RESAnything: Attribute Prompting for Arbitrary Referring Segmentation

Ruiqi Wang Hao Zhang
Simon Fraser University

Figure 1. **Open-vocabulary and zero-shot referring expression segmentation with RESAnything.** Our method produces accurate object or part masks from general- and free-form text expressions including, from left to right: object or part semantic label, material/style properties, function/design descriptions, or logos and packaging labels in textual or other graphical in an image. For visualization purposes, we overlay segmentation regions with red color in each example.

Abstract

We present an open-vocabulary *and* zero-shot *method for* arbitrary *referring expression segmentation (RES), targeting more general input expressions than those handled by prior works. Specifically, our inputs encompass both object- and* part-level *labels as well as* implicit *references pointing to* properties *or* qualities *of object/part function, design, style, material, etc. Our model, coined RESAnything, leverages* Chain-of-Thoughts *(CoT) reasoning, where the key idea is* attribute prompting. *We generate detailed descriptions of object/part attributes including shape, color, and location for potential segment proposals through systematic prompting of a large language model (LLM), where the proposals are produced by a foundational image segmentation model. Our approach encourages deep reasoning about object/part attributes related to function, style, design, etc., to handle implicit queries without any part annotations for training or fine-tuning. As the first zero-shot and LLM-based RES method, RESAnything achieves superior performance among zero-shot methods on traditional RES benchmarks and significantly outperforms existing methods on challenging scenarios involving implicit queries and complex part-level relations. We contribute a new benchmark dataset of ∼3K carefully curated RES instances to assess part-level, arbitrary RES solutions.*

1. Introduction

With rapid developments in Large Multimodal Models (LMMs), visual perception systems have evolved significantly, demonstrating remarkable capabilities in bridging vision and language tasks [15, 20, 29, 35]. Recent advancements in LMMs have enabled sophisticated understanding of visual content, from object detection to semantic segmentation [5, 9, 42]. One of the emerging segmentation tasks that has drawn a great deal of attention lately is the so-called Referring Expression Segmentation (RES) which aims at obtaining a segmentation mask in an image or video that represents an object instance referred to by a natural language expression [11, 18, 24, 32, 57, 60, 71, 75, 77].

Despite much progress made on RES, two common limitations are often observed. First, while existing approaches excel at identifying and segmenting objects as whole entities, they often fall short when the input expressions refer to specific object *parts*. Such situations arise frequently in applications such as eCommerce, where sellers and buy-

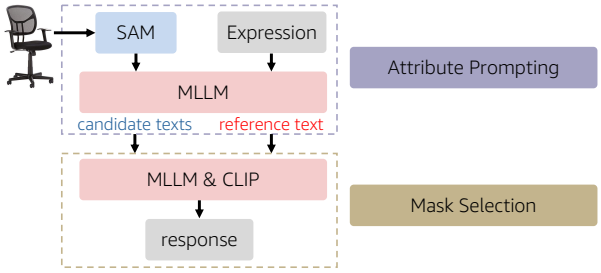


Figure 2. Overview of RESAnything: a two-stage framework for zero-shot arbitrary RES. The attribute prompting stage generates reference and candidate texts from input image and referring expression using SAM-generated proposals and an MLLM. The mask proposal selection stage leverages MLLM and CLIP to evaluate both candidates and proposals and produce the final response.

ers often promote or review product features referring to specific parts, and in robotics, human-computer interaction, and automated systems, where agents must interact with object parts. Second, most works to date on RES have focused on referring expressions that contain semantic labels in one way or another. Even the so-called generalized RES (GRES) [32] only extends the expression coverage to an arbitrary number of (including zero) target objects, *with labels*. On the other hand, object/part references are often *implicit*, without semantic labels. Such expressions can refer to *properties* or *qualities* related to object/part function, design, style, material, or they may appear in textual or other graphical forms as a logo or packaging label; see Fig. 1 for some samples expressions and segmentations.

In this paper, we present an *open-vocabulary* and *zero-shot* RES method to address both limitations. For lack of a better term, we call our task *arbitrary* referring segmentation and our model as *RESAnything*. Our goal is to allow input expressions to be more general than what prior works have been designed to handle, while solving our problem without any training or fine-tuning on specialized datasets. To this end, we leverage the generalization and zero-shot capabilities of modern-day foundational models such as Pixtral [4] and Claude [1] as Large Language Models (LLMs and SAM [25] for image segmentation. However, solving the arbitrary RES task demands a deeper understanding of object and part properties, moving beyond traditional object-level and label-centric referencing to more nuanced reasoning for part- and attribute-level perception.

There have been recent works [26, 27, 45] on reasoning-based segmentation through active LLM querying. An implicit query text, such as "the object containing the most Vitamin C," is first analyzed by a text LLM and then referenced to the "orange" object in the provided image. Nonetheless, such methods often fall short when the implicit connections between object/part properties (e.g., functional or stylistic ones) and their visual manifestations are cascadedly hidden. Even advanced LLMs, with their sophisticated reasoning capability, struggle to ground their understanding without explicit supervision at the part or attribute level. Additionally, existing methods, e.g., LISA [26], typically rely on fine-tuning on specially prepared or curated datasets — they are *not zero-shot*.

Our model for arbitrary RES is *training-free*. It leverages *Chain-of-Thoughts* (CoT) for comprehensive part-level understanding. Our key idea is *attribute prompting*, which generates detailed descriptions of object/part attributes including shape, color, and location for potential segment proposals through systematic prompting of LLMs [1, 4], where the proposals are produced by a foundational image segmentation model such as SAM [25]. Our approach encourages deep reasoning about object/part attributes related to function, style, design, etc., enabling the system to handle implicit queries without any part annotations for training or fine-tuning. By bridging abstract descriptions with concrete visual attributes through a *two-stage* evaluation framework (attribute prompting + grouping and selection of segment proposals), as illustrated in Fig. 2, RESAnything achieves robust performance on both traditional referring expressions and challenging implicit queries for arbitrary RES.

In summary, our contributions are as follows:

- The *first zero-shot* and *LLM-based open-vocabulary RES* method, targeting input expressions that are more general than those addressed by prior works.
- The novel idea of attribute prompting, as a means for Chain-of-Thoughts (CoT) reasoning, to achieve SOTA performance on both object- and part-level RES tasks.
- A new dataset, ABO-Image-ARES, built upon ABO [13], offering carefully curated RES instances as a benchmark to assess part-level, arbitrary RES solutions.
  Our dataset consists of 2,989 expression-segment pairs: 1,360 with object/part semantic labels, 742 depicting logos/packaging labels, 502 referring to functions/designs, and finally, 385 covering material/style properties.

We demonstrate by extensive experiments that RESAnything achieves superior performance among zero-shot methods on traditional RES benchmarks such as RefCOCO, RefCOCO+ [76], RefCOCOg [39, 41]. Our method also significantly outperforms existing methods on the recent reasoning segmentation dataset ReasonSeg [26], as well as RES tasks in challenging scenarios involving implicit queries and complex part-level relationships such as those from ABO-Image-ARES. With its zero-shot capabilities, the most important practical advantage of our method lies in the improved scalability and generalizability for real-world applications with diverse referring expressions. In contrast, current supervised methods, e.g., LISA [26] and GLaMM [45], require substantial training resources, with high data collection and annotation costs by humans. While performing well on vanilla RES benchmarks, they are not as scalable and are limited to scenarios in their training data.

2. Related Work

Recently, multimodal LLMs (MLLMs) has brought the success of LLMs to image understanding by integrating the visual and linguistic modalities. Example state-of-the-art proprietary models include Claude Sonnet [1], Gemini [2], GPT-4 series [3] etc. Most existing MLLM architectures connect a pre-trained vision encoder to the LLM decoder with a modality connector. For example, Flamingo [5] proposed the Perceiver Resample to bridge the modality gap, with follow-up works OpenFlamingo [6] and Otter [28] particularly developed for effective in-context instruction tuning. InstructBLIP [15] built upon the Querying Transformer as in BLIP2 [30]. The LLaVA models [33, 35] and MiniGPT4 [87] utilized a lightweight MLP and achieved appealing performances in various MLLM benchmarks. Recent developments include supporting high-resolution image inputs [34, 68, 82], optimizing model efficiency [7, 74, 85], and constructing higher-quality datasets [10, 16].

2.1. Open-Vocabulary and RES

RES [21, 24, 41] aims to segment target image regions based on textual descriptions. The core challenge lies in bridging the gap between image and language modalities. Typically, transformer-based text encoders [17, 44] are employed to extract textual embeddings, which are then integrated into segmentation architectures through cross-attention or feature alignment [12, 50, 59, 65, 73, 83] to achieve language-aware segmentation [31, 36, 56, 57, 66, 71]. Recently, SAM [25] has introduced text-guided segmentation [11, 38, 81]. For instance, Grounding-SAM [47] leverages bounding boxes returned by Grounding-DINO [37] to prompt SAM for mask prediction, while Fast-SAM [84] utilizes CLIP similarity scores [44] to select the final result from class-agnostic masks generated by SAM. However, the majority of these methods have been primarily designed for object-level segmentation based on explicit semantic expressions.

To address a broader range of segmentation targets and linguistic inputs beyond semantics, methods based on MLLMs have emerged, leveraging the powerful language understanding capabilities inherited from LLMs [9, 11, 14, 27, 42, 43, 55, 67, 75, 78–80]. One of the pioneering works in this area is LISA [26], which enables MLLMs to segment objects by using text embeddings from LLaVA to prompt a SAM [25] decoder to predict masks. LISA demonstrated promising performance on a new task called Reasoning Segmentation, similar to our Arbitrary Referring Segmentation. While improvements over LISA have been developed for extending it to generalized RES [63, 64] and grounded segmentation [45, 48], fine-tuning MLLMs on fixed segmentation datasets not only restricts the variety of referring expressions but also weakens the reasoning capability of pre-trained MLLMs. In contrast, our method operates in a training-free manner, preserving the complete ability of the MLLM to reason about the input images.

Some methods have demonstrated the feasibility of adopting pre-trained foundation models for RES without additional training [23, 52, 54, 77, 86]. MaskCLIP obtains pseudo masks by modifying the last attention layer of CLIP [86]. CaR couples CLIP and GradCAM to generate mask proposals, then employs a CLIP classifier to select the final masks, before a mask refinement [52] in post-processing. Global-Local CLIP [77] pioneered zero-shot RES using CLIP to extract visual features. Our approach follows a similar design, leveraging SAM for proposal generation and MLLMs for mask selection. Although MLLMs already exhibit superior reasoning abilities compared to CLIP, our novel attribute promoting technique further amplifies their inferential capabilities for arbitrary RES.

2.2. Visual Prompting

Prompting [49] has emerged as a powerful technique for adapting pre-trained language models to downstream applications. By incorporating additional hand-crafted instructions, prompt engineering methods effectively facilitate the adaptation process. For instance, Chain-of-Thought (CoT) prompting encourages models to explain their step-by-step reasoning while answering questions [58]. Recently, visual prompting [40, 51, 53, 70] has been proposed to enhance the adaptation of CLIP for open-vocabulary segmentation by overlaying ovals over segmentation targets [52]. SAM [25], on the other hand, allows users to provide points, boxes, masks as prompts for image segmentation, with the latest version supporting video segmentation [46]. Visual prompting has also been applied to MLLMs [62]. Overlaying image regions with bounding boxes, masks, circles, scribbles, and numeric markers has enhanced MLLMs' ability to perform region or pixel-level image understanding [8, 69, 72].

3. Method

Problem statement. Given an image I and a free-form expression E referring to a potential target region R in I, RESAnything first processes the image to generate and refine a set of segmentation proposals $M = \{m_1, \ldots, m_N\}$, from which it selects the most appropriate binary segmentation mask m_i representing R. The input expression E can be either an explicit referring expression (e.g., semantic label of an object/part) or an implicit expression (e.g., functional or material properties). For targets not directly visible, our method handles two scenarios: a) Irrelevant queries: indicate that the target does not exist in the image; b) Invisible targets: infer their location through their functional and spatial relationships, with explanatory reasoning.

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, indicating its similarity to

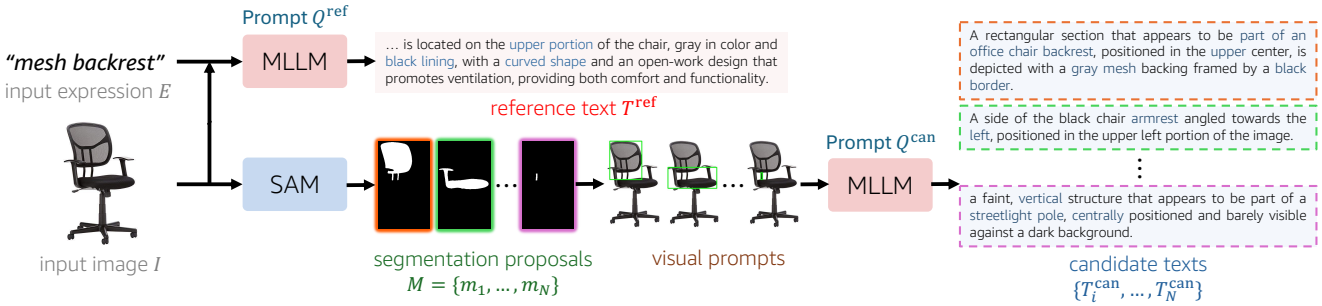


Figure 3. Attribute prompting using SAM and MLLM. Given the input image and referring expression, this stage produces two groups of predictions. The first output, a reference text T^{ref}, is generated from an MLLM with the text prompt Q^{ref}. It describes the visual attributes (e.g., color, shape, location) of the target region ("mesh backrest" in this example). The second group is a set of candidate texts T_i^{can}, generated by an MLLM with the text prompt Q^{can} and visual prompts derived from segmentation mask proposals. These texts describe the attributes of their corresponding segmentation region proposals, visualized with the same border color.

Figure 4. Example of different visual prompts V_i generated from a segmentation proposal m_i.

the input expression E. However, current MLLMs struggle with directly connecting the text description to the image region. It is possible to fine-tune a MLLM with many paired samples of texts and mask annotations, however, as mentioned earlier, this incurs significant computational cost during fine-tuning and human effort for data annotation.

Overview. Instead of fine-tuning, we propose a novel approach to facilitate reasoning between text descriptions and visual elements, by systematic "attribute prompting," which tasks the MLLMs with generating detailed text descriptions of visual properties including shape, color and location. By doing so, we not only encourages the MLLMs to perform in depth visual reasoning around the target regions, but also circumvents MLLMs weakness in handling image-text pairs, by creating additional intermediate text-text pairs that enable more robust comparison metrics.

Figure 2 provides an overview of RESAnything, which consists two main stages: 1) an attribute prompting stage that generates reference text for the target and candidate texts for generated segmentation proposals (Section 3.1); 2) a proposal selection stage that employs multiple metrics to robustly analyze the relationship between candidate and reference texts and produce the final response (Section 3.2).

3.1. Text Generation via Attribute Prompting

To facilitate reasoning between the input expression E and the segmentation proposals M, we first apply attribute prompting to generate detailed text descriptions: reference text T^{ref}, which describes the input expression E in relation to the image I, candidate texts $T^{\text{can}}_{1\ldots N}$, which describe each of the segmentation proposals in a format similar to that of the reference text. We apply MLLMs to generate these texts, carefully designing the input prompts to encourage the MLLMs to provide description that capture comprehensive object properties and inter-object relationships.

Reference text generation. The reference text T^{ref} functions as an extended visual description of the input expression E, providing more concrete visual attributes for challenging expressions such part-level semantic labels and functionality/feature-based descriptions. We task a MLLM to generate the reference text $T^{\text{ref}} = f_{\text{MLLM}}(I, E \mid Q^{\text{ref}})$, with a carefully designed reference text prompt Q^{ref} that instructs the MLLM to generate a single sentence with detailed visual attributes, such as shape, color and location, that describe the region R in I targeted by E. For invisible or irrelevant targets, the T^{ref} provides a reasoned explanation of why the target cannot be localized. We provide the full reference text prompt Q^{ref} in the supplementary. An example is shown in the top part of the Fig 3. Given the input "mesh backrest", the reference text describes its key attributes: "a *gray curved mesh* backrest with *black lining* located at the *upper portion* of the chair".

Candidate text generation. The candidate texts $T_1^{\text{can}}, \ldots, T_N^{\text{can}}$ describe the mask proposals $m_1, \ldots, m_N$ in a format similar to that of the reference text T^{ref}. Without requiring fine-tuning, our method can directly apply off-the-shelf SOTA image segmentation methods to obtain mask proposals. We adopt SAM [25] in this work. As SAM's raw outputs often contain duplicate or overlapping masks, as well as tiny segments, we configure SAM with sampling points at 0.015% of total image pixels and filter out segments smaller than 0.1% of the image area, preventing over-segmentation while maintaining meaningful region proposals. We also filter out duplicate proposals.

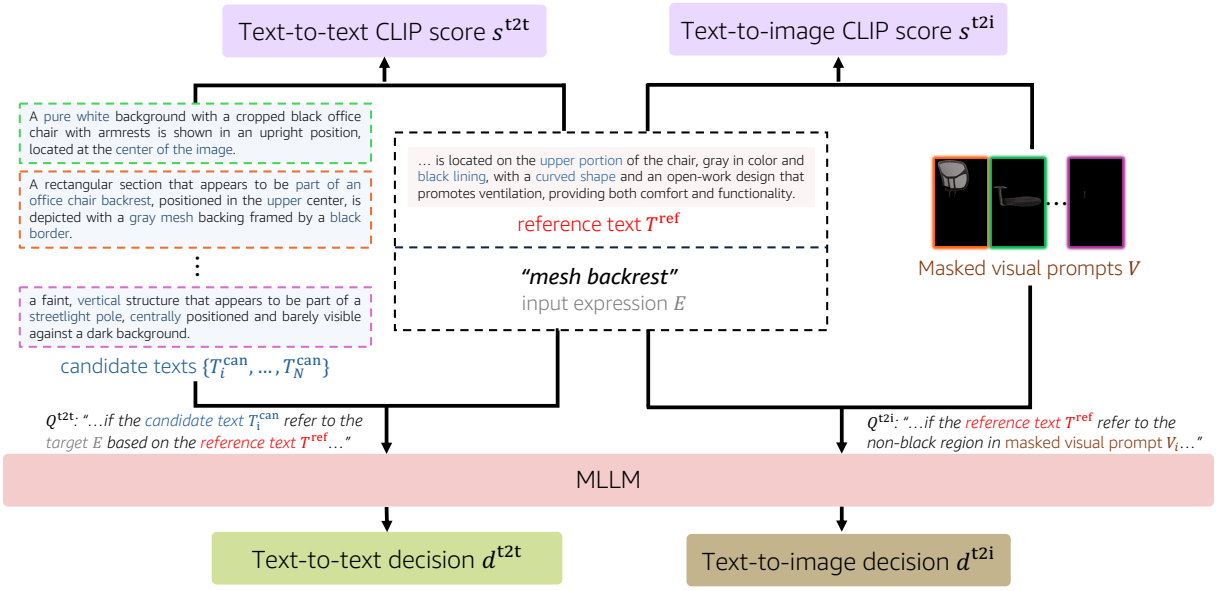


Figure 5. Multi-metric mask proposal selection using MLLM and CLIP. To select the final mask from mask proposals generated by SAM, we introduce four metrics computed across different modalities and models to evaluate the similarity between input expression E and the mask proposals. Specifically, the text-to-text MLLM-based binary decision d^{t2t} and CLIP score s^{t2t} match reference text to candidate texts. The text-to-image MLLM-based binary decision d^{t2i} and CLIP score s^{t2i} match reference text to masked visual prompts.

Given a mask proposal m_i, we generate a corresponding candidate text $T_i^{\text{can}} = f_{\text{MLLM}}(V_i^1, V_i^2 \dots V_i^K \mid Q^{\text{can}})$ using an MLLM, where Q^{can} is the candidate text prompt that similarly asks for visual attributes such as shape, color and location; and $V_i^1 \dots V_i^K$ are K visual prompts that provide distinct visual representations of the mask proposal M_i. A good visual prompt need to guide the MLLM to focus on the mask region, without removing attribute-related information or adding distractions. Figure 4 shows a few possible representations for visual prompts: *image* retains all information of the original image, but does not cover any mask-specific properties; *mask cropped* highlights the visual attributes of the masked region, but does not suggest the location of the masked region nor its relation with other parts of the image; in contrast, *bounding box*, *mask contour* and *blur background* provides such relational and locational information, but the bounding box outlines, the mask overlays, and blur background are distractions when it comes to visual properties such as color or shape. Using multiple visual prompts, intuitively, alleviate the issues of the respective prompting representation. In practice, we find using two visual prompts, *bounding box* (V^b) and *mask cropped* (V^m), is sufficient for our purpose. This is consistent with the observations of [52]. The complete candidate text prompt Q^{can} is provided in the supplementary. Fig 3, right part shows examples of generated candidate texts.

3.2. Multi-metric Mask Proposal Selection

The generated reference text and candidate texts allow us to assess the similarity between the input expression E and the mask proposals M much more effectively: the reference text T^{ref} provides more detailed information than the original expression E, thus facilitating in depth text-to-image comparisons; in addition, the candidate texts T^{can} enables an additional modality, allowing direct comparisons between two piece of texts. In this stage, we combine multiple evaluation metrics to perform both text-to-image and text-to-text comparisons to select the mask proposal (or none) that matches the input expression.

Text-to-text comparison. To compare a mask proposal m_i against the input expression E, we first evaluate the similarity between the reference text describing E, and the candidate text describing m_i. We first use the same MLLM to generate a binary decision $d_i^{\text{t2t}} = f_{\text{MLLM}}(T^{\text{ref}}, T_i^{\text{can}} \mid Q^{\text{t2t}}) \in \{0, 1\}$, where Q^{t2t} is the text-to-text comparison prompt, as shown in the lower left corner of Figure 5. The MLLM outputs a yes/no binary decision, as we observed empirically that it often struggles to output consistent scalar scores. However, there are cases where multiple mask proposals receive a "yes" response. To disambiguate such cases, we further employ CLIP to generate a scalar similarity score: $s_i^{\text{t2t}} = f_{\text{CLIP}}(T^{\text{ref}}, T_i^{\text{can}}) \in [0, 1]$. Although CLIP is generally more error-prone (as we show in the supplementary), its ability to output consistent scalar scores makes it well-suited for further disambiguating among the top candidates filtered by the binary MLLM decision.

Text-to-image comparison. While the text-to-text metrics already enable good candidate selection, potential errors during candidate text generation could degrade their

performance. To alleviate this, we further perform text-to-image comparisons between the reference text and the *mask cropped* visual prompt V_i^m. Similar to the text-to-text comparison, we use an MLLM-generated binary decision $d_i^{\text{t2i}} = f_{\text{MLLM}}(T^{\text{ref}}, V_i^m \mid Q^{\text{t2i}}) \in \{0, 1\}$, followed by a CLIP-generated scalar score $s_i^{\text{t2i}} = f_{\text{CLIP}}(T^{\text{ref}}, V_i^m) \in [0, 1]$, where Q^{t2i} is the text-to-image comparison prompt as shown in the lower right corner of Figure 5.

Grouping and selection. Given the computed metrics, we select the mask candidate that best matches the input expression E, or return the reference text T^{ref} if none is found. Algorithm 1 summarizes this process.

As MLLM decisions are prioritized over CLIP sores, we begin by checking whether any masks receive positive responses for both text-to-text and text-to-image MLLM decisions. In practice, we notice that the correct candidate is often the union of all the candidate masks that satisfy this condition, especially in cases where a single semantic entity spans multiple segments (e.g., all legs of a sofa). Therefore, we also include the union of these masks as another viable candidate. We then return the mask candidate with the highest combined CLIP score (sum of s_{t2t} and s_{t2i}). If no such masks exist, we then repeat this process, using only the text-to-text MLLM decisions as the filter, and then using only the text-to-image MLLM decisions as the filter.

We also prioritize text-to-text over text-to-image decisions, as empirically, we find the former more reliable. As a final verification step (lines 17-20 in Algorithm 1), when no candidates receive positive MLLM responses, we check if any of them has a combined CLIP score over a threshold (set to 1 for all experiments), and return the mask with the highest score. This threshold helps identify cases where the target is either invisible or irrelevant to the image, in which case we return the reference text T^{ref} explanation that describes why the target cannot be localized.

This algorithm enables our method to handle occlusion cases by combining parts segmentations, while also generalizing to multi-object scenarios. Additional discussions and results are available in the supplementary materials.

4. Experiment

We use Pixtral 12B [4] as the MLLM, SAM ViT-H [25] for generating segmentation proposals, and CLIP-ViT-B-32 for CLIP scores. Our experiments were conducted on a server with 8 NVIDIA 32GB V100 GPUs for parallel inference, but the entire inference process can run effectively on just a single NVIDIA 24GB 4090 GPU. Additional inference time details are provided in the supplementary materials.

Public datasets. Following the most previous works on referring segmentation [11, 26], we evaluate the performance of RESAnything on four public benchmark datasets:

Algorithm 1 Grouping and Selection Process

1: $conditions \leftarrow \{(True, True), (True, False), (False, True)\}$
2: **for** $(t2t, t2i)$ in $conditions$ **do**
3: **if** $t2t$ and $t2i$ **then**
4: $C \leftarrow \{m_i \mid d_i^{\text{t2t}} = 1 \wedge d_i^{\text{t2i}} = 1\}$
5: **else if** $t2t$ **then**
6: $C \leftarrow \{m_i \mid d_i^{\text{t2t}} = 1\}$
7: **else if** $t2i$ **then**
8: $C \leftarrow \{m_i \mid d_i^{\text{t2i}} = 1\}$
9: **if** $|C| = 1$ **then**
10: **return** $C[0]$
11: **else if** $|C| > 1$ **then**
12: $m_{\text{cmb}} \leftarrow \text{CombineMasks}(C)$
13: Compute $s_{\text{cmb}}^{\text{t2t}}$, $s_{\text{cmb}}^{\text{t2i}}$
14: **return** $\text{argmax}_{m \in \{C \cup m_{\text{cmb}}\}}(s_{\text{t2t}}^m + s_{\text{t2i}}^m)$
15: **else**
16: pass
17: **if** $\max_m(s_m^{\text{t2t}} + s_m^{\text{t2i}}) < 1$ **then**
18: **return** T^{ref}
19: **else**
20: **return** $\text{argmax}_{m \in M}(s_m^{\text{t2t}} + s_m^{\text{t2i}})$

Figure 6. Examples of different expressions in ABO-Image-ARES. Best viewed with zoom-in.

RefCOCO, RefCOCO+ [76], RefCOCOg [39, 41] and ReasonSeg [26]. Being a zero-shot method, we directly evaluate on the validation and test sets without any fine-tuning.

ABO-Image-ARES benchmark. To further evaluate the capability of RESAnything in handling implicit expressions (e.g., part-level materials, features, and functionalities), we establish the ABO-Image-ARES benchmark for complex reasoning segmentation tasks. We build upon the ABO dataset, which contains product listings with rich metadata, images, and 3D models from Amazon.com. Our benchmark comprises 2,482 high-resolution catalog images spanning 565 product types, with 2,989 referring expressions targeting part-level regions that describe specific materials, features, functionalities, or packaging elements. Fig. 6 shows representative examples, with detailed refer extraction procedures and data annotation provided in the supplementary.

Evaluation metrics. We evaluate our method using two standard metrics following prior works [26, 45]: generalized IoU (gIoU) and cumulative IOU (cIoU). gIoU computes the average of per-image Intersection-over-Union scores, while cIoU measures the ratio of cumulative intersection to cumulative union across all images. We report

Table 1. Quantitative results on standard RES benchmarks refCOCO/+/g, reported as cIoU values.

Method	refCOCO			refCOCO+			refCOCOg		
	val	testA	testB	val	testA	testB	val(U)	val(G)	test(U)
fully-supervised on the training set									
VLT [18]	67.5	70.5	65.2	56.3	61.0	50.1	55.0	-	57.7
CRIS [57]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	-	60.4
LAVT [71]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	-	62.1
GRES [32]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	-	66.0
pre-trained on the same task									
UniRES [56]	71.2	74.8	66.0	59.9	66.7	51.4	62.3	-	63.2
LISA-7B [26]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	-	70.6
GSVA [64]	77.2	78.9	73.5	65.9	69.6	59.8	72.7	-	73.3
GLaMM [45]	79.5	**83.2**	**76.9**	72.6	78.7	64.6	74.2	-	74.9
SAM4MLLM [11]	**79.8**	82.7	74.7	**74.6**	**80.0**	**67.2**	**75.5**	-	**76.4**
training-free zero-shot									
GLCLIP [77]	26.2	24.9	26.6	27.8	25.6	27.8	33.5	33.6	33.7
CaR [52]	33.6	35.4	30.5	34.2	36.0	31.0	36.7	36.6	36.6
RESAnything	**68.5**	**72.2**	**70.3**	**60.7**	**65.6**	**52.2**	**60.1**	**60.5**	**60.9**

Table 2. Quantitative results on ReasonSeg.

Method	val	
	gIoU	cIoU
GLaMM [45]	47.4	47.2
LISA-7B-LLaVA1.5 [26]	53.6	52.3
LISA-13B-LLaVA1.5 [26]	57.7	60.3
SAM4MLLM [11]	58.4	60.4
RESAnything	74.6	72.5

Table 3. Quantitative results on ABO-Image-ARES.

Method	test	
	gIoU	cIoU
LISA-13B-LLaVA1.5 [26]	43.3	34.0
GLaMM [45]	46.2	38.7
RESAnything	78.2	72.4

gIOU for RefCOCO, RefCOCO+, and RefCOCOg, and both metrics for ReasonSeg and ABO-Image-ARES.

4.1. Evaluation on Vanilla RES

We evaluate RESAnything on standard referring segmentation benchmarks, as shown in Table 1. Our method significantly outperforms existing zero-shot approaches, more than doubling the performance of GLCLIP (68.5% vs 26.2% on refCOCO val set) and achieving comparable results with early supervised methods like VLT. Despite UniRES [56] being described as a zero-shot method, it was pre-trained on their proposed MRES-32M dataset, which remains unavailable to the public. Furthermore, due to UniRES being closed source, our comparisons are limited to the accuracy figures reported in their paper. The performance gap compared to recent supervised methods can be attributed to our segmentation strategy with smaller mask proposals, which faces challenges when handling large complete objects that are common in these datasets. Qualitative results are provided in the supplementary. Furthermore, we evaluate RESAnything with competing methods on more general part-level and multi-object referring segmentation tasks as detailed in our supplementary.

4.2. Evaluation on Reasoning Segmentation

We evaluate RESAnything on the ReasonSeg benchmark (Table 2), where our method achieves state-of-the-art performance of 74.6% gIoU and 72.5% cIoU, surpassing LISA-13B by 17% and SAM4MLLM by 16%. Notably, while LISA variants require fine-tuning on reasoning tasks and GLaMM & SAM4MLLM rely on extensive training data, RESAnything achieves this superior performance without any task-specific training, demonstrating the effectiveness of leveraging MLLMs for deep reasoning. Qualitative comparisons are shown in Fig 7.

ABO-Image-ARES contains more challenging referring expressions targeting materials, features, functionalities or package elements. On this benchmark, RESAnything achieves 78.2% gIoU and 72.4% cIoU, significantly outperforming GLaMM by over 30% in both metrics, demonstrating our method's strong capability in handling complex reasoning queries. Qualitative comparisons are shown in Fig 8.

4.3. Ablation Study

Visual prompts. As shown in Fig 4, we explore different types of visual prompts for generating candidate texts T^{can} and performing text-to-image comparison. Table 4 compares their performance on RefCOCO test A set. The combination of mask-cropped and bounding box prompts achieves the best performance (72.2% gIoU), while using mask alone yields the lowest (47.2% gIoU) as it obscures contextual relationships. This demonstrates the importance of preserving spatial context through bounding box while maintaining region-specific details through mask cropping.

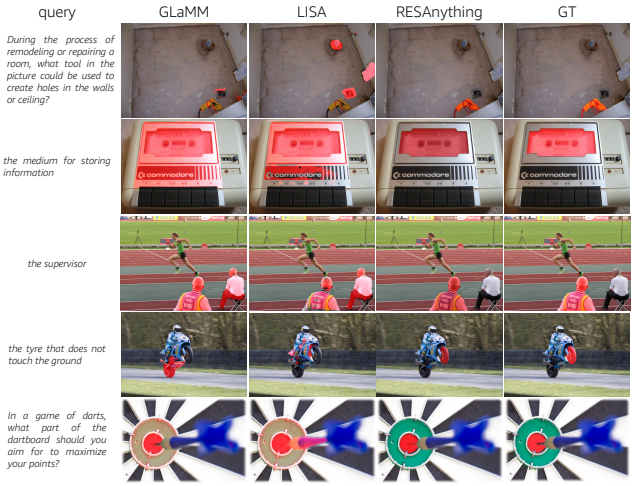


Figure 7. Qualitative comparisons on ReasonSeg. Our method demonstrates superior performance in both object localization accuracy (rows 1, 3, 4) and segmentation precision (rows 2, 5).

Table 4. Ablation study on different visual prompts.

Dataset	Visual Prompts					gIoU	cIoU
	image	mask	bbox	contour	blur		
RefCOCO test A		✓				47.2	42.3
	✓	✓				56.2	53.3
	✓			✓		48.4	44.2
					✓	43.5	39.2
		✓			✓	67.4	64.1
		✓	✓			72.2	69.5
		✓		✓		68.5	64.4
			✓	✓		50.4	46.6

Table 5. Ablation study on MLLM backbone.

LLM	gIoU	cIoU
Pixtral 12B[4]	74.6	72.5
Claude3.5 Sonnet[1]	76.2	73.4
Qwen 2-VL[7]	74.2	72.1

Additional analysis is provided in the supplement.

MLLM backbone. To analyze the impact of varying the MLLM backbone, we compare the performance of different MLLMs on ReasonSeg. Table 5 summarizes the results. While Pixtral-12B is our default choice, both Qwen2-VL and Claude 3.5 Sonnet achieve comparable or slightly better performance (74.2-76.2% gIoU), demonstrating our method's robustness across different MLLMs. See supplementary materials for extended analysis.

5. Conclusion, limitation, and future work

We present RESAnything, a zero-shot approach to advance open-vocabulary RES by supporting language expressions referring to highly general concepts. Our method comprises two key components: a novel attribute prompting technique to extract detailed attributes as text descriptions by synergizing SAM and MLLM for CoT analysis, and a multi-metric mask selection module based on CLIP and MLLM to select the optimal mask from SAM proposals.

Figure 8. Qualitative comparisons on ABO-Image-ARES. RESAnything demonstrates superior generalization ability across diverse queries, producing more fine-grained segmentation.

Our method demonstrates superior performance over prior zero-shot methods on standard RES benchmarks (RefCOCO/+/g). More importantly, our training-free approach substantially outperforms existing fine-tuned MLLM methods on both ReasonSeg [26] for reasoning segmentation and our newly augmented ABO dataset, underscoring its comprehensive reasoning capabilities. While RESAnything also performs well on object-level RES, attribute prompting excels especially at part-level reasoning since the attributes considered (color, shape, and location) tend to exhibit more consistency over parts, than objects, that share similar functions, styles, material, etc. It would be interesting to explore other attributes for CoT or automate the prompts.

Our method has substantial room for inference efficiency optimization in future work, particularly through RoI filtering and size-based mask proposal pruning to reduce candidate text generation overhead. RESAnything also inevitably inherits limitations common to foundation model-based approaches. Notably, SAM occasionally fails to produce the best mask candidates, potentially degrading RES accuracy, as shown in the supplementary materials. In addition, the effectiveness of RESAnything depends on the specific MLLMs employed. Future work could focus on improving the mask proposal generation process and exploring the integration of more advanced LLMs/MLLMs.

Acknowledgements

We thank Yiming Qian, Kai Wang, Fenggen Yu for their invaluable contributions in the early stage of this project.

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 similarity 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^{ref} to generated reference text T^{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^{ref} to obtain reference text T^{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^{can} to obtain candidate text T^{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^{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^{t2t}, we use following Q^{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^{ref}. Here is the candidate text to evaluate: T^{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^{t2i}, we use following Q^{t2i}:

You are evaluating if the following reference text describes the non-black region of the cropped mask image: T^{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^{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 RefCOCOm for part-level RES. Table 8 shows the quantitative results on *part-only* RefCOCOm. 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 RefCOCOm **Part-only** set.

Method	val	testA	testB
supervised / pre-trained			
UniRES [56]	19.6	16.4	25.2
LISA [26]	21.2	19.1	27.4
GLaMM [45]	30.0	27.2	31.8
training-free zero-shot			
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
pre-trained on vanilla RES tasks			
LISA [26]	48.4	45.1	46.3
GLaMM [45]	46.2	46.7	47.2
supervised (trained on gRefCOCO)			
GRES [32]	62.4	69.3	59.9
training-free zero-shot			
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 reasonaly 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
supervised (trained on training set)		
RefSegformer [61]	68.8	-
DMMI [22]	-	68.2
pre-trained		
LISA [26]	71.1	45.0
GLaMM [45]	72.1	47.4
training-free zero-shot		
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:

For the region described as E in the image, provide a single detailed sentence describing an object or part of an object.

attribute prompt:

For the region described as E in the image, provide a single detailed sentence describing an object or part of an object 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.

image	input expression E	reference text T^{ref} (standard prompt)	reference text T^{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^{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^{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^{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^{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.

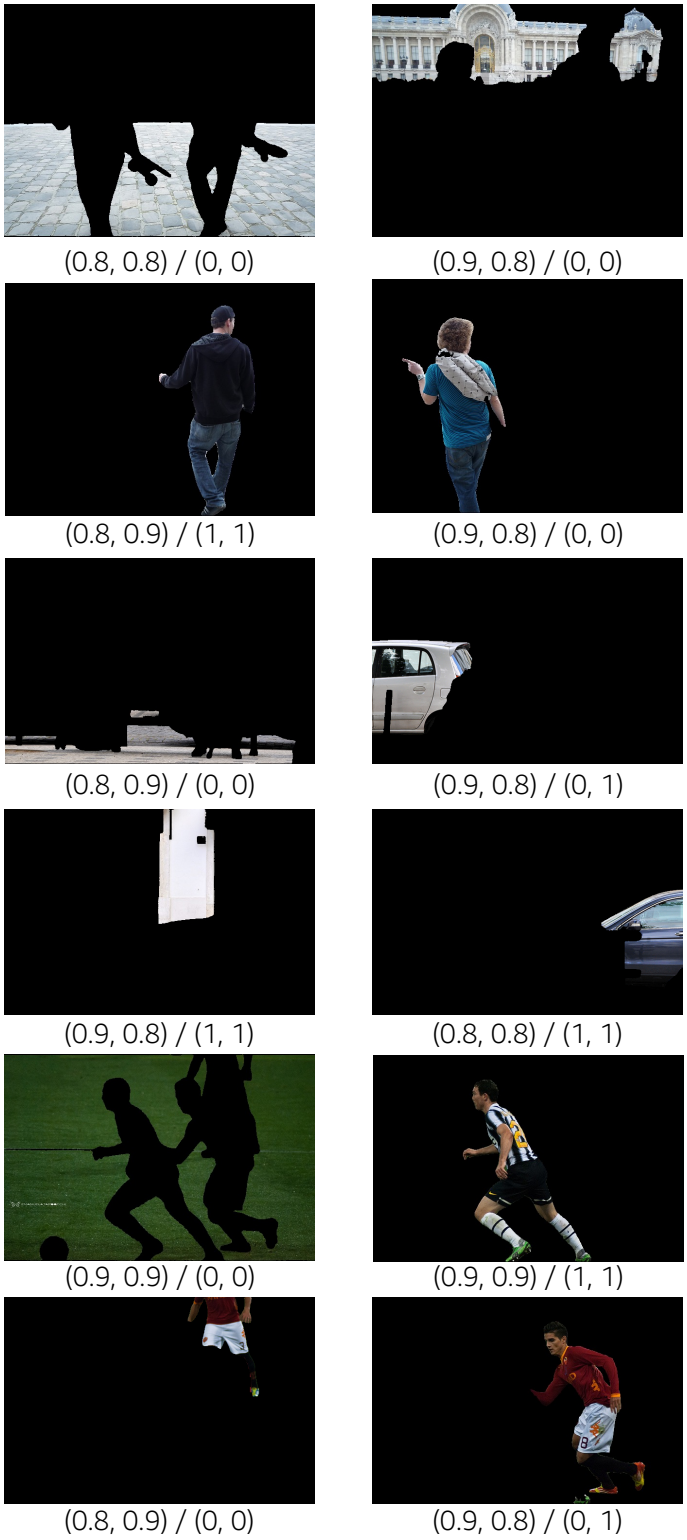
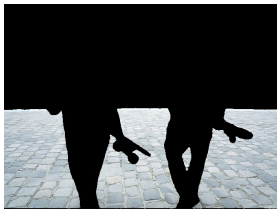


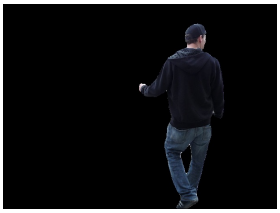
Figure 10. Analysis of MLLM's Rating and Binary Response Performance: For each mask-cropped region, we compare two types of outputs: numerical scores (score 1, score 2) and binary responses (d^{t2t}, d^{t2i}) (0='no', 1='yes'). The results reveal that MLLM struggles to generate meaningful similarity scores when comparing the input expression E and reference text T^{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^{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.



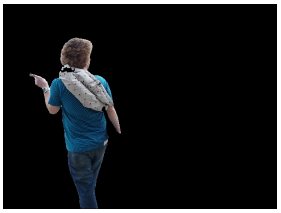
Candidate text T^{can}:
The gray ground region at the bottom of the image with two individual's silhouette walking on light-colored cobblestone pavement

0.234 / 0.241



Candidate text T^{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^{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^{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^{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^{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^{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^{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

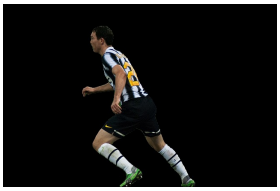
Candidate text T^{can}:
The blue car is parked near a restaurant on a cobblestone street, reflecting the bustling street life around it. In the foreground, a clear, sleek metallic surface meticulously mirrors the surrounding urban environment.

0.668 / 0.243

Reference text T^{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^{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

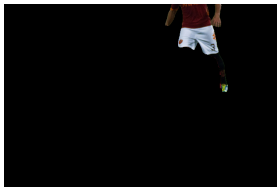


Candidate text T^{can}:
The athlete, dressed in a black and yellow uniform, is in motion on the soccer field.

0.539 / 0.313

Candidate text T^{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^{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^{t2t}) and text-to-image (s^{t2i}) 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.

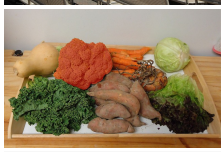
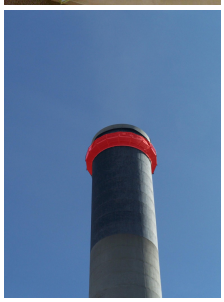
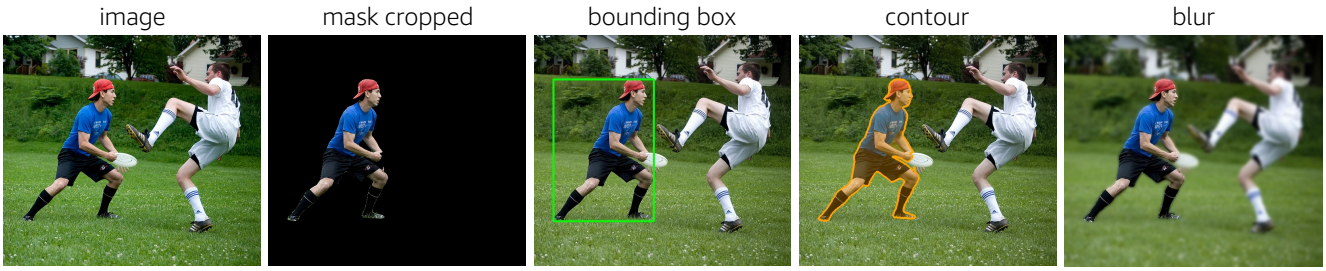
query	image	reference text	candidate text (selected as output)	output
blue and blue umbrella		In the center-right portion of the image, a large light blue umbrella with a darker blue underside stands out prominently, shading a group of seated individuals and contrasting with the surrounding multicolored umbrellas	A large beach umbrella with a two-toned blue canopy, featuring a darker blue center panel and lighter blue outer panels	
green color vegetable in between potato and carrot		The green vegetable located between the potato and carrots is a large head of broccoli, tightly clustered florets forming a rounded, textured dome of deep forest green color.	The vibrant head of broccoli with deep forest green 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 round, pink and white cupcake with colorful sprinkles, sitting on lower right corner of a black plastic tray.	A colorful donut covered in rainbow sprinkles, positioned in the lower right corner of the image, with a distinctive round shape and hole in the center.	
something that the animals are tied to		The target in the image is a wooden 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 tethering the animals.	A long, slender lance with a pointed tip and a wooden post stands prominently amidst a vibrant, bustling scene, flanked by two stately and ornately equipped horses in front of a distinguished building with blue walls and a terracotta roof.	
pink skirt		The pink skirt is worn by an older woman standing on the right side of the image, featuring a floral pattern 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 older woman wearing a beige jacket over a floral-patterned pink dress stands in the right 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 bookmark positioned between the pages of a book on the right side of the image, which stands out due to its narrow, rectangular shape and its placement marking a specific page.	A colorful bookmark lies on the right side of a book titled "Weekend Sewing". It is a small rectangular card featuring some patterns.	
the area where people can walk		The target region in the image is an elevated, circular platform with railings situated at the top of the tall, cylindrical tower, designed to be a designated area where people can walk and observe.	The metal grate at the top of the tall chimney appears rusted and slightly rounded, 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 right end of a white metal rod, featuring light to down.	A small, cylindrical, white plastic component with multiple \u03cd small holes along its length is positioned at the right end of a horizontal rod, 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 front of the car, specifically the white front bumper with black air vents and the rounded headlight area, located at the right of the image.	The front side of a white toy car, featuring to be the front bumper 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.

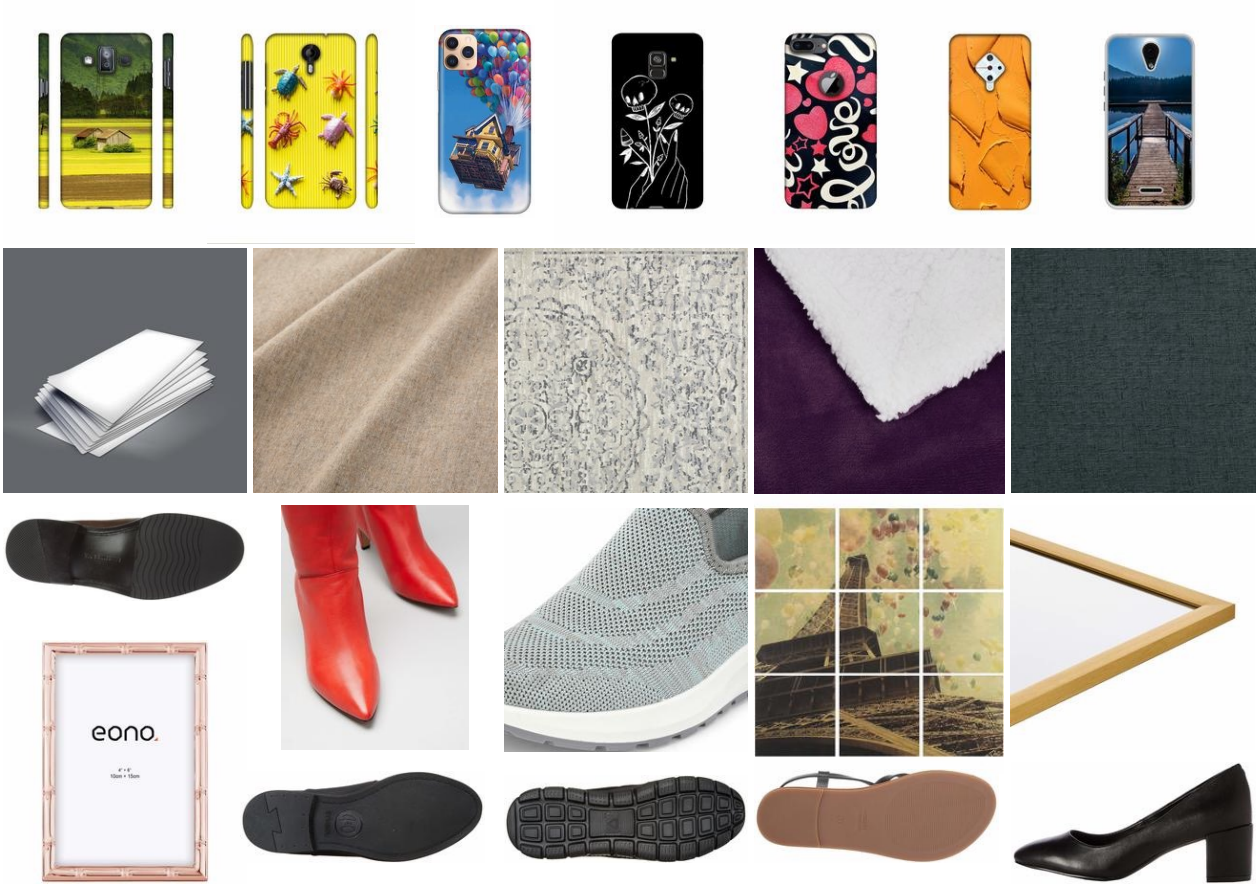


Figure 15. Images excluded from the ABO dataset typically lack meaningful or referrable parts. Row 1 shows phone related items that primarily consist of phone cases displaying only the back view of phones. Row 2 features images solely showing product textures or materials that fill the entire frame.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

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

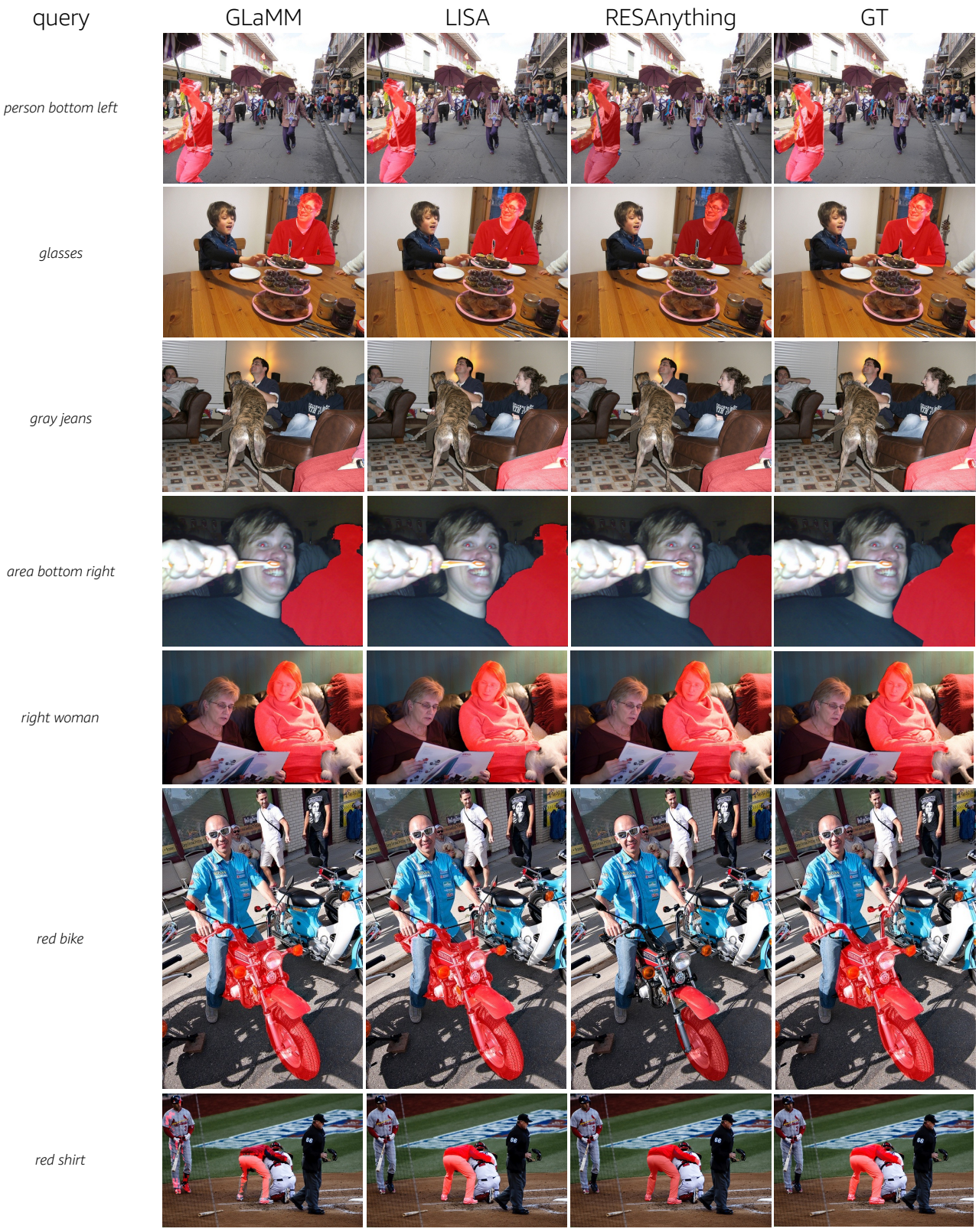


Figure 17. Qualitative results on RefCOCO test A (randomly selected). The ground truth annotations can be problematic - some queries refer only to an object/region while the GT marks an entire person (row 2: “glasses"; row 4: “area bottom right", row 7: “red shirt"). Row 6 shows a failure case of RESAnything.

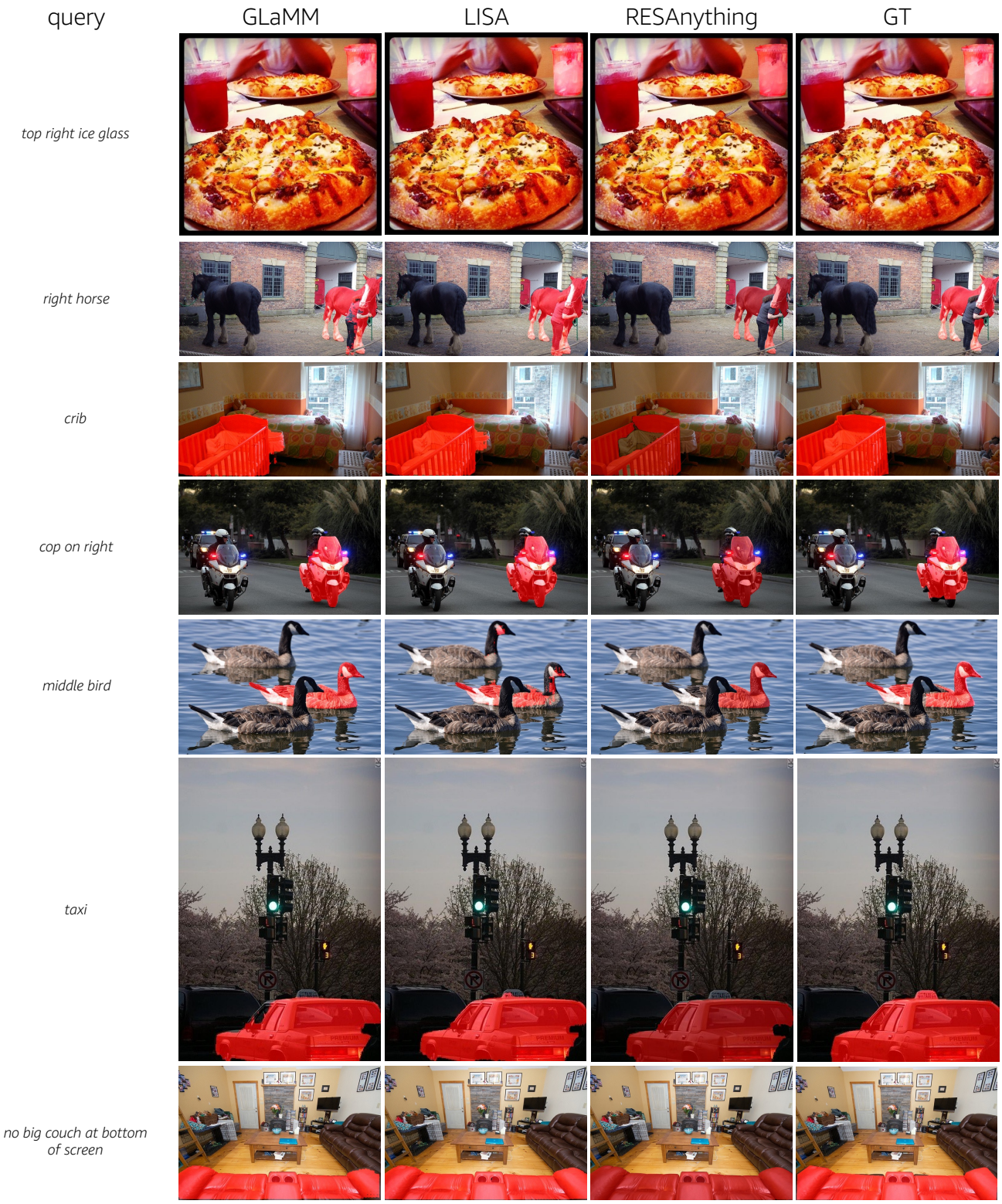


Figure 18. Qualitative results on RefCOCO test B (randomly selected). Despite being unsupervised, RESAnything achieves comparable results to supervised methods, particularly excelling at crowded regions (row 2). However, it occasionally misses parts when needing to combine multiple masks (row 3, 5).

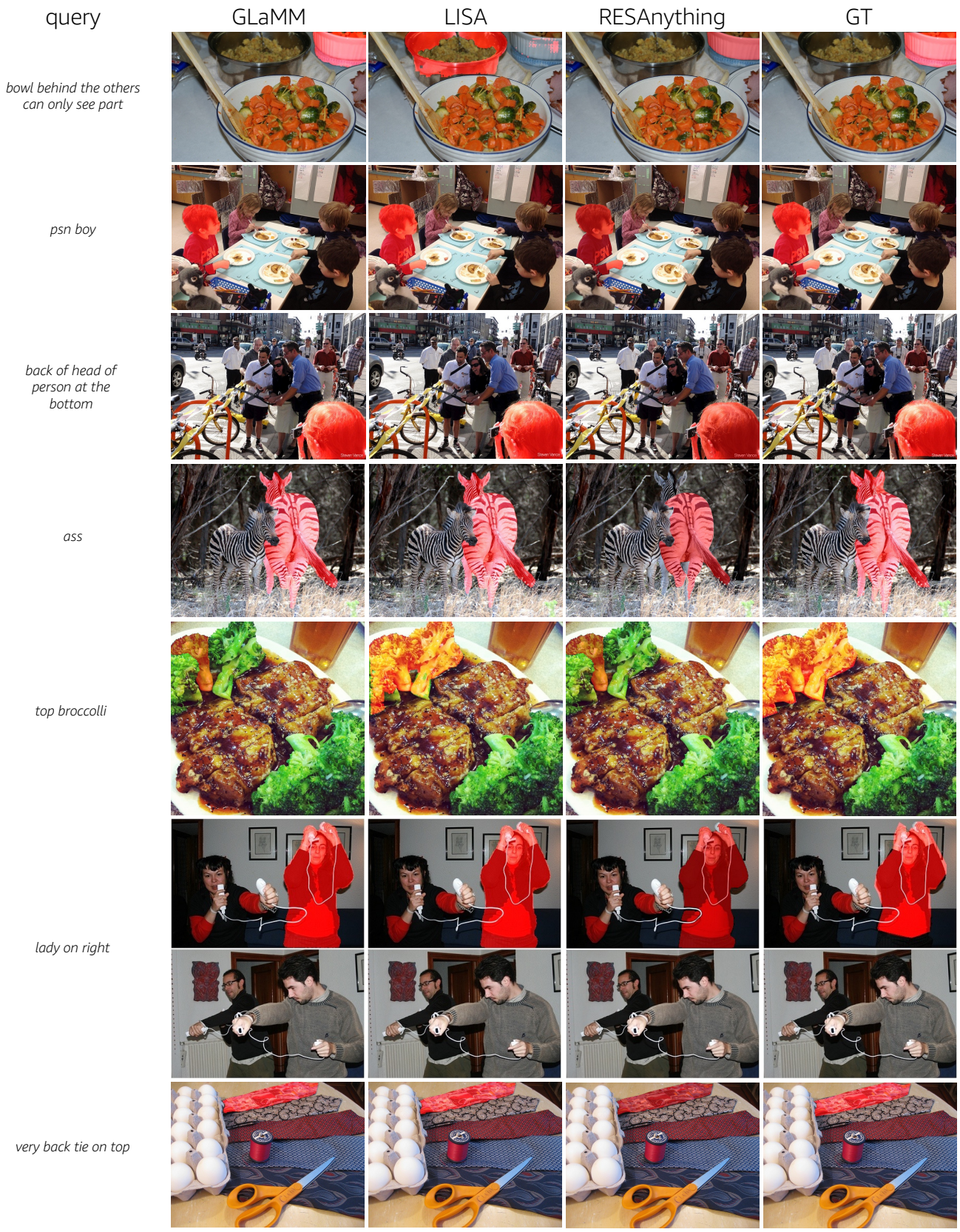


Figure 19. Qualitative results on RefCOCO val (randomly selected). RESAnything generates fine-grained segmentation of the input expression (row 1, 4). As mentioned in Fig 18, it may miss parts when combining multiple masks (row 5).

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).

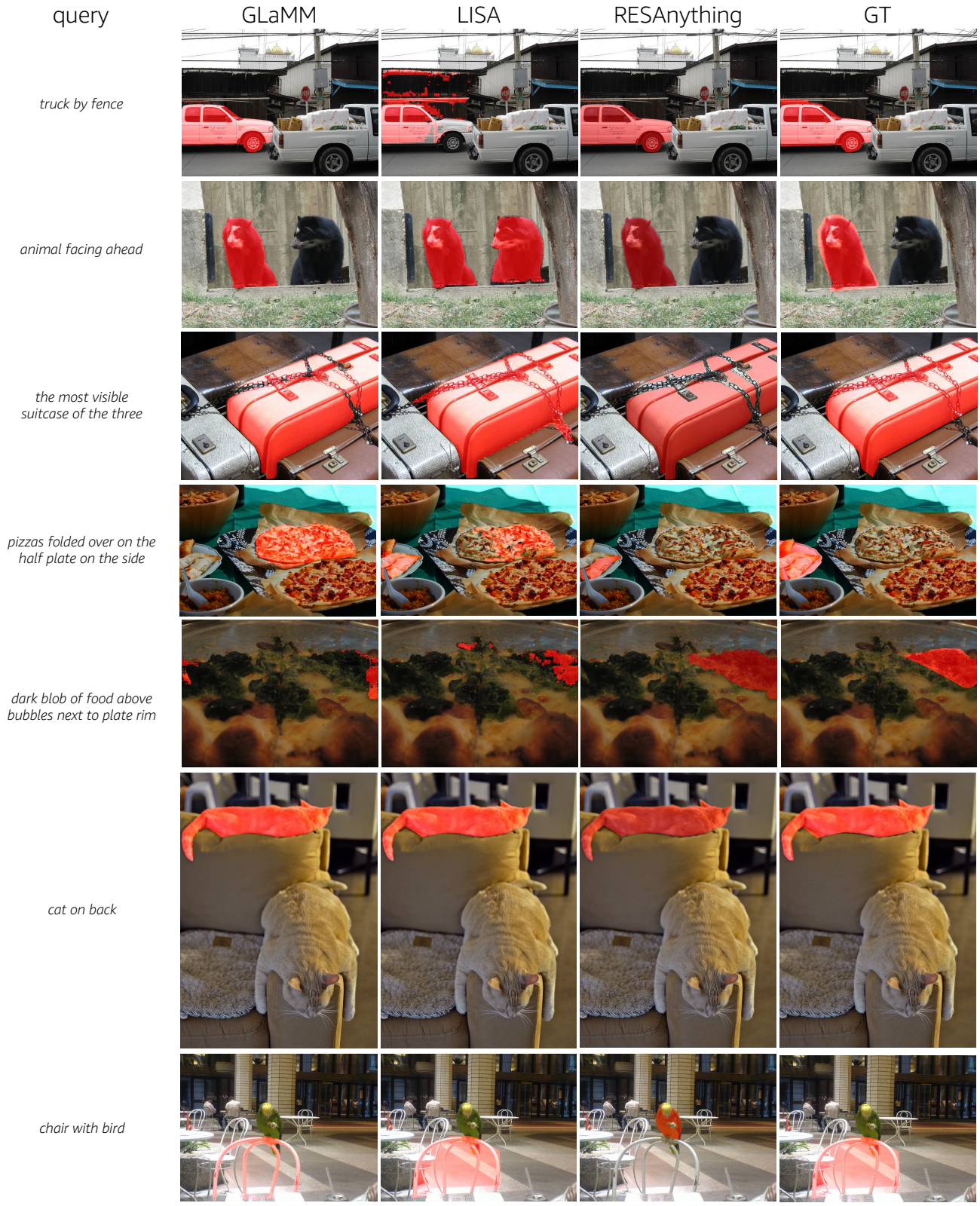


Figure 24. Qualitative results on RefCOCO+ test B (randomly selected).

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

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).

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

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

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

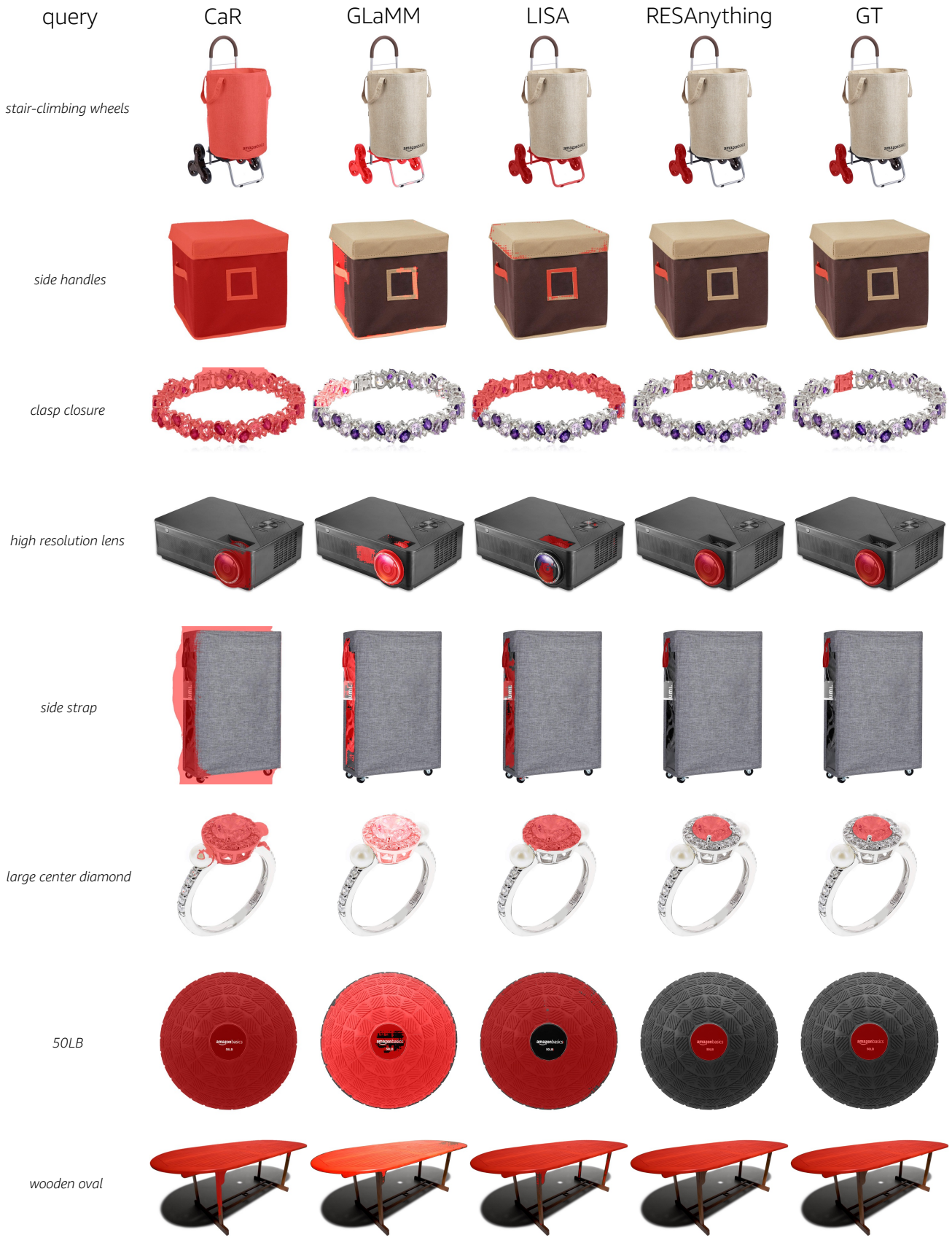


Figure 30. Qualitative results on ABO-Image-ARES (Part 2).

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

References

[1] Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet/. 2, 3, 8

[2] Google gemini. https://blog.google/technology/ai/google-gemini-ai/. 3

[3] Openai gpt-4o. https://openai.com/index/hello-gpt-4o/. 3

[4] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 2, 6, 8

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 3

[6] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 3

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3, 8

[8] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3

[9] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 3

[10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 3

[11] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 1, 3, 6, 7

[12] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26573–26583, 2024. 3

[13] 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. *CVPR*, 2022. 2, 3

[14] Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. *Advances in neural information processing systems*, 37:121670–121698, 2024. 3

[15] Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023. 1, 3

[16] 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 multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 3

[17] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[18] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 1, 7

[19] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971. 4

[20] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 1

[21] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 3

[22] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4067–4077, 2023. 5

[23] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*, 2023. 3

[24] 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)*, pages 787–798, 2014. 1, 3

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3, 4, 6

[26] 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*, pages 9579–9589, 2024. 2, 3, 6, 7, 8, 4, 5

[27] Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*, 2024. 2, 3

[28] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 3

[29] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: a multi-modal model with in-context instruction tuning. corr abs/2305.03726 (2023), 2023. 1

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[31] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021. 3

[32] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 1, 2, 7, 4, 5

[33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3

[34] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3

[36] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023. 3

[37] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[38] Sun-Ao Liu, Hongtao Xie, Jiannan Ge, and Yongdong Zhang. Refersam: Unleashing segment anything model for referring image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3

[39] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation

and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 2, 6

[40] Sayan Nag, Koustava Goswami, and Srikrishna Karanam. Safari: Adaptive s equence tr a ns f ormer for we a kly supervised r eferring expression segmentat i on. In *European Conference on Computer Vision*, pages 485–503. Springer, 2024. 3

[41] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 2, 3, 6

[42] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1, 3

[43] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14076–14088, 2024. 3

[44] 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*, pages 8748–8763. PMLR, 2021. 3

[45] 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*, pages 13009–13018, 2024. 2, 3, 6, 7, 4, 5

[46] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[47] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3

[48] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 3

[49] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024. 3

[50] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, 2018. 3

[51] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023. 3

[52] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13171–13182, 2024. 3, 5, 7, 4

[53] Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. Vrp-sam: Sam with visual reference prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23565–23574, 2024. 3

[54] Yucheng Suo, Linchao Zhu, and Yi Yang. Text augmented spatial-aware zero-shot referring image segmentation. *arXiv preprint arXiv:2310.18049*, 2023. 3

[55] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[56] Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12998–13008, 2024. 3, 7, 4

[57] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 1, 3, 7

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3

[59] Zhichao Wei, Xiaohao Chen, Mingqiang Chen, and Siyu Zhu. Linguistic query-guided mask generation for referring image segmentation. *arXiv preprint arXiv:2301.06429*, 2023. 3

[60] Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, and Ping Luo. Segment every reference object in spatial and temporal spaces. In *ICCV*, 2023. 1

[61] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *IEEE Transactions on Image Processing*, 2024. 5

[62] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A

Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024. 3

[63] Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E Gonzalez, and Trevor Darrell. See say and segment: Teaching lmms to overcome false premises. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13459–13469, 2024. 3

[64] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024. 3, 7

[65] 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*, pages 4818–4829, 2024. 3

[66] Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024. 3

[67] Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. Pixel-aligned language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13030–13039, 2024. 3

[68] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024. 3

[69] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 3

[70] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36:24993–25006, 2023. 3

[71] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 1, 3, 7

[72] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 3

[73] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019. 3

[74] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 3

[75] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 1, 3

[76] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2, 6

[77] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19456–19465, 2023. 1, 3, 7

[78] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211, 2024. 3

[79] Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023.

[80] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 3

[81] Yuxuan Zhang, Tianheng Cheng, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. 3

[82] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. 3

[83] Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, and Wei Ke. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. *Advances in Neural Information Processing Systems*, 35: 14729–14742, 2022. 3

[84] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 3

[85] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024. 3

[86] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 3

[87] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3