RESAnything: Attribute Prompting for Arbitrary Referring Segmentation

Ruiqi Wang **Yiming Qian** **Kai Wang** **Fenggen Yu** **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 $\sim$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 [16,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

20, 28, 36]. Recent advancements in LMMs have enabled sophisticated understanding of visual content, from object detection to semantic segmentation [5, 10, 43]. 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 [19, 60, 73, 33, 79, 23, 77, 63, 12].

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 buyers 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) [33] 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 [24] 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 [25, 46, 26] 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 [25], 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 [4, 1], where the proposals are produced by a foundational image segmentation model such as SAM [24]. 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 expressions and challenging implicit queries for arbitrary RES.

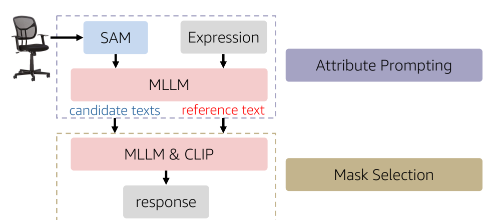

Figure 2: Overview of RESAnything: a two-stage framework for zero-shot arbitrary RES. The attribute prompting stage generates reference and candidate texts from input image and referring expression using SAM-generated proposals and an MLLM. The mask proposal selection stage leverages MLLM and CLIP to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

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 [14], 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+ [78], RefCOCOg [40, 42]. Our method also significantly outperforms existing methods on the recent reasoning segmentation dataset ReasonSeg [25], 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 [25] and GLaMM [46], 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 [27] particularly developed for effective in-context instruction tuning. InstructBLIP [16] built upon the Querying Transformer as in BLIP2 [29]. The LLaVA models [36, 34] and MiniGPT4 [89] utilized a lightweight MLP and achieved appealing performances in various MLLM benchmarks. Recent developments include supporting high-resolution image inputs [70, 35, 83], optimizing model efficiency [7, 76, 87], and constructing higher-quality datasets [11, 17].

2.1 Open-Vocabulary and RES

RES [23, 42, 21] 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 [18, 45] are employed to extract textual embeddings, which are then integrated into segmentation architectures through cross-attention or feature alignment [13, 52, 62, 75, 85, 67] to achieve language-aware segmentation [30, 73, 60, 37, 59, 68, 31, 32]. Recently, SAM [24] has introduced text-guided segmentation [84, 39, 12]. For instance, Grounding-SAM [48] leverages bounding boxes returned by Grounding-DINO [38] to prompt SAM for mask prediction, while Fast-SAM [86] utilizes CLIP similarity scores [45] 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 [81, 12, 26, 77, 80, 58, 10, 43, 82, 69, 15, 44]. One of the pioneering works in this area is LISA [25], which enables MLLMs to segment objects by using text embeddings from LLaVA to prompt a SAM [24] 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 [66, 65] and grounded segmentation [46, 49], 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 [79, 22, 56, 88, 54]. MaskCLIP obtains pseudo masks by modifying the last attention layer of CLIP [88]. CaR couples CLIP and GradCAM to generate mask proposals, then

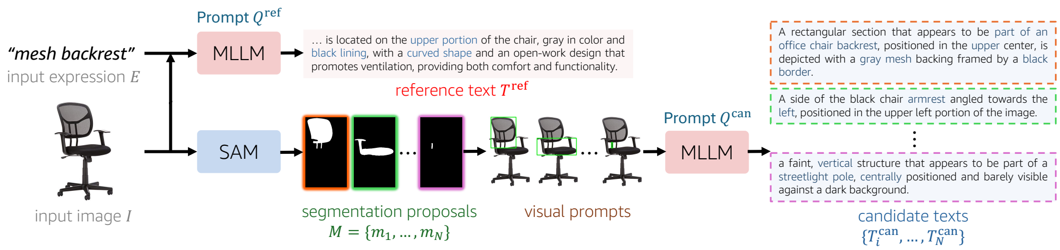

Figure 3: Attribute prompting using SAM and MLLM. Given the input image and referring expression, this stage produces two groups of predictions. The first output, a reference text T^{ref}, is generated from an MLLM with the text prompt Q^{ref}. It describes the visual attributes (e.g., color, shape, 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.

employs a CLIP classifier to select the final masks, before a mask refinement [54] in post-processing. Global-Local CLIP [79] 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 [50] 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 [61]. Recently, visual prompting [72, 41, 53, 55] has been proposed to enhance the adaptation of CLIP for open-vocabulary segmentation by overlaying ovals over segmentation targets [54]. SAM [24], on the other hand, allows users to provide points, boxes, masks as prompts for image segmentation, with the latest version supporting video segmentation [47]. Visual prompting has also been applied to MLLMs [64]. Overlaying image regions with bounding boxes, masks, circles, scribbles, etc has enhanced MLLMs' ability to perform region or pixel-level image understanding [74, 71, 8].

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 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.

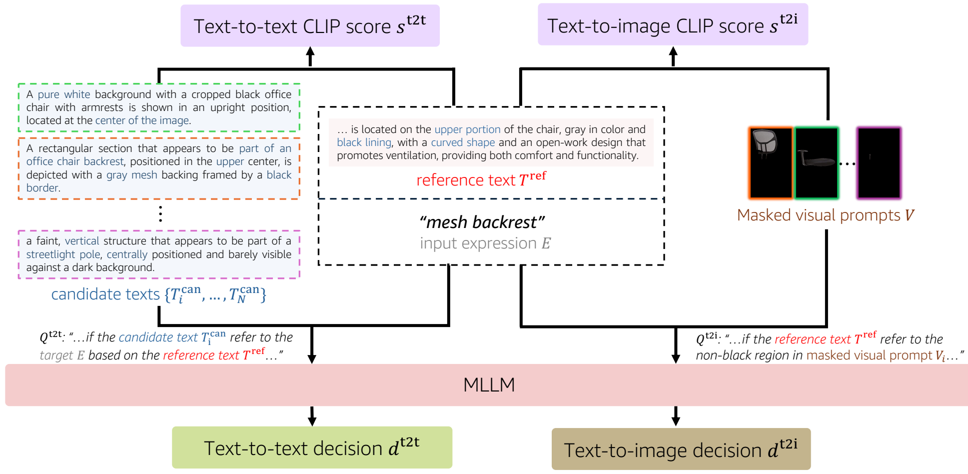

Figure 4: Multi-metric mask proposal selection using MLLM and CLIP. To select the final mask from mask proposals generated by SAM, we introduce four metrics computed across different modalities and models to evaluate the similarity between input expression E and the mask proposals. Specifically, the 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.

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 [24] 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.

Given a mask proposal m_i, we generate a corresponding candidate text $T_i^{\text{can}} = f_{\text{MLLM}}(V_i^1, V_i^2 \ldots V_i^K \mid Q^{\text{can}})$ using an MLLM, where Q^{can} is the candidate text prompt that similarly asks for visual attributes such as shape, color and location; and $V_i^1 \ldots V_i^K$ 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 5 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 [54]. The complete candidate text prompt Q^{can} is provided in the supplementary. Fig 3, right part shows examples of generated candidate texts.

Figure 5: Example of different visual prompts V_i generated from a segmentation proposal m_i.

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 4. 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 4.

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.

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 [19]	67.5	70.5	65.2	56.3	61.0	50.1	55.0	-	57.7
CRIS [60]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	-	60.4
LAVT [73]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	-	62.1
GRES [33]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	-	66.0
pre-trained on the same task									
UniRES [59]	71.2	74.8	66.0	59.9	66.7	51.4	62.3	-	63.2
LISA-7B [25]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	-	70.6
GSVA [66]	77.2	78.9	73.5	65.9	69.6	59.8	72.7	-	73.3
GLaMM [46]	79.5	**83.2**	**76.9**	72.6	78.7	64.6	74.2	-	74.9
SAM4MLLM [12]	**79.8**	82.7	74.7	**74.6**	**80.0**	**67.2**	**75.5**	-	**76.4**
training-free zero-shot									
GLCLIP [79]	26.2	24.9	26.6	27.8	25.6	27.8	33.5	33.6	33.7
CaR [54]	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**

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.

Algorithm 1 Grouping and Selection Process

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

4 Experiment

We use Pixtral 12B [4] as the MLLM, SAM ViT-H [24] 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 [25, 12], we evaluate the performance of RESAnything on four public benchmark datasets: RefCOCO, RefCOCO+ [78], RefCOCOg [40, 42] and ReasonSeg [25]. 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.

Table 2: Quantitative results on ReasonSeg (left) and ABO-Image-ARES(right).

Method	val	
	gIoU	cIoU
GLaMM [46]	47.4	47.2
LISA-7B-LLaVA1.5 [25]	53.6	52.3
LISA-13B-LLaVA1.5 [25]	57.7	60.3
SAM4MLLM [12]	58.4	60.4
RESAnything	74.6	72.5

Method	test	
	gIoU	cIoU
LISA-13B-LLaVA1.5 [25]	43.3	34.0
GLaMM [46]	46.2	38.7
RESAnything	78.2	72.4

Evaluation metrics. We evaluate our method using two standard metrics following prior works [25, 46]: 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 gIOU for RefCOCO, RefCOCO+, and RefCOCOg, and both metrics for ReasonSeg and ABO-Image-ARES.

4.1 Evaluation on Vanilla RES

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

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 [59] 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 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 (See Fig 8).

4.3 Evaluation on Affordance-based RES

We further evaluate our method on affordance-based referring expression segmentation tasks, as shown in Table 3. All compared methods are fully-supervised and fine-tuned on COCO-Tasks training data, while RESAnything operates in a zero-shot manner. As seen above, our method surpasses TOIST and TaskCLIP, demonstrating its strong generalization capability across affordance-based scenarios. When we incorporate task-specific prompt optimization, e.g., adding "human-object interaction" attributes, the performance improves from 51.2 to 54.6, approaching CoTDet. This demonstrates how prompt engineering can enhance the performance of our framework.

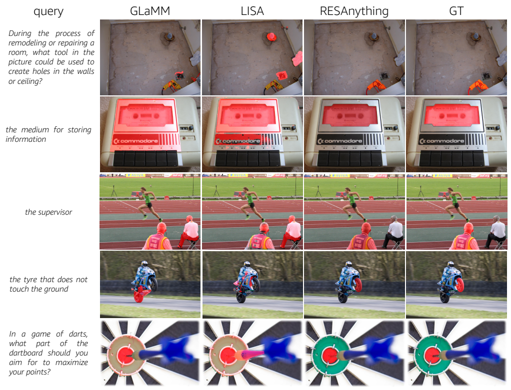

Figure 7: Qualitative comparisons on ReasonSeg. Our method demonstrates superior performance in both object localization accuracy (rows 1, 3, 4) and segmentation precision (rows 2, 5).

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

Table 3: Results on COCO-Tasks(mIoU)

Method	mIoU@0.5 (14 tasks average)
supervised (trained on training set)	
GGNN [51]	32.4
TOIST [31] (w distillation)	44.1
Taskclip [9]	50.3
CotDet [57]	56.9
training-free zero-shot	
RESAnything	51.2
RESAnything w prompt optimization	54.6

4.4 Ablation Study

Attribute prompts. Our core contribution is attribute prompting, a novel mechanism that emphasizes reasoning about part attributes to handle implicit queries and complex part-level relationships. To validate its effectiveness, we compare attribute prompting against conventional prompting baselines on the ReasonSeg test set, as shown in Table 4.

Table 4: Ablation study comparing prompting strategies on ReasonSeg test set.

Method	gIoU	cIoU
Standard prompt	50.8	49.3
Attribute prompt	**74.6**	**72.5**

Visual prompts. As shown in Fig 5, we explore different types of visual prompts for generating candidate texts T^{can} and performing text-to-image comparison. Table 5 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. 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.

Table 5: Ablation studies on visual prompts (left) and MLLM backbone (right).

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

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

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.

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 [25] 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.

References

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

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

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

[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.

[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.

[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.

[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.

[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.

[9] Hanning Chen, Wenjun Huang, Yang Ni, Sanggeon Yun, Yezi Liu, Fei Wen, Alvaro Velasquez, Hugo Latapie, and Mohsen Imani. Taskclip: Extend large vision-language model for task oriented object detection. In *European Conference on Computer Vision*, pages 401–418. Springer, 2024.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[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.

[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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Pengfei Li, Beiwen Tian, Yongliang Shi, Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Toist: Task oriented instance segmentation transformer with noun-pronoun distillation. *Advances in Neural Information Processing Systems*, 35:17597–17611, 2022.

[32] Yang Li, Xiaoxue Chen, Hao Zhao, Jiangtao Gong, Guyue Zhou, Federico Rossano, and Yixin Zhu. Understanding embodied reference with touch-line transformer. In *ICLR*, 2023.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

[40] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] Johann Sawatzky, Yaser Souri, Christian Grund, and Jurgen Gall. What object should i use?-task driven object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7605–7614, 2019.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

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

[57] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibei Yang. Cotdet: Affordance knowledge prompting for task driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3068–3078, 2023.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

6 NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Contributions and scope of the paper are clearly claimed in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Answer: [Yes]

Justification: Implementation details are presented in the main paper and supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Code and data are not included in the submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Implementation details including prompts used are presented in the main paper and supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are presented with details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details of computer workers are presented in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research was conducted win a paper conform in every respect, with the NeurIPS conde of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper introduces a method for referring expression segmentation, which has no societal impact regarding the negative presented in the guidelines.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper cite and state the original paper and dataset clearly.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper proposed a new dataset and benchmark in details.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: This paper leverages the power of LLMs to reason the relationship between targets in the images, which support the task in segmentation.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.