

# RESAnything: Attribute Prompting for Arbitrary Referring Segmentation

## Supplementary Material

The supplementary document provides (1) detailed analysis of limitation of current methods, including both MLLM and CLIP in our task, in Section 6; (2) comprehensive details of language and visual prompts used in RESAnything in Section 7; (3) additional information about the construction of ABO-Image-ARES in Section 8; (4) extended quantitative results on part level and multi-object GRES task and qualitative results, including failure cases in Section 9 and 10, respectively.

### 6. Limitation of Current Methods

Our method leverages Chain-of-Thought (CoT) attribute prompting for detailed descriptions and combines MLLMs and CLIP as mask selector to select optimal segmentation proposals. While this dual-model approach achieves strong performance, it arises from the inherent limitations of both components. In this section, we analyze the constraints of current MLLMs and CLIP that motivate our design choices in attribute prompting and the hybrid evaluation strategy.

#### 6.1. Limitation of MLLM

**Attribute Prompt.** While MLLMs exhibit strong reasoning capabilities, they often fail to perform systematic CoT reasoning without explicit prompting guidance. As shown in Fig 9, when asked to describe the details of input expression  $E$  without specific attribute requirements, MLLMs typically generate oversimplified descriptions that fail to capture the target’s essential characteristics and details effectively. Therefore, providing MLLMs with explicit attribute requirements is essential to guide their reasoning process effectively. RESAnything leverages this insight to generate more comprehensive and accurate descriptions, ensuring that all necessary details of the target expression are properly captured.

**Binary Response.** As mentioned in our main paper, a naive approach for applying MLLMs to solve our task would involve prompting the MLLMs to output a score for each segmentation proposal  $m_i$ , denoting its similarity with the input expression  $E$ . However, MLLMs are primarily designed to understand and generate text rather than compute precise numerical similarities. While they excel at comparing and reasoning about content qualitatively, they struggle to produce reliable numerical similarity scores. Our experiments reveal that MLLM-generated similarity scores exhibit high variance and poor correlation with actual contextual similarity, as the model essentially samples from its probability distribution rather than performing true sim-

ilarity computation. Therefore, we reformulate similarity assessment as binary classification queries, returning yes or no in our selection algorithm, which better aligns with MLLMs’ natural language understanding capabilities. As shown in Fig 10, our experiments reveal that MLLMs tend to generate similarity scores that appear arbitrary or biased by their training distribution, rather than computing true similarities between the given elements, and their binary responses prove to be more reliable indicators.

#### 6.2. Limitation of CLIP

The limitations of CLIP in analyzing contextual similarities become evident when dealing with complex descriptions and image content. As shown in Fig 11, while CLIP’s text-to-text similarity scores reveal meaningful comparison, they often fail to capture crucial contextual details like color attributes. Additionally, CLIP’s text-to-image similarity scores show limited discriminative power, consistently remaining below 0.3. These limitations underscore our decision to adopt MLLMs as our primary mask selector, as they demonstrate superior capability in understanding and comparing detailed contextual content.

#### 6.3. Ablation Study

We further evaluate the effectiveness of adopting both MLLM and CLIP as mask selectors in RESAnything. Table 6 compares the performance of RESAnything on ReasonSeg test set with different mask selectors configurations. Using CLIP as the sole mask selector results in poor performance due to its previously mentioned limitations in understanding complex relationships and abstract concepts. While MLLM demonstrates superior reasoning and contextual similarity capabilities compared to CLIP, using MLLM alone can lead to incomplete region selection, particularly for expressions targeting multiple parts (e.g., sofa legs or armrests). These results validate our design choice of incorporating both MLLM and CLIP as mask selectors to ensure robust region selection.

Table 6. Ablation study on different mask selectors.

Method	test	
	gIoU	cIoU
CLIP only	42.5	38.4
LLM only	70.5	64.6
both	74.6	72.5

## 7. Prompts

### 7.1. Language Prompts in Attribute Prompting

As mentioned in the main paper, we use reference text prompt  $Q^{\text{ref}}$  to generate reference text  $T^{\text{ref}}$  for each refer based on the input expression  $E$ . Given the input image  $I$  and referring expression  $E$ , we prompt the MLLM using following  $Q^{\text{ref}}$  to obtain reference text  $T^{\text{ref}}$ :

For the region described as  $\{E\}$  in the image, provide a single detailed sentence describing an object or part of a object by including its location, appearance (color, shape, location), and distinctive characteristics including relevant actions, state, or function. Focus on features that would help uniquely identify this specific region from others in the image. Be as succinct as possible and in English only.

Similarly, given the mask cropped  $V_i^m$  and bounding box image  $V_i^b$  as visual prompts of a segmentation proposal  $m_i$  we prompt the MLLM using following  $Q^{\text{can}}$  to obtain candidate text  $T^{\text{can}}$ :

You are presented with two complementary views of the same region: 1) A cropped masked view showing detailed visual properties; 2) A full view with a bounding box showing location and context. Generate a single detailed sentence following these guidelines:

FOR COMPLETE OBJECTS:

- Combine visual details and spatial context naturally;
- Visual properties (color, shape, texture, size);
- Location in the scene;
- Relationships with surroundings;
- State or action if relevant;

FOR PARTIAL REGIONS:

Describe the part while providing clear context:

- Part identification and its visual properties;
- Its position within the larger object/scene;
- Relevant contextual details;

Important Rules: Start directly with the subject: 'A [description]...' or 'The [description]...';

Describe only what is visible in the non-black regions for visual properties and the image with green bounding box is for location and relation analysis;

Never mention masks, boxes, or annotations;

Use confident language for clear identifications;

Use tentative language when inferring;

Create natural, flowing descriptions that

combine all information seamlessly;

Focus on creating cohesive descriptions that feel natural and informative without drawing attention to the source of the information.

We adjust the  $Q^{\text{can}}$  based on different visual prompts for ablation study, e.g. mask cropped  $V_i^m$  only: You are presented with a cropped masked view showing detailed visual properties; ...

Fig 12 shows examples of query (input expression) and generated reference & candidate text.

### 7.2. Language Prompts in Grouping and Selection

We employ MLLM as one of the mask selectors in our grouping and selection algorithm. Certainly, for text-to-text decision  $d^{\text{t2t}}$ , we use following  $Q^{\text{t2t}}$ :

You are evaluating if the following candidate text describes the input expression region:  $E$ . Reference information provided for context if the input expression text is not clear:  $T^{\text{ref}}$ . Here is the candidate text to evaluate:  $T^{\text{can}}$ . Evaluate if the candidate text refer to the target by checking:

- Spatial location match;
  - Visual characteristics match (color, shape, size);
  - Object/subject identity match;
  - State/action consistency (if applicable).
- Return 'yes' or 'no' ONLY: 'yes' if most aspects substantially match; 'no' if some significant aspect differs.

For text-to-image decision  $d^{\text{t2i}}$ , we use following  $Q^{\text{t2i}}$ :

You are evaluating if the following reference text describes the non-black region of the cropped mask image:  $T^{\text{ref}}$ . The target is  $E$  for context if the reference text is inaccurate. You have two images for context: 1) A cropped mask image showing a region in non-black color; 2) An image with a green bounding box surrounding the region showing the full scene and spatial relationships. Evaluate if the reference text describes the non-black region of the cropped mask image by checking:

- Spatial location match (the location is relative location, not absolute location);
- Visual characteristics match (color, shape, size)
- Object/subject identity match (the masked image could be only a part of the target);
- State/action consistency (if applicable).



Return 'yes' or 'no' ONLY: 'yes' if most aspects substantially match; 'no' if some significant aspect differs.

### 7.3. Visual Prompts Selection

We explore five visual prompts  $V_i$  in our method: (1) original image, (2) mask-cropped image, (3) bounding box overlaid on image, (4) mask contour overlaid on image and (5) blur background overlaid on image. We choose the combination of mask-cropped image and bounding box overlaid on image as the best visual prompts  $V_i$  to obtain candidate text  $T^{\text{can}}$ . Apart from quantitative results presented in the ablation study, we further analyze the effectiveness and limitation of different individual/combinations of these visual prompts, as shown in Fig 13:

- mask cropped only: with mask cropped as the only visual prompt, MLLM is usually failed to infer the action/relation of the region. Example in Fig 13 shows that from mask cropped image, MLLM generates incorrect description of the region regarding its location and action.
- blur only: similar to mask cropped only, using blurred background as the sole visual prompt creates challenges for MLLM in distinguishing boundaries between blurred and clear regions, resulting in inaccurate location identification. Critical action-related details may also be obscured by blurring, leading to incorrect classification of object activities.
- original image with mask-cropped: while adding the original image helps MLLM better understand location and relationships, the lack of explicit region guidance causes MLLM to be distracted by irrelevant regions outside the mask cropped area.
- mask cropped with mask contour overlay: adding contour helps MLLM focus on the target region's boundaries, but the choice of overlay color can inadvertently influence MLLM's perception of the region's visual attributes. Attempts to show contours without color overlay (Fig 14) often result in ambiguous or confusing visual prompts, particularly for intricate shapes or overlapping regions if the contour is a non-convex shape.
- bounding box with mask contour overlay: while both elements help localize the target region, their overlay colors can affect MLLM's understanding. Even when explicitly prompted to focus on either the bounding box or contour region, both colors influence MLLM's perception of visual attributes, leading to inconsistent descriptions.
- bounding box with mask-cropped (RESAnything): This combination achieves the best balance - the bounding box provides spatial context and relationship guidance, while the mask-cropped image offers detailed visual attributes without color interference. By instructing MLLM to focus on the mask-cropped region while using the bounding box for context, we avoid noise from overlay colors while

maintaining accurate spatial understanding.

## 8. ABO-Image-ARES Data Preparation

### 8.1. Image Data

Our dataset builds upon image data from ABO [13], a dataset collected from worldwide Amazon.com product listings, including their metadata, images, and 3D models. ABO encompasses 147,702 product listings across 576 product types from various Amazon-owned stores and websites (e.g., Amazon, PrimeNow, WholeFoods). Each listing is uniquely identified by an item ID and contains structured metadata from its public webpage, including product specifications such as type, material, color, and dimensions, along with associated media. The dataset contains 398,212 high-resolution catalog images in total. However, to better highlight product properties, we excluded images from 11 categories: phone-related items (phone accessories, cellular phone cases, cellular phones, phones, wireless locked phones), footwear (shoes, shoe inserts, technical sport shoes, boots, sandals), and picture frames. Most images from these categories have no meaningful or interesting groundable/referrable parts, as shown in Fig 15. We also selected only the main image of each product, as additional images often show material details or close-up views. As results, ABO-Image-ARES contains 2,482 high-resolution catalog images spanning 565 product types.

### 8.2. Referring Expression Generation

The referring expressions in ABO-Image-ARES were derived from product metadata, specifically the bulletpoint descriptions that accompany each product listing in ABO. These bulletpoints typically contain detailed information about product features, materials, and functionalities. We processed these descriptions through MLLM, instructing it to generate 2-3 referring expressions per product. Prompt for instruction is following:

Here is an image of a product. These are the product descriptions for it: {bulletpoints}. Please analyze the descriptions and list 2-3 most important features or functionality. Return key words only without any starting or ending statements. Do not include dimension or assembly information. Each feature should be informative. If you cannot extract any relevant product features from both the image and description, return 'N/A'.

To ensure quality and visual grounding, we manually filtered out expressions that is 'N/A' and could not be reliably mapped to specific regions in the product images. We also manually reviewed all generated expressions to

ensure the dataset’s quality. All manual processing was completed by 4 evaluators. Each evaluator was required to review all the image-expression pairs and judge each expression as either "good" or "bad." To quantify inter-annotator agreement, we employed Fleiss’ Kappa [19], which is suitable for measuring agreement among multiple raters beyond what would be expected by chance. For expressions with low agreement among evaluators (such as 2-2 splits), we either modified the expression manually or removed it from the dataset entirely. The final dataset consists only of expressions that received strong majority approval (3-1 or 4-0 votes) and demonstrated clear visual grounding in the product images. This rigorous curation process yielded 2,989 referring expressions, each targeting part-level regions and describing specific materials, features, functionalities, or packaging elements.

### 8.3. Mask Annotation

Our annotation process leverages SAM [25] to achieve efficient and accurate region segmentation. The annotation workflow consists of two stages: automatic segmentation and manual refinement. In the first stage, we utilize SAM’s automatic mode to generate a comprehensive set of candidate segmentation masks for each image. GT regions that correspond to our referring expressions are then selected from these candidates. For regions that SAM failed to identify automatically, we proceed to the second stage where we manually annotate them using SAM’s interactive mode with point supervision. This semi-automated approach significantly streamlines the annotation process while ensuring precise region segmentation for our dataset.

Similar to the evaluation of expressions, we also conducted quality assessment for the segmentation annotations. The same panel of 4 evaluators reviewed each segmented region and classified them as either "good" or "bad" based on their accuracy and alignment with the corresponding expressions. We applied Fleiss’ Kappa [19] to measure inter-annotator agreement for these segmentation evaluations as well. Regions with low agreement scores were flagged for re-annotation using more precise point supervision in SAM’s interactive mode. Only segmentations that received strong majority approval (3-1 or 4-0 votes) were retained in the final dataset, ensuring that our ground truth regions accurately represent the visual elements referenced in the expressions.

## 9. Quantitative Results

To ensure statistical robustness and account for potential variability in RESAnything’s performance, especially for the components involving LLM generation (reference text, candidate text, and similarity analysis), we conducted experiments with our approach 8 separate times and report the averaged results in both the main paper and supplementary

materials.

### 9.1. CLIP as RNN

We present quantitative results of CLIP as RNN, the current SOTA zero-shot method, on both ReasonSeg and ABO-Image-ARES in Table 7.

Table 7. Quantitative results of CLIP as RNN [52], with RESAnything’s results shown in parentheses for comparison.

Dataset	test	
	gIoU (ours)	cIoU (ours)
ReasonSeg[26]	35.2 (74.6)	26.4 (72.5)
ABO-Image-ARES	24.4 (78.2)	15.7 (72.4)

### 9.2. Part-only RES benchmark

We further evaluate the performance of RESAnything and competing methods on UniRES [56], which contains a subset RefCOCO<sub>m</sub> for part-level RES. Table 8 shows the quantitative results on *part-only* RefCOCO<sub>m</sub>. Since the code for UniRES [56] is not publicly available, we directly compare performances using the mIoUs reported in their paper. Although UniRES is claimed to be a zero-shot method, it is pre-trained on their proposed MRES-32M dataset, which is closed source. Our method significantly outperforms the training-free zero-shot CaR, and generally outforms the supervised UniRES and LISA, *even though* they were both pre-trained on related tasks. GLaMM is the same and is slightly ahead of ours, but this is attributable to its additional fine-tuning on their proposed Grand dataset.

Table 8. Quantitative results on RefCOCO<sub>m</sub> **Part-only** set.

Method	val	testA	testB
<i>supervised / pre-trained</i>			
UniRES [56]	19.6	16.4	25.2
LISA [26]	21.2	19.1	27.4
GLaMM [45]	30.0	27.2	31.8
<i>training-free zero-shot</i>			
CaR [52]	10.9	10.6	10.9
RESAnything	27.6	26.5	25.8

### 9.3. Multi-object GRES benchmarks

Although RESAnything is not specifically designed for multi-object RES task, it still effectively handles these cases through the grouping and selection algorithm, demonstrating the generalization on these tasks. Table 9 reports qualitative comparison on a GRES benchmark, g-RefCOCO [32]. Among the methods, only GRES is trained on g-RefCOCO. RESAnything achieves comparable results as LISA and GLaMM, while significantly outperforming the training-free zero-shot method CaR.

Table 9. Results on gRefCOCO (cIoU).

Method	val	testA	testB
<i>pre-trained on vanilla RES tasks</i>			
LISA [26]	48.4	45.1	46.3
GLaMM [45]	46.2	46.7	47.2
<i>supervised (trained on gRefCOCO)</i>			
GRES [32]	62.4	69.3	59.9
<i>training-free zero-shot</i>			
CaR [52]	25.6	22.0	21.5
RESAnything	52.7	46.2	46.3

We conducted additional evaluations of our method against competing methods on R-RefCOCO [61] and RefZOM [22]. Images in both datasets are extracted from the RefCOCO, with additional multi-object referring expressions. Table 10 shows the quantitative results on both benchmarks. Both RefSegformer [61] and DMMI [22] are fully supervised method trained on the training set of R-RefCOCO and RefZOM separately. LISA and GLaMM also pre-trained on image data from COCO, which serves as the based of both benchmarks. Our method reasonably underperformed against supervised methods that were explicitly exposed to the training set, but still outperforms the SOTA training-free zero-shot baseline.

Table 10. Results on R-RefCOCO and RefZOM(mIoU)

Method	R-RefCOCO	RefZOM
<i>supervised (trained on training set)</i>		
RefSegformer [61]	68.8	-
DMMI [22]	-	68.2
<i>pre-trained</i>		
LISA [26]	71.1	45.0
GLaMM [45]	72.1	47.4
<i>training-free zero-shot</i>		
CaR [52]	30.2	25.7
RESAnything	61.2	40.3

#### 9.4. Runtime Comparison

As stated in the main paper, our method’s entire inference process can run efficiently on a single NVIDIA 24GB 4090 GPU. For a fair comparison, we measured the execution times of all competing methods on the same hardware. The average per-image processing time was evaluated on the ReasonSeg test set, with detailed results provided in Table 11. While our main results in the main paper were conducted using 8 V100 GPUs for running multiple experiments in parallel during development, we optimized our method’s runtime for comparison experiments. These optimizations include: 1) utilizing the bfloat16 data format for the LLM, which is not supported on V100; 2) enabling flash attention for more efficient transformer operations; 3) implementing batch generation for LLM outputs rather than

sequential processing of each reference and candidate text; and 4) employing batch computation for CLIP similarity scores.

Table 11. Runtime comparison.

Method	Time/image (s)
CaR	5.3
LISA	7.0
GLaMM	8.6
RESAnything-Qwen 2-VL	12.1

## 10. Qualitative Results

Firstly, Fig 17 – 25 show qualitative results on RefCOCO test A, test B, RefCOCOg val (G), val (U), test (U), RefCOCO+ test A, test B, val set separately. **These examples are randomly selected to provide an unbiased assessment.** RESAnything achieves comparable results to supervised methods on vanilla referring segmentation tasks. While our approach effectively handles many occlusion cases, as shown in Fig 16, neither our method nor current SOTA approaches can guarantee perfect part detection in every inference. Some failure cases in these results demonstrate challenges in combining parts or handling occlusions. Detailed descriptions and analysis of these failure cases are provided in the individual figure captions.

Secondly, additional qualitative results on ReasonSeg [26] are shown in Figs. 26–28, including comparisons with CLIP as RNN (CaR) [52], the current SOTA zero-shot method. Compared to supervised methods, our Chain-of-Thoughts attribute prompting enables clearer input expression identification and more precise segmentation results. Detailed descriptions and failure cases are provided in the captions.

Lastly, we present qualitative comparisons on ABO-Image-ARES (Figs. 29–31) to demonstrate our method’s effectiveness on challenging expressions involving parts, materials, functionality, and package elements. See figure captions for detailed analysis.

standard prompt:		attribute prompt:	
For the region described as $E$ in the image, provide a single detailed sentence describing an object or part of an object.		For the region described as $E$ in the image, provide a single detailed sentence describing an object or part of an object <i>by including its location, appearance (color, shape, location), and distinctive characteristics including relevant actions or state. Focus on features that would help uniquely identify this specific region from others in the image.</i>	
image	input expression $E$	reference text $T^{\text{ref}}$ (standard prompt)	reference text $T^{\text{ref}}$ (attribute prompt)
	person on left	The person on the left is wearing black socks.	The person on the left is a young man in a crouched, athletic stance wearing a bright blue t-shirt, black shorts, black knee-high socks, and a distinctive red baseball cap, holding a white frisbee in his hands as he prepares to throw or catch it during what appears to be an Ultimate Frisbee game on a grassy field.
	right girl	The right girl is wearing a striped hat.	The right girl is wearing a black top with the lower part visible, extending from the midsection to the bottom of the image, featuring a slight flare.
	15	The sock of the player numbered 15 is black.	A player wearing a white jersey with the number 15, positioned near the goal net, with their back to the camera and hair tied up in a bun, actively engaged in the play as they appear to be moving towards or interacting with the soccer ball on the ground.
	front pizza	The front pizza has a charred crust edge with visible burn marks.	The 'front pizza' is a large, round Neapolitan-style pizza in the foreground of the image, featuring a thin crust with charred edges, topped with a vibrant red tomato sauce, scattered herbs, and what appear to be olives, sitting on a white plate with a fork beside it on a white paper napkin.
	far right dud	The far right man is holding a drink in his right hand.	The far right region of the image shows a person standing slightly apart from the group, wearing a white shirt with a lanyard, dark pants, and appearing to be observing the scene with their hands at their sides, positioned near what looks like an entryway or doorframe in the background.
	pillow over baby head	Pillow over baby head is a portion of a soft pillow that covers the top of the baby's head.	The pillow over the baby's head is located at the top of the image, partially covering the baby's head and neck, with a soft, padded appearance and gentle folds. It is white or light-colored, providing comfort and support while the baby drinks from the bottle.

Figure 9. Comparison of Text Generation With and Without Attribute Prompting: Our analysis demonstrates that when attribute prompting is not used, MLLM fails to accurately identify and reason about input expression attributes. The contrast between standard prompting and attribute-specific prompting highlights this significant limitation in attribute recognition.



prompt for rating:

*For the given mask-cropped image and texts ( $E$  and  $T^{\text{ref}}$ ), rate their semantic similarity. Provide two scores between 0-1, where 1 means perfect match.*

input expression  $E$ :  
*guy on right*

Reference text  $T^{\text{ref}}$ :

*The figure on the right is a person wearing a black hoodie and light blue jeans, standing with their back to the camera, holding a skateboard in their right hand while facing a grand, ornate building with a domed roof in the background.*

input expression  $E$ :  
*blue car right*

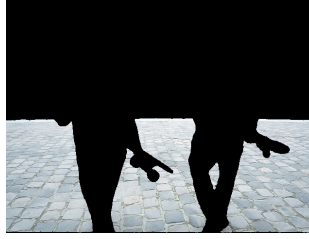
Reference text  $T^{\text{ref}}$ :

*On the right side of the image, a dark blue sedan is partially visible, parked alongside the curb in front of what appears to be a restaurant or bar, with only its rear quarter and taillight visible in the frame.*

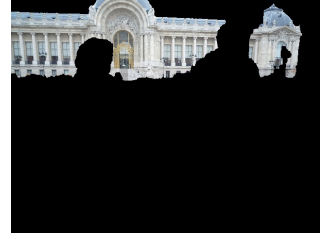
input expression  $E$ :  
*black shorts*

Reference text  $T^{\text{ref}}$ :

*The black shorts are worn by the player on the left, who is running forward with his body leaning slightly to his right, the shorts appearing snug-fitting and reaching to just above the knee, contrasting sharply with his black and white striped jersey and white socks.*



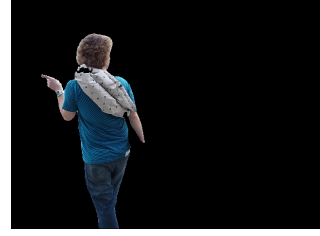
(0.8, 0.8) / (0, 0)



(0.9, 0.8) / (0, 0)



(0.8, 0.9) / (1, 1)



(0.9, 0.8) / (0, 0)



(0.8, 0.9) / (0, 0)



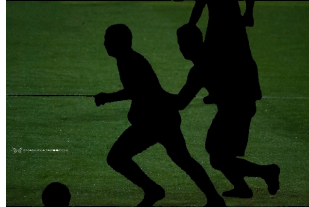
(0.9, 0.8) / (0, 1)



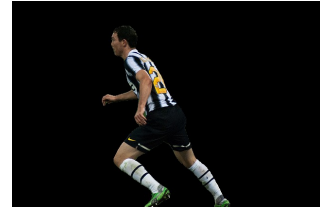
(0.9, 0.8) / (1, 1)



(0.8, 0.8) / (1, 1)



(0.9, 0.9) / (0, 0)



(0.9, 0.9) / (1, 1)



(0.8, 0.9) / (0, 0)



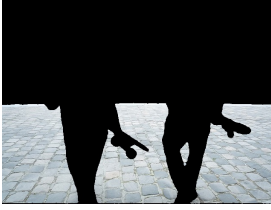
(0.9, 0.8) / (0, 1)

Figure 10. Analysis of MLLM’s Rating and Binary Response Performance: For each mask-cropped region, we compare two types of outputs: numerical scores (score 1, score 2) and binary responses ( $d^{12t}$ ,  $d^{2i}$ ) (0=’no’, 1=’yes’). The results reveal that MLLM struggles to generate meaningful similarity scores when comparing the input expression  $E$  and reference text  $T^{\text{ref}}$ . The assigned scores (typically around 0.8-0.9) appear arbitrary rather than reflecting accurate contextual similarities. In contrast, the model’s binary yes/no responses prove more reliable for assessment purposes.



Reference text  $T^{\text{ref}}$ :

*The figure on the right is a person wearing a black hoodie and light blue jeans, standing with their back to the camera, holding a skateboard in their right hand while facing a grand, ornate building with a domed roof in the background.*



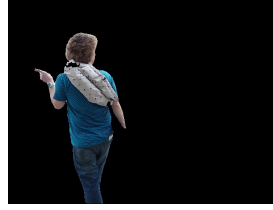
Candidate text  $T^{\text{can}}$ :  
*The gray ground region at the bottom of the image with two individual's silhouette walking on light-colored cobblestone pavement*

0.234 / 0.241



Candidate text  $T^{\text{can}}$ :  
*The man in the black hoodie and jeans appears focused while walking his skateboard through a grand, open public square.*

0.487 / 0.266



Candidate text  $T^{\text{can}}$ :  
*The young woman, carrying a large quilted courier bag over shoulder, stands in the middle of a bustling plaza, needing directions from the man beside her.*

0.543 / 0.223



Candidate text  $T^{\text{can}}$ :  
*A large, ornate building stands majestically in the background, its grand facade featuring large windows and classical architectural details.*

0.406 / 0.269

Reference text  $T^{\text{ref}}$ :

*On the right side of the image, a dark blue sedan is partially visible, parked alongside the curb in front of what appears to be a restaurant or bar, with only its rear quarter and taillight visible in the frame.*



Candidate text  $T^{\text{can}}$ :  
*The lower portions of the image that showing what appear to be bench legs with individual sitting on it on grayish pavement.*

0.476 / 0.217



Candidate text  $T^{\text{can}}$ :  
*The unpainted section at the rear door of the car appears to be made of plastic, contrasting with the smooth metallic surface of the rest of the vehicle.*

0.577 / 0.264



Candidate text  $T^{\text{can}}$ :  
*A white rectangle region, possibly a part of the wall on street beside a restaurant, with black sign on it.*

0.413 / 0.213



Candidate text  $T^{\text{can}}$ :  
*The blue car is parked near a restaurant on a cobblestone street, reflecting the bustling street life around it. In the foreground, a clear, sleek metallic surface meticulously mirrors the surrounding urban environment.*

0.668 / 0.243

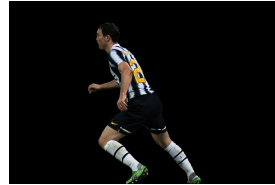
Reference text  $T^{\text{ref}}$ :

*The black shorts are worn by the player on the left, who is running forward with his body leaning slightly to his right, the shorts appearing snug-fitting and reaching to just above the knee, contrasting sharply with his black and white striped jersey and white socks.*



Candidate text  $T^{\text{can}}$ :  
*Silhouette of a central soccer player in a vibrant red jersey dribbling the ball across the lush green field, with two players closely tracking his movements on either side.*

0.455 / 0.233



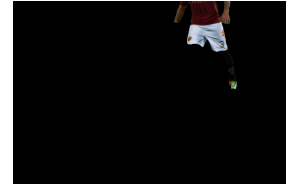
Candidate text  $T^{\text{can}}$ :  
*The athlete, dressed in a black and yellow uniform, is in motion on the soccer field.*

0.539 / 0.313



Candidate text  $T^{\text{can}}$ :  
*A soccer player wearing a red jersey with yellow accents and white shorts with an emblem on the right leg, appears to be in motion.*

0.609 / 0.241



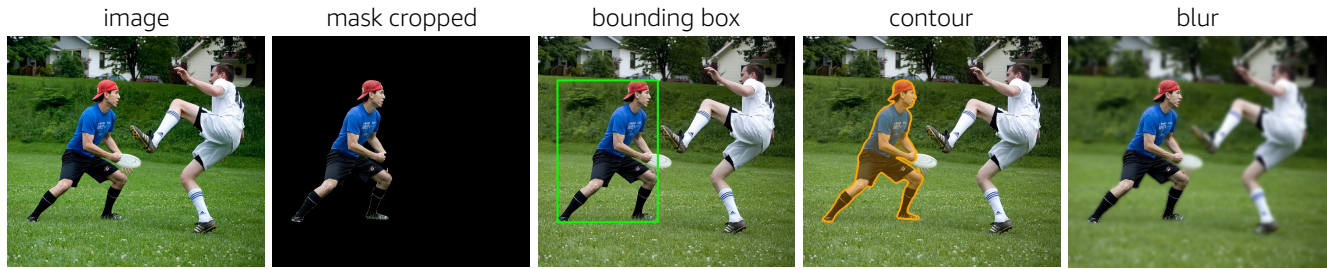
Candidate text  $T^{\text{can}}$ :  
*The player in the maroon jersey and white shorts with an embossed design appears to be in mid-air during a dynamic soccer match.*

0.456 / 0.246

Figure 11. Analysis of CLIP's Similarity Evaluation: For each mask-cropped region, we compare text-to-text ( $s^{\text{t}^{\text{t}}}$ ) and text-to-image ( $s^{\text{i}^{\text{t}}}$ ) CLIP scores. Text-to-text scores prove more reliable, while text-to-image scores consistently remain below 0.3, showing limited discriminative power. However, relying solely on text-to-text scores can be misleading, as demonstrated in the last row where a description containing "white shorts" receives a higher score despite incorrectly matching the reference image showing "black shorts". This highlights the limitation of text-to-text evaluation in capturing crucial contextual details.

query	image	reference text	candidate text (selected as output)	output
blue and blue umbrella		In the <b>center-right</b> portion of the image, a <b>large light blue</b> umbrella with a <b>darker blue</b> underside stands out prominently, shading a group of seated individuals and contrasting with the surrounding multicolored umbrellas	A <b>large</b> beach umbrella with a two-toned blue canopy, featuring a <b>darker blue center</b> panel and <b>lighter blue</b> outer panels	
green color vegetable in between potato and carrot		The green vegetable located between the potato and carrots is a large head of <b>broccoli</b> , tightly clustered florets forming a rounded, textured dome of <b>deep forest green</b> color.	The vibrant head of <b>broccoli</b> with <b>deep forest green</b> florets stands prominently in the center of a wooden tray, surrounded by an assortment of vegetables, including sweet potatoes and leafy greens.	
a frosted sprinkled cupcake, one out of four to the right and front of the others		The frosted sprinkled cupcake, located one out of four to the right and front of the others, is a <b>round</b> , pink and white cupcake with <b>colorful sprinkles</b> , sitting on <b>lower right corner</b> of a black plastic tray.	A colorful donut covered in <b>rainbow sprinkles</b> , positioned in the <b>lower right corner</b> of the image, with a distinctive <b>round</b> shape and hole in the center.	
something that the animals are tied to		The target in the image is a <b>wooden</b> post, located centrally between two donkeys, with a notable red frame around its base. It is vertical, weathered, with visible side openings, and appears to be used for <b>tethering the animals</b> .	A long, slender lance with a pointed tip and a <b>wooden post</b> stands prominently amidst a vibrant, bustling scene, flanked by two stately and ornately <b>equipped horses</b> in front of a distinguished building with blue walls and a terracotta roof.	
pink skirt		The pink skirt is worn by an <b>older woman</b> standing on the <b>right side</b> of the image, featuring a <b>floral pattern</b> and falling just below the knee, adding a pop of color to the predominantly blue and white color scheme of the wedding party gathered on the cobblestone street.	An <b>older woman</b> wearing a beige jacket over a <b>floral-patterned pink dress</b> stands in the <b>right</b> of the image against a black background, with their hands clasped in front of them.	
When someone is reading a book or a magazine and wants to take a break, they may need a specific object to mark their place. What item in the picture is commonly used for this purpose?		The target region is a white <b>bookmark</b> positioned between the pages of a book on the <b>right side</b> of the image, which stands out due to its narrow, <b>rectangular</b> shape and its placement marking a specific page.	A colorful <b>bookmark</b> lies on the <b>right side</b> of a book titled "Weekend Sewing". It is a small <b>rectangular</b> card featuring some patterns.	
the area where people can walk		The target region in the image is an elevated, <b>circular</b> platform with railings situated at the <b>top</b> of the tall, cylindrical tower, designed to be a designated area where people can walk and observe.	The metal grate at the <b>top</b> of the tall chimney appears rusted and slightly <b>rounded</b> , suggesting it has been exposed to the elements for an extended period.	
the lights that are placed in different directions		The lights that are placed in different directions are white, cylindrical, and mounted on the <b>right end</b> of a white metal <b>rod</b> , featuring light to down.	A small, cylindrical, white plastic component with multiple <b>small holes</b> along its length is positioned at the <b>right end</b> of a horizontal <b>rod</b> , holding a lamp in place amidst a row of similar lamps on a gallery display	
the area that is first impacted when a car is moving forward and crashes		The target region is the <b>front</b> of the car, specifically the white <b>front bumper</b> with black air vents and the rounded headlight area, located at the right of the image.	The <b>front side</b> of a white toy car, featuring to be the <b>front bumper</b> in light gray, providing protection of the car.	

Figure 12. Examples of query, reference, and candidate text. For each input expression query (column 1), RESAnything generates detailed reference text describing the input expression's attributes (column 3). Our grouping and selection algorithm identifies the most relevant segmentation from candidates. Columns 4 and 5 show RESAnything's output segmentation and its corresponding candidate text. Key words of attributes in both texts are highlighted in red color.



visual prompts used

candidate text

mask cropped

A man in a bright blue shirt and red cap is positioned in the *centre* of the image, possibly *holding a camera*, with his left leg extended forward as he looks ahead intently in a grassy park setting.

✗ incorrect location and action

blur

A person wearing a red hat and a blue t-shirt. The person is in a crouched position on the grassy field in the *centre* of the image, watching intently towards the right side of the image, likely focusing on a *soccer* game or activity.

✗ incorrect location and action

image  
+  
mask cropped

*Two players* are playing frisbee game in a lush green *grass field* with small white flowers scattered throughout, one in a bright blue t-shirt, black shorts and the another one in a *white uniform and blue socks* is leaping with one leg extended high in the air.

✗ distracted by irrelevant regions

image  
+  
contour

On a grassy field dotted with small white flowers, *two players* engage in a game of frisbee. One wears an *orange* t-shirt with black shorts, while their opponent, dressed in a white uniform and blue socks, leaps gracefully with one leg extended skyward.

✗ distracted by irrelevant regions;  
✗ affected by the contour overlay color

mask cropped  
+  
contour

A person in a bright *orange* t-shirt, black shorts, and an *orange* baseball cap is shown in a crouched, athletic stance, positioned in the left of the image.

✗ affected by the contour overlay color

bounding box  
+  
contour

Positioned on the left side of the frame, an athlete assumes a crouched stance, wearing a bright *green* t-shirt paired with black shorts and topped with an orange baseball cap.

✗ affected by both bounding box color and contour overlay color

bounding box  
+  
mask cropped

A person in a bright *blue* t-shirt, black shorts, and a red baseball cap is shown in a crouched, athletic stance, positioned in the *left* of the image against a black background.

✓ correct region, color and action

Figure 13. Comparison of different visual prompt combinations for attribute description generation. Top row shows the four basic visual prompts: original image, mask-cropped region, bounding box overlay, contour overlay and blur background. Bottom rows demonstrate how different combinations affect MLLM’s generated descriptions. Using mask-cropped or blur alone leads to incorrect location and action inference, while combining with original image causes distraction from irrelevant regions. Contour-based approaches (with either mask-cropped or bounding box) suffer from color overlay interference. Our chosen combination of bounding box and mask-cropped achieves the most accurate descriptions by leveraging spatial context while avoiding color interference.





Figure 14. When dealing with non-convex shapes, analyzing only the contour without considering the overlaid mask region can lead to ambiguous visual interpretations. This ambiguity often results in generated text descriptions that contain misleading information, where incorrectly identified objects are highlighted in red.

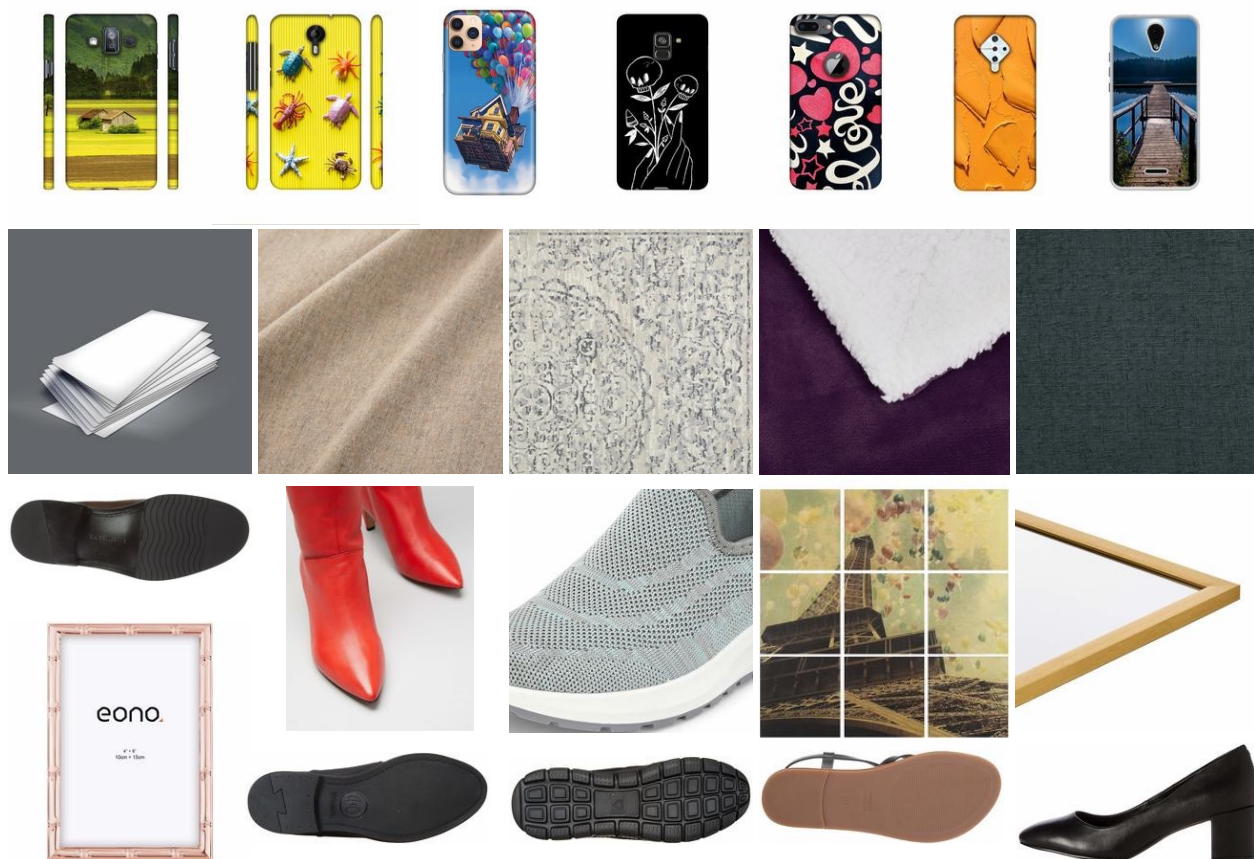


Figure 15. Images excluded from the ABO dataset typically lack meaningful or referrable parts. Row 1 shows phone related items that primarily consist of phone cases displaying only the back view of phones. Row 2 features images solely showing product textures or materials that fill the entire frame. Images from the footwear and picture frames categories in row 3 & 4 are commonly presented against plain white backgrounds without distinct parts for grounding.





*lady middle pink*



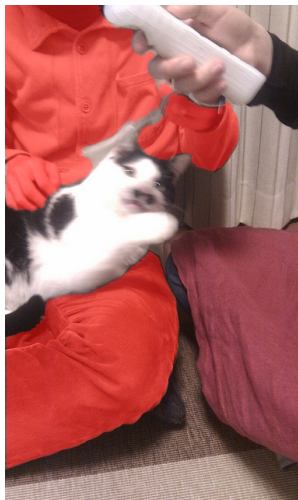
*second guy on right*



*girl in purple*



*catcher on left*



*guy leg*



*back cowboy*



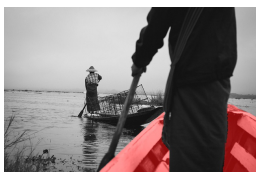
*the mom*



*dark brown horse*



*middle horse hidden*



*the boat in the foreground close to teh camera dont click the guy in the boat*



*middle bird*



*in middle*



*third from front on right*



*black car*



*only banana that is laying other way*



*white horse*



*cow in background*

Figure 16. RESAnything can handle occlusion cases by grouping and selection cases. Results from RefCOCO.



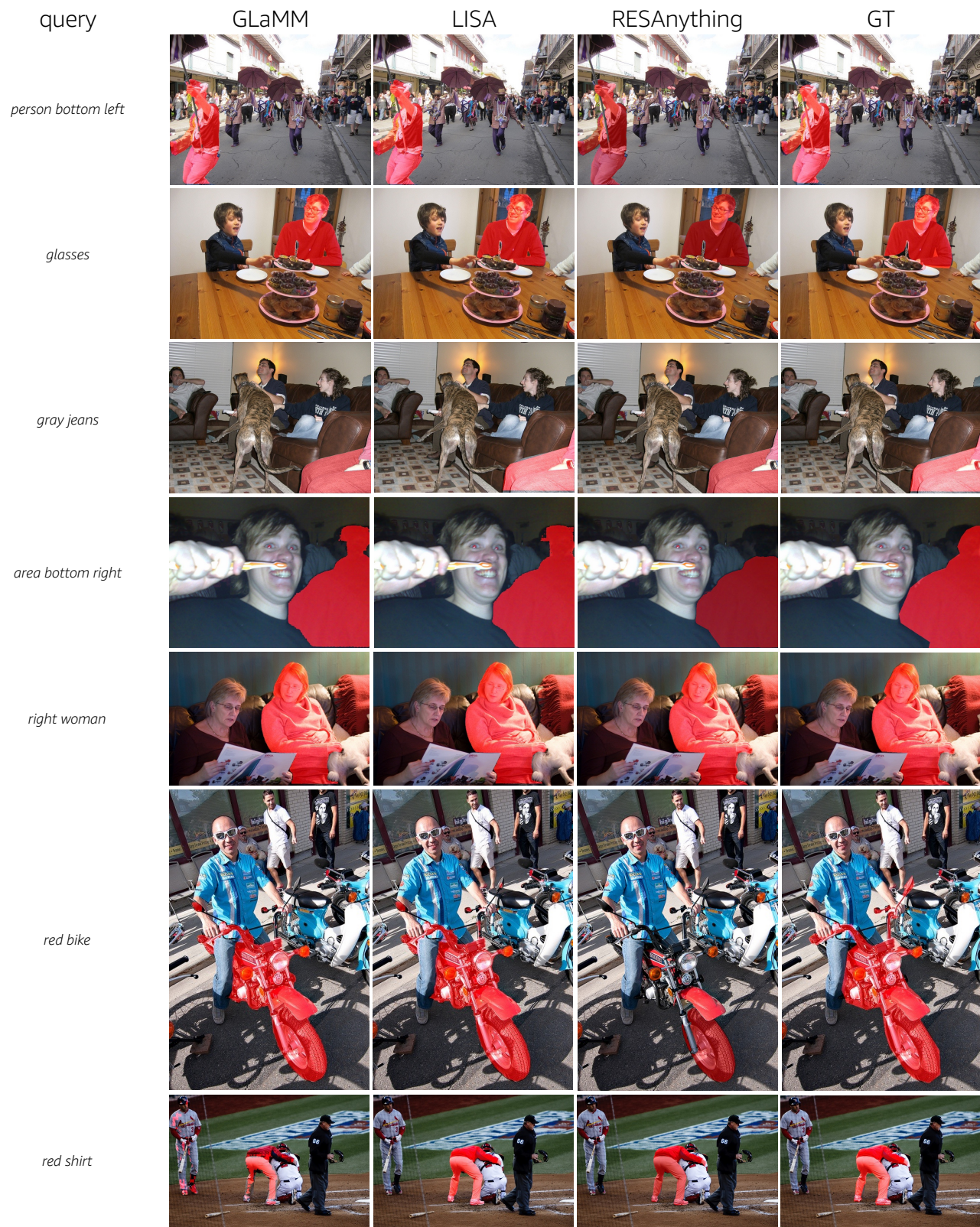


Figure 17. Qualitative results on RefCOCO test A (randomly selected). The ground truth annotations can be problematic - some queries refer only to an object/region while the GT marks an entire person (row 2: “glasses”; row 4: “area bottom right”, row 7: “red shirt”). Row 6 shows a failure case of RESAnything.



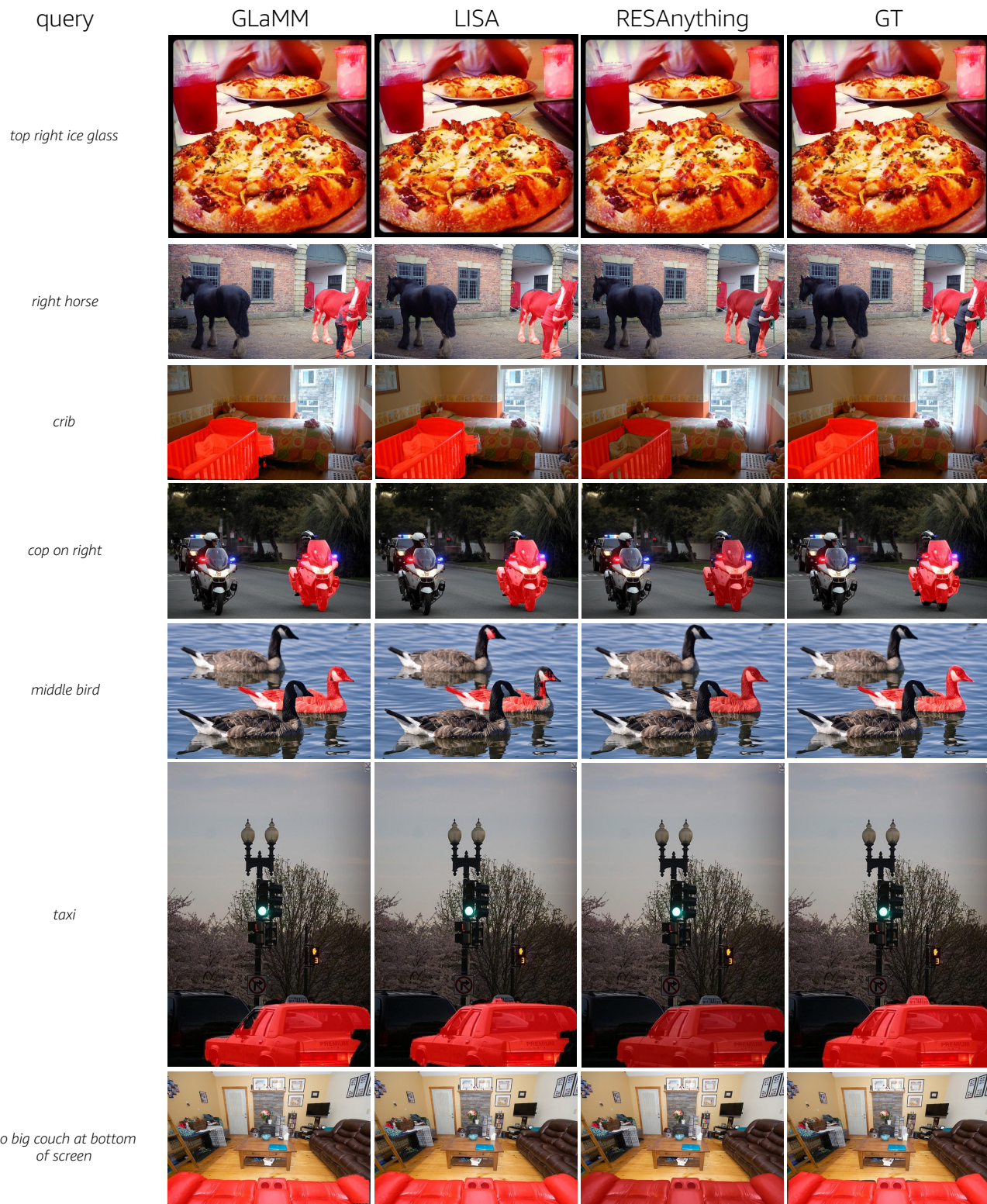


Figure 18. Qualitative results on RefCOCO test B (randomly selected). Despite being unsupervised, RESAnything achieves comparable results to supervised methods, particularly excelling at crowded regions (row 2). However, it occasionally misses parts when needing to combine multiple masks (row 3, 5).



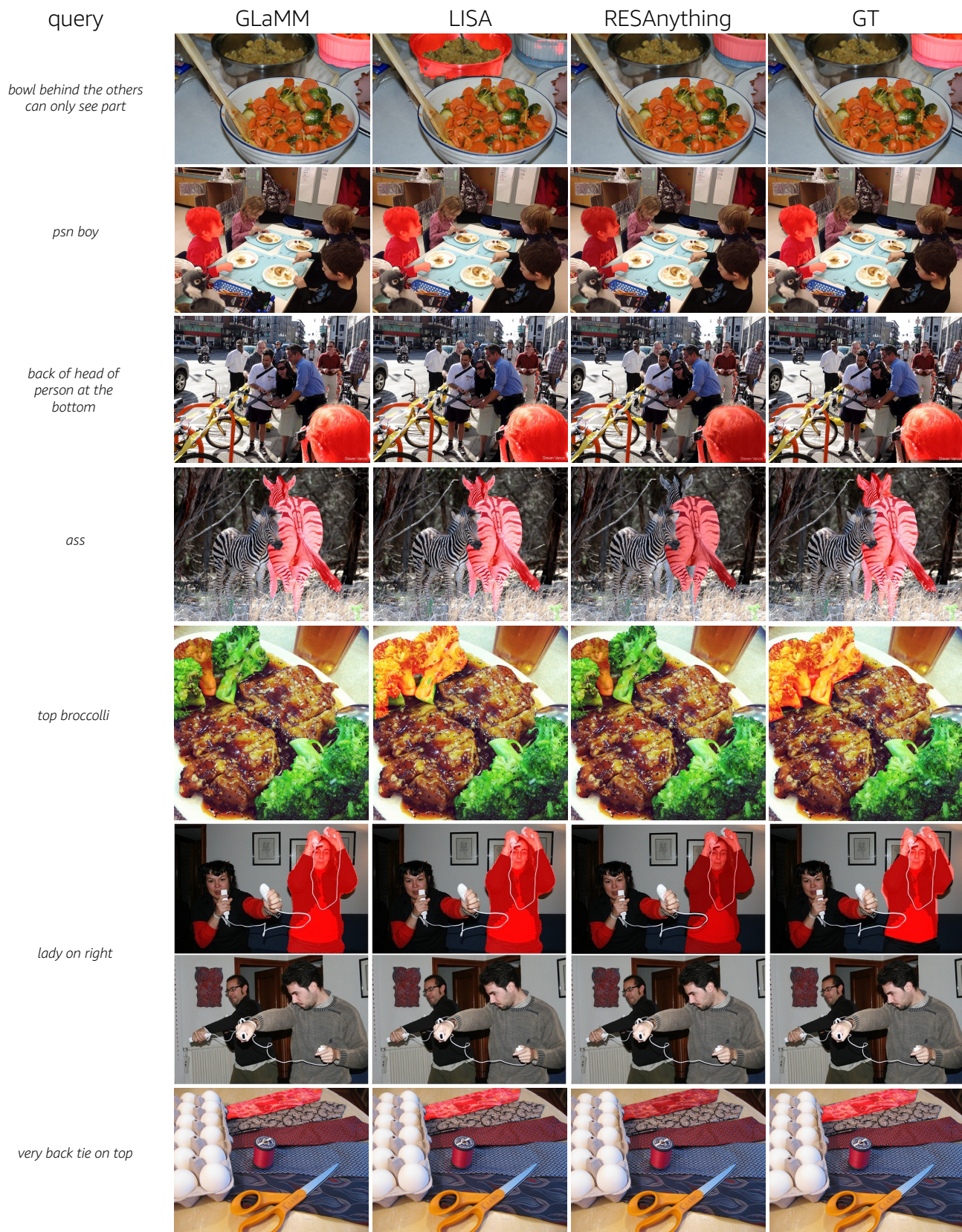


Figure 19. Qualitative results on RefCOCO val (randomly selected). RESAnything generates fine-grained segmentation of the input expression (row 1, 4). As mentioned in Fig 18, it may miss parts when combining multiple masks (row 5).





Figure 20. Qualitative results on RefCOCOg val (G) (randomly selected). RESAnything generalizes well on mask with hole (row 3, 5), but may suffering from over-segmentation (row 1, 4) or no good candidate found (row 7).





Figure 21. Qualitative results on RefCOCOg val (U) (randomly selected).





Figure 22. Qualitative results on RefCOCOg test (U) (randomly selected).





Figure 23. Qualitative results on RefCOCO+ test A (randomly selected).



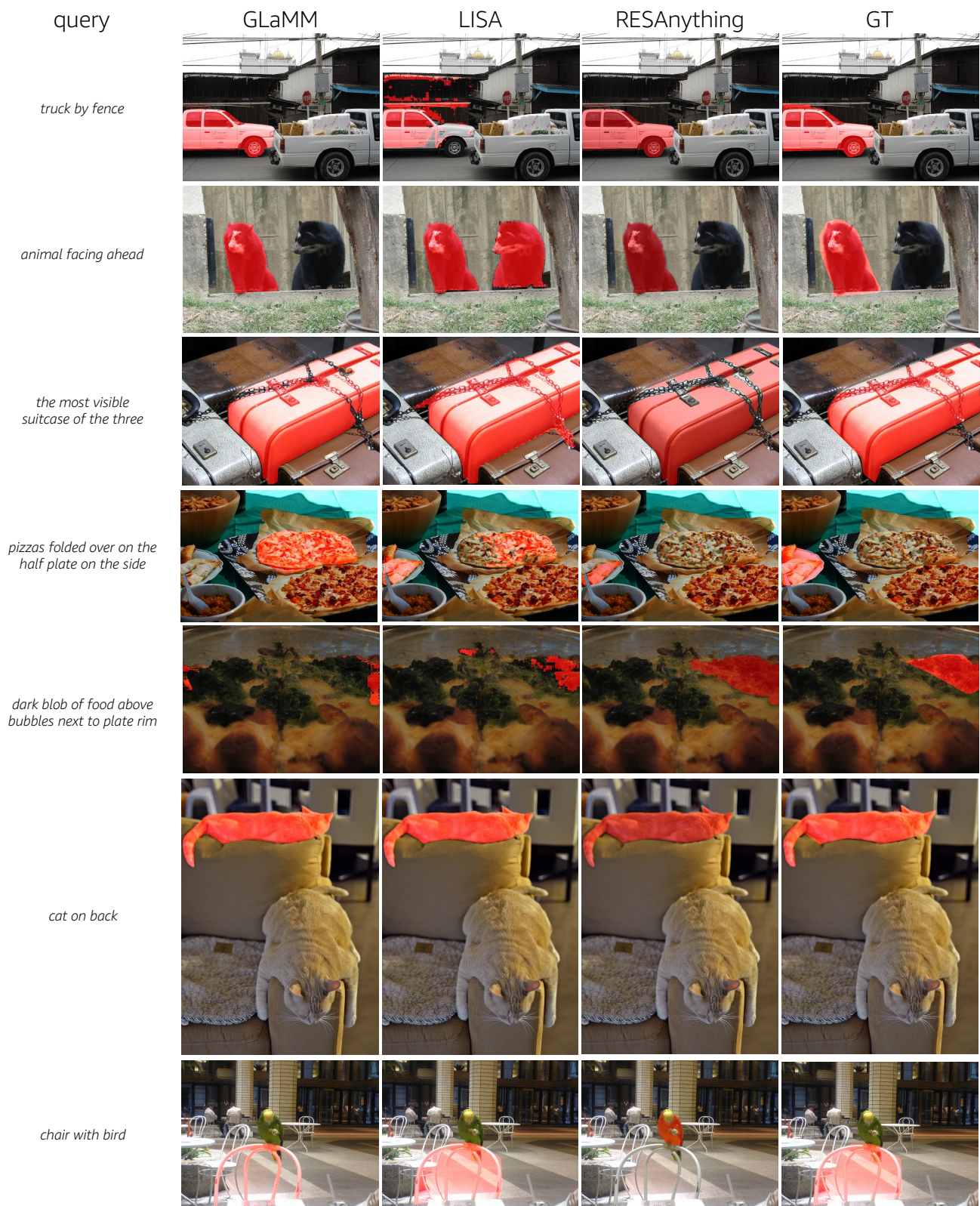


Figure 24. Qualitative results on RefCOCO+ test B (randomly selected).





Figure 25. Qualitative results on RefCOCO+ val (randomly selected).



query	CaR	GLaMM	LISA	RESAnything	GT
something indicating the brand of this car					
something that the animals are tied to					
In order to take a clear and stable photo or video, what equipment in the picture can be used to hold the camera steady?					
the object that helps to keep the neck warm					
Dogs are not typically allowed to freely roam in public spaces, and they usually need to be kept under control when walking indoors. What object in the picture could be used to keep the dog restrained in a hallway?					
Snails have a soft, fragile body that requires hard materials to protect them. What part in the picture can accomplish this task?					
the person who is pushing the baby carriage					
When it comes to water sports, people often use various tools to glide on the surface of the water. What item in the picture is designed specifically for this purpose?					
People can play beautiful music by pressing the keys on the piano keyboard. What part of the piano in the picture can be used for playing?					
the part of the vehicle that can be opened					

Figure 26. Qualitative results on ReasonSeg (Part 1). RESAnything outperforms others in correct localization (row 2, 3, 8, 9), refined segmentation (row 1, 4, 5, 6, 7) and part-level understanding (row 9, 10).



query	CaR	GLaMM	LISA	RESAnything	GT
<i>the food with high protein</i>					
<i>the area for passengers on the airship</i>					
<i>the area that is first impacted when a car is moving forward and crashes</i>					
<i>the damaged part of the silk stockings</i>					
<i>When someone is reading a book or a magazine and wants to take a break, they may need a specific object to mark their place. What item in the picture is commonly used for this purpose?</i>					
<i>the area where people can walk</i>	MISS				
<i>the household appliance used for heating food</i>					

Figure 27. Qualitative results on ReasonSeg (Part 2). “MISS” indicates that the method is failed to output a segmentation.



query	CaR	GLaMM	LISA	RESAnything	GT
We cannot breathe underwater, so diving requires additional equipment to help people breathe while underwater. What in the picture can help human accomplish this task?					
Backpacks are commonly used for carrying personal belongings during outdoor activities and travel. What part of the backpack in the picture can be used to store smaller items or accessories?					
In a formal event, such as a gala or award ceremony, what accessory in the picture can be worn around the neck to add a touch of elegance to a man's suit?					
In the picture, there is a type of bird with a distinctive feature on the top of its head, which usually indicates its gender. What part in the picture might have this characteristic?					
In a modern office, employees often have meetings to discuss work matters. What object in the picture can be used as a surface for employees to place documents or devices during a meeting?					
the musician					
the lights that are placed in different directions					
the container that is being held by a person and is about to pour liquid					
What object in this picture is often used to place coffee cup and pastry while sitting?					

Figure 28. Qualitative results on ReasonSeg (Part 3).



Figure 29. Qualitative results on ABO-Image-ARES (Part 1).



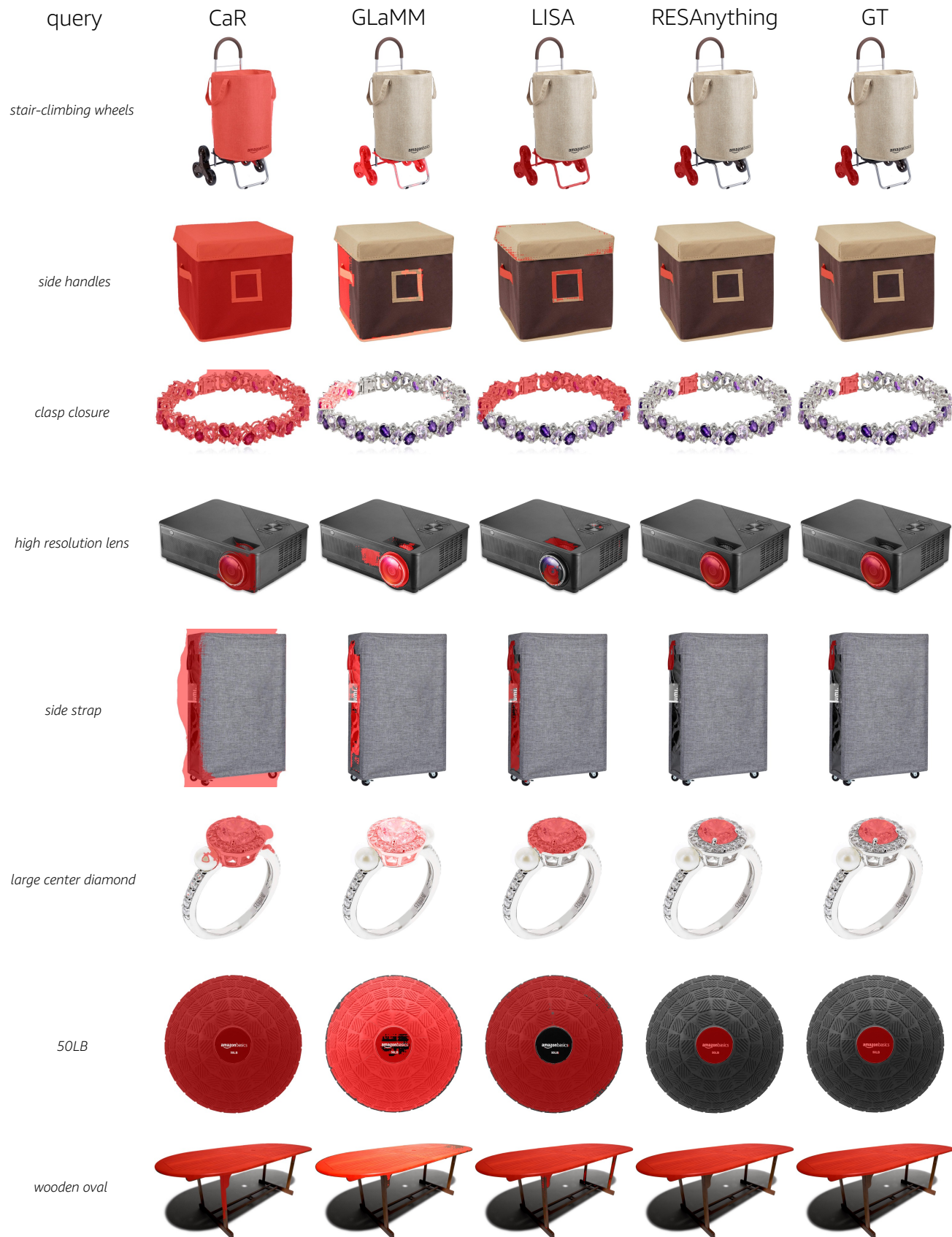


Figure 30. Qualitative results on ABO-Image-ARES (Part 2).

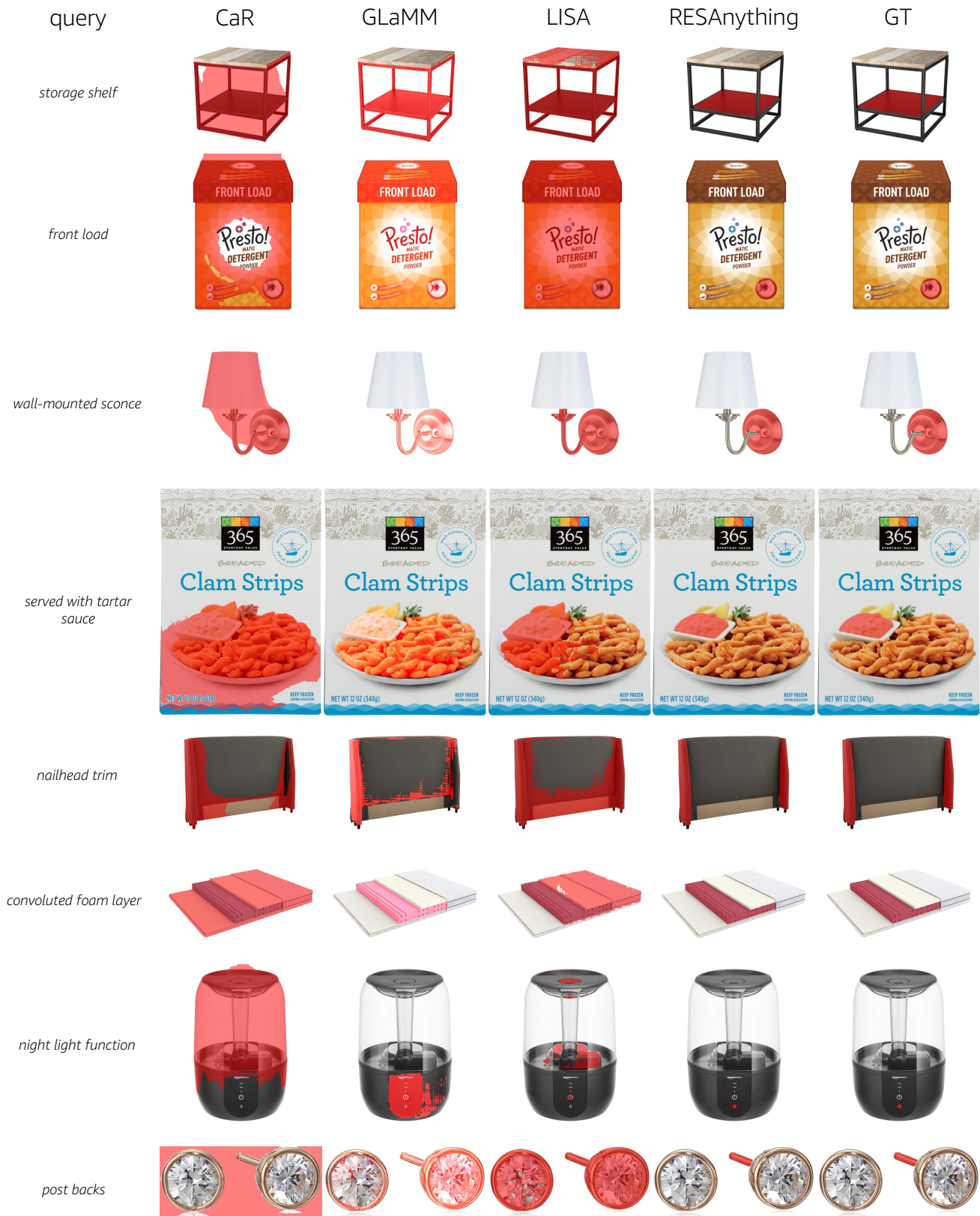


Figure 31. Qualitative results on ABO-Image-ARES (Part 3).