GEAR-Seg: A Grounded Explainable Agent for Reasoning Segmentation and Data Engine

Yanan Wang[1], Wen Li[1], Yibin Ying[1,2], and Zhenghao Fei[1,2]*

[1] College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China
[2] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 311215, China {mmwang, zfei}@zju.edu.cn

Abstract. Reasoning segmentation requires localizing targets based on complex, implicit queries. Current end-to-end models typically entangle perception and deduction into an opaque black box, severely limiting interpretability and scalability. To address this, we propose GEAR-Seg (**G**rounded **E**xplainable **A**gent for **R**easoning **S**egmentation), an explicitly decoupled agent that shifts the paradigm by translating visual pixels into dense, attribute-rich text. By decoupling class-agnostic segmentation, semantic description, and Large Language Model (LLM) deduction, GEAR-Seg transforms implicit reasoning into an explicit, trackable logic chain. As a zero-shot inference framework, it achieves highly competitive performance across diverse reasoning and fine-grained referring segmentation benchmarks. Furthermore, GEAR-Seg inherently functions as a highly scalable data engine. Utilizing this engine, we construct GEAR-131K, a massive benchmark (over 38k images, 656k QA-mask pairs) introducing a multifaceted taxonomy tailored for complex real-world manipulation-oriented reasoning. Finally, distillation experiments demonstrate that lightweight models supervised exclusively by our automated pipeline closely match the upper-bound performance of costly human-annotated baselines.

Keywords: Reasoning Segmentation · Explainable Vision Agent · Knowledge Distillation

1 Introduction

Visual segmentation has witnessed a remarkable evolution, transitioning from standard closed-set pixel classification [27,35,19] to open-vocabulary recognition [24,5], and most recently, to reasoning segmentation [18,3]. Unlike traditional referring expression comprehension [10,33] that relies on explicit noun phrases, reasoning segmentation requires models to accurately localize target regions based on complex, implicit textual queries. This task demands a deep fusion of visual perception and high-level cognitive deduction, making it a challenging benchmark for evaluating visual grounding.

* Corresponding author.

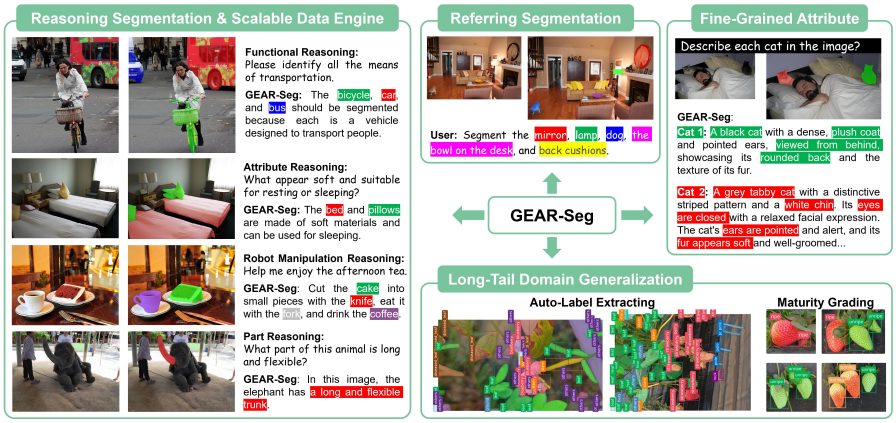


Fig. 1. Overview of GEAR-Seg's multifaceted capabilities. Serving as both a zero-shot inference agent and a scalable data engine, it explicitly translates pixels into text to seamlessly support complex reasoning segmentation, dense referring segmentation, and fine-grained attribute grounding in long-tail domains.

Despite rapid progress, current state-of-the-art (SOTA) architectures typically formulate reasoning segmentation as an end-to-end framework. By directly injecting visual features into LLMs, these methods [3,2,11] implicitly entangle perception and reasoning into an opaque black box. This architectural choice inherently restricts interpretability, limits zero-shot generalization, and creates a profound reliance on expensive human annotations. Consequently, existing benchmarks [3,2] remain severely limited in diversity, often confining queries to isolated instances or homogeneous categories (e.g., targeting "the single hat" or "all hats" in the image). Most critically, they lack a systematic taxonomy covering diverse reasoning scenarios beyond commonsense grounding. For instance, executing a cross-category instruction like *"divide the dessert into small pieces"* demands profound semantic comprehension to deduce the implicit targets (e.g., identifying the "dessert" as the cake and inferring the required tool as a knife), coupled with the precise spatial localization of these interacting entities—a complex deductive capability largely absent in current datasets.

To address these critical bottlenecks, we propose GEAR-Seg (**G**rounded **E**xplainable **A**gent for **R**easoning **S**egmentation), a novel framework that shifts the paradigm from implicit entanglement to explicit modular decoupling. As shown in fig. 1, at its core, GEAR-Seg pioneers a powerful pixel-to-text translation strategy. It decomposes the task into three transparent stages: universal region proposal via the Segment Anything 2 (SAM 2) [22], dense semantic description of every mask using the specialized Describe Anything Model (DAM) [15], and multi-step deduction driven by a pure LLM. By explicitly turning pixels into rich, human-readable text, and utilizing pretrained knowledge in the LLM, this decoupled architecture bypasses the information bottlenecks of previous

methods. It equips GEAR-Seg with dual operational modes: addressing complex reasoning segmentation via abstract deduction, and serving as a formidable dense referring segmentation engine via prompt-constrained grounding. Leveraging these unconstrained visual contexts and detailed physical attributes, GEAR-Seg achieves competitive zero-shot inference performance. Notably, it demonstrates strong generalization in long-tail applications, effectively handling fine-grained real-world tasks such as agricultural maturity grading without requiring manual annotations.

Beyond serving as a robust inference framework, the explicitly decoupled and transparent nature of GEAR-Seg uniquely positions it as a highly scalable data engine. To address the severe data scarcity for downstream deployment, we leverage GEAR-Seg to automatically construct GEAR-131K, a massive reasoning segmentation dataset comprising over 38k images and 656k diverse QA-mask pairs. Breaking away from restrictive "commonsense-only" norms, our dataset introduces a comprehensive taxonomy tailored for complex interactions: Commonsense, Functional, Manipulation-related, Part-based, and Attribute-based reasoning. To validate the high fidelity of our automated supervision, we train end-to-end networks [3] exclusively on our generated data. Extensive experiments demonstrate a profound data-centric advantage: models supervised solely by our low-cost GEAR-Seg pipeline easily surpass the performance of baselines trained on expensive human-annotated datasets.

In summary, our contributions are threefold:

1. **A Modular Agent with Pixel-to-Text Translation:** We propose GEAR-Seg, an explicitly decoupled agent that shifts the field from implicit end-to-end entanglement to transparent, modular reasoning. By systematically replacing sparse category prompts with dense, mask-level attribute descriptions via a *pixel-to-text* paradigm, GEAR-Seg achieves competitive zero-shot performance in reasoning segmentation and dense referring segmentation.
2. **A Scalable Data Engine and Comprehensive Benchmark:** Leveraging GEAR-Seg as an automated generation pipeline, we construct GEAR-131K, a massive, high-fidelity reasoning segmentation dataset (over 38k images, 656k QA-mask pairs). Breaking away from traditional single-target constraints, our dataset introduces a multi-dimensional taxonomy strictly tailored for real-world manipulation-oriented reasoning and functional affordances.
3. **Effective Knowledge Distillation Across Model Scales:** We propose a low-cost, data-centric distillation paradigm. By training end-to-end networks exclusively on our automatically generated dataset, we effectively distill the cognitive capabilities of the GEAR-Seg agent into downstream architectures. Extensive experiments demonstrate that student models—ranging from large VLMs (e.g., LISA [3]) to lightweight models (e.g., YOLOv8 [29])—supervised solely by our generated data achieve highly competitive performance, closely matching the rigorous upper-bound metrics of their counterparts trained on costly human-annotated datasets.

2 Related Work

2.1 Evolution from Referring to Reasoning Segmentation

Visual segmentation has evolved from predefined category recognition to open-ended, language-driven perception [12]. While referring segmentation methods (e.g., VLT [6], CRIS [31]) and open-vocabulary models (e.g., Grounded SAM [24], YOLO-World [5]) excel at localizing targets specified by explicit nouns, they fundamentally struggle with implicit queries requiring the deduction of human intents or physical affordances. To bridge this gap, recent works have introduced reasoning segmentation. Pioneering models like LISA [3], LLM-Seg [2], and PixelLM [25] address this by injecting visual tokens into LLMs to perform deduction and segmentation end-to-end. However, these methods implicitly entangle perception and reasoning within an opaque black box. This architectural limitation not only restricts interpretability but also degrades performance in long-tail domains that demand fine-grained attribute differentiation [14]. Consequently, there is an urgent need for paradigms that explicitly decouple visual perception from cognitive deduction.

2.2 Benchmarks for Reasoning Segmentation

Due to the emerging nature of reasoning segmentation, high-quality and comprehensive datasets remain relatively scarce [3,2,11]. Although recent benchmarks have expanded task boundaries from single-instance targets to multi-category scenarios, they predominantly lack designs tailored for embodied AI and complex robotic manipulation [25]. Early benchmarks, such as ReasonSeg [3], rely heavily on costly human annotations, severely limiting their scale. To alleviate this, recent works explore automated data generation pipelines; for instance, LLM-Seg [2] employs GPT for prompt generation, and PixelLM [25] utilizes LLaMA for brief target descriptions. Nevertheless, these methods are bottlenecked by their reliance on predefined vocabularies or existing dataset annotations, failing to provide the LLM with an unbiased, holistic scene understanding. Furthermore, existing benchmarks predominantly provide simple binary (Query, Mask) pairs or rudimentary category lists, completely lacking the structured, logical reasoning explanation chains crucial for training interpretable models.

2.3 Agentic AI and Knowledge Distillation

The rapid advancement of LLMs has catalyzed the emergence of Agentic AI [1], characterized by its ability to explicitly decompose complex, open-ended problems into manageable sub-tasks [34]. Visual agents leverage LLMs as central cognitive brains to orchestrate specialized tools via API calls [4,16], achieving impressive zero-shot capabilities. However, deploying such heavy, multi-step agents in real-time or resource-constrained robotic systems is highly impractical [13]. To overcome this, data-centric knowledge distillation [28] has emerged as an

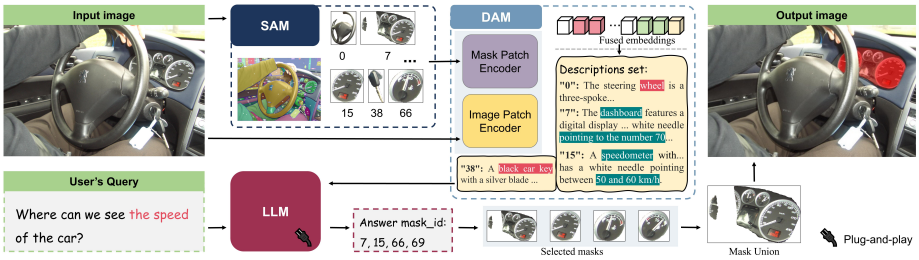


Fig. 2. Overview of the GEAR-Seg framework. The agent explicitly decouples the reasoning segmentation task into class-agnostic perception (SAM 2), dense semantic description (DAM), and logic-driven abstraction (LLM), serving as both a zero-shot inference engine and a scalable data generator.

effective paradigm. Instead of directly distilling model weights, this approach utilizes a powerful agent as a teacher to automatically generate high-fidelity training data, which subsequently supervises lightweight, end-to-end student models [30]. Building upon these insights, we introduce GEAR-Seg. Operating initially as a transparent, decoupled agent, it overcomes previous architectural limitations to achieve state-of-the-art zero-shot reasoning. Crucially, it doubles as a highly scalable data engine to automatically construct a massive benchmark—complete with dense explanation chains—thereby seamlessly distilling its advanced cognitive capabilities into various downstream architectures.

3 Method

3.1 The GEAR-Seg Framework

Given an input image x and an implicit reasoning query Q, our objective is to predict a precise segmentation mask M that satisfies the complex logical intent embedded within Q. To achieve this, we propose the GEAR-Seg framework, as illustrated in Fig. 2. Unlike end-to-end black-box approaches, GEAR-Seg operates across three explicitly decoupled modules:

1) Class-Agnostic Segmentation. We employ SAM 2 [22] in Everything Mode. A Mask Non-Maximum Suppression [30] yields valid instance masks $\mathcal{M} = \{m_1, m_2, \ldots, m_N\}$, ensuring no salient object or subtle background attribute is overlooked.

2) Dense Semantic Description. This is the core pixel-to-text paradigm shift. Using DAM [15], the global context z^G and localized features z^R are extracted. DAM fuses these to generate a context-aware description d_i for each mask, producing a structured sequence $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$. This fine-grained text is the interpretable bedrock for reasoning.

3) Logic-Driven Reasoning. Bypassing entangled VLMs, we deploy a plug-and-play LLM. Given Q and $\mathcal{D}$, the LLM explicitly evaluates semantic entailment to output valid mask indices $I = \mathcal{F}_{\text{LLM}}(Q, \mathcal{D})$. The final output is

$\widehat{M} = \bigcup_{i \in I} m_i$. And this decoupled architecture endows GEAR-Seg with two synergistic modes: **Reasoning Segmentation Mode**, where the LLM autonomously discovers targets from implicit queries via abstract deduction; and **Referring Segmentation Mode**, where fine-grained physical attributes in $\mathcal{D}$ are exploited to ground explicit instructions, significantly outperforming open-vocabulary SOTAs in long-tail tasks.

3.2 Data-Centric Knowledge Distillation

To facilitate edge deployment, GEAR-Seg functions as a scalable data engine via a pure data distillation paradigm. Crucially, the student and teacher models remain architecturally independent. Given unannotated images $\mathcal{X} = \{x_1, \ldots, x_K\}$, GEAR-Seg generates diverse queries Q_g^k alongside pseudo-labels Y_p^k (comprising both explicit logic chains and segmentation masks):

$$Y_p^k = \text{GEAR-Seg}(x_k, Q_g^k) \tag{1}$$

These labels supervise a lightweight student f_θ entirely through its *native* task-specific objective $\mathcal{L}$:

$$\theta^* = \arg\min_\theta \sum_{k=1}^{K} \mathcal{L}\left(f_\theta(x_k, Q_g^k), Y_p^k\right) \tag{2}$$

Crucially, the loss $\mathcal{L}$ refers entirely to the student's *native* objective. For instance, in student Vision-Language Models (e.g., LISA), this native loss inherently comprises both text and mask objectives ($\mathcal{L} = \mathcal{L}_{txt} + \mathcal{L}_{mask}$). This flexible design allows task-specific lightweight models to genuinely inherit GEAR-Seg's deductive strengths while achieving the rapid inference latency required for robotics.

4 The GEAR-131K Benchmark

Beyond zero-shot inference, GEAR-Seg inherently functions as a scalable data engine. To address the scarcity of complex interaction scenarios in existing benchmarks, we construct GEAR-131K, a massive dataset generated via our automated pipeline.

4.1 Data Engine Workflow

As illustrated in Fig. 3, the creation of GEAR-131K leverages the GEAR-Seg framework as an autonomous, closed-loop data engine. The generation pipeline operates by fusing multi-granularity visual semantics. First, macro-level scene contexts and global relationships are extracted using a lightweight VLM (LLaMA 3.2 [8]). Concurrently, at the micro-level, SAM 2 and the Dense Attribute Module (DAM) extract unconstrained, fine-grained semantic regions to capture intricate object details and attributes. Finally, a central LLM acts as the reasoning core,

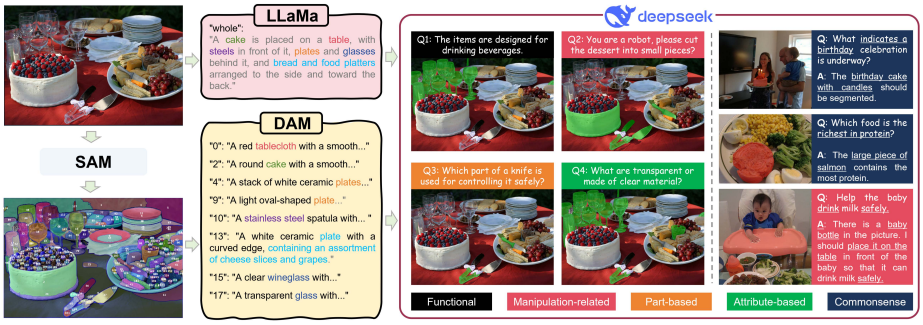


Fig. 3. Overview of the GEAR-Seg data generation pipeline and operational modes.

synthesizing these comprehensive modalities to autonomously generate a diverse set of annotations. For each image, the engine outputs a challenging base query, an explicit step-by-step logic chain, and the corresponding precise mask indices, thereby establishing a high-quality benchmark for reasoning segmentation.

4.2 Taxonomy of Reasoning Scenarios

To bridge the critical gap between standard academic evaluation and complex real-world physical interactions, GEAR-131K pioneers the explicit categorization of multifaceted reasoning scenarios. We systematically break traditional single-target limitations by defining five distinct reasoning categories:

- **Commonsense Reasoning:** Evaluates deductive capabilities based on contextual visual cues (e.g., identifying the birthday cake from *"What suggests a birthday celebration here?"*).
- **Functional Reasoning:** Targets object groups based on utility or affordances, breaking single-category limitations (e.g., retrieving cars, bicycles, and buses simultaneously given *"Identify all means of transportation"*).
- **Manipulation-related:** Designed for complex interaction tasks requiring multi-target coordination across distinct semantic categories (e.g., returning both the cake and knife for *"Divide the dessert into several pieces"*).
- **Part-based Reasoning:** Breaks the holistic object-level boundary by localizing fine-grained, sub-instance regions, demanding deeper structural understanding (e.g., isolating *"the safe-to-hold part of the knife"*).
- **Attribute-based Reasoning:** Uniquely enabled by the rich dense descriptions of GEAR-Seg, this category evaluates the fine-grained perception of physical properties (e.g., material, shape, state). This is crucial for embodied agents where physical attributes dictate safety and affordances (e.g., distinguishing a *"plastic toy knife"* from a *"sharp metallic blade"*).

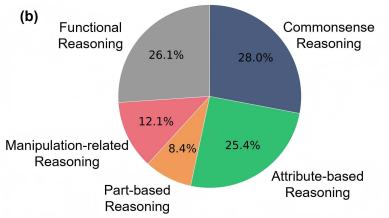
4.3 Dataset Statistics and Integrity Protocol

Scale and Distribution. As detailed in Fig. 4(a), GEAR-131K comprises 39,017 high-resolution images meticulously curated from LVIS [9], VOC [7],

(a)

Dataset	# Image
LVIS	6,462
VOC	11,316
Mapillary	18,001
ADE20K	3,238
Total	39,017

(b)



(c)

(d)

Dataset	Multi-target	Part-level	Attribute-based	Manipulation-related	Q&A pairs
ReasonSeg	×	×	×	×	1.2k
LLM-Seg40k	✓	×	×	×	40k
MUSE	✓	×	×	×	214k
MMR	✓	✓	×	×	194k
GEAR-Seg (ours)	✓	✓	✓	✓	**656k**

Fig. 4. Detailed statistics of the GEAR-131K benchmark. (a) Image distribution across source datasets. (b) Proportion of the five specialized reasoning categories. (c) Word cloud illustrating the semantic diversity of the targeted entities. (d) Comprehensive feature comparison against existing reasoning segmentation datasets.

Mapillary [20], and ADE20K [37]. Our automated engine initially generated 162k raw proposals. To guarantee high-fidelity supervision, we enforced rigorous automated filtering: discarding masks occupying $< 1/600$ of the image area and removing pairs with LLM confidence scores $< 6/10$. This refined the set to 131k base pairs (averaging 3.36 scenarios per image). Following a 5-fold linguistic expansion, the final dataset scales to over 656k QA-mask pairs, featuring a high-density average query length of 14.3 words. Furthermore, fig. 4(b) illustrates a balanced distribution across the five specialized reasoning scenarios, effectively mitigating the bias towards simplistic queries.

Semantic Diversity and Superiority. GEAR-131K encompasses 3,014 distinct target entities, as visualized in fig. 4(c), covering a vast spectrum of open-world categories, fine-grained parts, and unapparent physical attributes. Crucially, fig. 4(d) contextualizes our architectural advantages against existing benchmarks (ReasonSeg [3], LLM-Seg40k [2], MUSE [25], MMR [11]). GEAR-131K establishes a new comprehensive standard as the only benchmark simultaneously facilitating multi-target grounding, part-level reasoning, attribute-based differentiation, and complex manipulation-reltaed at scale.

Data Splits and Annotation Protocol. To prevent data leakage, we partition the dataset strictly at the image level into training (35.4k), validation (2k), and testing (1.8k) splits. Guaranteeing a rigorous, unbiased evaluation, the 1.8k-image test split underwent strict manual verification. Three independent experts reviewed the data following unified annotation guidelines, achieving a $> 95\%$ inter-annotator agreement. Illogical queries were actively filtered, and inaccurate masks were corrected via Labelme, with edge-cases resolved through consensus. This ensures the test set serves as a highly reliable gold standard for complex embodied perception.

5 Experiments

5.1 Zero-Shot Inference Performance of GEAR-Seg

To systematically validate our framework, we evaluate its zero-shot inference capabilities across two synergistic operational modes. First, we assess the Reasoning Segmentation Mode using complex, implicit commonsense queries. Second, we evaluate the Referring Segmentation Mode through explicit, fine-grained attribute grounding in specialized, long-tail domains.

Performance on Reasoning Segmentation Benchmarks. We evaluate the agent's complex deductive capabilities across two representative public benchmarks.

Evaluation Setup. We utilize ReasonSeg [3] and LLM-Seg40k [2] as evaluation datasets. We select SOTA reasoning models LISA [3] and LLM-Seg [2], and the referring method GRES [17] as baselines. Following standard protocols, we report gIoU [26] and cIoU [36]. Additionally, due to the significant variance in image sizes within the ReasonSeg dataset, we report the normalized cIoU (ncIoU) [36] to ensure a balanced evaluation.

Results and Analysis. As summarized in table 1, GEAR-Seg demonstrates highly competitive zero-shot performance. On ReasonSeg, our zero-shot approach achieves a gIoU of 57.5, outperforming the fully fine-tuned LISA-13B (56.2). While GEAR-Seg yields a lower standard cIoU, it establishes superior performance on the size-balanced ncIoU metric (54.0 vs. 52.8). This divergence is an expected byproduct of our decoupled architecture: SAM 2 excels at precise sub-instance localization but occasionally over-segments massive targets in high-resolution images, disproportionately degrading unnormalized cIoU. By neutralizing this scale bias, the superior ncIoU reliably reflects GEAR-Seg's precise instance-level capabilities. Furthermore, on LLM-Seg40k, GEAR-Seg achieves a gIoU of 52.2, establishing a substantial margin over the fine-tuned LLM-Seg-7B (45.5). Some qualitative results are shown in fig. 5(a).

Generalization on Long-Tail Tasks. To validate the Referring Segmentation Mode and inherent generalization in long-tail domains, we extend our zero-shot evaluation to complex agricultural scenarios.

Instance Segmentation. We evaluate zero-shot instance segmentation on three fruit datasets: StrawDI_Db1 [21], Mega_Blueberry [30], and Mega_Peach [30], reporting COCO metrics against Grounded SAM [24], YOLO-World [5], and SDM [30].

As shown in table 2, GEAR-Seg demonstrates robust perception. In the stringent $mAP_{50:95}$ metric, it outperforms the second-best method by 13.2, 10.0, and 4.8 percentage points on StrawDI_Db1, Mega_Blueberry, and Mega_Peach, respectively. Beyond standard prompting, GEAR-Seg autonomously discovers and assigns semantic labels to identifiable instances. As illustrated in fig. 5(b), it successfully discovers long-tail categories like "diseased leaf," underscoring its potential in unstructured environments.

Table 1. Quantitative comparison on the ReasonSeg and LLM-Seg40k validation splits. "Use LLM" indicates integration of a pretrained Large Language Model. "Zero-shot" denotes evaluation without task-specific fine-tuning. Dashes (-) indicate unreported metrics. The **best** and second-best results are highlighted.

Method	Use LLM	Zero-shot	ReasonSeg			LLM-Seg40k	
			gIoU	cIoU	ncIoU	gIoU	cIoU
GRES	×	✓	22.4	19.9	-	14.2	15.9
LISA-7B	✓	✓	44.4	46.0	-	33.2	38.0
LISA-13B	✓	✓	48.9	46.9	-	-	-
LLM-Seg-7B	✓	✓	47.4	35.4	43.5	36.0	39.4
LISA-7B	✓	×	52.9	54.0	-	37.6	48.5
LISA-13B	✓	×	56.2	**62.9**	52.8	-	-
LLM-Seg-7B	✓	×	55.4	45.6	51.0	45.5	**54.2**
GEAR-Seg (Ours)	✓	✓	**57.5**	47.5	**54.0**	**52.2**	52.9

Table 2. Zero-shot fruit instance segmentation results. The **best** and second-best results are highlighted.

Method	StrawDI_Db1			Mega_Blueberry			Mega_Peach		
	$mAP_{50:95}$	mAP_{50}	mAR	$mAP_{50:95}$	mAP_{50}	mAR	$mAP_{50:95}$	mAP_{50}	mAR
Grounded SAM	0.242	0.329	0.425	0.297	0.328	**0.706**	0.350	0.447	0.617
YOLO-World	0.108	0.210	0.211	0.303	0.380	0.542	0.163	0.230	**0.630**
SDM	0.374	0.429	0.586	0.320	0.348	0.607	0.320	0.460	0.577
GEAR-Seg (Ours)	**0.506**	**0.594**	**0.642**	**0.420**	**0.476**	0.595	**0.368**	**0.473**	0.585

Fine-Grained Maturity Grading. To test attribute-based reasoning, we introduce a fine-grained state classification task using 100 images (478 instances) from StrawDI. The model must classify targets into: unripe, turning, and ripe. This demands a synergy of language comprehension, physical perception, and robust dense segmentation. By seamlessly integrating LLM deduction with high-fidelity segmentation, GEAR-Seg overcomes the bottlenecks of existing open-vocabulary models, achieving a robust zero-shot accuracy of 83.2%. As visually validated in fig. 5(c), it performs reliably under severe occlusion.

5.2 Ablation on Modularity and Scalability

To systematically validate the structural advantages of our explicitly decoupled paradigm, we ablate the core text representation module and evaluate the framework's plug-and-play cognitive scalability.

Necessity of Dense Descriptions. A central premise of GEAR-Seg is that implicit reasoning requires fine-grained, attribute-rich representations. To prove this, we replace the DAM module with a general-purpose VLM (LLaMA) to generate mask descriptions. As shown in table 3, this substitution causes a drastic performance collapse on ReasonSeg (57.4 → 35.3 gIoU). Standard VLM captions inherently lack the granular physical and spatial details essential for resolving complex, manipulation-related queries, confirming that DAM's dense descriptions are the indispensable bedrock of our architecture.

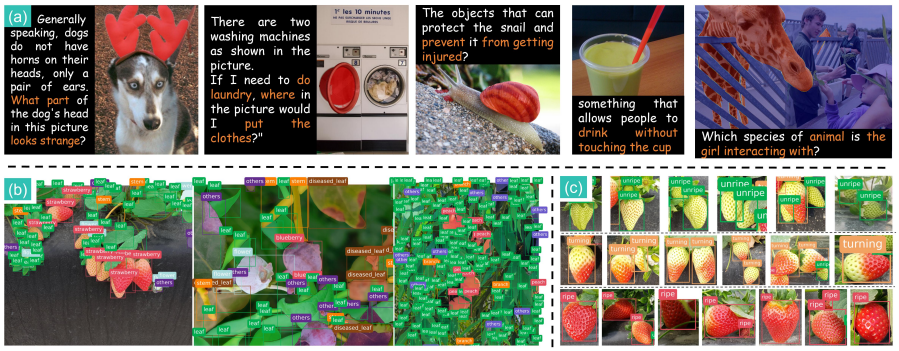


Fig. 5. Qualitative results of the GEAR-Seg agent. (a) Complex reasoning segmentation on ReasonSeg and LLM-Seg40k. (b) Open-world auto-label extraction across diverse agricultural scenes, showcasing the zero-shot discovery of long-tail categories. (c) Fine-grained maturity grading, demonstrating precise attribute-based grounding under severe occlusion.

Plug-and-Play Cognitive Flexibility. Unlike end-to-end black-box architectures that implicitly entangle perception and deduction, our decoupled nature allows for zero-cost cognitive upgrades without retraining. As detailed in table 3, effortlessly substituting the local Qwen3-32B reasoning core with the Gemini-3.1-Pro API directly boosts zero-shot performance to **58.2 gIoU** on ReasonSeg, establishing a highly flexible and competitive scaling path alongside concurrent SOTAs.

Scalability for Edge Devices. While heavy LLMs are indispensable for implicit commonsense reasoning, deploying them in resource-constrained environments is impractical. To validate cognitive on-demand capabilities, we replace the massive LLM with lightweight local models (Qwen-8B/30B [32]) and a non-generative text encoder (Sentence-Transformers [23]) for explicit agricultural referring tasks. As detailed in table 4, thanks to the high-quality dense descriptions provided by DAM, lightweight models seamlessly integrate into the framework. Remarkably, Qwen-8B marginally outperforms DeepSeek-R1 on the StrawDI dataset (e.g., 0.512 vs. 0.507 in $\text{mAP}_{50:95}$). This implies that for direct attribute-matching tasks, smaller models effectively bypass the parsing complexities introduced by the overly verbose reasoning chains of massive LLMs. This dynamic scalability ensures optimal deployment for edge robotics.

5.3 Cascading Error and Efficiency Analysis

Cascading Error Analysis. As a modular agent, GEAR-Seg inherently encounters cascading errors. To rigorously quantify dataset noise and algorithmic bottlenecks, we manually sampled and analyzed 2,000 QA-mask pairs (1,000 from the reasoning-heavy GEAR-131K and 1,000 from the long-tail StrawDI dataset). As detailed in table 5, SAM's class-agnostic perception remains highly

Table 3. Ablation on modularity and plug-and-play flexibility. We report zero-shot performance on ReasonSeg and LLM-Seg40k. The **best** results are highlighted.

Method Configuration	ReasonSeg			LLM-Seg40k	
	gIoU	cIoU	ncIoU	gIoU	cIoU
SAM + LLaMA + DeepSeek-R1	35.3	17.7	28.1	30.4	26.2
SAM + DAM + Qwen3-32B	53.9	50.2	52.5	48.9	46.4
SAM + DAM + DeepSeek-R1	57.4	47.5	54.0	**52.2**	**52.9**
SAM + DAM + Gemini-3.1-Pro	**58.2**	**49.6**	**57.9**	48.5	48.0

Table 4. Scalability analysis of the cognitive module on agricultural referring tasks. DeepSeek-R1 is substituted with lightweight local models to evaluate plug-and-play robustness. The **best** and second-best results are highlighted.

Cognitive Brain	StrawDI			Mega_Blueberry			Mega_Peach		
	$mAP_{50:95}$	mAP_{50}	mAR	$mAP_{50:95}$	mAP_{50}	mAR	$mAP_{50:95}$	mAP_{50}	mAR
DeepSeek-R1 (685B)	0.507	0.595	0.641	**0.420**	**0.476**	**0.595**	**0.368**	**0.472**	**0.585**
Qwen (8B, local)	**0.512**	**0.605**	**0.653**	0.395	0.442	0.543	0.345	0.439	0.564
Qwen (30B, local)	0.488	0.573	0.650	0.379	0.424	0.523	0.358	0.457	0.580
Sentence-TransF.	0.456	0.546	0.622	0.391	0.410	0.507	0.353	0.435	0.582

stable across domains. However, downstream failure modes diverge significantly based on task complexity. In GEAR-131K, LLM deduction and choice errors dominate (combined 51.25%), reflecting the inherent difficulty of complex semantic reasoning. Conversely, in long-tail dense segmentation (StrawDI), DAM hallucination becomes the primary bottleneck (78.93%). This explicitly underscores that mitigating Large Vision-Language Model (LVLM) hallucinations is critical for reliable fine-grained perception [38]. Typical failure modes are visualized in fig. 6.

Table 5. Quantitative cascading error analysis on 2,000 manually sampled pairs. The breakdown isolates the source of failure across the decoupled modules.

Dataset	Accuracy	Error Breakdown			
		SAM	DAM	LLM Question	LLM Choice
GEAR-131K	0.73 (gIoU)	12.46%	36.28%	15.03%	36.22%
StrawDI	0.66 (mAP_{50})	13.26%	78.93%	/	7.81%

Efficiency and Amortized Cost. A common critique of agentic frameworks is inference latency. For GEAR-Seg, reasoning a highly dense 60-mask image on an NVIDIA RTX 4090 requires ∼34.3s (SAM: 2.3s/img, DAM: 0.5s/mask). While seemingly intensive for real-time edge deployment, this explicitly decoupled translation functions as a *one-time* caching mechanism. Once the dense textual representation $\mathcal{D}$ is generated, the LLM can instantly resolve an unlimited number of diverse queries for the same image without any redundant visual re-computation. This renders the amortized computational cost highly efficient,

Evaluation across 5 reasoning scenes on 1k pairs

	Fun.	Rob.	Part.	Attr.	Com.	All
giou	0.80	0.76	0.65	0.63	0.81	0.73
ciou	0.84	0.72	0.59	0.66	0.84	0.78

Fig. 6. Accuracy evaluation and typical failure modes of the GEAR-Seg agent, illustrating cascading errors in complex reasoning and attribute hallucination.

making it a highly worthwhile tradeoff for offline dataset generation and complex multi-turn reasoning tasks.

5.4 Knowledge Distillation to End-to-End Models

To fully unleash the potential of the massive datasets generated by our data engine, we conduct knowledge distillation experiments. By transferring the dense cognitive capabilities of GEAR-Seg into end-to-end architectures, we aim to obtain lightweight models suitable for efficient real-time deployment.

Distillation and SFT Utility for Reasoning Segmentation. We rigorously evaluate the distillation efficacy across two dimensions: the capability to substitute human annotation, and the utility of our generated logic chains in teaching semantic comprehension. We select LISA-13B [3] as the student architecture.

Substituting Manual Annotation. We first evaluate on the public ReasonSeg dataset to benchmark label quality. As reported in table 6, when trained on identical 239 images, the GEAR-Seg pseudo-label supervised model recovers a remarkable 92.4% of the gIoU achieved by the strictly human-annotated baseline (51.4 vs. 55.6). Crucially, by effortlessly expanding the training data to 1.2k generated images, the distilled model substantially surpasses the human upper-bound across all metrics (e.g., 57.2 gIoU). This proves that high-fidelity automated supervision from our engine can effectively substitute costly manual annotation via scaled data volumes.

SFT Utility of Logic Explanations. To address whether our explicit reasoning chains actively teach semantic deduction, we ablate the textual loss ($\mathcal{L}_{txt}$) on the full 37.4k GEAR-131K training set. For a more rigorous evaluation, we expanded our manual verification to a highly challenging 1.8k test split. As detailed in table 7, training LISA-13B solely with mask supervision ($\mathcal{L}_{mask}$)

yields a baseline gIoU of 26.6. However, incorporating our explicit explanation chains to co-optimize the native text objective ($\mathcal{L}_{txt}$) triggers a massive performance leap, reaching **37.2 gIoU**. This unequivocally demonstrates that our automated pipeline teaches genuine logical reasoning rather than merely fitting spatial pseudo-masks.

Table 6. Distillation on ReasonSeg val set. Comparing GEAR-Seg pseudo-labels vs. Human annotations using LISA-13B.

Training Images	Label Source	gIoU	cIoU	ncIoU
239	Human	55.6	57.0	53.4
239	GEAR-Seg	51.4	55.4	50.8
1.2k	GEAR-Seg	**57.2**	**63.6**	**55.4**

Table 7. SFT Utility on GEAR-131K test set. Explanations ($\mathcal{L}_{txt}$) significantly boost reasoning capabilities.

Training Data	Explanations ($\mathcal{L}_{txt}$)	gIoU	cIoU	ncIoU
37.4k	× (Mask Only)	26.6	24.6	24.1
37.4k	✓ (Full Loss)	**37.2**	**36.3**	**36.1**

Distillation for Instance Segmentation. To validate practical utility for resource-constrained robotics, we deploy YOLOv8s for long-tail fruit instance segmentation. We compare lightweight student models supervised entirely by pseudo-labels from various automated pipelines against a rigorous human-annotated upper bound.

As detailed in table 8, the GEAR-Seg supervised model substantially outperforms competing automated methods (e.g., Grounded SAM and YOLO-World). Remarkably, without any human intervention, our distilled model recovers up to 93.2% of the rigorous manual baseline's performance (e.g., $mAP_{50:95}$ on Mega_Blueberry). These results confirm GEAR-Seg's capability to supply highly accurate, scalable supervision for ultra-efficient edge models.

Table 8. Zero-shot distillation results for fruit instance segmentation using YOLOv8s. The manual label baseline serves as the upper bound (*). The **best** and <u>second-best</u> results among automated methods are highlighted.

Label Source	StrawDI_Db1			Mega_Blueberry			Mega_Peach		
	$mAP_{50:95}$	mAP_{50}	mAR	$mAP_{50:95}$	mAP_{50}	mAR	$mAP_{50:95}$	mAP_{50}	mAR
Manual *	0.773	0.935	0.790	0.766	0.897	0.815	0.680	0.912	0.741
Grounded SAM	0.443	0.567	0.600	<u>0.662</u>	0.776	0.750	0.506	0.705	0.632
YOLO-World	0.278	0.396	0.377	0.623	<u>0.805</u>	<u>0.754</u>	0.545	0.761	0.665
SDM	<u>0.636</u>	<u>0.774</u>	**0.699**	0.596	0.699	0.730	<u>0.580</u>	<u>0.792</u>	<u>0.680</u>
GEAR-Seg (Ours)	**0.671**	**0.788**	<u>0.685</u>	**0.714**	**0.873**	**0.855**	**0.611**	**0.809**	**0.688**

6 Conclusion

We introduced GEAR-Seg, a grounded and explainable agent-based framework shifting reasoning segmentation from black-box entanglement to trackable, logic-

driven deduction. GEAR-Seg delivers highly competitive zero-shot performance across five challenging benchmarks and serves as a scalable automated engine for GEAR-131K. This multifaceted dataset is specifically tailored for complex real-world interactions and diverse manipulation-related tasks. Furthermore, our distillation experiments demonstrate that GEAR-Seg effectively empowers lightweight, edge-deployable models with dense cognitive capabilities, successfully bridging the gap between sophisticated high-level reasoning and the rigorous practical constraints of resource-limited real-world applications.

Limitations and Future Work. As a modular agent, GEAR-Seg inevitably encounters cascading errors—a characteristic challenge inherent to the agentic paradigm—where initial perception inaccuracies regarding small or ambiguous targets can propagate through the subsequent reasoning chain. Addressing these fine-grained perception bottlenecks and further enhancing the robustness of pixel-to-text modality translation represent critical avenues for future research. Ultimately, we hope the GEAR-Seg framework and the GEAR-131K benchmark will serve as foundational catalysts for the next generation of interpretable vision systems in complex real-world applications and human-centric interactions.

Code and Data Availability

The source code and the GEAR-131K dataset are publicly available at `https://github.com/AgRoboticsResearch/GEAR-Seg.git`.

Acknowledgements

We sincerely thank the anonymous reviewers and Area Chairs for their constructive comments that helped improve this paper. This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. 226-2025-00074).

References

1. Acharya, D.B., Kuppan, K., Divya, B.: Agentic AI: Autonomous intelligence for complex goals—a comprehensive survey. IEEE Access **13**, 18912–18936 (2025). https://doi.org/10.1109/ACCESS.2025.3532853
2. Chen, R., Li, C., Wu, Q., Zhong, Y.Z., Han, P., Li, W., Wei, Y., Zhao, Y.: LLM-Seg: Bridging image segmentation and large language model reasoning. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 1765–1774 (2024). https://doi.org/10.1109/CVPRW63382.2024.00183
3. Chen, X., Hu, J., Chen, Z., Li, Y., Darrell, T., Yu, F., Gao, J.: LISA: Reasoning segmentation via large language models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9579–9589 (2024). https://doi.org/10.1109/CVPR52733.2024.00915
4. Chen, Y.C., Li, W.H., Sun, C., Wang, Y.C.F., Chen, C.S.: SAM4MLLM: Enhance multi-modal large language model for referring expression segmentation. In: Eur. Conf. Comput. Vis. pp. 323–340 (2024). https://doi.org/10.1007/978-3-031-73004-7_19

5. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: YOLO-World: Real-time open-vocabulary object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16901–16911 (2024). https://doi.org/10.1109/CVPR52733.2024.01599
6. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: Int. Conf. Comput. Vis. pp. 16301–16310 (2021). https://doi.org/10.1109/ICCV48922.2021.01601
7. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010). https://doi.org/10.1007/s11263-009-0275-4
8. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024). https://doi.org/10.48550/arXiv.2407.21783
9. Gupta, A., Dollár, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5356–5364 (2019). https://doi.org/10.1109/CVPR.2019.00550
10. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Eur. Conf. Comput. Vis. pp. 108–124 (2016). https://doi.org/10.1007/978-3-319-46448-0_7
11. Jang, D., Cho, Y., Lee, S., Kim, T., Kim, D.: MMR: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. In: Int. Conf. Learn. Represent. (2025), https://openreview.net/forum?id=mzL19kKE3r
12. Kirillov, A., Girshick, R.M., Dollár, P., Mahajan, D.R., et al.: Segment anything. In: Int. Conf. Comput. Vis. pp. 4015–4026 (2023). https://doi.org/10.1109/ICCV51070.2023.00371
13. Kozlov, A., Lazarevich, I., Shamporov, V., Lyalyushkin, N., Gorbachev, Y.: Neural network compression framework for fast model inference. In: Lecture Notes in Networks and Systems. vol. 285, pp. 240–253 (2021). https://doi.org/10.1007/978-3-030-80129-8_17
14. Li, Y., Chen, C., Dai, X., Chen, H.: Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10988–10997 (2020). https://doi.org/10.1109/CVPR42600.2020.01100
15. Lian, L., Ding, Y., Ge, Y., Cui, Y., Yala, A., Darrell, T.: DAM: Describe anything model for detailed localized image and video captioning. In: Int. Conf. Comput. Vis. pp. 21766–21777 (2025)
16. Liang, Y., Li, C., Zhang, D., Yang, Z., Wang, B., Mei, T.: CogAgent: A visual language model for GUI agents. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14281–14290 (2024). https://doi.org/10.1109/CVPR52733.2024.01354
17. Liu, C., Ding, H., Jiang, X.: GRES: Generalized referring expression segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23592–23601 (2023). https://doi.org/10.1109/CVPR52729.2023.02259
18. Liu, Y., Zhang, J., Han, J., Yang, Y., Li, C., Gao, J.: LAVT: Language-aware vision transformer for referring image segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18134–18144 (2022). https://doi.org/10.1109/CVPR52688.2022.01762
19. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple Open-Vocabulary object detection. In: Eur. Conf. Comput. Vis. pp. 728–755 (2022). https://doi.org/10.1007/978-3-031-20080-9_42

20. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The Mapillary Vistas dataset for semantic understanding of street scenes. In: Int. Conf. Comput. Vis. pp. 5122–5130 (2017). https://doi.org/10.1109/ICCV.2017.534
21. Pérez-Borrero, I., Marín-Santos, D., Gegúndez-Arias, M.E., Cortés-Ancos, E.: A fast and accurate deep learning method for strawberry instance segmentation. Comput. Electron. Agric. **178**, 105736 (2020). https://doi.org/10.1016/j.compag.2020.105736
22. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. In: Int. Conf. Learn. Represent. (2024). https://doi.org/10.48550/arXiv.2408.00714
23. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. pp. 3982–3992 (2019). https://doi.org/10.18653/v1/D19-1410
24. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded SAM: Assembling open-world models for diverse visual tasks. In: arXiv preprint arXiv:2401.14159 (2024). https://doi.org/10.48550/arXiv.2401.14159
25. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: PixelLM: Pixel reasoning with large multimodal model. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 26374–26383 (2024). https://doi.org/10.1109/CVPR52733.2024.02491
26. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized Intersection Over Union: A metric and a loss for bounding box regression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 658–666 (2019). https://doi.org/10.1109/CVPR.2019.00075
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Med. Image Comput. Comput.-Assist. Intervent. pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28
28. Sachdeva, N., Dhaliwal, M., Wu, C.J., McAuley, J.: Infinite Recommendation Networks: A data-centric approach. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 31292–31305 (2022)
29. Varghese, R., S. M., S.: YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In: International Conference on Artificial Intelligence and Data Sciences. pp. 1–6 (2024). https://doi.org/10.1109/ADICS58448.2024.10533619
30. Wang, Y., Fei, Z., Li, R., Ying, Y.: Learn from foundation model: Fruit detection model without manual annotation. Pattern Recognition **174**, 112799 (2026). https://doi.org/10.1016/j.patcog.2025.112799
31. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: CRIS: CLIP-driven referring image segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11676–11685 (2022). https://doi.org/10.1109/CVPR52688.2022.01139
32. Yang, A., Li, A., Yang, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025). https://doi.org/10.48550/arXiv.2505.09388
33. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: MAttNet: Modular attention network for referring expression comprehension. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1307–1315 (2018)
34. Zhang, L., et al.: AI agents vs. Agentic AI: A conceptual taxonomy, applications and challenges. Information Fusion **122**, 103599 (2025). https://doi.org/10.1016/j.inffus.2025.103599
35. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., et al.: Rethinking semantic segmentation from a sequence-to-sequence

perspective with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6877–6886 (2021). https://doi.org/10.1109/CVPR46437.2021.00681
36. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: Faster and better learning for bounding box regression. In: AAAI. pp. 12993–13000 (2020). https://doi.org/10.1609/aaai.v34i07.6999
37. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5122–5130 (2017). https://doi.org/10.1109/CVPR.2017.544
38. Zhu, L., Chen, T., Xu, Q., Liu, X., Ji, D., Wu, H., Soh, D.W., Liu, J.: Popen: Preference-based optimization and ensemble for LVLM-based reasoning segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)