Visual Grounding for Object Questions (Supplementary Material)

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

9. Architecture of the lightweight model

Our lightweight grounding model is based on CLIPSeg [1]. 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 [3] 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 [2] 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.

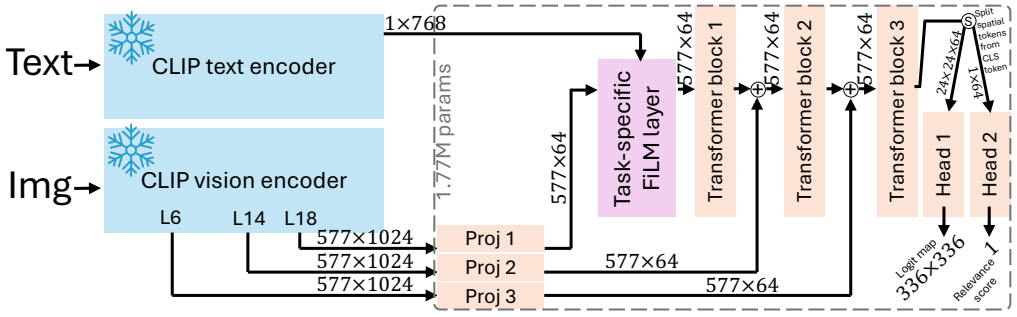


Figure 6. **Architecture of the lightweight model**

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.

*Work done during an internship at Amazon.

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.

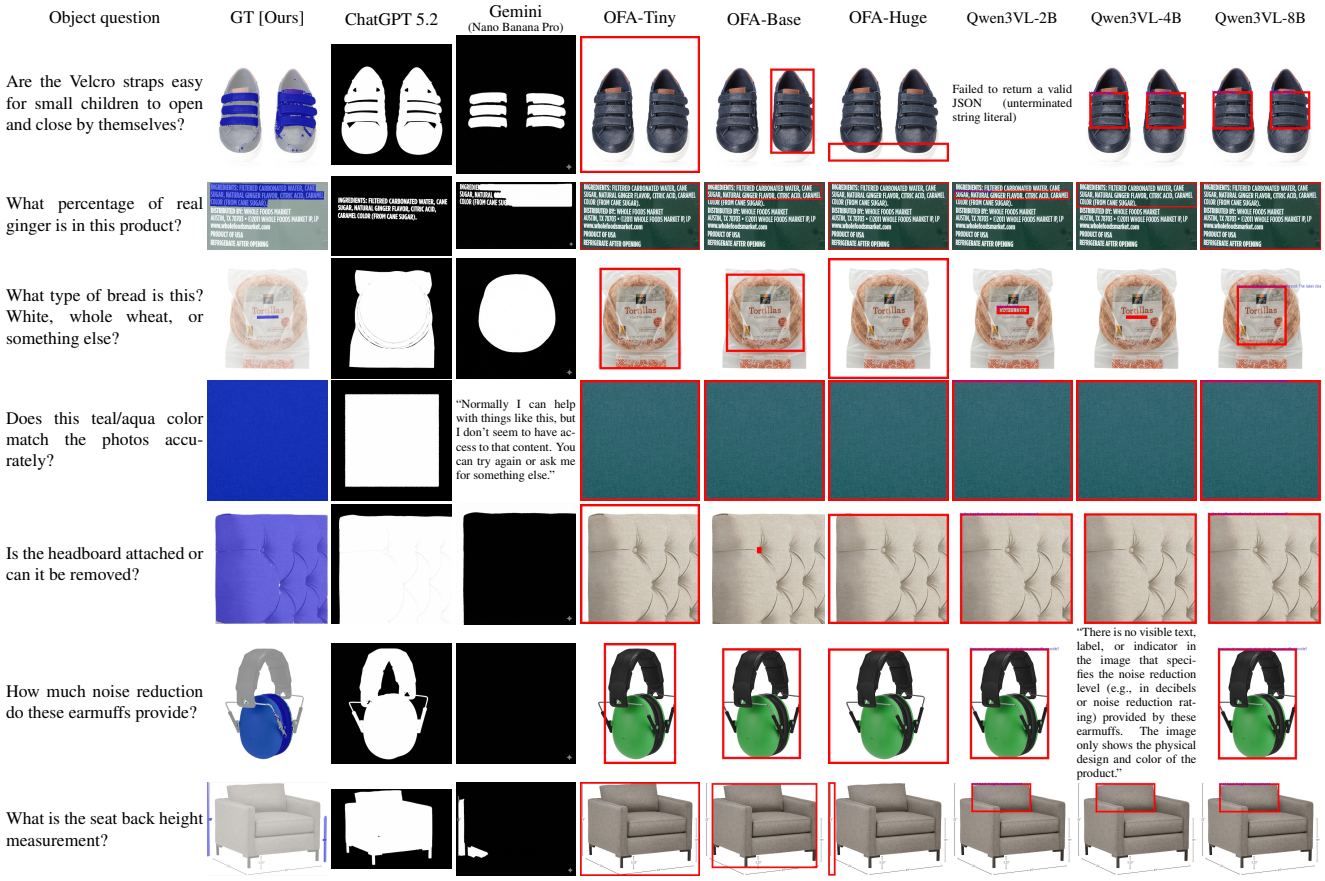


Figure 7. **Comparison of additional models on VGOQ**.

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.

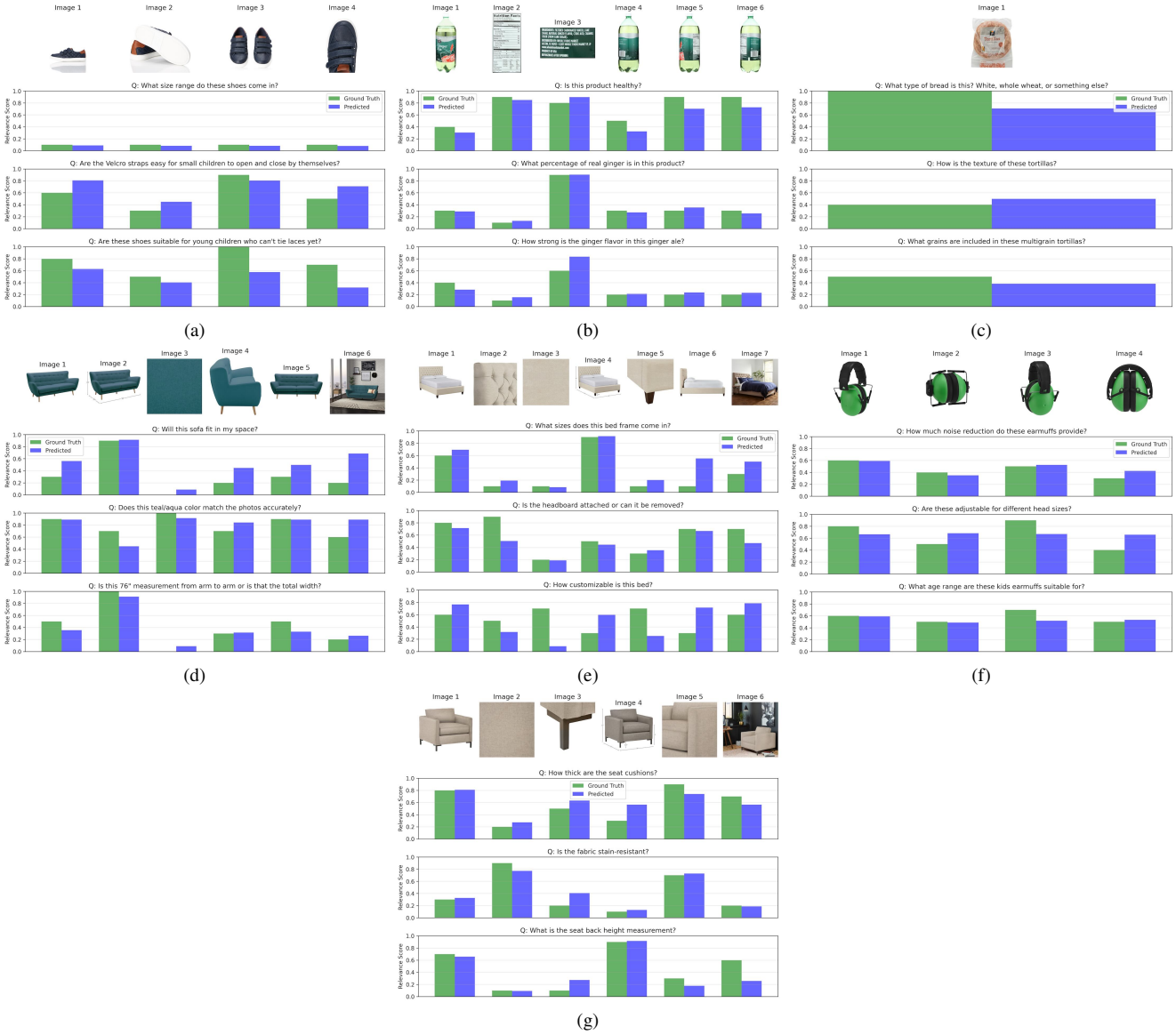


Figure 8. Relevance score predictions on ABO-VGOQ validation set. For each product, we show multiple images (top row) and three corresponding generated questions (steps 1 and 2 of our data generation pipeline) below. Bar charts compare our lightweight model's predicted relevance scores (blue) with ground 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.

- **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 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?

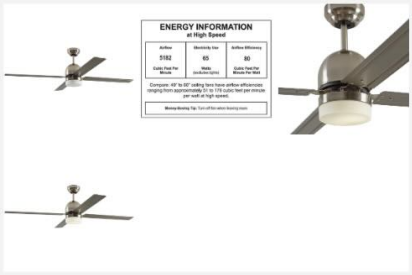
Product Image:

(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)

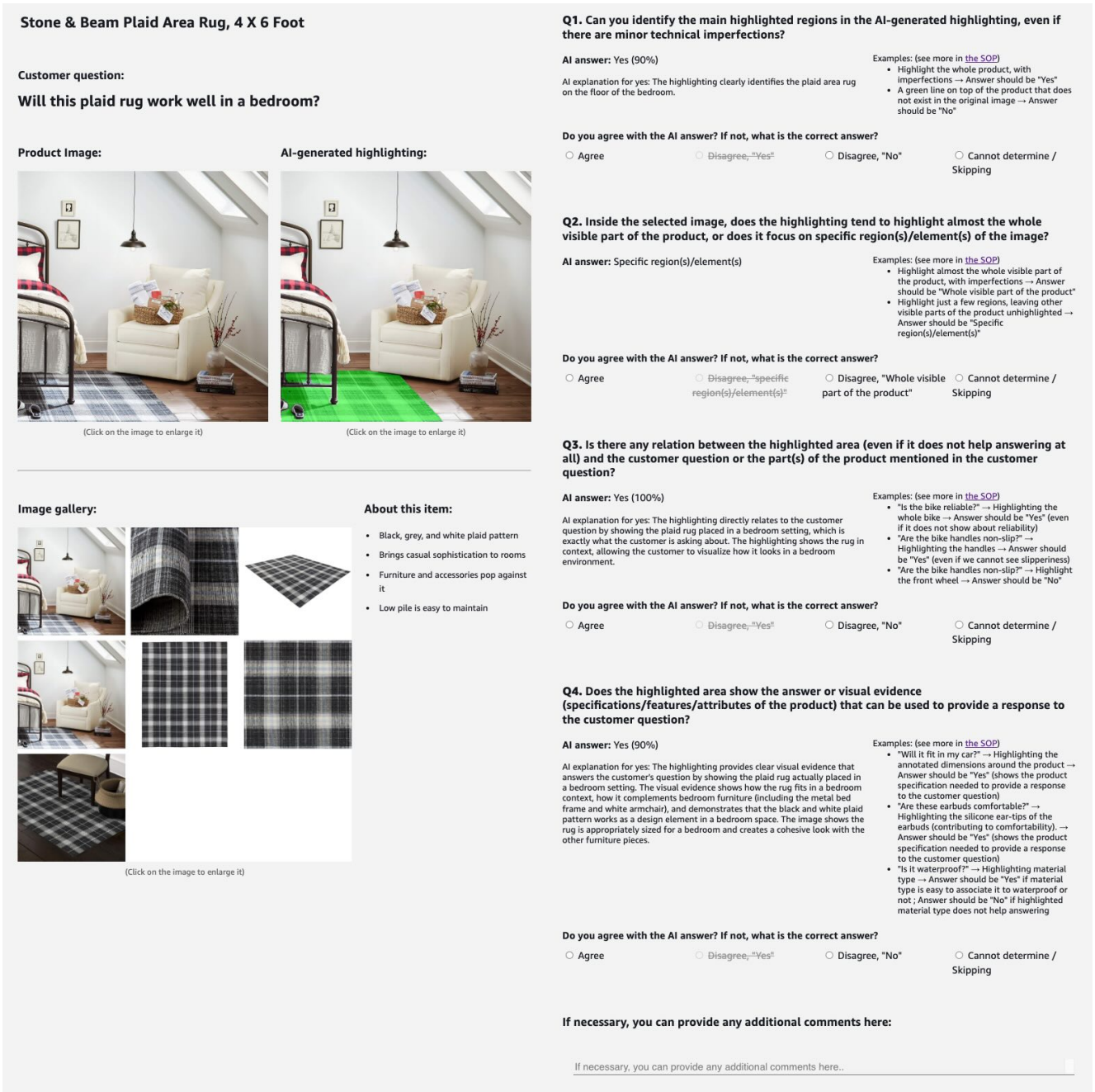


Figure 10. Screenshots from the annotation tool (2/2)

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).

References

S1 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. 1

S2 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. 1

S3 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. 1

```
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