaches, and is competitive with our concurrent work, Qwen3-VL [3], while being much smaller and faster. Our main contributions are:

- We introduce **Visual Grounding for Object Questions (VGOQ)**, a novel task that extends visual grounding be-

Figure 2. **Creating the synthetic ABO-VGOQ dataset:** Our data generation pipeline for ABO-VGOQ combines multiple product images and textual metadata from e-commerce listings. Each example shows: (left) product images from the Amazon Berkeley Objects dataset (ABO [9]), (right) generated customer question, generated groundtruth visual grounding in the most relevant image, and generated evidential quality category. Notice how the VGOQ task requires multiple skills: material recognition (identifying fabric textures, ingredient lists), spatial reasoning (understanding dimensions and proportions), inference from visual context (connecting observable features to functional properties), multi-modal integration (combining textual and visual information), and infographics understanding (reading text within images).

yond direct visual queries to handle questions about objects, potentially abstract, requiring identification of visual evidence or context rather than direct answers.
- We develop **two automated data generation techniques** and create **two new datasets**: VizWiz-VGOQ (7469 samples) and ABO-VGOQ (6571 samples), allowing us to train and evaluate models for VGOQ. We released the evaluation sets as benchmarks.
- We train a simple **lightweight visual grounding model** that outperforms prior SoTA visual grounding models on VGOQ task, while being much smaller (1.77M parameters) and trained on much less data.
- We highlight some **practical applications** of VGOQ, particularly for e-commerce stores, *e.g.*, enhanced shopping assistants with visual evidence, automatic infographics generation, and automated feedback systems for vendor-provided product imagery.

2. Related work

2.1. Visual Grounding

Visual Grounding is the task of localizing specific parts of an image based on textual descriptions. Existing approaches have focused primarily on traditional *Image segmentation* [6, 14, 22], *Referring Expression Segmentation* (RES) [15] and *Visual Question Answering* (VQA) grounding [8, 27]. While these approaches have shown success in their respective domains, they are often limited to directly observable visual elements.

Traditional segmentation approaches associate classes to segmentation maps. They typically use a predefined list of classes, *e.g.*, 80 object classes for COCO dataset [20]. **Open vocabulary segmentation** [6, 14, 22] extends this to arbitrary class names. However, these methods cannot handle longer descriptions, describing what to segment precisely, or requiring more reasoning, such as segmenting the visual evidence that could be used to answer a customer question about properties of a product shown in the images.

Referring Expression Segmentation (RES) focuses on localizing objects in images based on longer natural language descriptions [15], beyond simple class names. Traditional RES datasets like RefCOCO contain annotations for simple referring expressions (*e.g.*, "the red car on the left"). More recent work like ABO-Image-ARES [35] has extended this to product imagery, including references to object parts and attributes of the object parts (*e.g.*, "frosted glass"). However, these approaches still primarily handle direct visual descriptions rather than general open-ended questions about objects.

Visual Question Answering (VQA) systems aim to answer Visual Questions (VQ) about images [2, 4, 12, 30]. Traditional VQA focuses on questions that can be answered by direct observation of the image, such as counting objects or describing visible attributes. While powerful, these systems typically don't handle questions requiring inference beyond what's directly visible. Newer datasets like VQA v2.0 [11], or OK-VQA [23], have expanded to include questions requiring external knowledge and more complex reasoning. VQA is typically done with a single image as input, and datasets and works with multiple images as input remain underexplored [4].

The emergence of **multimodal large language models (MLLMs)** like GPT-4V [24], GPT-4o [13], Claude 3.7 Sonnet [1], and Gemini 1.5 [32] has expanded VQA capabilities to include more abstract reasoning and world knowledge. These models often show strong answering capabilities, but typically lack explicit visual grounding for their answers, limiting transparency and verifiability. Recent work [31] also shows that MLLMs struggle with abstract-oriented language, due to concrete-expression-biased training datasets.

VQA grounding datasets like TextVQA-X [27] and VizWiz-VQA-Grounding [8] combine VQA with Visual Grounding capabilities. They include ground truth segmentation indicating the answer to the questions. However, these datasets still primarily focus on questions about directly visible elements (*e.g.*, "What text is written here?", "What is this object?", "What color is this object?") rather than general questions about objects (*e.g.*, "How comfortable is this object?"). Figure 1 illustrates this difference.

2.2. Visual grounding models

Large Multimodal Models. Several large multimodal models have been developed for visual grounding tasks. To mention a few, **UnifiedIO** [21] converts heterogeneous inputs and outputs (images, text, bounding boxes, masks, keypoints) into unified token sequences. The vocabulary of the LLM is expanded from only text tokens with 1000 coordinates tokens, to allow input/output like bounding boxes, and 16384 visual tokens, to allow outputs like generated images, segmentation masks, depth maps, etc. UnifiedIO is trained on 95 diverse datasets, including VizWiz-VQA-Grounding [8]. **LISA** [19] combines VQA, segmentation, and RES, and outputs special <SEG> tokens that are decoded into segmentation masks for segmentation and RES. **GLaMM** [28] is trained to generate image captions, where the captions are interleaved with segmentation masks (using the special <SEG> token) of the phrases mentioned in the caption. Interestingly, GLaMM is trained on unannotated images, and annotation is obtained automatically by combining outputs of several existing state of the art models. **Molmo** [10] provides pointing capabilities that can be combined with **SAM** [29] to obtain visual grounding [5, 7, 37].

Lightweight visual grounding models. Several approaches have explored efficient architectures for visual grounding, often building upon pre-trained vision-language

models like CLIP [26]. Works such as [16, 17, 38] or CLIPSeg [22] have demonstrated that a lightweight transformer model (or simply self-attention / cross-attention layer) on top of the image CLIP features can achieve relatively competitive performance while maintaining computational efficiency. These lightweight models are relevant for practical applications requiring deployment at scale, particularly in e-commerce stores with millions of products.

3. Problem statement

We focus on Visual Grounding for Object Questions (VGOQ), where the input consists of an open-ended general question q about an object, images $(I_i)_{i=1,\ldots,j}$ of this object, and optional textual information t. In the setting of e-commerce stores, q could be an abstract customer question about a product, $(I_i)_{i=1,\ldots,j}$ represents available product images, and t represents the product listing. The output consists of segmentation masks $(V_i)_{i=1,\ldots,j}$ highlighting visual evidence or relevant context that would support an answer to the object question q:

$$q, t, (I_i)_{i=1,\ldots,j} \rightarrow (V_i)_{i=1,\ldots,j} \quad (1)$$

When a single image I is available, the problem reduces to $q, t, I \rightarrow V$, where V is a segmentation mask of visual evidence in image I. This formulation enables evaluation of existing multimodal models, which typically accept only one image as input. It also makes VGOQ similar to existing VQA grounding tasks, which follow the pattern $q, I \rightarrow V$.

When multiple images $(I_i)_{i=1,\ldots,j}$ are available for an object, we may first select the most relevant image for answering question q, then perform visual grounding only in this selected image. We approach this by computing relevance scores between the question q and each image I_i. For questions like "how comfortable are these earbuds?", we prioritize images that show the relevant visual features (*e.g.*, ear tips, cushioning materials).

4. Automated data generation for Visual Grounding of Object Questions

One main challenge of this new task is the lack of existing datasets for developing, training, and evaluating VGOQ models. We address this by designing two complementary automated data generation pipelines that create training data with different characteristics and coverage. VizWiz-VGOQ offers diverse real-world scenarios, while ABO-VGOQ provides structured e-commerce data with multiple images and rich metadata per product.

4.1. VizWiz-VGOQ: Rewriting Visual Questions into Object Questions

Data generation. Our first method transforms existing Visual Question grounding data to create VGOQ samples.

Figure 3. **Wordcloud** of the words in the questions in the VizWiz-VQA-Grounding dataset [8] (visual questions, left) and in our new VizWiz-VGOQ dataset (general object questions, right). Larger font indicates more frequent words. Blue indicates more concrete words, while red indicates more abstract words [31].

(1)	(2)	(3)	(4)	Evidential quality category	Train	Validation
✗	-	-	-	Cannot identify	25	4
✓	✗	-	-	Not valuable	111	14
✓	✓	✗	-	Related, no visual evidence	2212	390
✓	✓	✓	✗	Non-specific visual evidence	2700	411
✓	✓	✓	✓	Specific visual evidence	1446	312
✓	✓	-	-	Total (related to question)	6356	1113
-	-	-	-	Total (all)	6494	1131

Table 1. **Evidential quality classification logic (left) and number of samples in each category for the VizWiz-VGOQ training and validation sets (right)**. We ask Claude [1] the following 4 yes-no questions: (1) Can we identify which elements of the image are highlighted? (2) Do highlighted areas relate to the question? (3) Does the highlighting provide clear visual evidence for answering the question? (4) Does the highlighting focus on specific elements of the image? Based on these evaluations, we classify samples into categories from *not valuable* to *specific visual evidence*.

We use Claude [1] to rewrite visual questions from the VizWiz-VQA-Grounding dataset [8, 12] into general questions about objects, such that the existing segmentation masks contain useful evidence for answering the general question rather than being the direct answer. Figures 1 and 3 show this transformation from visual question answers to more natural object questions.

Evidential quality of the segmentation. We use Claude [1] to automatically categorize the segmentation based on its evidential relationship to the corresponding newly generated question. This allows us to evaluate visual grounding models on the different evidential quality categories. We classify samples into categories from *not valuable* to *specific visual evidence* according to Table 1.

4.2. ABO-VGOQ, and zero-shot pipeline for visual grounding of abstract customer queries in product imagery

Our second approach, which we illustrate in Figure 4, generates realistic customer questions and corresponding visual

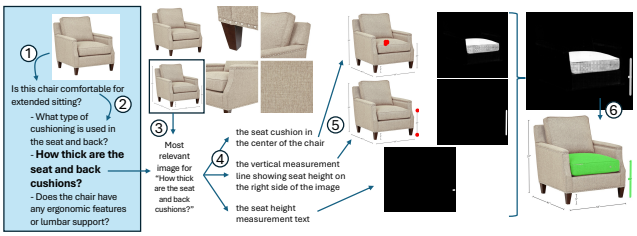


Figure 4. **Overview of our zero-shot pipeline for ABO-VGOQ generation**. Our six-step pipeline: (1) Generate initial questions using Claude [1], (2) Rewrite the initial question into concrete or abstract questions, for diversity, (3) Generate a draft answer and identify the most relevant image for grounding the answer, (4) Identify concrete visual elements of the image to ground the answer, (5) Combine Claude with Molmo 7B-D [10] (pointing), Florence-2 [36] (OCR), and SAM-2 [29] (points to segmentation map), for visual grounding of these concrete visual elements, (6) Combine outputs into final segmentation masks, highlighting visual context or evidence to accompany the answer.

(1)	(2)	(3)	(4)	Evidential quality category	Train	Val.	Test
✗	-	-	-	Cannot identify	36	5	3
✓	✗	-	-	Not valuable	100	19	13
✓	✓	✗	-	Related, no visual evidence	1848	362	208
✓	✓	✓	✗	Non-specific visual evidence	1015	201	113
✓	✓	✓	✓	Specific visual evidence	2205	434	203

Table 2. **Evidential quality classification logic (left) and number of samples in each category for the ABO-VGOQ training, validation, and test sets (right)**. See description in Table 1.

Split	Products	Questions	Visual Grounded Questions	Related to question (Claude/Human)	Specific visual evidence (Claude/Human)
Training	1000	6873	5204	5068/.	2205/.
Val.	200	1337	1025	997/977	434/346
Test	100	700	540	524/526	203/169
Total	1300	8910	6769	6589/6571	2842/2720

Table 3. **Statistics of our ABO-VGOQ generation**.

grounding for e-commerce products from product metadata (listing) and catalog images provided in the Amazon-Berkeley Objects dataset [9]. We sample 1300 products (1000 train, 200 validation, 100 test) and develop a six-step zero-shot pipeline, with human verification for the validation and test sets. Our pipeline generates 8910 object questions, creating 6769 visual groundings of object questions. More detailed dataset statistics are shown in Table 3.

Step 1: Initial question generation. For each product, we prompt Claude [1] to generate plausible customer questions, that might be asked at different stages of the shopping journey: before seeing the product, after viewing the product image, and after seeing both the image and product title. We also suggest various question categories (Table 1 in [18]). These approaches aim to simulate a range of customer information-seeking behaviors.

Step 2: Rewrite the initial question into abstract and concrete questions. We then employ Claude to rewrite the initial question into an abstract and one to three concrete questions. Abstract questions capture high-level customer concerns (*e.g.*, "Is it comfortable?", "How easy is it for...?"), while concrete questions focus on more specific product features.

Step 3: Relevance assessment, draft answer, and selection of most relevant image for grounding. For each question (both abstract and concrete versions), Claude computes relevance scores (0-1) for each product metadata field and image, identifying which elements are most pertinent to answering the customer's query. See examples of relevance scores for the images in Figure 8 of the supplementary material. Simultaneously, Claude generates a draft answer based on the most relevant product information and selects the most appropriate image for visual grounding.

Step 4: From question-answer to concrete visual elements of the image. For the selected most relevant image, Claude determines if the answer should be grounded and generates detailed descriptions of concrete visual elements where it should be grounded. These descriptions specify whether to target bounded areas of the image, points, lines, text regions, the whole image, etc, providing guidance for the subsequent grounding step.

Step 5: Visual grounding of the concrete visual elements. We combine multiple models to ground each concrete visual element: Molmo 7B-D [10] for pointing to relevant locations (*e.g.*, multiple points in an area of the image, extremities of lines, etc), Florence-2 [36] for optical character recognition of text regions, and SAM-2 [29] for converting points into segmentation masks. This multi-model approach handles diverse grounding requirements, covering many cases often encountered in product imagery where existing visual grounding models fail, *e.g.*, lines and text.

Step 6: Final processing, evidential quality assessment. We combine individual segmentations (one segmentation mask per concrete visual element) into a single final segmentation mask per customer question. We first apply evidential quality assessment with Claude (Table 2), as for VizWiz-VGOQ, and additional human validation for validation and test sets. We obtain human annotations through Amazon SageMaker GroundTruth. Each sample is annotated by 4 expert annotators who review Claude's assessments and provide corrections. We provided the annotation task batch by batch (1558 samples in total, times 4 annotators) to the 4 annotators over a span of 2 weeks. We reviewed the annotations, and regularly provided feedback on the samples where we disagreed through shared documents and online meetings. We use a consensus mechanism: if at least 3 annotators agree with Claude, we keep Claude's annotation; if at least 2 disagree, we flip the annotation; otherwise (when annotators select "Cannot determine") we default to "No ✗", *e.g.*, for the second criterion: highlighting is

not related to the question. For each of the 4 questions, the 4 annotators give the same answer between 79% (*Highlights specific elements?*) and 97% (*Can identify highlighted elements?*) of samples, and they agree with Claude's annotation between 84% (*Highlights specific elements?*) and 98% (*Can identify highlighted elements?* and *Related to question?*) of samples.

4.3. Datasets limitations

Automated generation and evaluation subjectivity. Both our datasets are automatically generated rather than manually created, lacking the quality and precision of human-created segmentation (for ABO-VGOQ) and naturally occurring question distribution. Visual evidence assessment is subjective, but that subjectivity is inherent to the needs of real applications. For instance, it is unclear if showing an image where the size is only conveyed relatively to other visual elements is acceptable as visual evidence for a question about the product size. We try to mitigate this subjectivity by assessing evidential quality with Claude and expert annotator consensus, but further work is needed to categorize question types and define what counts as valid evidence.

Data biases. For VizWiz-VGOQ, our backwards approach reuses segmentation masks originally designed for direct visual answers, potentially misrepresenting optimal visual evidence for general object questions. The transformed question distribution may not reflect natural user queries, and single real-world photos differ from e-commerce scenarios with multiple product images and metadata. For ABO-VGOQ, our automated pipeline may introduce biases from underlying models (Claude, Molmo, SAM-2, Florence-2). The focus on product imagery also limits generalization to other domains.

4.4. Skills required for VGOQ

Our VGOQ task requires several key capabilities, as illustrated in Figures 1 and 2, including: **Material recognition**: identifying materials from visual texture and appearance cues; **Spatial Reasoning**: understanding relative proportions and dimensional relationships for size and fit questions; **Inference from visual context**: connecting visible properties to functional characteristics (*e.g.*, texture of the shoe to water absorption / waterproof characteristics); **Multi-modal integration**: combining textual information from product listing and visual information from images effectively; **Infographics understanding**: reading text within images for specifications (*e.g.*, product size annotations) and instructions (*e.g.*, product label); **Context assessment**: determining whether to highlight specific elements or treat images as complete evidence, *e.g.*, if an image is already a closeup on a feature of interest.

5. Lightweight model for Visual Grounding of Object Questions

While our zero-shot data generation pipeline effectively creates visual grounding data, it involves multiple large models (Claude [1], Molmo [10], Florence-2 [36], SAM-2 [29]) and is computationally expensive for real-time deployment. For practical e-commerce applications that need to process millions of product queries efficiently, we train a lightweight model that can perform visual grounding directly without the multi-step pipeline overhead. The input consists of a single image and either: a referring expression describing what to segment (*e.g.*, RefCOCO/+/g [15], concrete visual elements from intermediate output of our zero-shot pipeline), a general question about objects in the image (*e.g.*, data from our two data generation techniques), or a visual question (*e.g.*, Toloka [33], TextVQA-X [27], VizWiz-VQA-Grounding [8]), visual-question-answer (*e.g.*, TextVQA-X [27], VizWiz-VQA-Grounding [8]), or a visual-question-visual-question-answer pair (*e.g.*, TextVQA-X [27], VizWiz-VQA-Grounding [8]). The output is a segmentation mask highlighting the relevant visual evidence or context.

Architecture. Our lightweight grounding model is based on CLIPSeg [22] and consists of three main components. The **vision encoder**, a frozen CLIP [26] vision transformer, extracts visual features at different layers, capturing low-level details and high-level semantic information, necessary for grounding abstract queries. The **text encoder**, frozen from CLIP, processes the textual input (object questions, visual questions, referring expression, etc.) and generates contextual text embeddings. The **grounding transformer** (1.77M trainable parameters) takes the visual features and text embeddings as input, and outputs (1st head) a segmentation heatmap (336×336) and (2nd head) a relevance score (how relevant the image is for this question). See the supplementary material, Section 9, for a diagram and more details about the architecture.

Training strategy. We train our lightweight model using a multi-task learning approach that combines several visual grounding tasks. We train jointly on traditional RES datasets (RefCOCO/+/g [15]), VQA grounding datasets (VizWiz-VQA-Grounding [8], TextVQA-X [27]), and our two new VGOQ datasets (VizWiz-VGOQ and ABO-VGOQ). We use Dice and binary cross-entropy losses for training. Our training data includes:

- VizWiz-VGOQ: 6,356 triplets of image, general question, and corresponding segmentation.
- ABO-VGOQ: 5,068 triplets of image, general question, and corresponding segmentation. This also includes 10,713 triplets of image, description of a concrete image

element, and corresponding segmentation (intermediate output from our zero-shot data generation pipeline).

- VizWiz-VQA-Grounding: 6,494 triplets of image, visual question and answer, and corresponding segmentation.
- TextVQA-X: 14,476 triplets of image, visual question and answer, and corresponding segmentation.
- RefCOCO/+/g: 120,624 + 120,191 + 80,512 triplets of image, referring expression, and corresponding segmentation.

The model is trained for 10,000 steps with a batch size of 8 and learning rate of 0.001, corresponding to just under one epoch over the largest dataset (RefCOCO) and multiple epochs for smaller datasets. For VizWiz-VGOQ and ABO-VGOQ, we use all samples categorized as 'related to the question' for this training. We use different FiLM [25] conditioning for each task type: referring expressions, concrete visual elements, visual questions, visual question answers, visual question-answer pairs, and general object questions. This unified training enables the model to handle diverse grounding scenarios within a single architecture. Finetuning only on samples categorized as 'specific visual evidence' can further improve performance for specific visual evidence visual grounding task (+1.7 to +7.4 percentage points of gIoU, see Table 5 of the supplementary material).

6. Results

For evaluation, we use the evaluation sets of our synthetic datasets (ABO-VGOQ, VizWiz-VGOQ), as well as existing visual grounding datasets (VizWiz-VQA-Grounding [8], TextVQA-X [27], and Toloka [33]). Our evaluation covers Visual Grounding for Object Questions (VGOQ) with single image input (image + object question $\rightarrow$ segmentation), VQA grounding (image + visual question + answer $\rightarrow$ segmentation), and VQ grounding (image + visual question $\rightarrow$ segmentation). We evaluate our lightweight model, prior visual grounding models, namely GLaMM [28] (FullScope and RefSeg variants), UnifiedIO [21] (Small, Base, Large, XL variants), and OFA [34], as well as a contemporary / concurrent work of ours, Qwen3-VL [3]. Since these models are vision-language LLMs, we follow the original implementations to prompt these models for visual grounding (details in supplementary material, Section 10).

Qualitative results. Qualitative results on ABO-VGOQ validation set are shown in Figure 5. Our lightweight model (LW) produces more accurate visual grounding compared to other models. Still, there remains a gap between our lightweight model's output and the ground truth generated by our zero-shot pipeline, indicating room for further improvement in training strategies, architectural design, and potentially dataset size. See also Figure 7 in supplementary material.

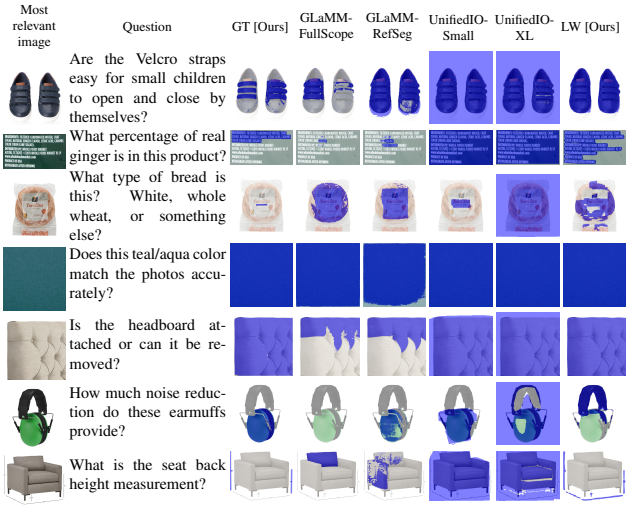


Figure 5. **Qualitative evaluation on ABO-VGOQ.** We compare visual grounding results across diverse product questions requiring different reasoning skills: functionality assessment (Velcro straps), ingredient analysis (ginger content), material identification (bread type), color matching, structural features (headboard attachment), performance specifications (noise reduction), and dimensional measurements. Our lightweight model (LW) produces more accurate and focused grounding compared to existing methods, though a gap remains with our zero-shot pipeline ground truth (GT). Existing methods often highlight irrelevant regions or fail to identify specific visual evidence needed to answer the questions. See also Figure 2 for more context on these 7 samples.

Quantitative results. In Table 4, we report gIoU $\pm$ standard error where an Intersection over Union (IoU) score is computed for each sample, and the IoU scores are averaged. For reference of quantitative metrics, we also report results of a simple baseline that directly segment the whole image **uniform**ly. We also report (in parentheses) the ratio of the gIoU of the method with the gIoU of this reference.

Discussion. We notice a performance drop from traditional VQ/VQA grounding to VGOQ. For instance, on the 312 samples categorized as "Specific visual evidence" in our ABO-VGOQ validation set, the performance of SoTA models drops from 52.2% gIoU when the input is a Visual Question, to 37.2% gIoU when the input is an Object Question. This drop highlights the novelty and difficulty of our proposed problem, and demonstrates that existing models struggle to identify visual evidence rather than direct answers. Our lightweight model achieves competitive or superior performance compared to much larger models, despite being much smaller (1.77M parameters), much faster, and trained for only 10,000 steps. The results also suggest that models do not generalize well to data and tasks that differ from their training data, emphasizing the importance of our work in filling these data/tasks gaps. Surpris-

	task:	— VQA grounding —		——— VQ grounding ———				——— Visual Grounding for Object Questions (VGOQ) ———					
		Visible answer to the question						Specific visual evidence (SVE)			Context related to the question (RTQ)		
	data:	TextVQA-X	VizWiz-VQA-Ground.	Toloka	TextVQA-X	VizWiz-VQA-Ground.	VizWiz-VQA-Ground.	VizWiz-VGOQ	ABO-VGOQ		VizWiz-VGOQ	ABO-VGOQ	
Model	Params.	val ($n = 3620$)	val ($n = 1131$)	test-public ($n = 1705$)	val ($n = 3620$)	val ($n = 1131$)	val ($n = 312$)	val-SVE ($n = 312$)	val-SVE ($n = 346$)	test-SVE ($n = 169$)	val-RTQ ($n = 1113$)	val-RTQ ($n = 977$)	test-RTQ ($n = 526$)
Uniform	0	3.2 ± 0.1	33.4 ± 1.0	4.3 ± 0.2	3.2 ± 0.1	33.4 ± 1.0	15.6 ± 1.1	15.6 ± 1.1	12.9 ± 0.9	15.1 ± 1.4	33.4 ± 1.0	30.5 ± 1.0	34.4 ± 1.4
Qwen3-VL-8B-Instruct [3]	8B	41.1 ± 0.5 ($\times 12.9$)	59.5 ± 0.9 ($\times 1.8$)	$\mathbf{71.4 \pm 0.8}$ ($\times \mathbf{16.6}$)	$\mathbf{38.5 \pm 0.5}$ ($\times \mathbf{12.0}$)	57.4 ± 0.9 ($\times 1.7$)	47.0 ± 1.7 ($\times 3.0$)	36.0 ± 1.7 ($\times 2.3$)	30.3 ± 1.4 ($\times 2.4$)	30.3 ± 2.2 ($\times 2.0$)	49.4 ± 1.0 ($\times 1.5$)	43.9 ± 1.0 ($\times 1.4$)	44.8 ± 1.5 ($\times 1.3$)
Qwen3-VL-4B-Instruct [3]	4B	38.6 ± 0.5 ($\times 12.1$)	59.0 ± 0.9 ($\times 1.8$)	69.7 ± 0.8 ($\times 16.3$)	36.9 ± 0.5 ($\times 11.5$)	57.2 ± 0.9 ($\times 1.7$)	47.8 ± 1.7 ($\times 3.1$)	36.9 ± 1.7 ($\times 2.4$)	27.5 ± 1.5 ($\times 2.1$)	$\mathbf{32.7 \pm 2.3}$ ($\times \mathbf{2.2}$)	48.9 ± 1.0 ($\times 1.5$)	41.4 ± 1.1 ($\times 1.4$)	45.4 ± 1.5 ($\times 1.3$)
Qwen3-VL-2B-Instruct [3]	2B	36.2 ± 0.5 ($\times 11.3$)	55.4 ± 0.9 ($\times 1.7$)	64.4 ± 0.8 ($\times 15.0$)	32.7 ± 0.5 ($\times 10.2$)	53.3 ± 1.0 ($\times 1.6$)	39.8 ± 1.7 ($\times 2.5$)	34.9 ± 1.7 ($\times 2.2$)	25.8 ± 1.3 ($\times 2.0$)	26.6 ± 2.1 ($\times 1.8$)	49.4 ± 1.0 ($\times 1.5$)	42.5 ± 1.0 ($\times 1.4$)	43.8 ± 1.4 ($\times 1.3$)
GLaMM-FullScope [28]	7B	8.5 ± 0.2 ($\times 2.7$)	55.4 ± 1.2 ($\times 1.7$)	24.0 ± 0.7 ($\times 5.6$)	10.9 ± 0.3 ($\times 3.4$)	48.9 ± 1.1 ($\times 1.5$)	30.2 ± 1.8 ($\times 1.9$)	28.1 ± 1.7 ($\times 1.8$)	20.2 ± 1.5 ($\times 1.6$)	22.3 ± 2.1 ($\times 1.5$)	50.3 ± 1.2 ($\times 1.5$)	44.8 ± 1.2 ($\times 1.5$)	47.0 ± 1.6 ($\times 1.4$)
GLaMM-RefSeg [28]	7B	9.2 ± 0.2 ($\times 2.9$)	51.1 ± 1.1 ($\times 1.5$)	15.5 ± 0.6 ($\times 3.6$)	9.6 ± 0.2 ($\times 3.0$)	45.2 ± 1.1 ($\times 1.4$)	29.6 ± 1.7 ($\times 1.9$)	26.0 ± 1.7 ($\times 1.7$)	19.4 ± 1.5 ($\times 1.5$)	20.0 ± 2.0 ($\times 1.3$)	39.0 ± 1.1 ($\times 1.2$)	37.9 ± 1.1 ($\times 1.2$)	37.8 ± 1.5 ($\times 1.1$)
UnifiedIO-XL [21]	3B	5.7 ± 0.2 ($\times 1.8$)	65.0 ± 1.1 ($\times 1.9$)	3.7 ± 0.3 ($\times 0.9$)	7.1 ± 0.3 ($\times 2.3$)	$\mathbf{68.1 \pm 1.0}$ ($\times \mathbf{2.0}$)	$\mathbf{52.2 \pm 1.9}$ ($\times \mathbf{3.3}$)	37.2 ± 1.8 ($\times 2.4$)	12.4 ± 0.9 ($\times 1.0$)	12.3 ± 1.4 ($\times 0.8$)	57.4 ± 1.1 ($\times 1.7$)	30.2 ± 1.0 ($\times 1.0$)	32.7 ± 1.4 ($\times 1.0$)
UnifiedIO-Large [21]	776M	8.7 ± 0.3 ($\times 2.7$)	54.2 ± 1.1 ($\times 1.6$)	5.3 ± 0.4 ($\times 1.2$)	8.6 ± 0.3 ($\times 2.7$)	62.5 ± 1.1 ($\times 1.9$)	47.1 ± 1.9 ($\times 3.0$)	34.5 ± 1.8 ($\times 2.2$)	13.5 ± 1.1 ($\times 1.1$)	15.6 ± 1.7 ($\times 1.0$)	53.0 ± 1.2 ($\times 1.6$)	34.8 ± 1.1 ($\times 1.1$)	38.0 ± 1.5 ($\times 1.1$)
UnifiedIO-Base [21]	241M	5.9 ± 0.2 ($\times 1.8$)	51.0 ± 1.1 ($\times 1.5$)	4.2 ± 0.3 ($\times 1.0$)	7.1 ± 0.3 ($\times 2.2$)	58.0 ± 1.1 ($\times 1.7$)	38.0 ± 1.8 ($\times 2.4$)	29.7 ± 1.6 ($\times 1.9$)	13.6 ± 1.1 ($\times 1.1$)	12.9 ± 1.4 ($\times 0.9$)	47.5 ± 1.1 ($\times 1.4$)	33.1 ± 1.1 ($\times 1.1$)	34.8 ± 1.5 ($\times 1.0$)
UnifiedIO-Small [21]	71M	5.7 ± 0.2 ($\times 1.8$)	41.0 ± 1.1 ($\times 1.2$)	4.4 ± 0.4 ($\times 1.0$)	4.9 ± 0.2 ($\times 1.6$)	48.2 ± 1.1 ($\times 1.4$)	29.2 ± 1.7 ($\times 1.9$)	22.6 ± 1.4 ($\times 1.4$)	13.6 ± 1.1 ($\times 1.1$)	12.3 ± 1.4 ($\times 0.8$)	39.0 ± 1.1 ($\times 1.2$)	30.9 ± 1.1 ($\times 1.0$)	31.6 ± 1.5 ($\times 0.9$)
OFA-Huge [34]	930M	3.6 ± 0.2 ($\times 1.1$)	30.1 ± 1.0 ($\times 0.9$)	4.4 ± 0.2 ($\times 1.0$)	3.3 ± 0.2 ($\times 1.0$)	30.6 ± 1.0 ($\times 0.9$)	13.6 ± 1.1 ($\times 0.9$)	14.3 ± 1.1 ($\times 0.9$)	11.9 ± 0.9 ($\times 0.9$)	13.4 ± 1.4 ($\times 0.9$)	30.3 ± 1.0 ($\times 0.9$)	25.4 ± 1.0 ($\times 0.8$)	28.8 ± 1.4 ($\times 0.8$)
OFA-Large [34]	470M	17.5 ± 0.4 ($\times 5.5$)	32.8 ± 1.0 ($\times 1.0$)	17.9 ± 0.7 ($\times 4.2$)	14.6 ± 0.4 ($\times 4.6$)	27.2 ± 1.0 ($\times 0.8$)	17.0 ± 1.4 ($\times 1.1$)	16.5 ± 1.4 ($\times 1.1$)	17.9 ± 1.1 ($\times 1.4$)	16.5 ± 1.6 ($\times 1.1$)	31.8 ± 1.0 ($\times 1.0$)	27.0 ± 0.9 ($\times 0.9$)	28.9 ± 1.3 ($\times 0.8$)
OFA-Base [34]	180M	10.5 ± 0.4 ($\times 3.3$)	35.9 ± 1.1 ($\times 1.1$)	20.1 ± 0.8 ($\times 4.7$)	10.2 ± 0.3 ($\times 3.2$)	35.0 ± 1.0 ($\times 1.0$)	21.3 ± 1.5 ($\times 1.4$)	19.9 ± 1.5 ($\times 1.3$)	15.5 ± 1.1 ($\times 1.2$)	16.1 ± 1.6 ($\times 1.1$)	35.4 ± 1.1 ($\times 1.1$)	33.3 ± 1.0 ($\times 1.1$)	35.2 ± 1.4 ($\times 1.0$)
OFA-Medium [34]	93M	8.2 ± 0.3 ($\times 2.6$)	34.2 ± 1.0 ($\times 1.0$)	13.3 ± 0.6 ($\times 3.1$)	8.5 ± 0.3 ($\times 2.7$)	32.9 ± 1.0 ($\times 1.0$)	16.8 ± 1.3 ($\times 1.1$)	15.1 ± 1.2 ($\times 1.0$)	15.4 ± 1.0 ($\times 1.2$)	15.0 ± 1.5 ($\times 1.0$)	33.2 ± 1.0 ($\times 1.0$)	33.2 ± 1.0 ($\times 1.1$)	34.9 ± 1.4 ($\times 1.0$)
OFA-Tiny [34]	33M	4.0 ± 0.2 ($\times 1.3$)	34.8 ± 1.0 ($\times 1.0$)	4.8 ± 0.2 ($\times 1.1$)	4.1 ± 0.2 ($\times 1.3$)	34.1 ± 1.0 ($\times 1.0$)	16.4 ± 1.1 ($\times 1.0$)	16.2 ± 1.1 ($\times 1.0$)	14.2 ± 0.9 ($\times 1.1$)	15.6 ± 1.4 ($\times 1.0$)	34.7 ± 1.0 ($\times 1.0$)	34.8 ± 1.0 ($\times 1.1$)	36.9 ± 1.4 ($\times 1.1$)
Our LW	1.77M	$\mathbf{38.6 \pm 0.5}$ ($\times \mathbf{12.2}$)	$\mathbf{68.7 \pm 1.0}$ ($\times \mathbf{2.1}$)	12.6 ± 0.5 ($\times 2.9$)	34.9 ± 0.5 ($\times 11.0$)	67.2 ± 1.0 ($\times 2.0$)	51.5 ± 1.8 ($\times 3.3$)	$\mathbf{47.0 \pm 1.8}$ ($\times \mathbf{3.0}$)	$\mathbf{39.5 \pm 1.5}$ ($\times \mathbf{3.1}$)	32.5 ± 1.9 ($\times 2.1$)	$\mathbf{64.1 \pm 1.0}$ ($\times \mathbf{1.9}$)	$\mathbf{56.0 \pm 1.0}$ ($\times \mathbf{1.8}$)	$\mathbf{56.1 \pm 1.5}$ ($\times \mathbf{1.6}$)

Table 4. Evaluation (gIoU $\pm$ standard error, higher is better) on VQA grounding, VQ grounding, and VGOQ benchmarks. Our lightweight model (1.77M parameters) outperforms most others on the VGOQ tasks. The performance gap demonstrates that prior visual grounding models (GLaMM, UnifiedIO, OFA) struggle with identifying specific visual evidence or related context for general object questions, validating the need for specialized approaches like ours.

ingly, our model also outperforms all UnifiedIO variants but UnifiedIO-XL on VizWiz VQ and VQA grounding, even though UnifiedIO was trained on this data, demonstrating the effectiveness of our training approach.

7. Application of our work

VGOQ has direct practical applications in e-commerce stores.

Enhanced shopping assistants: Modern AI shopping assistants lack visual grounding capabilities—they cannot show customers which aspects of product images support their answers.

Automatic infographics generation: VGOQ can be used to identify when product images lack visual information needed to answer common customer questions, and create additional infographics imagery, *e.g.*, using information extracted from product listing.

Vendor Feedback Systems: VGOQ can be used to alert sellers when their image sets fail to address frequent customer queries, enabling them to improve their product presentations.

Beyond e-commerce, VGOQ has a lot of other practical implications. To mention a few, "Does this ad convey the desired product features?" (for marketing), "Does this component show wear?" (highlighting defect regions), "Does this item look fresh?" (highlighting spoiled area), "Does this car appear well-maintained?" (highlighting poorly maintained area), etc.

8. Conclusion

We introduced Visual Grounding for Object Questions (VGOQ), a task that extends visual grounding to handle general object questions. While traditional visual grounding focuses on locating directly visible answers, VGOQ requires identifying visual evidence or context that supports answering general questions about objects. Our approach includes two automated data generation pipelines that enable training and evaluating models for this task. We developed a lightweight grounding model that, despite its small size, shows promising results on our benchmarks, though significant challenges remain. The performance gap with prior models on VGOQ tasks (*e.g.*, on ABO-VGOQ 'Specific Visual Evidence': 32.5-39.5% for our model and 25.8-32.7% for Qwen3-VL, vs 12.3%-22.3% gIoU for prior visual grounding models) suggests that prior visual grounding models do not generalize well to tasks requiring identification of visual evidence rather than direct answers.

Our work opens new directions for practical visual grounding applications, particularly in e-commerce where understanding abstract customer queries is crucial. Future work should focus on developing more sophisticated models that can better understand the relationship between abstract concepts and visual evidence, and expanding the approach to handle more complex multi-modal reasoning scenarios. Categorizing question types and defining what counts as valid evidence (including direct visual evidence vs visual evidence with external knowledge) is also an important direction for future work.

Acknowledgments

We thank the Amazon Machine Learning Data Operations team for data annotation support, and Ganesh Iyer for assistance with tools and early ideation.

References

[1] Anthropic. Introducing the next generation of Claude, 2024. https://www.anthropic.com/news/claude-3-family. 2, 3, 4, 5, 6

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2015)*, pages 2425–2433, 2015. 3

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. 2, 7, 8

[4] Ankan Bansal, Yuting Zhang, and Rama Chellappa. Visual Question Answering on Image Sets. In *European Conference on Computer Vision (ECCV 2020)*, pages 51–67. Springer, 2020. 3

[5] Anand Bhattad, Konpat Preechakul, and Alexei A Efros. Visual Jenga: Discovering Object Dependencies via Counterfactual Inpainting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. 3

[6] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding Everything: Emerging Localization Properties in Vision-Language Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pages 3828–3837, 2024. 3

[7] Shaofei Cai, Zihao Wang, Kewei Lian, Zhancun Mu, Xiaojian Ma, Anji Liu, and Yitao Liang. ROCKET-1: Mastering Open-World Interaction with Visual-Temporal Context Prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, pages 12122–12131, 2025. 3

[8] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding Answers for Visual Questions Asked by Visually Impaired People. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 19098–19107, 2022. 1, 2, 3, 4, 6, 7

[9] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and Benchmarks for Real-World 3D Object Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 21126–21136, 2022. 2, 5

[10] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, pages 91–104, 2025. 2, 3, 5, 6

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 6904–6913, 2017. 3

[12] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 3608–3617, 2018. 2, 3, 4

[13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[14] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. DINOv2 Meets Text: A Unified Framework for Image- and Pixel-Level Vision-Language Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, pages 24905–24916, 2025. 3

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, 2014. 2, 3, 6

[16] Seyedalireza Khoshsirat and Chandra Kambhamettu. Sentence Attention Blocks for Answer Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023)*, pages 6080–6090, 2023. 4

[17] Seyedalireza Khoshsirat and Chandra Kambhamettu. Embedding Attention Blocks For Answer Grounding. In *2024 IEEE International Conference on Image Processing (ICIP 2024)*, pages 521–527. IEEE, 2024. 4

[18] Ashish Kulkarni, Kartik Mehta, Shweta Garg, Vidit Bansal, Nikhil Rasiwasia, and Srinivasan Sengamedu. ProductQnA: Answering User Questions on E-Commerce Product Pages. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 354–360, 2019. 5

[19] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pages 9579–9589, 2024. 3

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV 2014)*, pages 740–755. Springer, 2014. 3

[21] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. UNIFIED-IO: A Unified

Model for Vision, Language, and Multi-modal Tasks. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023. 3, 7, 8

[22] Timo Lüddecke and Alexander Ecker. Image Segmentation Using Text and Image Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 7086–7096, 2022. 2, 3, 4, 6, 11

[23] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pages 3195–3204, 2019. 3

[24] OpenAI. GPT-4V(ision) System Card, 2023. https://openai.com/index/gpt-4v-system-card/. 3

[25] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI conference on Artificial Intelligence*, 2018. 7, 11

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML 2021)*, pages 8748–8763. PmLR, 2021. 2, 4, 6, 11

[27] Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian, and Mingwei Shen. A First Look: Towards Explainable TextVQA Models via Visual and Textual Explanations. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 19–29, 2021. 2, 3, 6, 7

[28] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. GLaMM: Pixel Grounding Large Multimodal Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pages 13009–13018, 2024. 3, 7, 8

[29] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment Anything in Images and Videos. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. 2, 3, 5, 6

[30] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pages 8317–8326, 2019. 3

[31] Davide Talon, Federico Girella, Ziyue Liu, Marco Cristani, and Yiming Wang. Seeing the Abstract: Translating the Abstract Language for Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, pages 9253–9262, 2025. 3, 4

[32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[33] Dmitry Ustalov, Nikita Pavlichenko, Sergey Koshelev, Daniil Likhobaba, and Alisa Smirnova. Toloka Visual Question Answering Benchmark. *arXiv preprint arXiv:2309.16511*, 2023. 2, 6, 7

[34] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *International Conference on Machine Learning (ICML 2022)*, pages 23318–23340. PMLR, 2022. 7, 8

[35] Ruiqi Wang and Hao Zhang. RESAnything: Attribute Prompting for Arbitrary Referring Segmentation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. 3

[36] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pages 4818–4829, 2024. 2, 5, 6

[37] Rongtao Xu, Jian Zhang, Minghao Guo, Youpeng Wen, Haoting Yang, Min Lin, Jianzheng Huang, Zhe Li, Kaidong Zhang, Liqiong Wang, et al. A0: An Affordance-Aware Hierarchical Model for General Robotic Manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2025)*, pages 13491–13501, 2025. 3

[38] Jihee Yoon, Seunga Lee, Haesol Jeong, and Junseok Kwon. Lightweight Grounding Model Combining a CLIP-based Encoder With an Upsampling Decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025) Workshops*, Nashville, USA, 2025. 4

9. Architecture of the lightweight model

Our lightweight grounding model is based on CLIPSeg [22]. We modify the architecture to handle diverse query types (object questions, visual questions, referring expressions, etc) via task-specific FiLM conditioning. We also augment the decoder with a second output head that predicts an image-level relevance score, enabling the model to assess whether the image is pertinent to the input object question (see Section 11.2).

Vision encoder: We use a pre-trained CLIP [26] vision transformer (ViT-L/14-336px) that processes images at 336×336 resolution. We extract visual features (resolution 24×24) from layers 6, 14, and 18 to capture both low-level details and high-level semantic information, necessary for grounding abstract queries. We freeze the pre-trained weights to preserve learned semantic representations.

Text encoder: We use the pre-trained text encoder from CLIP, to process the textual input (object questions, visual questions, referring expression, etc) and generate contextual text embeddings. We freeze the pre-trained weights to preserve learned semantic representations.

Grounding transformer: The decoder takes CLIP visual features and text embeddings as input. It consists of (1) input projection layers, that reduce the 1024-dimensional CLIP visual features to a compact 64-dimensional representation, (2) a 3-layer transformer with width 64, that processes the visual features in a U-Net-like fashion, (3) task-specific FiLM conditioning [25] for different query types (object questions, visual questions, referring expression, etc), to modulate the transformer input to the desired task, and (4) two output heads, that take as input the transformer output (resolution 24×24) and give (1^{st} head) a segmentation heatmap (336×336) and (2^{nd} head) a relevance score (how relevant the image is for this question). The grounding transformer contains 1.77M trainable parameters in total.

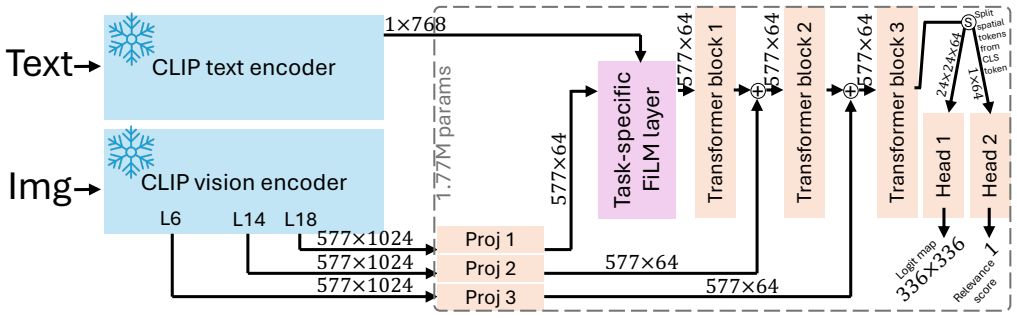


Figure 6. **Architecture of the lightweight model**

10. Evaluation

Because the models expect only one image as input, we provide, for ABO-VGOQ, as input the most relevant image selected by our data generation pipeline.

GLaMM models are prompted with "`Q: {question} Please identify and segment the relevant area for answering this question in the image.`" for VQ grounding and VGOQ and "`Q: {question} A: {answer} Please identify and segment the relevant area for answering this question in the image.`" for VQA grounding.

UnifiedIO models are prompted directly with questions for VQ grounding and VGOQ, and with "`Q: {question} A: {answer}.`" for VQA grounding. We use the flag `generate_image=True` to obtain a segmentation mask image as output.

OFA models are prompted with "`Which region does the text " the relevant area for answering this question: {question} " describe?`" for VQ grounding and VGOQ, and "`Which region does the text " the relevant area for answering this question: {question} and answer: {answer} " describe?`" for VQA grounding.

Qwen3-VL models are prompted with "`Q: question Please identify and locate the relevant area for answering this question in the image. Report bbox coordinates in JSON format.`" for VQ grounding and VGOQ, and "`Q: question A: answer Please identify and`

`locate the relevant area for answering this question in the image. Report bbox coordinates in JSON format.`" for VQA grounding.

For ChatGPT and Gemini (Figure 7), we prompt the model by defining what a segmentation mask is, asking to "`identify and segment the relevant area for answering the question`" and "`Generate the image (binary, black and white) of the segmentation mask`". Inference can fail or generate segmentation results that are misaligned with the original image (*e.g.*, different aspect ratio).

For all models, when inference fails (*e.g.*, output text not in the expected correct JSON format), the sample is excluded from the gIoU computation.

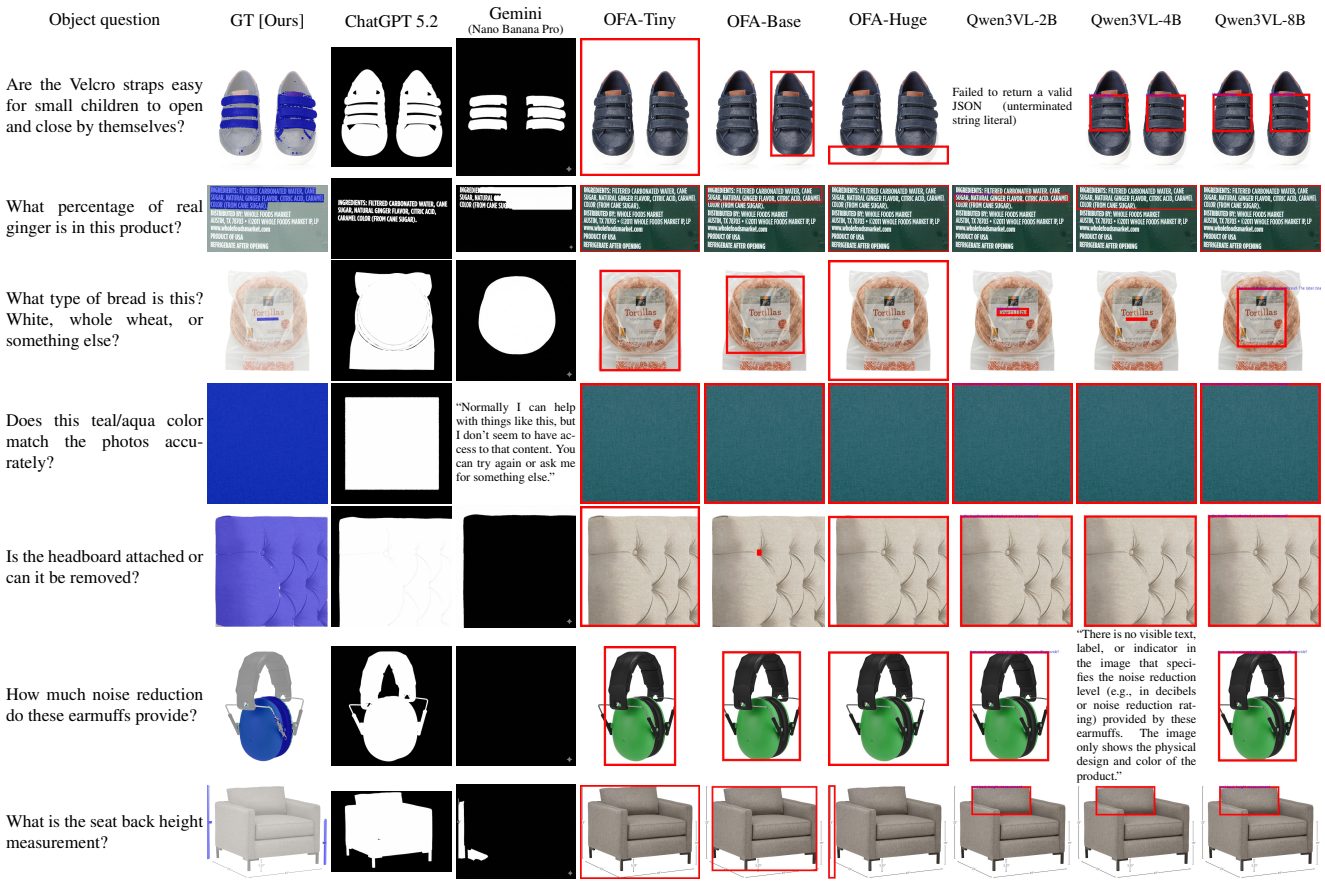


Figure 7. **Comparison of additional models on VGOQ**.

11. Ablation studies

11.1. Fine-tuning on "specific visual evidence" only

This section provides additional analysis for the model training approach described in Section 5 of the main paper. We conduct an ablation study to evaluate the impact of training exclusively on data categorized as "specific visual evidence" (SVE) versus the broader category of "related to the question" (RTQ), as defined in Tables 1, 2, and 3 of the main paper. The "specific visual evidence" category contains more localized annotations where the highlighted regions contain visual evidence that can be used to answer the question, while the broader "related to the question" category also includes regions that are contextually related.

Data characteristics. The size distribution of masks in the "specific visual evidence" category shows a mean coverage of 12.9% of the image area for ABO-VGOQ (15.6% for VizWiz-VGOQ), compared to 30.5% for the broader "related to the

question" category for ABO-VGOQ (33.4% for VizWiz-VGOQ). This reflects the more targeted nature of specific visual evidence annotations, which focus on precise object parts or features rather than entire objects.

Setup of the ablation study. We train a specialized version of our lightweight (LW) model on samples annotated as "specific visual evidence" (SVE) of the VizWiz-VGOQ and ABO-VGOQ training sets, *i.e.* 1,446 + 2,205 samples. The specialized model uses the same architecture but is initialized from the version described in the main document section 5 (trained on multiple tasks) and fine-tuned for 10 epochs on this data only.

Results. Table 5 shows the comparative results between the generic and specialized lightweight VGOQ models. When training only on specific visual evidence data, we observe improved performance on the specific visual evidence visual grounding task: +7.4 percentage points of gIoU on VizWiz-VGOQ SVE validation set and +3.3 percentage points of gIoU on ABO-VGOQ SVE validation set. This demonstrates that the model learns more precise localization when trained on higher-quality annotations. However, this specialization comes at the cost of reduced performance on the broader "related to question" categories (-2.0 to -3.6 percentage points of gIoU), indicating the model becomes more conservative and misses some relevant contextual information. This trade-off has practical implications for downstream applications: while the specialized model provides more precise visual evidence identification, the broader model is more suitable for applications requiring comprehensive contextual visual grounding (*e.g.*, infographics generation where related areas should be highlighted even when direct visual evidence is absent).

	– Specific visual evidence (SVE) –			Context related to the question (RTQ)		
	VizWiz-VGOQ	ABO-VGOQ		VizWiz-VGOQ	ABO-VGOQ	
	val-SVE	val-SVE	test-SVE	val-RTQ	val-RTQ	test-RTQ
Model	($n = 312$)	($n = 346$)	($n = 169$)	($n = 1113$)	($n = 977$)	($n = 526$)
Uniform	15.6 ± 1.1	12.9 ± 0.9	15.1 ± 1.4	33.4 ± 1.0	30.5 ± 1.0	34.4 ± 1.4
Our LW	47.0 ± 1.8 ($\times 3.0$)	39.5 ± 1.5 ($\times 3.1$)	32.5 ± 1.9 ($\times 2.1$)	$\mathbf{64.1 \pm 1.0}$ ($\times \mathbf{1.9}$)	$\mathbf{56.0 \pm 1.0}$ ($\times \mathbf{1.8}$)	$\mathbf{56.1 \pm 1.5}$ ($\times \mathbf{1.6}$)
Our LW SVE	$\mathbf{54.4 \pm 1.8}$ ($\times \mathbf{3.5}$)	$\mathbf{42.8 \pm 1.6}$ ($\times \mathbf{3.3}$)	$\mathbf{34.2 \pm 2.2}$ ($\times \mathbf{2.3}$)	62.1 ± 1.0 ($\times 1.9$)	53.4 ± 1.1 ($\times 1.8$)	52.5 ± 1.5 ($\times 1.5$)

Table 5. Evaluation (gIoU $\pm$ standard error, higher is better $\uparrow$) on Visual Grounding for Object Questions (VGOQ). The "Our LW SVE" model, fine-tuned only on "specific visual evidence" samples, achieves superior performance on specific visual evidence tasks (+1.7 to +7.4 pp gIoU) but shows reduced performance on broader contextual grounding (-2.0 to -3.6 pp gIoU).

11.2. Predicting relevance scores

This section provides additional details on the relevance score prediction capability mentioned in Section 5 of the main paper. Our model architecture extends the visual grounding framework to predict the relevance score of an image given a question. The architecture uses a dual-head approach where the transformer encoder outputs are processed through two separate pathways:

- **Spatial grounding head:** The spatial tokens (dimension $24 \times 24 \times 64$) are processed through transposed convolutions to generate pixel-level heatmaps.
- **Relevance score head:** The CLS token from the grounding transformer output (dimension 1×64) is processed through a three-layer MLP with ReLU activations and sigmoid output to produce a relevance score between 0 and 1. This output represents how relevant the entire image is to the given question.

The relevance prediction task uses the sum of L1 and MSE losses to train the model to match Claude-generated relevance scores (step 3 of our zero-shot pipeline described in Section 4.2 of the main document), which are our ground-truth data for this task. We evaluate against a baseline that outputs a constant relevance score of 0.4657, the average relevance score from the ABO-VGOQ training set (samples with relevance scores: 981 products, 5344 images, 6808 questions, 37510 relevance scores).

Results. Our lightweight model achieves a Mean Absolute Error (MAE) of 0.17 and Root Mean Square Error (RMSE) of 0.23 on the ABO-VGOQ validation set (samples with relevance scores: 195 products, 1070 images, 1336 questions, 7387 relevance scores), significantly outperforming the constant baseline (MAE: 0.25, RMSE: 0.28). The model shows strong correlation (Pearson $r = 0.62$) with Claude-generated ground-truth (step 3 of our zero-shot pipeline), demonstrating its

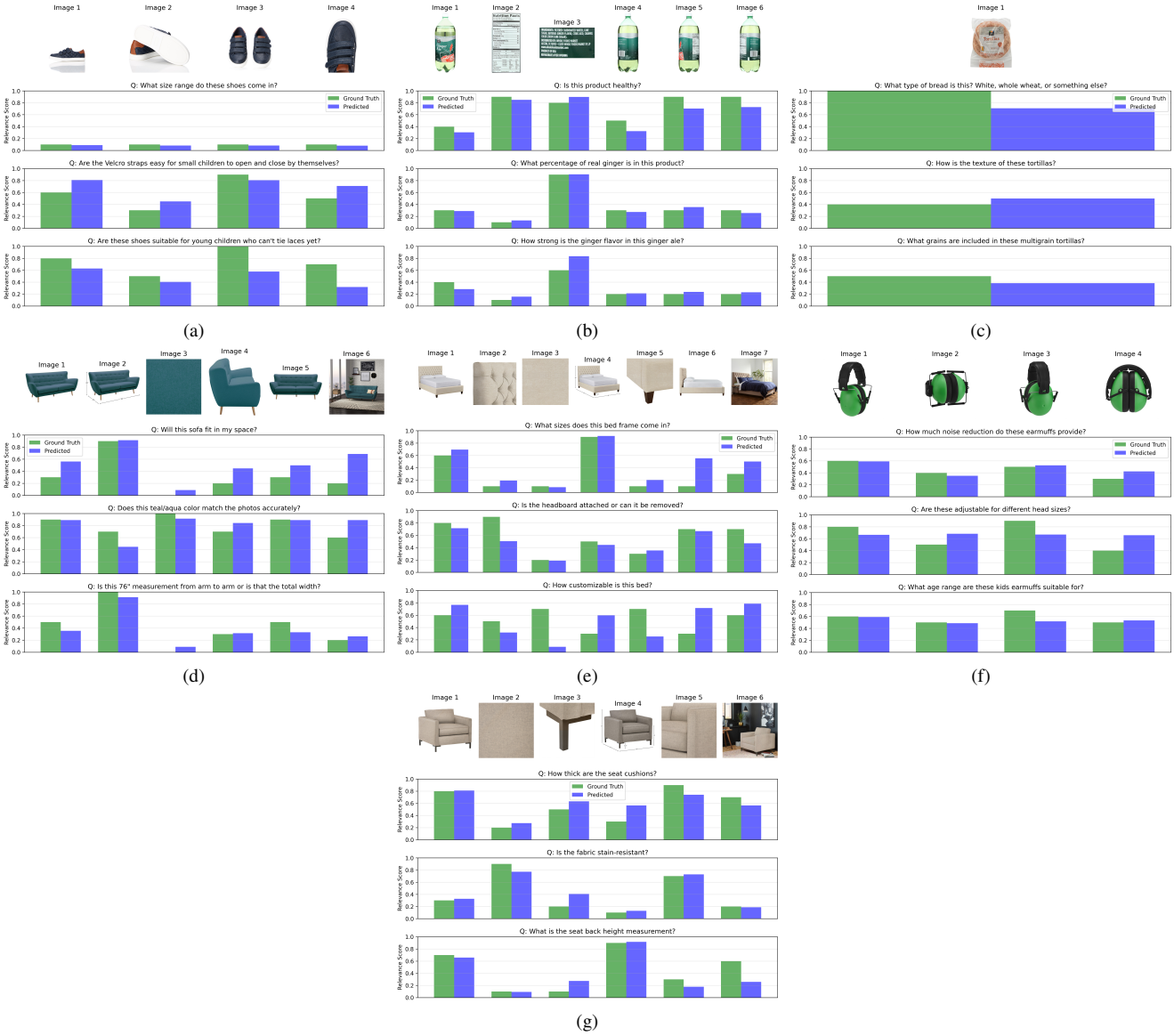


Figure 8. Relevance score predictions on ABO-VGOQ validation set. For each product, we show multiple images (top row) and three corresponding generated questions (steps 1 and 2 of our data generation pipeline) below. Bar charts compare our lightweight model's predicted relevance scores (blue) with ground truth annotations from Claude (green, step 3 of our zero-shot data generation pipeline). The lightweight model successfully captures varying degrees of relevance across images, with higher scores for images containing visual evidence or context relevant to answering the question. See also Figure 2 of the main document for more context on these 7 samples.

ability to capture nuanced relevance relationships between images and questions. Figure 8 shows qualitative results across different product categories.

12. Annotation tool for ABO-VGOQ validation and test sets

This section provides details on the evidential quality annotation process used for the validation and test sets for ABO-VGOQ, as mentioned in Section 4.2 of the main paper. The annotation tool was designed to correct the evidential quality annotations of Claude.

Annotation framework. Expert annotators evaluate AI-generated visual grounding through a 4-question sequential framework:

- **Q1, Technical quality:** "Can you identify the main highlighted regions in the AI-generated highlighting, even if there are minor technical imperfections?" This filters out samples with severe technical issues like missing highlights, phantom elements, or completely distorted masks.
- **Q2, Specificity assessment:** "Does the highlighting focus on specific region(s)/element(s) or highlight almost the whole visible part of the product?" This categorizes the granularity of the visual grounding.
- **Q3, Relevance to the question:** "Is there any relation between the highlighted area and the customer question?" This ensures the highlighted regions relate to the question topic, even if they don't provide useful information.
- **Q4, Visual evidence:** "Does the highlighted area show visual evidence that can be used to provide a response to the customer question?" This evaluates whether the highlighting contains specifications, features, or attributes useful for answering the question.

The annotation interface shows AI-generated answers and explanations for each question, allowing annotators to quickly accept correct assessments or focus their attention on cases where the AI reasoning is incorrect.

Annotation interface. See Figures 9-10 for screenshots of the annotation interface. The tool allows annotators to view multiple product images simultaneously, read the generated question, and verify or modify the automatically generated segmentation masks through this structured evaluation process.

Rivet Modern Ceiling Fan With LED Light

Customer question:

Does this fan provide lighting functionality?

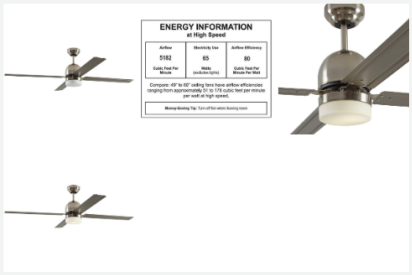
Product Image:

(Click on the image to enlarge it)

AI-generated highlighting:

(Click on the image to enlarge it)

Image gallery:



(Click on the image to enlarge it)

About this item:

- Integrated LED light included
- 53 x 53 x 16 Inches, Brushed Nickel
- Remote control included
- For indoor use only
- Professional installation recommended

Q1. Can you identify the main highlighted regions in the AI-generated highlighting, even if there are minor technical imperfections?

AI answer: Yes (90%)

AI explanation for yes: The highlighting clearly identifies the light fixture/lamp at the bottom center of the ceiling fan. This is the cylindrical frosted glass or plastic light cover that houses the integrated LED light mentioned in the product description.

Examples: (see more in the SOP)
- Highlight the whole product, with imperfections → Answer should be "Yes"
- A green line on top of the product that does not exist in the original image → Answer should be "No"

Do you agree with the AI answer? If not, what is the correct answer?

○ Agree ○ ~~Disagree, "Yes"~~ ○ Disagree, "No" ○ Cannot determine / Skipping

Q2. Inside the selected image, does the highlighting tend to highlight almost the whole visible part of the product, or does it focus on specific region(s)/element(s) of the image?

AI answer: Specific region(s)/element(s)

Examples: (see more in the SOP)
- Highlight almost the whole visible part of the product, with imperfections → Answer should be "Whole visible part of the product"
- Highlight just a few regions, leaving other visible parts of the product unhighlighted → Answer should be "Specific region(s)/element(s)"

Do you agree with the AI answer? If not, what is the correct answer?

○ Agree ○ ~~Disagree, "specific region(s)/element(s)"~~ ○ Disagree, "Whole visible part of the product" ○ Cannot determine / Skipping

Q3. Is there any relation between the highlighted area (even if it does not help answering at all) and the customer question or the part(s) of the product mentioned in the customer question?

AI answer: Yes (100%)

AI explanation for yes: The highlighting directly relates to the customer question by focusing on the light fixture component of the ceiling fan. Since the customer is asking specifically about lighting functionality, highlighting the actual light fixture is perfectly relevant and shows that the system understands what part of the product the question is referring to.

Examples: (see more in the SOP)
- "Is the bike reliable?" → Highlighting the whole bike → Answer should be "Yes" (even if it does not show about reliability)
- "Are the bike handles non-slip?" → Highlighting the handles → Answer should be "Yes" (even if we cannot see slipperiness)
- "Are the bike handles non-slip?" → Highlight the front wheel → Answer should be "No"

Do you agree with the AI answer? If not, what is the correct answer?

○ Agree ○ ~~Disagree, "Yes"~~ ○ Disagree, "No" ○ Cannot determine / Skipping

Q4. Does the highlighted area show the answer or visual evidence (specifications/features/attributes of the product) that can be used to provide a response to the customer question?

AI answer: Yes (100%)

AI explanation for yes: The highlighting provides clear visual evidence that answers the customer's question by focusing on the light fixture component of the ceiling fan. The highlighted area shows a frosted white cylindrical light cover attached to the bottom of the fan, which is unmistakably a light fixture. This visual evidence directly confirms that yes, this ceiling fan does indeed provide lighting functionality, as also stated in the product information ('Integrated LED light included').

Examples: (see more in the SOP)
- "Will it fit in my car?" → Highlighting the annotated dimensions around the product → Answer should be "Yes" (shows the product specification needed to provide a response to the customer question)
- "Are these earbuds comfortable?" → Highlighting the silicone ear-tips of the earbuds (contributing to comfortability). → Answer should be "Yes" (shows the product specification needed to provide a response to the customer question)
- "Is it waterproof?" → Highlighting material type → Answer should be "Yes" if material type is easy to associate it to waterproof or not ; Answer should be "No" if highlighted material type does not help answering

Do you agree with the AI answer? If not, what is the correct answer?

○ Agree ○ ~~Disagree, "Yes"~~ ○ Disagree, "No" ○ Cannot determine / Skipping

If necessary, you can provide any additional comments here:

If necessary, you can provide any additional comments here..

Figure 9. Screenshots from the annotation tool (1/2)

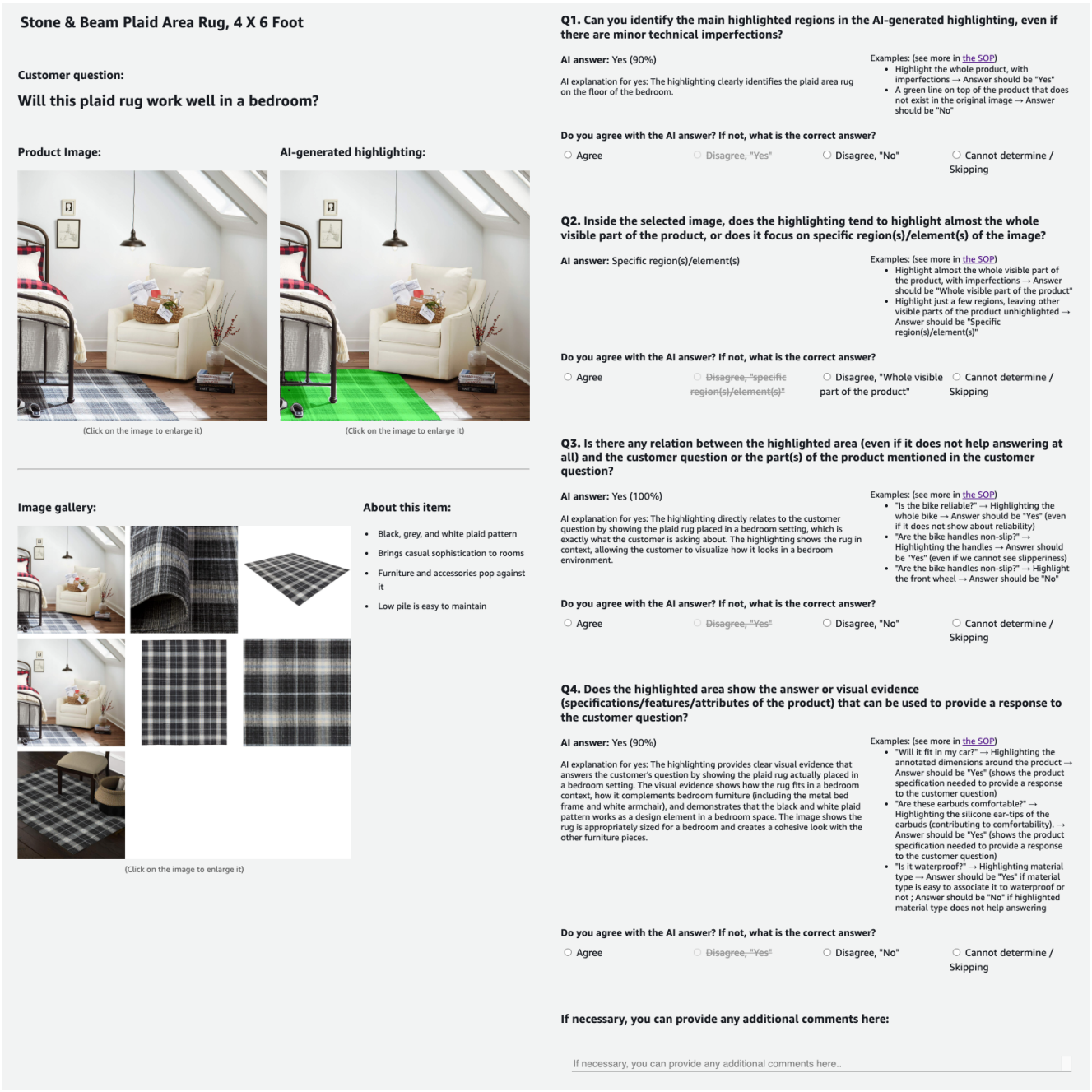


Figure 10. Screenshots from the annotation tool (2/2)

13. Prompt for creating the VizWiz-VGOQ dataset

This section provides the prompt used to create the VizWiz-VGOQ dataset, as described in Section 4.1 of the main document. Our prompt is composed of 2 parts. The first part describes the expected task, and the second part provides the current input (VQA data, and necessary images).

```
You are an AI assistant specialized in analyzing Question Answering data, where the data consists of images,
     questions, and segmentation masks.
Your task is to transform existing data containing Visual Questions (where the question can be directly answered and
     the segmentation mask is the answer to the question) into General Questions (where the segmentation mask
     contains useful evidence to provide a reply to the general question).

INPUT FORMAT:
- Original image
- Original Visual Question (directly about visible elements of the image)
- Answer to the Original Visual Question (text)
- Answer to the Original Visual Question (segmentation mask)

IMAGE DESCRIPTIONS:
Image 1: The original image, without highlighting
Image 2: The same image with the green highlighting
  - Visible in original colors: The original image.
  - Green transparent overlay: Highlighted area (segmentation mask).
Image 3: The same image where the non-highlighted area has been blacked out.
  - Visible in original colors: Highlighted area (segmentation mask).
  - Masked in black: Regions that are not highlighted (not part of the segmentation mask).

ANALYSIS STEPS:
Follow these steps:
  - What is the main object of the image?
  - What specifications/features of the object are visible in the segmentation mask (i.e. highlighted area)?
  - Think about general questions for which these specifications would be useful for answering.
  - Output the new General Question.
The new General Question must meet these criteria:
  - The new question must be about the object (eg, its use, properties, suitability, etc), not about the image itself.
  - The segmentation mask (highlighted area) must show relevant features/specifications of the object to provide a
       reply to the question
  - It might require combining information from highlighted area with general knowledge to provide a reply

EXAMPLES:
1. Original: "What is this?" -> "dog"
   (Highlighted area shows a dog with rather long hair)
   New: "Would this pet need regular grooming?"
   (Hair length is useful for knowing about regular grooming)
2. Original: "What color is this shirt?" -> "purple"
   (Highlighted area shows the purple shirt)
   New: "Would this show stains easily?"
   (Color information useful for stain visibility)
3. Original: "How many tablets in box?" -> "8"
   (Highlighted area shows the label 8 Tablets per box)
   New: "Is this enough for a two-week treatment?"
   (Count information useful but needs dosage knowledge)
4. Original: "What does label say?" -> "USB-C"
   (Highlighted area shows the label with the text USB-C)
   New: "Can this charge the latest smartphones?"
   (Port type useful but needs compatibility knowledge)

OUTPUT FORMAT:
Provide your output using YAML format. Use double-quoted strings for all fields.
main_object_thinking: "Think about what is the main object shown in the image"
main_object: "Name the main object"
segmentation_mask_analysis: "Look at the Answer to the Original Visual Question (text+segmentation mask). Think about
     what specifications/features of the objects are visible in the segmentation mask (i.e. highlighted area)"
specifications:
  - "First specification/feature of the object that is visible in the segmentation mask (i.e. highlighted area)."
  - "Second specification/feature of the object that is visible in the segmentation mask (i.e. highlighted area)."
  - "Third specification/feature of the object that is visible in the segmentation mask (i.e. highlighted area)."
  ...
new_question_thinking: "Think about a general question for which these specifications would be useful for answering.
     Look again at the criteria the new question must meet"
new_question: "Output the new General Question. If no suitable question can be generated where the highlighting would
     be useful, output 'No suitable question possible'"
new_answer: "Output a response to the new General Question, that uses information from the segmentation mask."
```

Listing 1. Prompt (part 1) for creating the VizWiz-VGOQ dataset

```
Original image:
Image 1

Original Visual Question (directly about visible elements of the image):
{visual_question}

Answer to the Original Visual Question:
Text: {visual_question_answer}
Segmentation mask: Images 2 and 3

Provide your output in YAML format. Be sure to output a valid YAML output. Do not use quotes when quoting the text
     shown in the image, or as an abbreviation of inch, so that it is valid YAML format.
```

Listing 2. Prompt (part 2) for creating the VizWiz-VGOQ dataset