# VISION HARNESSING AGENT FOR OPEN AD-HOC SEGMENTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Segmentation has become easy when the concept is known, requiring retrieval of a learned visual grounding from text. It remains hard for open ad-hoc concepts, where the grounding may not exist as one learned mask and must often be constructed from image evidence through parts, relations, exclusions, and collections. We propose a Vision-guided Ad-hoc Segmentation Agent (VASA), the first vision harnessing agent for open ad-hoc segmentation. VASA is training-free and couples a VLM agent, a segmentation foundation model, and a visual harness that maintains a working mask to make visual progress persistent, inspectable, and editable. Rather than revising text prompts alone, it plans visual operations, invokes segmentation tools, inspects results, edits the mask, and recovers from errors. We construct PARS, a new benchmark that turns part-level labels into open ad-hoc concepts through long-form definition queries. We show that VASA is consistently effective across six VLMs with varying capabilities. Using Qwen3-VL 32B Thinking as the VLM, VASA outperforms various baselines on PARS, surpassing SAM3 Agent by 13.5%–25.3%. On RefCOCOm, VASA improves over SAM3 Agent by 4.8%–8.8% and over other agentic baselines by more. VASA also remains competitive with SAM3 Agent on ReasonSeg for common, named concepts. These results validate VASA's agentic visual construction for open ad-hoc segmentation.

## 1 INTRODUCTION

Segmentation has become easy when the concept is known. Modern foundation models can segment familiar visual wholes, such as *cats* and *cars*, because these concepts are visually coherent, commonly named, visually grounded, and richly represented in training data. SAM3 (Carion et al., 2026) builds on the success of SAM (Kirillov et al., 2023) and SAM2 (Ravi et al., 2024) by using large-scale text-region supervision, achieving strong performance across diverse segmentation benchmarks.

However, segmentation remains hard when the concept is open and ad-hoc (Fig. 1). A user may ask not for a named concept, but for an arbitrary one involving parts, relations, exclusions, or collections, constructed on the fly for the purpose at hand. We call this setting *open ad-hoc segmentation*: Segmenting open-ended, task-contingent visual concepts beyond retrieval of established concepts. This setting extends *open ad-hoc categorization* (Wang et al., 2025), which contextualizes recognition on demand, but imposes a stricter demand: The ad-hoc concept must be grounded at the pixel level.

Large-scale language and vision-language models (LLM/VLMs) have shown strong capabilities in text-conditioned visual reasoning, making it possible to pair segmentation foundation models with language-based reasoning beyond direct concept matching. One direction finetunes VLMs to produce intermediate segmentation representations, such as the `[SEG]` token in LISA (Lai et al., 2023) and geometric prompts in Seg-Zero (Liu et al., 2025). Another direction uses VLM agents to decompose hard tasks into executable steps, treating an external segmenter as a tool. The agent interprets a complex query, proposes geometric prompts as in SegAgent (Zhu et al., 2025b) or text prompts as in SAM3 Agent (Carion et al., 2026), examines the resulting masks, and tries again. Their reasoning advances through revised prompts, but *their visual solution does not build up accordingly* (Fig. 1, 2).

This disconnect between textual reasoning and visual progress hinders their performance on open ad-hoc segmentation, but does not necessarily imply that the models lack the required perceptual or reasoning capabilities. They may instead lack an effective harness to guide their use. This motivates our research question: *How can we harness VLM agents for effective open ad-hoc segmentation?*



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

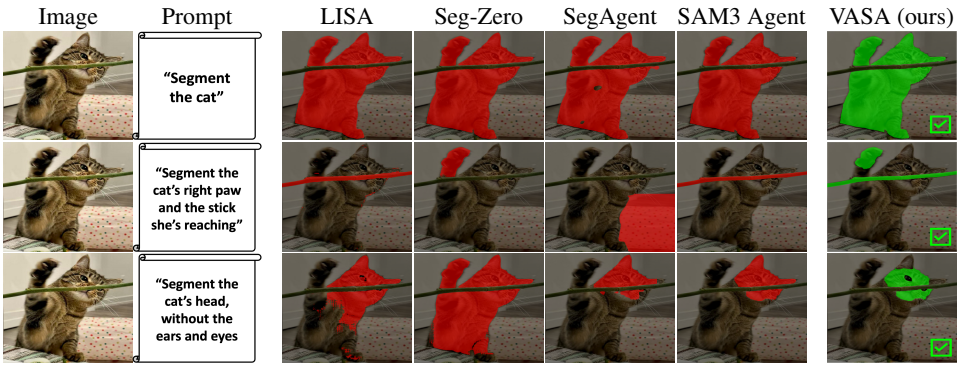



Figure 1: **Open ad-hoc segmentation requires constructing visual concepts on the fly, not merely retrieving common ones.** We compare LISA (Lai et al., 2023), Seg-Zero (Liu et al., 2025), SegAgent (Zhu et al., 2025b), SAM3 Agent (Carion et al., 2026), and VASA (our proposed vision harnessing agent), on three prompts for the same image. **Row 1** asks for a known visual whole, *cat*; all methods succeed. **Row 2** asks for an ad-hoc collection, *the cat's right paw and the stick she is reaching for*; baselines retrieve only one part of the requested composition, while VASA captures both. **Row 3** asks for an exclusion-defined concept, *the cat's head without the ears and eyes*; baselines confuse it with *cat head* or *cat*, while VASA constructs the requested mask.

In this work, we introduce a *Vision-guided Ad-hoc Segmentation Agent* (**VASA**), the first vision harnessing agent for open ad-hoc segmentation. VASA is a training-free framework that couples a VLM agent, a segmentation foundation model, and a visual harness. Unlike the previous state-of-the-art, SAM3 Agent (Carion et al., 2026), which retains textual reasoning history but not the evolving visual construction, VASA maintains a working mask that *makes visual progress persistent, inspectable, and editable*. The agent can add newly found regions, remove incorrect regions, replace coarse masks with refined ones, and verify progress against the user query. Unlike SAM3 Agent that are built with prompt and context engineering, VASA is built with a complete visual harness over state management, tool invocation, action constraints, planning, scrutiny, and error recovery.

Fig. 2 compares the workflows of VASA and SAM3 Agent. The side-by-side comparison reveals the key difference: SAM3 Agent searches over prompts, advancing text reasoning without accumulating visual progress, whereas VASA incrementally constructs a visual solution. Text reasoning and visual construction proceed together, enabling VASA to successfully segment this open ad-hoc concept.

For evaluation, we construct **PARS**: *PartImageNet Ad-hoc Referring Segmentation*, an open ad-hoc segmentation benchmark that turns part-level labels in PartImageNet (He et al., 2022) into open ad-hoc concepts through long-form definition queries. Starting from annotated examples, we prompt a VLM to write segmentation instructions that specify what to include, what to exclude, and how the concept is structurally or contextually identified; the generated queries are then manually verified. Instead of short, often ambiguous category names, PARS provides detailed, unambiguous queries.

We first assess VASA's generalization across six VLMs spanning open-source and proprietary model families with varying capabilities. Our results show that VASA works effectively on older VLMs and scales well to the latest VLMs. On PARS, VASA consistently outperforms open-vocabulary, reasoning-based, and agentic segmentation baselines, surpassing SAM3 Agent by 13.5%-25.3% using Qwen3-VL 32B Thinking (Bai et al., 2025). On the multi-granularity referring segmentation benchmark RefCOCOm (Wang et al., 2024c), VASA outperforms SAM3 Agent by 4.8%-8.8% and other agentic baselines by more. We further present ablation studies, additional analysis, qualitative results, and show that VASA is still on par with SAM3 Agent on ReasonSeg (Lai et al., 2023), a standard reasoning segmentation benchmark on common, named concepts.

**Contributions: 1)** We introduce open ad-hoc segmentation, a challenging segmentation setting for open-ended, ad-hoc visual concepts. **2)** We develop VASA, the first vision harnessing agent for open ad-hoc segmentation, with persistent visual construction and complete visual harness. **3)** We curate PARS, a benchmark that uses detailed long-form queries to evaluate open ad-hoc segmentation. **4)** We show that VASA generalizes across VLMs, significantly outperforms the state of the art on PARS and RefCOCOm, and outperforms SAM3 Agent on ReasonSeg.

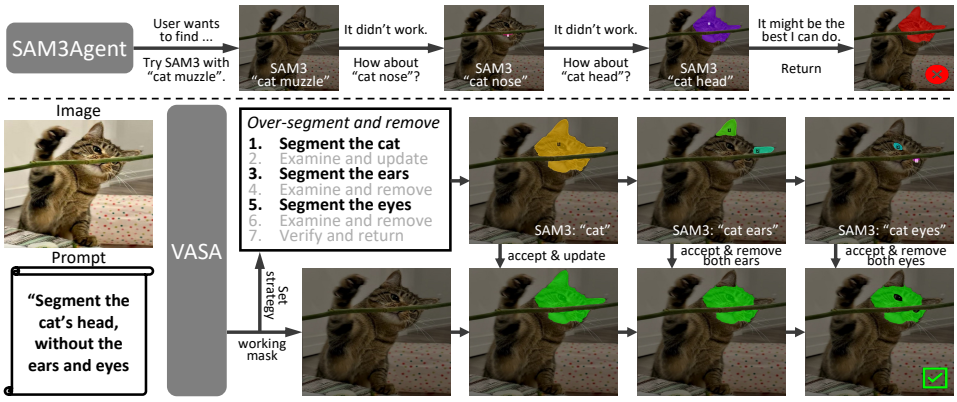



Figure 2: **VASA reasons, constructs, and validates an open ad-hoc solution, while SAM3 Agent searches over text prompts.** We consider the query *the cat's head, without the ears and eyes*. **Top:** SAM3 Agent revises text queries to SAM3, trying prompts like *cat muzzle*. Each trial is a fresh retrieval attempt: Intermediate masks are not preserved for the final answer. **Bottom:** VASA instead plans a visual strategy, maintains a persistent working mask, and updates it via executable operations.

## 2 Related Work

Additional related work on **referring segmentation** and **reasoning segmentation** is in Appendix A.

**Referring segmentation** focuses on segmenting image regions specified by explicit text descriptions. Much of its success comes from scaling region-text alignment supervision, like the semi-automatic data engine in SAM3 (Carion et al., 2026). Such scaling works well for visually coherent, commonly named objects, but cannot exhaust the space of open ad-hoc concepts users may define on demand.

**Reasoning segmentation** focuses on segmenting typically well-defined targets from *implicit* text queries through commonsense and spatial reasoning. Most existing works use VLMs for reasoning and segmenters for mask decoding. Three types of reasoning paradigms prevail. **1)** Implicit reasoning, where the VLM predicts the target segment representation, keeping the reasoning process inaccessible, such as the `[SEG]` token in LISA (Lai et al., 2023). **2)** Text-based visual reasoning, where the VLM is tuned through RL (Shao et al., 2024b) to reason in text, such as the target's coordinate prediction in Seg-Zero (Liu et al., 2025). **3)** Grounded visual chain-of-thought, where each reasoning step is visually grounded, and subsequent steps build on both prior reasoning and grounding results, such as SegLLM (Wang et al., 2024d). Open ad-hoc segmentation is an orthogonal direction, reasoning about what constitutes a visual concept rather than how to find the existing concept hinted by the user.

**Agentic segmentation** pairs VLMs with segmenters through *iterative tool use*. Beyond reasoning segmentation, they also refine how the target is delineated or composed. Prior works like CoReS (Bao et al., 2024) decomposes segmentation into coarse-to-fine stages, while SegAgent (Zhu et al., 2025b), RSAgent (He et al., 2025), and SAM Veteran (Du et al., 2026) use finetuned VLMs to iteratively refine geometric prompts. Evol-SAM3 (Ye et al., 2025) and SAM3 Agent (Carion et al., 2026) instead operate without finetuning, using evolutionary search and sequential revision of text prompts, respectively. GuidelineSeg (Vats et al., 2025) considers a related setting in which the targets are defined by detailed guidelines, with a pipeline designed around only driving scenes that coordinates multiple pretrained models through Worker-Supervisor refinement and an RL-based stopping policy. Our PARS broadens it with a more diverse benchmark of object and part compositions. These methods approach open ad-hoc segmentation but remain limited by disconnected textual reasoning and visual progress. In contrast, VASA is the first vision harnessing agent to make visual progress persistent, inspectable, and editable, integrating other harness components for open ad-hoc segmentation.

**Harness engineering** (Hashimoto, 2026; OpenAI, 2026c) designs the full environment around an agent for reliable execution. It extends prompt and context engineering to govern actions, constraints, verification, and error recovery. Beyond software engineering, it has been used in visual games in VISTA (Han et al., 2026), robot policy training in HARBOR (Li et al., 2026b) and ENPIRE (Xiao et al., 2026), and robot task execution in Thea (Wang et al., 2026b) and HarnessVLA (Zhang et al., 2026). VASA is the first vision harnessing agent for open ad-hoc segmentation.

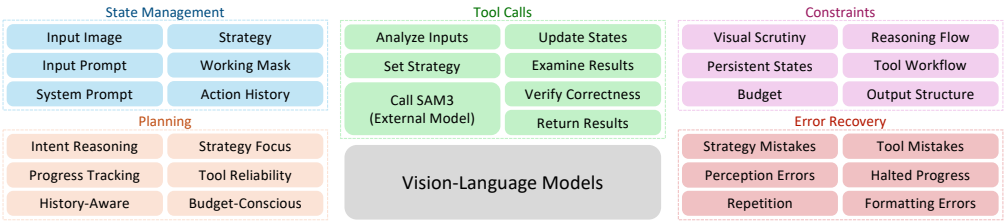



Figure 3: **VASA provides a complete visual harness around VLMs for open ad-hoc segmentation.** The five harness components form the full environment that supports steady visual progress, enabling complex open ad-hoc concepts to be progressively constructed from simpler visual primitives. The harness structures execution while leaving the VLM free to choose strategies, prompts, and edits.

## 3 VISION HARNESSING AGENT

Open ad-hoc segmentation calls for visual construction in which intermediate masks support further reasoning and refinement. We develop VASA (Vision-guided Ad-hoc Segmentation Agent) through a visual harness that preserves this visual progress and guides subsequent actions toward satisfying the user query, allowing reasoning and construction to advance together (Fig. 3).

**1. Overview.** VASA couples a VLM agent with a segmenter through a visual harness. The segmenter provides pixel-level grounding, returning candidate masks for the VLM's prompts. The VLM interprets the query, sets a strategy, chooses prompts, inspects candidate masks, and decides how to compose them into the target. The harness executes and constrains these decisions, maintains visual state, and guides planning, inspection, and recovery. Each result informs the VLM's next decision.

**2. State Management.** Exploring new regions should not erase existing visual progress. We therefore maintain an initially empty working mask separately from the latest segmentation candidates. Segmentation calls refresh the candidates without committing them to the solution; only explicit edits change the working mask, which is rendered on the image after each edit for inspection. We retain the image, query, system instructions, strategy, and compact action history to help the agent track what has been achieved and what remains unresolved in the evolving construction.

**3. Tool Calls and Mask Editing.** Visual construction requires translating the VLM's decisions into executable operations. We expose these operations through structured tool calls, spanning input analysis, strategy recording, segmentation, state updates, inspection, verification, and final output. Following SAM3 Agent, `segment_phrase` prompts SAM3 with a short noun phrase and returns numbered candidate masks, while `examine_each_mask` supports closer inspection. Our `update_working_mask` executes VLM-selected ADD, REMOVE, or REPLACE edits as deterministic pixel-level operations, adding or subtracting regions, or overwriting the working mask. Candidates thus need not match the entire query. Additional tools are introduced below.

**4. Long-Horizon Planning.** Visual construction is a long-horizon planning problem in which intermediate masks may only partially satisfy the query. Setting a strategy beforehand gives these intermediate results a clear role and coordinates subsequent steps toward the target. We instruct the agent to begin with `set_strategy`, analyzing the image and query to plan direct segmentation, undersegment-and-add, or oversegment-and-remove. The plan specifies initial prompts and subsequent edits, including preserving a base mask before seeking additions or exclusions. We retain the plan throughout inference to guide decisions alongside action history, tool reliability, progress, and the remaining budget, while allowing revisions when intermediate results reveal limitations.

**5. Constraints and Visual Scrutiny.** Open ad-hoc segmentation requires deliberate observation and reasoning, supported by constraints that guide the VLM without prescribing every decision. We encourage close inspection before action by instructing the agent to examine fine-grained boundaries, attributes, and relations against the query, alongside the input image and working mask. For small or overlapping masks, `examine_each_mask` provides individual views for closer scrutiny. To carry these observations into actions, we design workflow instructions to guide reasoning, preserve accumulated state, restrict selection to current candidates, and discourage unnecessary replacement. We constrain tool use through structured output requirements and execution checks, leaving the VLM free to choose and revise strategies, prompts, and edits as visual evidence accumulates.

**6. Error Recovery and Stopping.** In a long construction sequence, a failed step can mislead subsequent actions. We therefore design the harness to support recovery while preserving accumulated progress, rather than restarting the construction. When inspection reveals incorrect regions or stalled progress, VASA can revise its strategy, change prompts or concept granularity, and correct the working mask. Our execution checks return corrective feedback for repeated prompts, formatting errors, and invalid tool calls, enabling retries within the existing interaction. We instruct the agent to stop when the query is satisfied or further refinement appears unlikely to help, and impose a generation budget to limit execution. The harness returns the current working mask when available.

The **VASA inference algorithm** and **implementation details** are provided in Appendix B and C.

## 4 Experiments on Open Ad-hoc Segmentation

We evaluate VASA's visual harness on three settings: **1)** Our new PARS benchmark for open ad-hoc segmentation across six VLMs spanning different model families of various capabilities; **2)** The standard benchmark RefCOCOm for a related task of fine-grained multi-granularity referring segmentation; **3)** Ablation studies, further analysis, qualitative results, and results on ReasonSeg.

**1. Datasets.** PartImageNet Ad-hoc Referring Segmentation (**PARS**) is a new benchmark we construct from PartImageNet (He et al., 2022) for open ad-hoc segmentation. PartImageNet provides part annotations for different objects, such as the *"head/arm/foot/tail/body"* of a *"gorilla"* and the *"body/tire/side mirror"* of a *"car"*. We focus on the *"body"* class because it is shared across objects but defined differently, making it ad-hoc and challenging. We replace the original short class labels with detailed long-form descriptions that precisely specify how the target concept should be constructed (Fig. 4). We then divide them into 937 *open ad-hoc concepts* and 1,084 *common concepts* across 2,021 images based on VLM's familiarity. We provide more details on PARS construction in Appendix D.




Figure 4: **PARS provides detailed long-form descriptions for open ad-hoc concepts via a semi-automatic pipeline.** Given a small set of annotated examples for each concept, we prompt a VLM to disambiguate the target concept and what to include or exclude, followed by manual correction.

We additionally evaluate on RefCOCOm (Wang et al., 2024c) and ReasonSeg (Lai et al., 2023). RefCOCOm is a multi-granularity referring segmentation benchmark containing both part- and object-level concepts. ReasonSeg is a reasoning segmentation benchmark with implicit and long queries for typically otherwise common concepts. For both datasets, we use all validation and test splits and keep the original text prompts without further processing, unlike in PARS. This evaluation assesses whether our harness generalizes beyond open ad-hoc segmentation to common fine-grained part- and object-level localization and maintains its reasoning ability on common concepts.

**2. Evaluation metrics.** Following common practice, we adopt generalized Intersection-over-Union (gIoU) and cumulative Intersection-over-Union (cIoU), and introduce **xIoU** for evaluating *cross-concept confusion* in datasets like PARS, where all relevant parts of objects are annotated.

Formally, given predicted mask $P_i$, ground-truth target mask $G_i$, and $O_i$, the union of all other annotated concept masks in the same image excluding the target concept, the metrics are defined as:

$$\text{gIoU} = \frac{1}{N}\sum_i \frac{|P_i \cap G_i|}{|P_i \cup G_i|}, \quad \text{cIoU} = \frac{\sum_i |P_i \cap G_i|}{\sum_i |P_i \cup G_i|}, \quad \text{xIoU} = \frac{1}{N}\sum_i \frac{|P_i \cap O_i|}{|P_i|}. \tag{1}$$

Intuitively, xIoU measures how often predictions incorrectly include regions belonging to other close concepts in the same image. The *"x"* in xIoU denotes both *"cross"*-concept confusion and the *"error"* rate from overlapping incorrect concepts. xIoU captures fine-grained mistakes that cause only small IoU drops but are critical for perfect segmentation and conceptual correctness. Since xIoU can be trivially reduced by conservative masks, we use it as a secondary metric alongside gIoU and cIoU.

Table 1: **VASA consistently outperforms prior methods across VLMs and remains competitive with a fully supervised segmenter on PARS.** All results are reproduced by the authors. We use xIoU as the secondary metric for cross-concept confusion, alongside gIoU and cIoU. *Semantic-SAM relies on a ground-truth-derived point prompt. †GuidelineSeg uses Gemma 3 4B as a second VLM.

| Method | VLM Version | Train. Free | Ad-hoc Concepts | | | Common Concepts | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | gIoU | cIoU | xIoU↓ | gIoU | cIoU | xIoU↓ | gIoU | cIoU | xIoU↓ |
| *Fully Supervised Segmenter* | | | | | | | | | | | |
| VLPart (Sun et al., 2023) | N/A | N/A | 55.2 | 53.8 | 35.1 | 60.5 | 73.3 | 10.7 | 57.5 | 63.2 | 23.9 |
| *Open-Vocabulary Segmenter* | | | | | | | | | | | |
| OVSeg (Liang et al., 2023) | N/A | N/A | 34.4 | 27.6 | 42.2 | 46.0 | 43.0 | 20.7 | 39.3 | 34.2 | 33.5 |
| Semantic-SAM* (Li et al., 2023a) | N/A | N/A | 39.2 | 30.4 | 45.2 | 51.7 | 53.7 | 25.7 | 44.5 | 39.4 | 38.4 |
| SAM3 (Carion et al., 2026) | N/A | N/A | 34.1 | 36.1 | 59.8 | 41.8 | 57.8 | 30.6 | 37.3 | 45.0 | 48.9 |
| SAM3-I (Li et al., 2026a) | N/A | N/A | 45.8 | 47.1 | 26.2 | 38.3 | 46.0 | 30.4 | 42.6 | 46.5 | 28.4 |
| *VLM as Visual Reasoner* | | | | | | | | | | | |
| LISA (Lai et al., 2023) | Llama2 13B | ✗ | 37.9 | 39.9 | 45.0 | 49.0 | 59.7 | 19.2 | 42.6 | 49.4 | 32.9 |
| LISA++ (Yang et al., 2024) | LLaVA1.5 7B | ✗ | 35.2 | 33.7 | 52.6 | 51.8 | 50.4 | 25.4 | 42.3 | 41.5 | 41.0 |
| VisionReasoner (Liu et al., 2026) | Qwen2.5-VL 7B | ✗ | 27.2 | 28.8 | 47.9 | 29.4 | 41.9 | 16.5 | 28.1 | 35.0 | 34.5 |
| Seg-Zero (Liu et al., 2025) | Qwen2.5-VL 7B | ✗ | 40.8 | 39.6 | 58.2 | 57.2 | 66.9 | 30.3 | 47.8 | 50.8 | 47.4 |
| RESAnything (Wang & Zhang, 2025) | Qwen2.5-VL 7B | ✓ | 33.1 | 34.0 | 49.7 | 45.2 | 51.0 | 22.8 | 38.2 | 41.8 | 38.1 |
| *Context Agent for Segmentation* | | | | | | | | | | | |
| CoReS (Bao et al., 2024) | LLaVA 7B | ✗ | 33.0 | 34.1 | 36.2 | 42.2 | 53.4 | 19.7 | 36.9 | 44.0 | 27.5 |
| SegAgent-SC (Zhu et al., 2025b) | Qwen-VL 7B | ✗ | 34.6 | 34.8 | 46.6 | 48.2 | 61.8 | 20.9 | 40.4 | 47.9 | 33.8 |
| SegAgent-SAM (Zhu et al., 2025b) | Qwen-VL 7B | ✗ | 33.1 | 31.7 | 48.2 | 45.9 | 56.0 | 19.8 | 38.5 | 43.1 | 35.0 |
| GuidelineSeg† (Vats et al., 2025) | Gemini 2.5 Flash | ✗ | 33.4 | 33.8 | 46.5 | 46.5 | 56.3 | 19.7 | 39.0 | 44.1 | 34.6 |
| Evol-SAM3 (Ye et al., 2025) | Qwen2.5-VL 7B | ✓ | 41.4 | 44.4 | 43.9 | 50.7 | 57.1 | 22.3 | 45.3 | 50.3 | 35.0 |
| SAM3 Agent (Carion et al., 2026) | Gemini 2.5 Flash Lite | ✓ | 39.6 | 38.8 | 58.6 | 54.0 | 60.9 | 30.6 | 45.7 | 47.7 | 48.4 |
| | Qwen3-VL 32B Think | ✓ | 40.5 | 39.1 | 58.8 | 55.0 | 62.1 | 27.5 | 46.4 | 48.1 | 48.0 |
| | GPT-5.6 Luna | ✓ | 40.8 | 39.2 | 58.8 | 57.2 | 68.3 | 27.4 | 47.7 | 50.7 | 47.3 |
| | GPT-6 Luna | ✓ | 40.0 | 39.1 | 58.7 | 55.7 | 65.5 | 28.2 | 46.7 | 49.7 | 47.5 |
| | GLM 5.3 Flash | ✓ | 40.5 | 39.5 | 58.4 | 57.3 | 67.3 | 26.7 | 47.6 | 50.7 | 46.7 |
| | DeepSeek V4.1 Flash | ✓ | 40.7 | 39.2 | 59.0 | 57.3 | 68.2 | 26.5 | 47.7 | 50.6 | 47.3 |
| *Vision Harnessing Agent for Segmentation* | | | | | | | | | | | |
| **VASA (ours)** | Gemini 2.5 Flash Lite | ✓ | 48.0 | 51.6 | 39.7 | 58.5 | 75.5 | 11.0 | 52.4 | 62.6 | 27.4 |
| | Qwen3-VL 32B Think | ✓ | 54.0 | 56.9 | 33.5 | 60.8 | 77.0 | 10.0 | 56.9 | 66.5 | 22.9 |
| | GPT-5.6 Luna | ✓ | 54.6 | 56.5 | 37.2 | 61.8 | 73.9 | 11.6 | 57.7 | 64.5 | 26.5 |
| | GPT-6 Luna | ✓ | 54.5 | 54.9 | 40.5 | 67.4 | 78.0 | 12.5 | 60.0 | 65.3 | 28.8 |
| | GLM 5.3 Flash | ✓ | 55.0 | 57.8 | 36.3 | 65.5 | 78.6 | 11.2 | 59.5 | 67.5 | 25.4 |
| | DeepSeek V4.1 Flash | ✓ | **59.8** | **62.7** | **30.4** | **68.6** | **80.6** | **10.3** | **63.5** | **71.4** | **21.2** |
| ***largest gain v.s. prior SOTA* (SAM3 Agent)** | | | **+19.1** | **+23.5** | **+28.6** | **+11.3** | **+12.4** | **+16.2** | **+15.8** | **+20.8** | **+26.1** |

**3. Experiment Settings.** We consider three groups of baselines with increasing VLM involvement: 1) *Open-vocabulary segmenters* for text-prompted segmentation without VLMs. 2) *VLMs as Visual Reasoners*, which integrate segmenters into VLMs and are fine-tuned to reason before producing a segmentation mask. 3) *Context agents for segmentation*, which involve more interaction between VLMs and segmenters and typically fine-tune VLMs to call segmenters iteratively as tools. We evaluate VASA with SAM3 and six VLMs spanning open-source and proprietary families across various capabilities: Gemini 2.5 Flash Lite (Comanici et al., 2025), Qwen3-VL 32B Thinking (Bai et al., 2025), GPT-5.6 Luna (OpenAI, 2026a), GPT-6 Luna (OpenAI, 2026b), GLM 5.3 Flash (GLM-5-Team et al., 2026), and DeepSeek V4.1 Flash (DeepSeek-AI et al., 2026), all using the same harness. We evaluate our closest competitor, SAM3 Agent, with the same VLMs to make the comparison *independent of VLM choice*. We use the open-source Qwen model for ablation studies and analysis. We also train a supervised segmenter, VLPart, on the training set as a closed-world reference.

**4. Results on PARS.** We evaluate VASA and the baselines on the PARS benchmark for open ad-hoc segmentation (Table 1). Our results show that *the effectiveness of a segmentation agent depends strongly on how it is harnessed.* Across all six VLMs, SAM3 Agent's performance saturates within a narrow range of 39.6%-40.8% ad-hoc gIoU. This suggests that, even with stronger VLMs, SAM3 Agent's design fundamentally limits its ability to segment rarer and more complex concepts, leading it to confuse common concepts with the open ad-hoc concepts specified by the query. In contrast, VASA with the oldest VLM, Gemini 2.5 Flash Lite, already outperforms SAM3 Agent with the latest DeepSeek V4.1 Flash by 7.2%. This gap grows to 19.1% when VASA uses the same DeepSeek V4.1 Flash backbone. The reduction in ad-hoc xIoU from 59.0% to 30.4% further shows that VASA more clearly distinguishes the requested concepts from related but incorrect regions and excludes those regions from its predictions. Our visual harness enables VASA to generalize more effectively to stronger VLMs, translating greater VLM capability into better downstream performance. These gains extend to the broader set of baselines, including a supervised segmenter trained directly on the benchmark's training set when using DeepSeek V4.1 Flash without any task-specific fine-tuning.

| Method | val | | testA | | testB | |
|---|---|---|---|---|---|---|
| | Part | +Obj | Part | +Obj | Part | +Obj |
| X-Decoder | 16.2 | 29.5 | 13.6 | 23.6 | 20.3 | 33.8 |
| SEEM | 16.1 | 29.4 | 13.6 | 23.4 | 20.4 | 33.9 |
| UniRES | 19.6 | 34.3 | 16.4 | 27.8 | 25.2 | 41.7 |
| SAM3 | 12.3 | 17.3 | 9.0 | 13.0 | 20.7 | 26.1 |
| SAM3-I | 17.0 | 29.5 | 14.7 | 24.7 | 22.3 | 36.1 |
| LISA | 21.3 | 34.3 | 18.5 | 28.6 | 25.7 | 40.1 |
| GSVA | 11.4 | 23.1 | 9.2 | 19.2 | 16.8 | 28.2 |
| GLaMM | 21.4 | 35.3 | 18.6 | 29.5 | 26.9 | 41.1 |
| $M^2$SA | 22.4 | 35.5 | 19.9 | 30.1 | 27.1 | 41.4 |
| RESAnything | 27.6 | - | 26.5 | - | 25.8 | - |
| CoReS | 21.2 | 29.5 | 17.3 | 24.0 | 20.6 | 30.6 |
| SegAgent-SC | 19.7 | 33.4 | 16.8 | 27.3 | 22.9 | 38.0 |
| SegAgent-SAM | 19.7 | 32.9 | 16.9 | 27.1 | 22.6 | 36.7 |
| Evol-SAM3 | 24.3 | 34.0 | 20.6 | 28.8 | 30.8 | 41.0 |
| SAM3 Agent | 40.2 | 48.1 | 34.4 | 41.5 | 41.3 | 50.4 |
| **VASA (ours)** | **45.0** | **51.1** | **43.2** | **47.4** | **47.6** | **54.1** |
| ***v.s. prior SOTA*** | **+4.8** | **+3.0** | **+8.8** | **+5.9** | **+6.3** | **+3.7** |

Table 2: **VASA achieves state-of-the-art performance on the standard multi-granularity referring segmentation benchmark RefCOCOm**, with the largest gains on the fine-grained parts. We report gIoU for part-only (Part) and combined part-and-object (+Obj), with Qwen3-VL 32B Thinking as the VLM. We compare VASA against *open-vocabulary segmenters*, *VLMs for visual reasoning*, and *context agents*, with additional works: X-Decoder (Zou et al., 2022), SEEM (Zou et al., 2023), GSVA (Xia et al., 2024), GLaMM (Rasheed et al., 2024), $M^2$SA (Jang et al., 2025). We reproduce results for SAM3, SAM3-I, and all context agent methods in the third group; the remaining results are reported from Wang & Zhang (2025) and Jang et al. (2025).

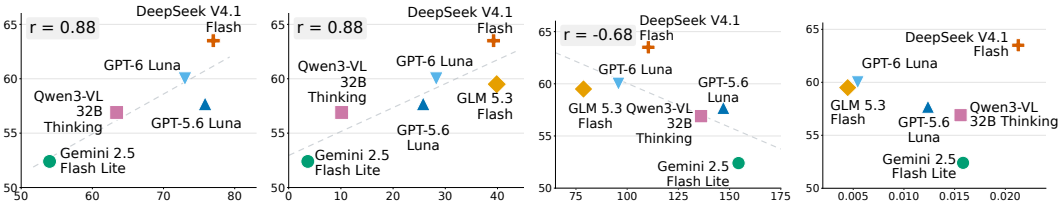
| Method | W.M. | Strategy | C&S | E.R. | gIoU | cIoU | xIoU↓ |
|---|---|---|---|---|---|---|---|
| SAM3 Agent | – | – | – | – | 46.4 | 48.1 | 48.0 |
| W.M. Only | ✓ | – | – | – | 40.0 | 46.1 | 37.5 |
| + Strategy | ✓ | ✓ | – | – | 44.9 | 55.1 | 29.3 |
| + C&S | ✓ | ✓ | ✓ | – | 48.3 | 59.9 | 27.3 |
| **VASA (ours)** | ✓ | ✓ | ✓ | ✓ | **56.9** | **66.5** | **22.9** |

Table 3: **Harness component ablation on PARS.** We start with working masks only and progressively add to the preceding setting. We use Qwen3-VL 32B Thinking and report SAM3 Agent here as a reference. W.M.: working mask; C&S: constraints & scrutiny; E.R.: error recovery.

**5. Results on RefCOCOm.** We further evaluate VASA on a related task of *fine-grained multi-granularity referring segmentation* on RefCOCOm (Table 2). Unlike our PARS for open ad-hoc segmentation, RefCOCOm uses shorter referring expressions than PARS, such as *the left leg of the man facing us*, which we use without modifications. Compared with all baselines, VASA achieves the best performance across all splits and evaluation settings by large margins. Compared with SAM3 Agent, it improves gIoU by up to 8.8% in the part-only setting and 5.9% in the combined part-and-object setting. The gains are consistently larger in the part-only setting, where fine-grained targets can be confused with neighboring regions. Such targets call for the same capabilities emphasized by our harness: maintaining visual state, inspecting intermediate masks, and correcting errors before returning the final prediction. The improvements across both conventional referring segmentation methods and agentic baselines demonstrate the broader effectiveness of VASA's visual harness.

**6. Harness component ablation.** Table 3 examines how the harness components jointly support VASA's effective visual construction. The *working-mask-only* variant maintains a persistent mask and allows the agent to update and inspect it, but lacks support from other harness components. This gives the agent the mechanism for accumulating visual progress without the guidance to use it reliably, improving xIoU on PARS from 48.0% to 37.5% but dropping the gIoU from 46.4% to 40.0%. Adding strategy partially addresses this limitation, raising gIoU to 44.9%. Constraints and scrutiny further improve gIoU to 48.3%, while the complete harness reaches 56.9% with the lowest xIoU of 22.9%. These results show that a working mask alone does not establish an effective visual construction process; it benefits from coordinated guidance from VASA's complete harness.

**7. Which VLM capabilities matter most for VASA?** Visual construction requires both *multimodal understanding* and *broader knowledge and reasoning*. Fig. 5(a,b) relates VASA's overall PARS gIoU to MMMU-Pro and HLE scores across different VLMs. GPT-6 Luna scores higher on HLE but lower on MMMU-Pro than GPT-5.6 Luna. On PARS, it improves *common*-concept segmentation but leaves *ad-hoc* gIoU unchanged and increases cross-concept confusion (xIoU↓). This suggests that broader reasoning gains may not compensate for weaker multimodal understanding when grounding ad-hoc concepts. Both benchmarks correlate strongly with gIoU with Pearson $r = 0.88$, re-verifying that VASA translates stronger VLM capabilities into better open ad-hoc segmentation performance.




a) MMMU-Pro score b) HLE score c) tokens/query (k) d) cost/query ($)

Figure 5: **VASA's performance *v.s.* VLM capability and inference cost.** Each panel plots overall PARS gIoU against **a)** *MMMU-Pro* score, which measures multimodal understanding and reasoning, **b)** *Humanity's Last Exam* (HLE) score, measuring expert-level academic knowledge and reasoning, **c)** average token usage per query, and **d)** average VLM API cost per query. Dashed lines indicate linear fits, with Pearson's coefficient $r$ reported where shown. Benchmark scores are from Artificial Analysis (Artificial Analysis, 2026); GLM 5.3 Flash's MMMU-Pro score is missing from the source.

Table 4: **Detailed long-form queries improve VASA but overwhelm SAM3 Agent.** Results are reported on the total split of PARS. *"short/long"* denote the original class labels and the long queries used in PARS. Long queries improve VASA by 5.9% in gIoU, but reduce SAM3 Agent by 22.1% in gIoU.

| Method | Text | gIoU | cIoU | xIoU↓ |
|---|---|---|---|---|
| SAM3 Agent | short | **46.4** | **48.1** | 48.0 |
| SAM3 Agent | long | 24.3 | 34.5 | **30.7** |
| **VASA (ours)** | short | 51.0 | 51.3 | 41.7 |
| **VASA (ours)** | long | **56.9** | **66.5** | **22.9** |

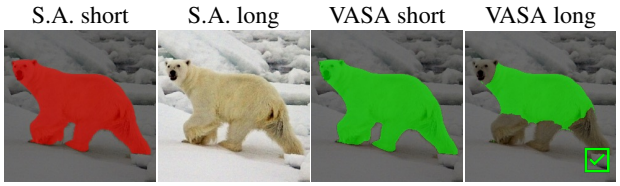



Figure 6: **Effect of detailed long-form queries.** Short query: *"the body of polar bear"*. Long query (shortened): *"Segment the torso of the polar bear, starting from below the head to just above the legs."*. The short query is ambiguous, leading both methods to segment the entire bear. The long-form query resolves the ambiguity, but SAM3 Agent (S.A.) completely fails. In contrast, VASA successfully follows the query and isolates the requested torso region.

**8. Effect of detailed queries for open ad-hoc segmentation.** In PARS, we replace the original class names from PartImageNet with detailed queries because short names often leave the intended open ad-hoc concepts ambiguous. Here, we revert to the original names while keeping the images and target masks unchanged to examine the effect of query detail. When using Qwen3-VL 32B Thinking as the VLM, detailed queries improve VASA's gIoU from 51.0% to 56.9% and reduce xIoU from 41.7% to 22.9%, whereas SAM3 Agent's gIoU drops from 46.4% to 24.3% (Table 4). Fig. 6 illustrates this distinction. This contrast shows that additional query detail benefits segmentation only when the agent can effectively act on it, and also confirms the need for detailed queries in PARS.

**9. Reasoning steps, token usage, and inference cost.** We examine whether VASA's gains come simply from a *larger inference budget* or from *more effective use of computation*. Both VASA and SAM3 Agent are allowed up to 20 reasoning rounds. Yet SAM3 Agent stops after just 3.8 rounds on average despite its lower segmentation results, struggling to turn the remaining budget into further improvements. VASA instead uses 7.9 rounds, while making only slightly more segmentation calls (3.4 versus 3.1; Table 5). These additional rounds support planning, mask inspection, and refinement of a persistent visual solution, rather than repeated attempts to obtain the complete target from a single segmentation call. Consistent with this interpretation, larger increases in reasoning steps over SAM3 Agent are associated with larger gIoU gains (Fig. 7). Although average runtime rises from 5.2 to 9.3 seconds per image, the additional computation supports productive visual construction, enabling VASA to make better use of the same available budget.

Across VLM backbones, however, better segmentation is often obtained using fewer tokens ($r = -0.68$; Fig. 5(c)), suggesting that more capable models can reach better solutions with fewer tokens. API cost has a less clear relationship with performance, as it also reflects provider-specific pricing (Fig. 5(d)). Among the evaluated configurations, GLM 5.3 Flash and GPT-6 Luna offer the strongest cost-performance tradeoffs, while DeepSeek V4.1 Flash achieves the highest gIoU at a relatively higher cost than the other models but still remains low compared to other frontier models.

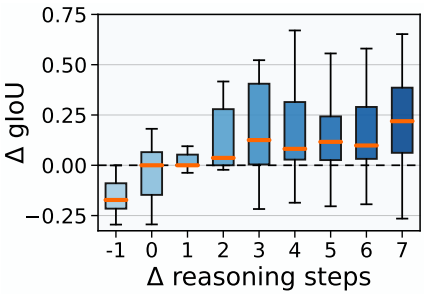



Figure 7: **Additional reasoning steps correlate with improved segmentation quality.** The x-axis shows the number of more reasoning steps used by VASA than SAM3 Agent, while the y-axis shows the corresponding gIoU improvement. Darker blue colors indicate bins containing more instances. The positive correlations suggest that additional long-horizon reasoning and refinement are often beneficial for challenging open ad-hoc concepts. VASA naturally has more reasoning steps than SAM3 Agent due to the iterative visual construction.

| Method | SAM3 calls | Reason. rounds | Time per image |
|---|---|---|---|
| SAM3 Agent | 3.1 | 3.8 | 5.2 s |
| **VASA** | 3.4 | 7.9 | 9.3 s |

Table 5: **The additional reasoning rounds of VASA are effectively used for visual inspection and refinement.** We report average inference statistics on PARS with Qwen3-VL 32B Thinking hosted on an 8× NVIDIA A40 server under the same 20-round budget.




Figure 8: **Qualitative comparisons between SAM3 Agent and VASA on PARS (left) and RefCOCOm (right).** SAM3 Agent often oversegments semantically related regions or confuses nearby concepts, while VASA precisely follows the text query through visual reasoning and iterative mask refinement. PARS prompts are shortened for clarity, while RefCOCOm prompts are original.

**10. Qualitative results.** Fig. 8 shows how VASA follows concept boundaries beyond familiar object masks. On PARS, SAM3 Agent includes excluded regions, whereas VASA more closely isolates the requested *taxi body*, *shark trunk*, and *dog torso*. On RefCOCOm, it better separates the specified *legs*, *torso*, or *arm from the surrounding person*. These examples illustrate more faithful grounding of both detailed definitions and short referring expressions. Additional examples are provided in Appendix E.

**11. Results on ReasonSeg.** With Qwen3-VL 32B Thinking, VASA achieves 72.1% validation and 70.6% test gIoU, compared with 71.9% and 69.4% for SAM3 Agent. While intended for challenging open ad-hoc concepts, VASA remains competitive on conventional reasoning segmentation.

**Summary.** We present VASA, the first vision harnessing agent for open ad-hoc segmentation, and PARS, a new benchmark with detailed queries defining open ad-hoc concepts. VASA features a complete harness that coordinates strategy, constraints, scrutiny, and recovery around a persistent working mask, enabling the agent to inspect and refine its visual progress. We demonstrate VASA on PARS across six different VLMs, as well as on related standard benchmarks like RefCOCOm and ReasonSeg. We conduct ablation studies and detailed analysis to understand where VASA's gain comes from. Our work points toward AI agents that go beyond invoking tools, using structured harnesses to maintain visual state, guide reasoning, inspect progress, and recover from errors.

### AI USE STATEMENT

We used Qwen3-VL 8B Instruct to construct our PARS benchmark, which the authors manually verified and corrected. (Sec. 4 and Appendix D). We also used AI tools for manuscript writing, polishing, and presentation. The authors take responsibility for the final content and results.

### REPRODUCIBILITY STATEMENT

We will publicly release our code and data to enable reproduction of all results reported in this paper. The full VASA inference algorithm and additional implementation settings are provided in Appendix B and C.

## REFERENCES


Artificial Analysis. Model evaluations. https://artificialanalysis.ai/evaluations, 2026. 8

Shuai Bai et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631. 2, 6, 20

Xiaoyi Bao, Siyang Sun, Shuailei Ma, Kecheng Zheng, Yuxin Guo, Guosheng Zhao, Yun Zheng, and Xingang Wang. Cores: Orchestrating the dance of reasoning and segmentation. In *European Conference on Computer Vision*, pp. 187–204. Springer, 2024. 3, 6

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw. 1, 2, 3, 6, 18, 20, 21

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts, 2014. URL https://arxiv.org/abs/1406.2031. 18

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, 2024. URL https://arxiv.org/abs/2409.10542. 18

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023. 18

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261. 6

DeepSeek-AI et al. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026. URL https://arxiv.org/abs/2609.19969. 6

Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. In *International Conference on Machine Learning*, 2023. 18

Qihua Dong, Luis Figueroa, Handong Zhao, Kushal Kafle, Jason Kuen, Zhihong Ding, Scott Cohen, and Yun Fu. Cot referring: Improving referring expression tasks with grounded reasoning, 2026. URL https://arxiv.org/abs/2510.06243. 18

Tianyuan Du, Haopeng Li, Zhen Fan, Jiarui Zhang, Panwang Pan, and Yang Zhang. SAM-veteran: An MLLM-based human-like SAM agent for reasoning segmentation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=oN55r8iJJW. 3

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels, 2022. URL https://arxiv.org/abs/2112.12143. 18

GLM-5-Team et al. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763. 6

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=lL3lnMbR4WU. 18

Qiushi Han, Keya Hu, Linlu Qiu, Cathy Wu, and Kaiming He. VISTA: A visual harness for reasoning in an interactive world, aug 2026. URL https://vista-research.github.io/. 3

Mitchell Hashimoto. My ai adoption journey. https://mitchellh.com/writing/my-ai-adoption-journey, 2026. 3

Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts, 2022. URL https://arxiv.org/abs/2112.00933. 2, 5, 18, 20

Xingqi He, Yujie Zhang, Shuyong Gao, Wenjie Li, Lingyi Hong, Mingxi Chen, Kaixun Jiang, Jiyuan Fu, and Wenqiang Zhang. Rsagent: Learning to reason and act for text-guided segmentation via multi-turn tool invocations, 2025. URL https://arxiv.org/abs/2512.24023. 3

Sandesh Hegde, Jaison Saji Chacko, Debarshi Banerjee, and Uma Mahesh. Genseg-r1: Rl-driven vision-language grounding for fine-grained referring segmentation, 2026. URL https://arxiv.org/abs/2602.09701. 18

Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning, 2026. URL https://arxiv.org/abs/2505.22596. 18

Donggon Jang, Yucheol Cho, Suin Lee, Taehyeon Kim, and Dae-Shik Kim. Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation, 2025. URL https://arxiv.org/abs/2503.13881. 7, 18

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL https://arxiv.org/abs/2102.05918. 18

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy, 2024. 18

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021. 18

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1, 18

Weicheng Kuo, Fred Bertsch, Wei Li, AJ Piergiovanni, Mohammad Saffar, and Anelia Angelova. Findit: Generalized localization with natural language queries, 2022. URL https://arxiv.org/abs/2203.17273. 18

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1, 2, 3, 5, 6, 18, 21

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RriDjddCLN. 18

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023a. 6, 18

Jingjing Li, Yue Feng, Yuchen Guo, Jincai Huang, Wei Ji, Qi Bi, Yongri Piao, Miao Zhang, Xiaoqi Zhao, Qiang Chen, Shihao Zou, Huchuan Lu, and Li Cheng. Sam3-i: Segment anything with instructions, 2026a. URL https://arxiv.org/abs/2512.04585. 6, 18

Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 18

Liunian Harold Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. Desco: Learning object recognition with rich language descriptions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=J2Cso0wWZX. 18

Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation?, 2024. URL https://arxiv.org/abs/2401.10229. 18

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.111409. URL https://www.sciencedirect.com/science/article/pii/S003132032500069X. 18

Zechu Li, Yufeng Jin, Xiaoyang Liu, Puze Liu, Vignesh Prasad, Carlo D'Eramo, and Georgia Chalvatzaki. Harbor: A harness framework for agentic robot reinforcement learning, 2026b. URL https://arxiv.org/abs/2606.08610. 3

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023. 6, 18

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation, 2023a. URL https://arxiv.org/abs/2306.00968. 18

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023b. 18

Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. *arXiv preprint arXiv:2312.01623*, 2023c. 18

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. 1, 2, 3, 6, 18, 21

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. 6, 18

Yi Lu, Jiawang Cao, Yongliang Wu, Bozheng Li, Licheng Tang, Yangguang Ji, Chong Wu, Jay Wu, and Wenbo Zhu. Rsvp: Reasoning segmentation via visual prompting and multi-modal chain-of-thought, 2025. URL https://arxiv.org/abs/2506.04277. 18

Zhenyu Lu, Liupeng Li, Jinpeng Wang, Yan Feng, Bin Chen, Ke Chen, and Yaowei Wang. CoPRS: Learning positional prior from chain-of-thought for reasoning segmentation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Fcsop01h40. 18

Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *ICML*, 2023. 18

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought, 2025. URL https://arxiv.org/abs/2505.23766. 18

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022. URL https://arxiv.org/abs/2205.06230. 18

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection, 2024. URL https://arxiv.org/abs/2306.09683. 18

Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip H. S. Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning, 2022. URL https://arxiv.org/abs/2212.04994. 18

Minheng Ni, Yabo Zhang, Kailai Feng, Xiaoming Li, Yiwen Guo, and Wangmeng Zuo. Ref-diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*, 2023. 18

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026a. 6

OpenAI. Introducing GPT-6 Sol and Luna. https://openai.com/index/introducing-gpt-6-sol-and-luna/, 2026b. 6

OpenAI. Harness engineering. https://openai.com/index/harness-engineering/, 2026c. 3

Seulki Park, Zilin Wang, and Stella X. Yu. Free-grained hierarchical visual recognition, 2026. URL https://arxiv.org/abs/2510.14737. 18

Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how <seg> token works, 2026. URL https://arxiv.org/abs/2412.17741. 18

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020. 18

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: Parts and attributes of common objects. In *arXiv preprint arXiv:2301.01795*, 2023. 18

Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models, 2023. URL https://arxiv.org/abs/2210.09996. 18

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. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7, 18

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. URL https://arxiv.org/abs/2408.00714. 1, 18

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, Xingyu Chen, Zhuheng Song, Yuhong Zhang, Hongjie Huang, Han Gao, Shilong Liu, Hao Zhang, Feng Li, Kent Yu, and Lei Zhang. Dino-x: A unified vision model for open-world object detection and understanding, 2024a. URL https://arxiv.org/abs/2411.14347. 18

Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding dino 1.5: Advance the "edge" of open-set object detection, 2024b. URL https://arxiv.org/abs/2405.10300. 18

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024c. 18

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*, 2023. 18

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024a. 18

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024b. URL https://arxiv.org/abs/2402.03300. 3, 18

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 18

Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. *arXiv preprint arXiv:2305.11173*, 2023. 6, 18

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 18

Vanshika Vats, Ashwani Rathee, and James Davis. Guideline-consistent segmentation via multi-agent refinement, 2025. URL https://arxiv.org/abs/2509.04687. 3, 6

Zifu Wan, Yaqi Xie, Ce Zhang, Zhiqiu Lin, Zihan Wang, Simon Stepputtis, Deva Ramanan, and Katia P. Sycara. InstructPart: Task-oriented part segmentation with instruction reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24202–24227, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1179. URL https://aclanthology.org/2025.acl-long.1179/. 18

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference, 2024a. URL https://arxiv.org/abs/2312.01597. 18

Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X-sam: From segment anything to any segmentation, 2026a. URL https://arxiv.org/abs/2508.04655. 18

Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding, 2024b. URL https://arxiv.org/abs/2310.15308. 18

Junchi Wang and Lei Ke. Llm-seg: Bridging image segmentation and large language model reasoning. *arXiv preprint arXiv:2404.08767*, 2024. 18

Qi Wang, Tianyi Wang, Chengyang Li, Shikun Ban, Yurun Chen, Yizhong Ge, Jason Qin, Chengtai Li, and Wentao Zhu. Towards the harness of embodied agents, 2026b. URL https://arxiv.org/abs/2608.11246. 3

Ruiqi Wang and Hao Zhang. Resanything: Attribute prompting for arbitrary referring segmentation, 2025. URL https://arxiv.org/abs/2505.02867. 6, 7, 18

Song Wang, Gongfan Fang, Lingdong Kong, Xiangtai Li, Jianyun Xu, Sheng Yang, Qiang Li, Jianke Zhu, and Xinchao Wang. Pixelthink: Towards efficient chain-of-pixel reasoning, 2026c. URL https://openreview.net/forum?id=wvxq3qNzHB. 18

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, 2024c. URL https://arxiv.org/abs/2312.08007. 2, 5, 18, 21

XuDong Wang, Shaolun Zhang, Shufan Li, Konstantinos Kallidromitis, Kehan Li, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Segllm: Multi-round reasoning segmentation. *arXiv preprint arXiv:2410.18923*, 2024d. 3, 18

Zilin Wang, Sangwoo Mo, Stella X. Yu, Sima Behpour, and Liu Ren. Open ad-hoc categorization with contextualized feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1

Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model, 2024. URL https://arxiv.org/abs/2411.17606. 18

Meng Wei, Xiaoyu Yue, Wenwei Zhang, Shu Kong, Xihui Liu, and Jiangmiao Pang. Ov-parts: Towards open-vocabulary part segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 18

Byeongju Woo, Zilin Wang, Byeonghyun Pak, Sangwoo Mo, and Stella X. Yu. Aligning forest and trees in images and long captions for visually grounded understanding, 2026. URL https://arxiv.org/abs/2602.02977. 18

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models, 2025. URL https://arxiv.org/abs/2503.12799. 18

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 (CVPR)*, pp. 3858–3869, June 2024. 7, 18

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. In *CVPR*, 2025. 18

Wenli Xiao, Jia Xie, Tonghe Zhang, Haotian Lin, Letian "Max" Fu, Haoru Xue, Jalen Lu, Yi Yang, Cunxi Dai, Zi Wang, Jimmy Wu, Guanzhi Wang, S. Shankar Sastry, Ken Goldberg, Linxi "Jim" Fan, Yuke Zhu, and Guanya Shi. Enpire: Agentic robot policy self-improvement in the real world, 2026. URL https://arxiv.org/abs/2606.19980. 3

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *arXiv preprint arXiv:2202.11094*, 2022. 18

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models, 2023a. URL https://arxiv.org/abs/2303.04803. 18

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b. 18

Sihan Xu, Ziqiao Ma, Wenhao Chai, Xuweiyi Chen, Weiyang Jin, Joyce Chai, Saining Xie, and Stella X. Yu. Next-embedding prediction makes strong vision learners, 2025. URL https://arxiv.org/abs/2512.16922. 18

Yifan Xu, Mengdan Zhang, Chaoyou Fu, Peixian Chen, Xiaoshan Yang, Ke Li, and Changsheng Xu. Multi-modal queried object detection in the wild. *Advances in Neural Information Processing Systems*, 36, 2024. 18

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model, 2024. URL https://arxiv.org/abs/2312.17240. 6, 18

Kai Ye, Xiaotong You, Jianghang Lin, Jiayi Ji, Pingyang Dai, and Liujuan Cao. Evolving, not training: Zero-shot reasoning segmentation via evolutionary prompting, 2025. URL https://arxiv.org/abs/2512.24702. 3, 6

Muyang Yi, Quan Cui, Hao Wu, Cheng Yang, Osamu Yoshie, and Hongtao Lu. A simple framework for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7071–7080, 2023. 18

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14358–14368, 2025. doi: 10.1109/CVPR52734.2025.01339. 18

Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. In *ECCV*, 2024. 18

Seokju Yun, Dongheon Lee, Noori Bae, Jaesung Jun, Chanseul Cho, and Youngmin Ro. Star: Segment anything reasoner, 2026. URL https://arxiv.org/abs/2603.14382. 18

Quan-Sheng Zeng, Yunheng Li, Daquan Zhou, Guanbin Li, Qibin Hou, and Ming-Ming Cheng. High-quality mask tuning matters for open-vocabulary segmentation, 2025. URL https://arxiv.org/abs/2412.11464. 18

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343. 18

Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 18

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng YAN. OMG-LLaVA: Bridging image-level, object-level, pixel-level reasoning and understanding. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=WeoNd6PRqS. 18

Yixian Zhang, Huanming Zhang, Feng Gao, Xiao Li, Zhihao Liu, Chunyang Zhu, Jiaxing Qiu, Yuchen Yan, Jiyuan Liu, Wenhao Tang, Zhengru Fang, Yi Nie, Changxu Wei, Yu Wang, Wenbo Ding, and Chao Yu. Harness vla: Steering frozen vlas into reliable manipulation primitives via memory-guided agents, 2026. URL https://arxiv.org/abs/2607.08448. 3

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*, 2024b. 18

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2025. 18

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip, 2022. URL https://arxiv.org/abs/2112.01071. 18

Lianghui Zhu, Bin Ouyang, Yuxuan Zhang, Tianheng Cheng, Rui Hu, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Li Yu, Wenyu Liu, and Xinggang Wang. Lens: Learning to segment anything with unified reinforced reasoning, 2025a. URL https://arxiv.org/abs/2508.14153. 18

Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. *arXiv preprint arXiv:2503.08625*, 2025b. URL https://arxiv.org/abs/2503.08625. 1, 2, 3, 6

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language, 2022. URL https://arxiv.org/abs/2212.11270. 7, 18

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=UHBrWeFWlL. 7, 18

# Vision Harnessing Agent for Open Ad-hoc Segmentation

## Appendix

CONTENTS

**A Extended Related Work** **18**

**B VASA Inference Algorithm** **19**

**C Implementation Details** **20**

**D PARS Construction Details** **20**

**E Additional Qualitative Results** **21**

# A Extended Related Work

**Referring segmentation** focuses on segmenting image regions by explicit text descriptions. Most prior works either derive segmentation as a byproduct of image–text alignment models (Radford et al., 2021; Cherti et al., 2023; Zhai et al., 2023; Tschannen et al., 2025; Jia et al., 2021), leveraging attention maps (Woo et al., 2026; Zhou et al., 2022; Li et al., 2025; Wang et al., 2024a; Xiao et al., 2025; Park et al., 2026; Xu et al., 2025) or patch-level similarity (Zeng et al., 2025; Xu et al., 2022; Ding et al., 2023; Mukhoti et al., 2022; Ranasinghe et al., 2023; Yi et al., 2023; Luo et al., 2023; Ghiasi et al., 2022; Ni et al., 2023), or directly train models with region–text alignment supervision (Kuo et al., 2022; Li et al., 2022; Yuan et al., 2024; Li et al., 2024; Xu et al., 2023b; Gu et al., 2022; Kamath et al., 2021; Jiang et al., 2024; Li et al., 2023b; Li* et al., 2022; Zhang et al., 2022; Liang et al., 2023; Minderer et al., 2022; 2024; Liu et al., 2023b; Ren et al., 2024b;a;c; Shen et al., 2024; Xu et al., 2024; Yin et al., 2025; Zhang et al., 2024b; Zou et al., 2022; 2023; Xu et al., 2023a; Wang et al., 2024b; Carion et al., 2026). Open ad-hoc segmentation generalizes it from grounding existing visual entities toward constructing arbitrary user-specified visual concepts through composition and exclusion reasoning. GRES and the gRefCOCO dataset (Liu et al., 2023a) consider a special case of open ad-hoc segmentation, where the target concept is composed of multiple commonly seen objects.

**Referring part segmentation** focuses on segmenting existing part concepts. Semantic-SAM (Li et al., 2023a) and RESAnything (Wang & Zhang, 2025) leverages SAM (Kirillov et al., 2023; Ravi et al., 2024) to generate part-level masks for training and inference. Other works (Liu et al., 2023c; Wang et al., 2024c; Jang et al., 2025; Wei et al., 2023; Sun et al., 2023; Wan et al., 2025) have also explored closed-set part datasets like the PartImageNet (He et al., 2022) and introduced new part referring segmentation datasets like the RefCOCOm (Wang et al., 2024c), along with others (Chen et al., 2014; Ramanathan et al., 2023; Wang & Zhang, 2025; Jang et al., 2025; Wei et al., 2023). Like with objects, these methods mainly cover the common parts. Since there are far more parts than objects, open ad-hoc segmentation becomes even more critical in the part domain.

**Reasoning segmentation** focuses on segmenting typically well-defined targets from *implicit* text queries through commonsense and spatial reasoning. Most existing works use VLMs for reasoning and segmenters for mask decoding. Three types of reasoning paradigms prevail. **1)** Implicit reasoning, where the VLM predicts the target segment representation, keeping the reasoning process inaccessible (Qian et al., 2026; Lai et al., 2023; Yang et al., 2024; Xia et al., 2024; Rasheed et al., 2024; Zhang et al., 2024a; Wei et al., 2024; Wang & Ke, 2024; Chen et al., 2024; Zhang et al., 2025; Wang et al., 2026a; Li et al., 2026a), such as the `[SEG]` token in LISA (Lai et al., 2023). **2)** Text-based visual reasoning, where the VLM is tuned through RL (Shao et al., 2024b) to reason in text (Huang et al., 2026; Yun et al., 2026; Liu et al., 2026; 2025; Wang et al., 2026c; Ren et al., 2023; Zhu et al., 2025a; Hegde et al., 2026), such as the target's coordinates prediction in Seg-Zero (Liu et al., 2025). **3)** Grounded visual chain-of-thought, where each reasoning step is visually grounded, and subsequent steps build on both prior reasoning and grounding results (Lu et al., 2025; Dong et al., 2026; Lu et al., 2026; Wu et al., 2025; Man et al., 2025; Shao et al., 2024a; Wang et al., 2024d), such as SegLLM (Wang et al., 2024d). Open ad-hoc segmentation is an orthogonal direction, reasoning about what constitutes a visual concept rather than how to find the existing concept hinted by the user.

## B VASA INFERENCE ALGORITHM

Algorithm 1 describes the interaction between the VLM, segmentation model, and harness. Each controller turn selects one tool call; it need not contain a segmentation call or a mask edit. Planning, candidate generation, optional individual-mask inspection, and editing can therefore occupy separate turns. The working mask changes only through an explicit edit.

**Algorithm 1:** VASA controller loop (one tool call per turn)

**Input:** Image $I$, query $q$, VLM $A$, segmenter SAM3, threshold $T$
**Output:** Constructed mask, or fallback prediction
Initialize working mask $M \leftarrow \emptyset$, candidates $\mathcal{C} \leftarrow \emptyset$, strategy $\pi$, history $h$, warning $w$, and counter $n \leftarrow 0$;
**while** *the interaction is active* **do**
  Assemble context from $I, q, M, \mathcal{C}, \pi, h, w$ and system instructions;
  Request one controller response from $A$; increment $n$ and record the response;
  Parse the tool call and apply the supported execution checks;
  **if** *a formatting or validation error is detected* **then**
    Set corrective warning $w$; **continue**;
  **if** `set_strategy` **then**
    Store the supplied plan as $\pi$;
  **else if** `segment_phrase` **then**
    Generate new candidates $\mathcal{C}$ from the supplied phrase; keep $M$ unchanged;
    **if** *no candidates are returned* **then**
      Set warning $w$; **continue**;
    Render numbered candidate masks;
  **else if** `examine_each_mask` **then**
    Inspect candidates individually with auxiliary VLM calls;
    Retain accepted candidates and return an overlay or warning; keep $M$ unchanged;
  **else if** `update_working_mask` **then**
    Form $C$ as the union of selected valid candidates;
    Apply $M \leftarrow M \cup C$, $M \leftarrow M \setminus C$, or $M \leftarrow C$;
    Mark $M$ as updated and render its overlay;
  **else if** `return_final_output` *or* `report_no_mask` **then**
    **break**;
  **if** $n > T$ **then**
    **break**;
**return** $M$ *if explicitly updated; otherwise latest candidates, or an empty prediction*;

## C Implementation Details

We use pretrained SAM3 (Carion et al., 2026) as the segmentation tool and Qwen3-VL 32B Thinking (Bai et al., 2025) for the ablations and additional benchmarks, without task-specific fine-tuning. We host Qwen locally on eight A40 GPUs. Images sent to the VLM are downscaled by a factor of two in each dimension. We leave decoding parameters at the serving endpoint's defaults, except for GLM 5.3 Flash, for which we set reasoning effort to low to avoid exhausting the generation budget before producing a tool call. Both VASA and SAM3 Agent use a 20-round controller interaction budget. Auxiliary per-mask inspection calls count toward this budget, as do error-recovery turns other than network and rate-limit retries. Per-response token limits vary by model. A round can be used for planning, inspection, or mask editing, so it need not involve a segmentation call.

At each turn, the controller receives the original image and query, recorded strategy, current working mask, latest segmentation candidates, and a compact history of tool calls. Earlier intermediate images and free-form responses are omitted. Mask edits are deterministic binary operations at the original image resolution; obtaining new candidates does not overwrite the working mask. Optional per-mask inspection uses the original image, a mask overlay, and a zoomed view to check candidates against the latest segmentation phrase. Execution errors return feedback while preserving the working mask. All semantic judgments are made from the query and visual observations, without access to ground-truth masks.

## D PARS Construction Details

We construct PARS from the validation split of PartImageNet (He et al., 2022), retaining its images and human-annotated part masks. Masks of the same part category within an image are combined by union. We focus on the *body* category because its extent varies across objects, making the short class name insufficient to specify the intended region.

For each object–part concept, we provide Qwen3-VL 8B Instruct with five randomly sampled examples per class, each showing the original image alongside its annotated target. The model produces a single description specifying which regions to include, which neighboring parts to exclude, and how to identify their boundaries. The generation prompt asks for a definition that applies across the examples and avoids references to annotation overlays or bounding boxes. Each description is reused for the corresponding concept across evaluation images. Thus, the VLM supplies the textual definition, while the target masks remain human annotations from PartImageNet. The demonstration images overlap the evaluation set. The authors manually verify and correct the generated descriptions.

We use Qwen3-VL 8B Instruct again to classify concepts as common or open ad-hoc from images overlaid with their ground-truth masks, yielding 1,084 common and 937 ad-hoc concepts.

The short-query ablation uses the original class names with the same images and target masks. Detailed queries therefore change the specification of the target, not the target itself. During segmentation, the agent receives the image and query, without the annotated demonstrations or ground-truth masks.

## E Additional Qualitative Results

We provide additional qualitative comparisons among LISA (Lai et al., 2023), Seg-Zero (Liu et al., 2025), SAM3 Agent (Carion et al., 2026), and VASA (ours) on PARS (Figs. 9 and 10) and RefCOCOm (Wang et al., 2024c) (Figs. 11 and 12). The qualitative results show that baselines often produce overly broad masks or confuse the target with nearby related regions. In contrast, VASA more precisely isolates the requested regions across both PARS and RefCOCOm, suggesting that its visual construction workflow generalizes beyond open ad-hoc segmentation to fine-grained referring segmentation.

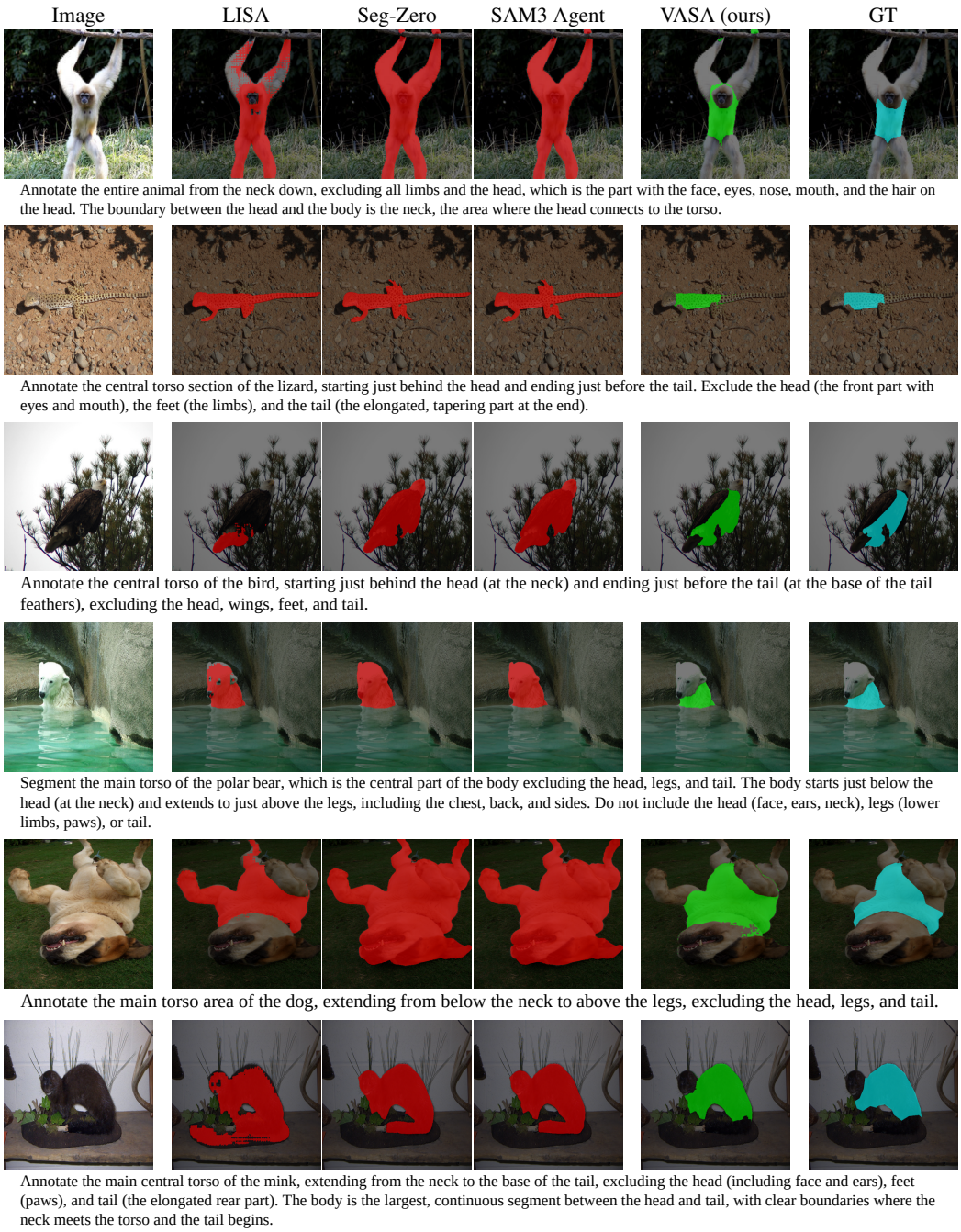



Figure 9: **Additional qualitative results of VASA and baselines on PARS.**

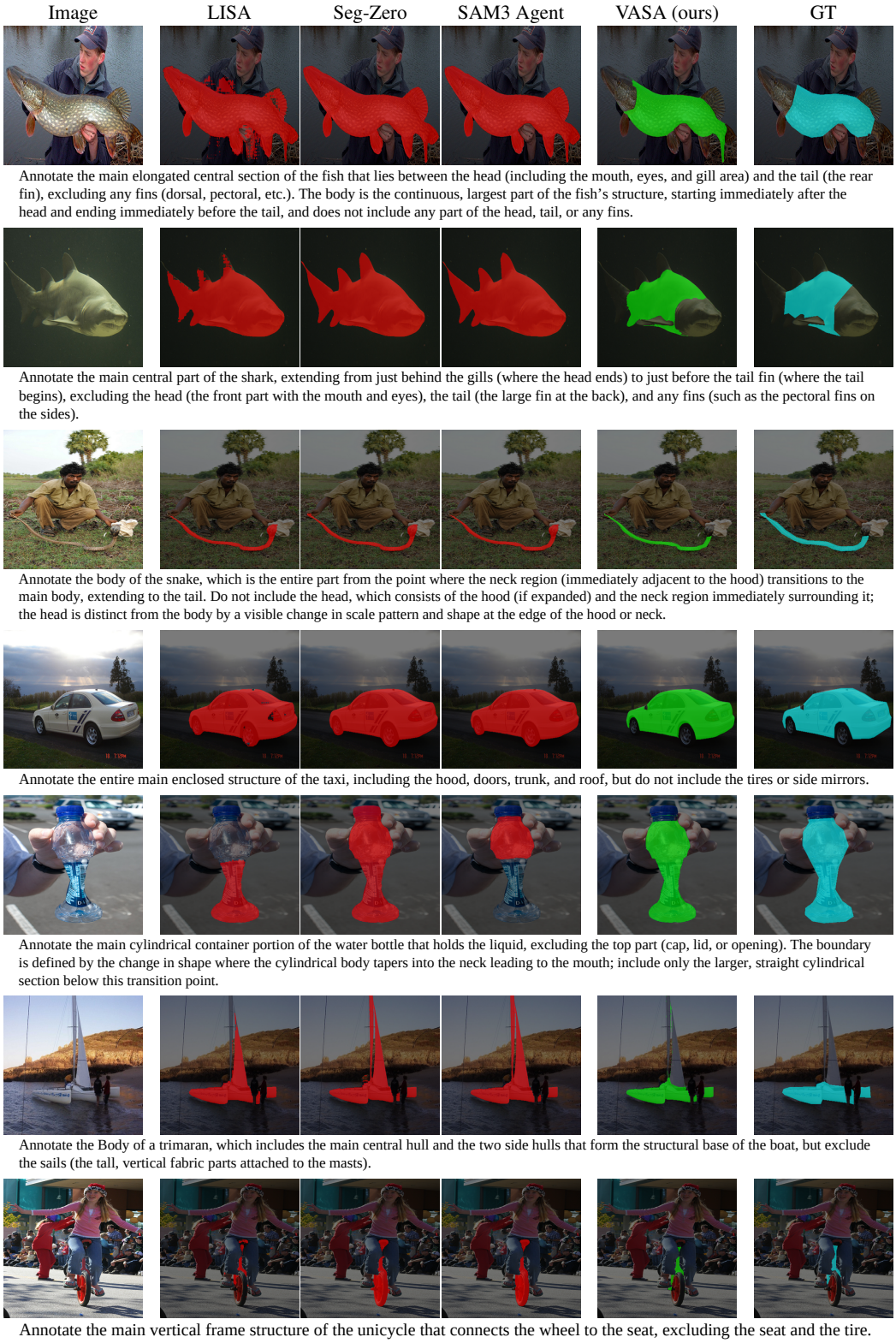



Figure 10: **Additional qualitative results of VASA and baselines on PARS** (continued).

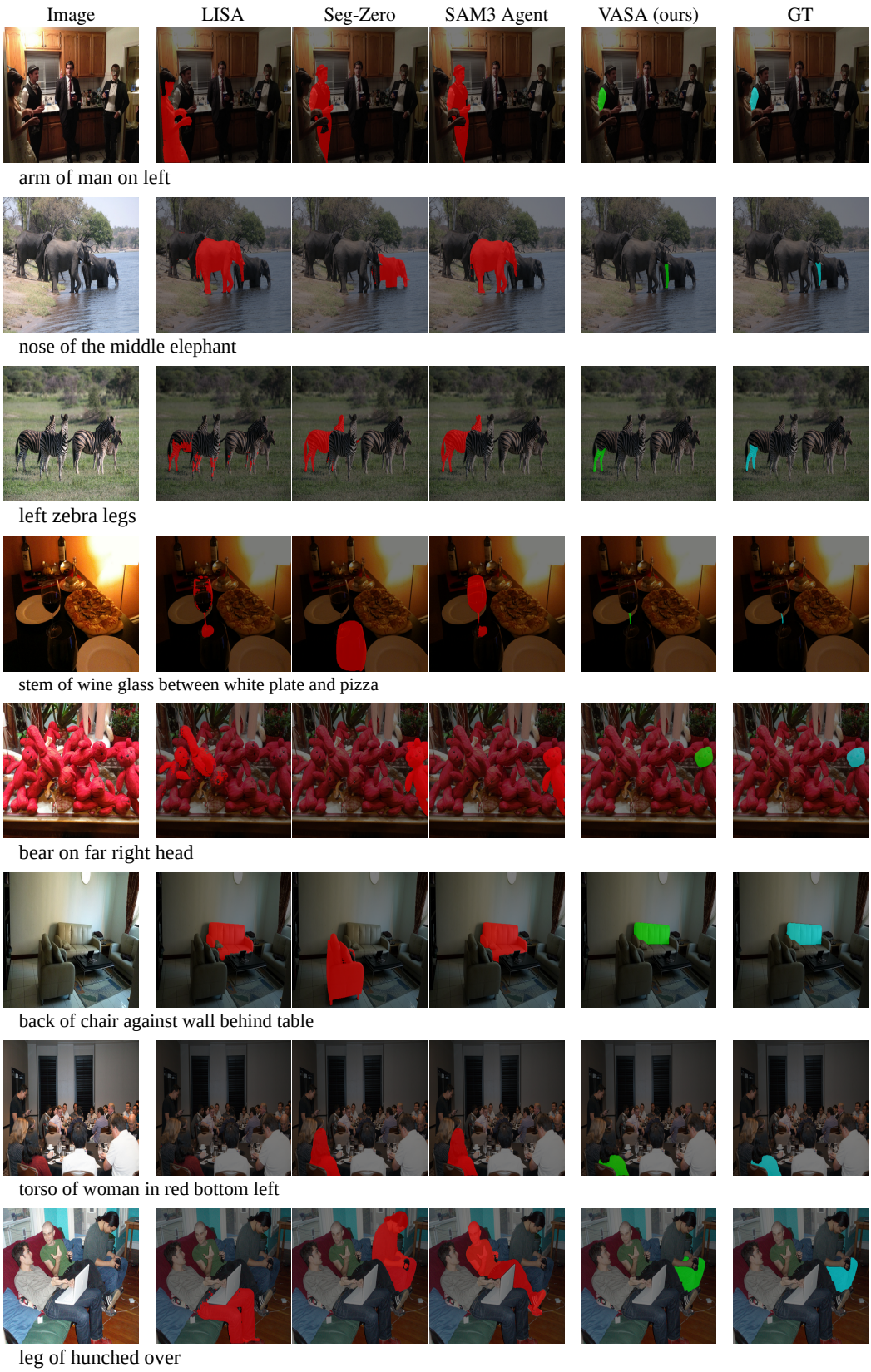



Figure 11: **Additional results of VASA and baselines on RefCOCOm.**




Figure 12: **Additional results of VASA and baselines on RefCOCOm** (continued).