# From Words to Pixels: A Comprehensive Survey on Large Language Models in Visual Segmentation

**Yizhou Wang**[†‡*] **Mang Tik Chiu**[‡] **Lingzhi Zhang**[‡] **Xuan Shen**[†]
**Sohrab Amirghodsi**[‡] **Yun Fu**[†]
[†]Northeastern University, [‡]Adobe
wyzjack990122@gmail.com, yunfu@ece.neu.edu

## Abstract

Visual segmentation, the task of segmenting an image into semantically meaningful regions, is a cornerstone in machine learning and has widespread applications in industry. Nevertheless, visual segmentation with instruction has been a challenging task for many years. This largely stems from the cross-modal discrepancy between language and image domains, resulting in difficulty in relating the instruction semantics and the pixel-level predictions. In recent years, the remarkable reasoning capabilities of Large Language Models (LLMs) and Large Multimodal Models (LMMs) have spurred a new wave of research aiming to bridge the disparity between natural language instructions and pixel-level understanding. This survey offers the **first** comprehensive overview of the rapidly evolving field of LLM-driven visual segmentation. We categorize existing approaches based on their core objectives and methodologies, including reasoning-based segmentation, open-vocabulary segmentation, grounding techniques connecting language to pixels, and extensions to video domains. We review recent seminal works in LLM-based visual segmentation, analyzing their architectural innovations, training strategies, and benchmark performance. Furthermore, we discuss the common datasets, evaluation metrics, and identify key challenges and promising future directions at the intersection of language and visual segmentation. We hope this survey serves as a valuable resource for researchers and practitioners seeking to understand the current landscape and future directions of leveraging LLMs for sophisticated visual segmentation tasks and applications. The resource summary is available at https://github.com/wyzjack/Awesome-LLM-Visual-Segmentation.

## 1 Introduction

Visual segmentation, the process of assigning a label to every pixel in an image corresponding to the object class or region it belongs to, is a fundamental task in computer vision with broad applications spanning autonomous driving, medical image analysis, and robotics (Kirillov et al., 2023). Traditional segmentation methods often rely on predefined, fixed category labels and struggle with understanding complex, nuanced instructions or segmenting objects described in natural language, especially in open-world scenarios. The advent of Large Language Models (LLMs) (Zhao et al., 2023; Dong et al., 2024) and Large Multimodal Models (LMMs) (Yin et al., 2024; Zhang et al., 2024a) has revolutionized natural language processing and demonstrated unprecedented capabilities in understanding, reasoning, and generation. This success has inspired researchers to explore their potential in narrowing the gap between high-level semantic understanding, often expressed in language, and low-level pixel perception required for segmentation. Integrating LLMs into segmentation frameworks promises systems that interpret complex, free-form language queries, perform reasoning about object relationships and attributes, and segment corresponding regions in images or videos with greater flexibility and accuracy.

This survey offers a comprehensive review of the recent advancements in leveraging LLMs for visual segmentation tasks. The field has witnessed explosive growth, with numerous approaches emerging that utilize language models in diverse ways – from providing high-level guidance and reasoning chains (Lai et al., 2024; Liu et al., 2025a; Qian et al., 2024) to enabling direct pixel-level grounding and open-vocabulary understanding (Zhang et al., 2024c; Rasheed et al., 2024; Liang et al., 2023). Foundational models like the Segment Anything Model (SAM) (Kirillov et al., 2023) have

[*]Work was done while Yizhou Wang was an intern at Adobe.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46447–46461

also served as a crucial building blocks for many subsequent works aiming at language-driven segmentation (Chen et al., 2024). The remainder of this survey is organized as follows: We first cover background concepts (Sec. 2) and propose a taxonomy of LLM-based segmentation methods (Sec. 3). We then provide a technical deep dive into architectures (Sec. 4) before discussing representative works in each category (Sections 5-9). Finally, we analyze datasets, evaluation, and performance (Sec. 10), and outline future challenges (Sec. 11).

## 2 Background

### 2.1 Visual Segmentation Paradigms

Visual segmentation aims to partition an image into distinct regions corresponding to multiple segments or regions, often associating each pixel with a specific label. Key paradigms include: **Semantic Segmentation:** Assigns each pixel to a pre-specified semantic category (e.g., car, person, road, building). Models like FCN (Long et al., 2015), U-Net (Ronneberger et al., 2015), and DeepLab (Chen et al., 2017) are foundational. **Instance Segmentation:** Extends semantic segmentation by distinguishing among distinct instances belonging to the same object category (e.g., labeling each individual car separately). Mask R-CNN (He et al., 2017) is a seminal work. **Panoptic Segmentation:** Unifies semantic and instance segmentation, assigning both a semantic label and an instance ID (if applicable) to every pixel, covering both "stuff" (amorphous regions such as the sky, road) and "things" (countable objects such as cars and pedestrians). See (Kirillov et al., 2019). **Referring Expression Segmentation (RES):** Segments the specific object instance referred to by a natural language description (e.g., Liu et al., 2023a).

Traditional methods typically rely on manually annotated datasets with fixed category labels and often struggle with zero-shot generalization to unseen classes or understanding complex, free-form language instructions. For a quantitative comparison of representative pre-LLM RES methods with recent LLM-based approaches, see Appendix C.

### 2.2 Large Language and Multimodal Models

**Large Language Models (LLMs)** (Yin et al., 2024; Dong et al., 2024) are deep learning models, typically based on the Transformer architecture (Vaswani et al., 2017), trained on vast amounts of text data. They demonstrate remarkable capabilities in natural language understanding, generation, and reasoning. Key characteristics relevant to segmentation include: **Instruction Following:** Ability to interpret and follow complex instructions provided in natural language. **Reasoning Capabilities:** Capacity for multi-step logical deduction, spatial reasoning (inferred from text), and commonsense knowledge. **In-Context Learning:** Ability to adapt to novel tasks or contexts based on examples provided within the prompt.

**LMMs** or Vision-Language Models (VLMs) (Yin et al., 2024; Zhang et al., 2024a) extend LLMs by integrating visual information. They are trained on large datasets consisting of paired images and text (e.g., CLIP (Radford et al., 2021a), ALIGN (Jia et al., 2021), BLIP (Li et al., 2022)). LMMs aim to bridge the divide between vision and language, thereby enabling tasks like: **Image/Video Captioning:** Generating textual descriptions of visual content. **Visual Question Answering (VQA):** Answering questions about images or videos. **Image-Text Retrieval:** Finding images corresponding to text queries, and vice-versa. **Vision-Language Grounding:** Associating phrases or concepts in text with specific regions in an image. The ability of LLMs/LMMs to understand nuanced language and reason about content makes them powerful tools for enhancing traditional segmentation approaches, leading to the developments surveyed in the following sections. While our primary focus is on methods that integrate modern LLM backbones for their advanced reasoning capabilities, we also discuss foundational Vision-Language Models like CLIP. These are not only crucial as components within larger architectures but are also featured in some standalone approaches for completeness.

## 3 Taxonomy of LLM-Based Segmentation

To provide a structured overview of the field of LLM-driven visual segmentation, we propose a taxonomy based on the primary task formulation and the role the language model plays within the segmentation pipeline. Our main categories are:

1. **Reasoning Segmentation (Sec. 5):** Leverages LLM reasoning for complex instructions involving spatial relationships and multi-step inference. Works include LISA (Lai et al., 2024), Seg-Zero (Liu et al., 2025a), and CoReS (Bao et al., 2024).

2. **Pixel Grounding & RES (Sec. 6):** Focuses on precise localization by directly mapping textual descriptions to pixel masks. Examples: GroundHog (Zhang et al., 2024c), GLaMM (Rasheed et al., 2024).

3. **Open-Vocabulary Segmentation (Sec. 7):** Aims to segment arbitrary object categories beyond training classes from free-form text. Key works: Mask-adapted CLIP (Liang et al., 2023), SegPrompt (Zhu et al., 2023).

4. **Video Segmentation (Sec. 8):** Extends language-guided segmentation to videos, handling temporal consistency and events. Notable works: ViSA (Yan et al., 2024), VideoGLaMM (Munasinghe et al., 2024).

5. **LLM-Enhanced Architectures (Sec. 9):** Focuses on novel integration strategies, such as for few-shot learning (LLaFS (Zhu et al., 2024a)) or building unified models (OMG-LLaVA (Zhang et al., 2024b)).

## 4 Technical Deep Dive: Architectures and Components

While the taxonomy in Sec. 3 categorizes models by their primary objective, a deeper understanding requires analyzing their underlying technical components and architectural patterns. This section provides a technical deep dive into the structural differences of how LLMs are integrated, the fusion strategies employed, and the core components that constitute these models.

### 4.1 LLM Integration Architectures

The central challenge in this field is bridging the semantic divide between high-level language instructions and low-level pixel representations. Different architectural patterns have emerged to address this, as summarized in Tab. 1. These patterns reveal how the role of the LLM and its connection to the vision pipeline vary across different approaches.

### 4.2 Technical Component Analysis

Beyond high-level patterns, the specific choice of encoders, decoders, and fusion mechanisms defines a model's capabilities. Tab. 2 provides a detailed breakdown of the core technical components for a range of representative models, illustrating the substantial architectural diversity in the field. This matrix shows that while some models build upon common backbones like LLaMA and ViT, their fusion strategies and mask decoders are highly specialized, reflecting the distinct technical challenges of their target tasks (e.g., reasoning vs. grounding).

### 4.3 Information Flow and Innovation Patterns

The architectural differences also manifest in how information flows between the language and vision modalities and how different categories of models innovate to solve core technical challenges. Fig. 1 outlines common information-flow patterns, from sequential pipelines to more complex hierarchical and adaptive interactions. Tab. 3 provides a cross-category look at innovation, mapping technical challenges like language grounding and multi-step processing to the specific solutions developed within different research paradigms. Fig. 2 further links our taxonomy in Section 3 to the orthogonal technical axes introduced in Section 4, providing a visual roadmap to navigate how each task category draws upon different architectural mechanisms.

## 5 Reasoning Segmentation

Reasoning segmentation represents a significant advancement where models are tasked not just with recognizing objects, but with understanding complex user intentions, spatial relationships, and contextual information conveyed through natural language to perform targeted segmentation. This often involves interpreting multi-turn dialogues or intricate instructions that require logical inference or world knowledge.

A seminal work in this area is LISA (Lai et al., 2024). LISA introduces a novel paradigm where the LLM generates a reasoning description based on the input query and image context. This description is then embedded into the vision model using a specific embedding token, '<SEG>' and a MLP-based projection layer, guiding the segmentation process. This approach effectively translates the LLM's understanding into actionable segmentation guidance, enabling it to handle complex queries like "Segment the largest animal" or "Mask the object left of the red car". The illstration of this approach is shown in Fig. 3. The later works on reasoning segmentation using LLMs are more or less based on this foundation. For example, LISA++ (Yang et al., 2023) builds upon this foundation, proposing improvements such as curriculum learning and enhanced data augmentation strategies to further boost performance and robustness. Seg-Zero (Liu et al., 2025a) explores

| LLM Integration Pattern | Methods | LLM Role | Vision-Language Bridge | Training Strategy |
|---|---|---|---|---|
| **SEG Token + Reasoning Chain** | LISA, READ, CoReS, Seg-Zero, VisionReasoner | Multi-step reasoning generator | Special token embedding → mask decoder | End-to-end LLM reasoning training |
| **Dense Cross-modal Alignment** | PixelLM, PSALM, GLaMM | Dense pixel-text correspondence | Pixel-level feature alignment | Dense prediction training |
| **LLM as Query Processor** | GroundHog, SegAgent | Query/instruction processor | Attention-based fusion | Progressive LLM→vision training |
| **Unified Multimodal Architecture** | OMG-LLaVA, HyperSeg | Unified reasoning + segmentation | Hierarchical multimodal fusion | Joint multimodal pre-training |
| **Foundation Model Enhancement** | SAM4MLLM | SAM capability enhancer | LLM → SAM prompt generation | LLM-enhanced SAM training |
| **Temporal LLM Extension** | VISA | Video reasoning coordinator | Temporal-aware LLM processing | Video-specific LLM adaptation |

Table 1: LLM Integration Structural Patterns across representative methods. This analysis highlights the diverse ways models assign roles to the LLM and structure the vision-language bridge.

| Method | Base Language Backbone | Vision Encoder | Fusion Strategy | Mask Decoder | Training Paradigm |
|---|---|---|---|---|---|
| LISA (Lai et al., 2024) | Llama 2 | ViT + CNN | Late Fusion + SEG token | SAM-style | End-to-end + curriculum |
| PixelLM (Ren et al., 2024) | LLaMA | ViT | Dense alignment | Custom dense | Pixel-level reasoning |
| GroundHog (Zhang et al., 2024c) | Llama 2 | CLIP ViT | Cross-attention | Transformer | Progressive training |
| GLaMM (Rasheed et al., 2024) | LLaMA | ConvNeXt | Early Fusion | Custom dense | Multi-task joint |
| VISA (Yan et al., 2024) | Llama 2 | ViT | Video-aware fusion | Temporal decoder | Video-specific training |
| CoReS (Bao et al., 2024) | Llama 2 | ViT + CNN | Coordinated reasoning | SAM-style | Coordination learning |
| SAM4MLLM (Chen et al., 2024) | Llama 2 | SAM backbone | SAM integration | SAM decoder | SAM-enhanced training |
| OMG-LLaVA (Zhang et al., 2024b) | InternLM2 | Multi-scale ViT | Unified fusion | Multi-task decoder | Unified pre-training |
| PSALM (Zhang et al., 2024d) | Phi-1.5 | ViT | Spatial attention | Dense prediction | Attention-based training |
| READ (Qian et al., 2024) | Llama 2 | ViT + CNN | Reasoning chain | SAM-style | Reasoning-focused training |
| SegAgent (Zhu et al., 2025b) | Qwen | ViT | Agentic fusion | Agent decoder | Agent-based training |
| HyperSeg (Wei et al., 2024b) | Phi-2 | Multi-scale ViT | Hierarchical | Multi-resolution | Universal pre-training |
| Seg-Zero (Liu et al., 2025a) | Qwen2.5 | ViT | Zero-shot fusion | Adaptive decoder | Zero-shot learning |
| VisionReasoner (Liu et al., 2025b) | Qwen2.5 | ViT | RL-based fusion | Adaptive decoder | Cognitive RL |
| POPEN (Zhu et al., 2025a) | LLaMA | ViT + CNN | Late Fusion + SEG token | SAM-style | Preference optimization |

Table 2: Technical Component Matrix across various methods. This highlights the architectural diversity, showing different choices for encoders, fusion strategies, and decoders tailored to specific tasks.

a different avenue by framing reasoning-guided segmentation as a cognitive reinforcement learning problem. It employs a reasoning chain approach where the LLM decomposes the task and generates intermediate steps, guiding the segmentation model progressively. This allows for more explicit step-by-step reasoning, potentially handling more complex or ambiguous instructions. VisionReasoner (Liu et al., 2025b) also utilizes reinforcement learning to create a unified framework for multiple visual perception tasks, including reasoning segmentation. It introduces format and accuracy rewards to promote structured reasoning and precise localization. CoReS (Bao et al., 2024) focuses on better coordination between the LLM's reasoning module and the visual segmentation module. It proposes mechanisms to ensure that the high-level reasoning aligns effectively with the low-level pixel predictions, addressing potential inconsistencies between the two modalities. More recently, POPEN (Zhu et al., 2025a) introduced preference-based optimization technique to further refine the reasoning process. Understanding the mechanisms behind these reasoning capabilities is also an active research area. For instance, Qian et al. (2024) investigate the role and behavior of the special '<SEG>' token introduced by LISA, aiming to shed light on how the LLM's reasoning is encoded and utilized by the segmentation backbone.

Furthermore, addressing the potential pitfalls of LLM reasoning, such as hallucination or sensitivity to false premises in the input query, is crucial. Wu et al. (2024) tackle this challenge by proposing methods to teach LMMs to identify and overcome false premises in the segmentation instructions, leading to more reliable outcomes. These approaches collectively demonstrate the power of integrating LLM reasoning into segmentation frameworks, enabling models to move beyond simple pattern matching towards a deeper, context-aware understanding of visual scenes guided by language.

## 6 Pixel Grounding and Referring Expression Segmentation (RES)

While reasoning segmentation focuses on interpreting complex instructions, Pixel Grounding and Referring Expression Segmentation (RES) primarily target the challenge of accurately localizing and segmenting the specific image region(s) described by a given text prompt. This requires a tight coupling between language understanding and pixel-level prediction.

Several recent LMMs have been specifically designed for dense, pixel-level grounding tasks. GroundHog (Zhang et al., 2024c) proposes a method to ground LLMs to holistic segmentation, enabling the model to understand and segment based on descriptions involving multiple objects or complex relationships within the scene. GLaMM (Rasheed et al., 2024) introduces an LMM architecture capable of fine-grained pixel-level

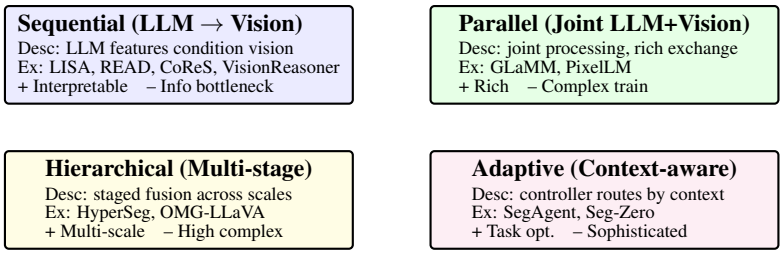


Figure 1: LLM–Vision information-flow patterns.

| Challenge | Grounding/RES | Reasoning | Unified/Architecture | Video |
|---|---|---|---|---|
| **Language Grounding** | Dense align. (PixelLM), Cross-attn (GroundHog), Early fusion (GLaMM) | SEG token (LISA, READ), Coord. reasoning (CoReS), RL-based rewards (VisionReasoner) | SAM integration (SAM4MLLM), Hierarchical (HyperSeg) | Video-aware (VISA) |
| **Multi-step Process** | Single-shot w/ attention | Explicit steps (LISA), Chains (READ), RL-guided chain (VisionReasoner) | Multi-task (OMG-LLaVA), Universal (HyperSeg) | Temporal (VISA) |
| **Novel Concepts** | Agent-based (SegAgent), Dense pred. (PSALM) | LLM injection (LISA), Zero-shot (Seg-Zero, VisionReasoner) | Foundation integration (SAM4MLLM) | Concept transfer (VISA) |

Table 3: Innovation Matrix by Technical Challenge.

grounding for various segmentation tasks, including referring segmentation, salient object detection, and phrase grounding. It emphasizes generating high-quality masks directly based on textual input. Similarly, PSALM (Zhang et al., 2024d) and PixelLM (Ren et al., 2024) develop LMMs focused on pixel-level reasoning and segmentation. They aim to bridge the divide between high-level language instructions and low-level pixel predictions by integrating segmentation capabilities directly into the multimodal architecture. Referring Expression Segmentation (RES) is a specific instance of this grounding task where the objective is to segment the object instance uniquely identified by a natural language expression. Traditional RES methods existed before the recent LMM wave, but integrating large models has pushed the boundaries. For instance, Kim et al. (2023) explore how to extend the image-text alignment capabilities of CLIP (Radford et al., 2021b) specifically for the RES task. Other works like GRES (Liu et al., 2023a), text-augmented spatial-aware approaches (Suo et al., 2023), and Huang and Satoh (2023) also contribute to improving referring segmentation, often by better integrating textual cues with spatial information, which LMM-based approaches implicitly or explicitly leverage. These grounding-focused methods are crucial for applications requiring precise object identification and segmentation based on textual descriptions, such as human-robot interaction, image editing, and detailed image retrieval.

## 7 Open-Vocabulary Segmentation

Traditional semantic segmentation models are typically constrained by a predefined, closed set of object categories seen during training. Open-vocabulary segmentation aims to overcome this limitation by enabling models to segment objects based on arbitrary textual descriptions, even for categories not encountered during the training phase.

Many successful open-vocabulary segmentation approaches leverage the powerful joint image-text embeddings learned by large-scale Vision-Language Models (VLMs) like CLIP (Radford et al., 2021a). These models learn aligned representations where similar concepts in text and images are close in the embedding space. Liang et al. (2023) propose adapting CLIP masks for open-vocabulary semantic segmentation. Their approach focuses on transferring the knowledge from CLIP's image-level understanding to pixel-level segmentation tasks, allowing the segmentation of novel categories described by text. Prompt learning has emerged as a key technique in this domain. SegPrompt (Zhu et al., 2023) introduces category-level prompt learning to boost open-world segmentation. By learning appropriate prompts, the model can better generalize its segmentation capabilities to new textual concepts. Building on this, Li et al. (2024) argue that learning prompts representing relationships between concepts is highly effective for open-vocabulary semantic segmentation. Their Relationship Prompt Learning approach further enhances the model's capacity of understanding and segmenting based on nuanced textual descriptions involving object relationships. Other works focus on improving domain generalization and adaptation. For instance, Zhang and Tan (2025) explore domain-generalized semantic segmentation by combining vision foundation models and vision-

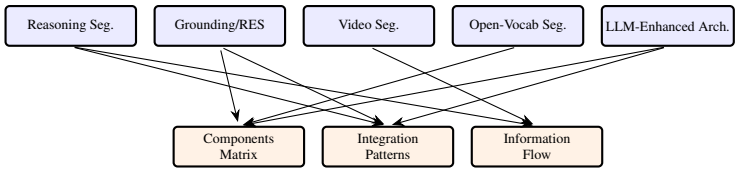


Figure 2: Taxonomy-to-technical map showing how task categories in Section 3 map to technical axes in Section 4.

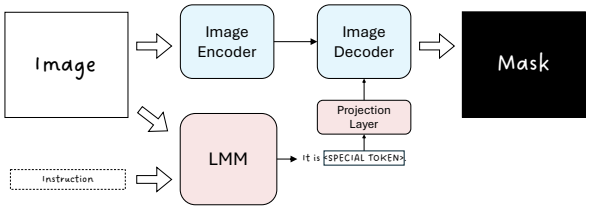


Figure 3: The structure of LISA (Lai et al., 2024) and LISA-based methods for Image segmentation. The hidden embeddings of Special Token (s) in MLLM output serve as the bridge to connect the pixel-level prediction branch and the visual language reasoning branch.

language models, while Basak and Yin (2025) propose a semi-supervised domain adaptation method using vision-language guidance.

## 8 Video Segmentation with LLMs

Extending language-guided segmentation from static images to dynamic videos introduces significant challenges, primarily related to capturing temporal dependencies, understanding motion, and maintaining consistency across frames. Recent works have begun to tackle these challenges by adapting LLM-based approaches for the video domain. ViSA (Yan et al., 2024) adapts the reasoning segmentation paradigm to videos. It leverages LLMs to understand complex language instructions referring to objects and actions over time, generating segmentation masks for the specified entities throughout the video sequence. Similarly, ViLLa (Zheng et al., 2024) proposes a framework for video reasoning segmentation using LLMs, focusing on interpreting instructions that require understanding temporal relationships and events within the video. Capturing fine-grained temporal details is crucial for high-quality video segmentation. Devil (Gong et al., 2025) emphasizes the importance of temporal tokens, proposing methods to effectively integrate temporal information within the LLM reasoning process for improved video segmentation quality. Bai et al. (2024) introduce a method using a single token (‘<SEG>‘) to perform language-instructed reasoning segmentation in videos, simplifying the interface between the language and vision components while handling temporal aspects.Grounding language in video pixels is addressed by VideoGLaMM (Munasinghe et al., 2024), which extends the principles of GLaMM (Rasheed et al., 2024) to the video domain. It aims to provide pixel-level visual grounding for objects and actions described in text across video frames. GLUS (Lin et al., 2025) proposes a unified approach combining global video-level understanding with local frame-level reasoning within a single LLM framework for robust video segmentation.

## 9 LLM-Enhanced Architectures and Strategies

Beyond the specific task formulations discussed previously, several works focus on novel architectural designs and strategies for integrating LLMs/LMMs with visual segmentation backbones, or leverage foundational models in unique ways.

The Segment Anything Model (SAM) (Kirillov et al., 2023), while primarily promptable via points or boxes, has become a foundational model in this space. Its ability to generate high-quality masks for arbitrary objects given simple prompts makes it a powerful component in more complex language-driven systems. For example, Chen et al. (2024) explicitly investigate enhancing LMMs for referring expression segmentation by leveraging SAM's capabilities, demonstrating how foundational segmentation models can be synergistically combined with LLMs. Adapting segmentation models to new tasks or domains with limited data is another area where LLMs show promise. LLaFS (Zhu et al., 2024a) explores the intersection of LLMs and few-shot segmentation, and this direction is further advanced by works like DSV-LFS (Karimi and Poullis, 2025), which unifies LLM-driven semantic cues with visual features for more robust few-shot performance.

Some works aim to build unified LMM architectures that handle not only segmentation but also other vision-language tasks like VQA, captioning, or even generation and editing, all within a single model. OMG-LLaVA (Zhang et al., 2024b) proposes bridging image-level, object-level, and pixel-

level reasoning and understanding in one LMM. Similarly, ViTroni (Fei et al., 2024) presents a unified pixel-level vision LLM designed for understanding, generating, segmenting, and editing images based on language instructions. Other approaches focus on the interface between the LLM and the segmentation model. LLM-Seg (Wang and Ke, 2024) investigates methods for effectively bridging image segmentation modules and LLM reasoning components. LASAGNA (Wei et al., 2024a) proposes a language-based segmentation assistant designed specifically to handle complex, compositional queries that might challenge simpler architectures. HyperSeg (Wei et al., 2024b) aims towards universal visual segmentation by leveraging LLMs, potentially handling a wide variety of segmentation tasks and instructions within a single framework. More recent approaches also explore agent-based systems for segmentation; for example, SegAgent (Zhu et al., 2025b) introduces an agentic methodology for referring segmentation, showcasing another direction in leveraging LLM capabilities for more autonomous visual understanding.

## 10 Analysis and Evaluation

### 10.1 Datasets and Evaluation Metrics

**Datasets:** Commonly used datasets on LLM-based visual segmentation include:

- **Referring Expression Segmentation (RES) Datasets:** These are the most common benchmarks for grounding textual descriptions. Prominent examples include RefCOCO, RefCOCO+, and RefCOCOg (Yu et al., 2016), all built upon the MS COCO (Lin et al., 2014a) dataset. They provide images paired with textual expressions uniquely identifying specific object instances. PhraseCut (Wu et al., 2020), based on Krishna et al. (2017), offers more complex, free-form phrases.

- **Reasoning Segmentation Datasets:** Evaluating complex reasoning often requires specialized datasets. (Lai et al., 2024) construct ReasonSeg evaluation sets by augmenting existing segmentation or VQA datasets with multi-step reasoning questions or complex instructions that require spatial, logical, or commonsense reasoning to arrive at the target mask.

- **Video Segmentation Datasets:** For video tasks, datasets like Refer-YouTube-VOS (Xu et al., 2018) or benchmarks derived from ActivityNet (Heilbron et al., 2015) or EPIC-KITCHENS (Damen et al., 2020) are often used. Recently, ViCaS (Athar et al., 2025) was introduced to support both holistic and pixel-level video understanding.

- **General Segmentation Datasets:** Foundational datasets like PASCAL VOC (Everingham et al., 2010), MS COCO (Lin et al., 2014b), ADE20K (Zhou et al., 2017), and Cityscapes (Cordts et al., 2016) are often used for pre-training segmentation backbones or evaluating open-vocabulary capabilities.

**Evaluation Metrics:** The primary metric for evaluating segmentation performance is the **Intersection over Union (IoU)**. It measures the divide between the predicted segmentation mask ($M_{pred}$) and the ground truth mask ($M_{gt}$):

$$\text{IoU} = \frac{|M_{pred} \cap M_{gt}|}{|M_{pred} \cup M_{gt}|} \quad (1)$$

Commonly used IoU metrics include:

- **Mean IoU (mIoU or gIoU):** The average IoU calculated over all object instances or classes in the dataset.

- **Cumulative IoU/ Overall IoU (cIoU or oIoU):** Total intersection over total union across all pixels in the dataset.

While IoU remains the primary mask-quality metric, it does not fully capture correctness for reasoning-heavy tasks where intermediate steps and constraints matter. For a discussion of IoU limitations and proposed complementary indicators, see Appendix B.

### 10.2 Performance and Scalability Analysis

Tab. 4 and 5 summarize the performance of various models on standard RES and reasoning segmentation benchmarks, respectively. Scalability is another key dimension. As shown in Tab. 6, larger models (13B) consistently outperform their smaller (7B) counterparts in both gIoU and cIoU on ReasonSeg set. The performance gap suggests that the emergent capabilities of larger LLMs are crucial for this more complex task requiring reasoning.

## 11 Challenges and Future Directions

Despite rapid progress, LLM-driven segmentation faces significant challenges, with top models still

| Method | LLM/LMM Backbone | RefCOCO | | | RefCOCO+ | | | RefCOCOg | |
|---|---|---|---|---|---|---|---|---|---|
| | | val | testA | testB | val | testA | testB | val | test |
| LISA (Lai et al., 2024) | LLaVA-1.5 (2024a) | 74.9 | 79.1 | 72.3 | 65.1 | 70.8 | 58.1 | 67.9 | 70.6 |
| PixelLM (Ren et al., 2024) | LLaVA (2023b) | 73.0 | 76.5 | 68.2 | 66.3 | 71.7 | 58.3 | 69.3 | 70.5 |
| GroundHog (Zhang et al., 2024c) | LLaVA-1.5 (2024a) | 78.5 | 79.9 | 75.7 | 70.5 | 75.0 | 64.9 | 74.1 | 74.6 |
| GLaMM (Rasheed et al., 2024) | Vicuna (2023) | 79.5 | 83.2 | 76.9 | 72.6 | 78.7 | 64.6 | 74.2 | 74.9 |
| VISA (Yan et al., 2024) | Chat-UniVi (2024) | 72.4 | 75.5 | 68.1 | 59.8 | 64.8 | 53.1 | 65.5 | 66.4 |
| CoReS (Bao et al., 2024) | LLaVA-v1.5 (2024a) | 76.0 | 78.6 | 72.5 | 65.1 | 70.0 | 58.6 | 69.0 | 70.7 |
| SAM4MLLM (Chen et al., 2024) | LLaVA-v1.6 (2024b) | 79.8 | 82.7 | 74.7 | 74.6 | 80.0 | 67.2 | 75.5 | 76.4 |
| OMG-LLaVA (Zhang et al., 2024b) | InternLM2 (2024) | 77.2 | 79.8 | 74.1 | 68.7 | 73.0 | 61.6 | 71.7 | 71.9 |
| PSALM (Zhang et al., 2024d) | Phi-1.5 (2023) | 83.6 | 84.7 | 81.6 | 72.9 | 75.5 | 70.1 | 73.8 | 74.4 |
| READ (Qian et al., 2024) | LLaVA-1.5 (2024a) | 78.1 | 80.2 | 73.2 | 68.4 | 73.7 | 60.4 | 70.1 | 71.4 |
| SegAgent (Zhu et al., 2025b) | Qwen-VL (2023) | 79.7 | 81.4 | 76.6 | 72.5 | 75.8 | 66.9 | 75.1 | 75.2 |
| HyperSeg (Wei et al., 2024b) | Mipha (2024b) | 84.8 | 85.7 | 83.4 | 79.0 | 83.5 | 75.2 | 79.4 | 78.9 |
| Seg-Zero (Liu et al., 2025a) | Qwen2.5-VL (2025) | - | 80.3 | - | - | 76.2 | - | - | 72.6 |
| VisionReasoner (Liu et al., 2025b) | Qwen2.5-VL (2025) | - | 78.9 | - | - | 74.9 | - | - | 71.3 |
| POPEN (Zhu et al., 2025a) | LLaVA (2023b) | 79.3 | 82.0 | 74.1 | 73.1 | 77.0 | 65.1 | 75.4 | 75.6 |

Table 4: Performance comparison (cIoU) on Referring Expression Segmentation (RES) benchmarks.

| Method | val | | test | |
|---|---|---|---|---|
| | gIoU | cIoU | gIoU | cIoU |
| LISA (Lai et al., 2024) | 65.0 | 72.9 | 61.3 | 62.2 |
| LISA++ (Yang et al., 2023) | 64.2 | 68.1 | 57.0 | 59.5 |
| VISA (Yan et al., 2024) | - | - | 52.7 | 57.8 |
| CoReS (Bao et al., 2024) | 68.1 | - | 65.5 | - |
| LLM-Seg (Wang and Ke, 2024) | - | - | 62.2 | 62.8 |
| HyperSeg (Wei et al., 2024b) | - | - | 59.2 | 56.7 |
| READ (Qian et al., 2024) | - | - | 62.2 | 62.8 |
| Seg-Zero (Liu et al., 2025a) | 62.6 | 62.0 | 57.5 | 52.0 |
| VisionReasoner (Liu et al., 2025b) | 66.3 | - | 63.6 | - |

Table 5: Performance comparison on the ReasonSeg (Lai et al., 2024) val and test sets.

| Metric | Small (7B) | Large (13B) | Gap |
|---|---|---|---|
| gIoU | 55.5 | 63.0 | +7.5 |
| cIoU | 57.8 | 62.5 | +4.7 |

Table 6: Performance of small vs. large models averaged across methods on ReasonSeg test set.

below 70% IoU on reasoning benchmarks like ReasonSeg (Lai et al., 2024) and exhibiting persistent failure modes in small-region, irregular-shape, and spatial-reasoning scenarios (Fig. 4). Here, we discuss these challenges and future directions.

### 11.1 Challenges

- While progress has been made (Lai et al., 2024; Liu et al., 2025a), handling deeply compositional instructions involving multiple steps of reasoning, complex spatial/temporal relationships, negation, and conditional logic remains difficult (as shown in Fig. 4).
- Large-scale datasets with high-quality pixel masks annotated for complex, multi-step reasoning instructions or conversational interactions are lacking.

Qualitative examples on ReasonSeg further highlight three recurring failure patterns as in Fig. 4. First, both LISA and READ often fail when the target occupies only a tiny region, suggesting a mismatch between training data dominated by common object masks and the fine-grained localization required by reasoning queries. Second, both models struggle with irregularly shaped targets such as nets or mixed-food regions, where SAM-style prompting is especially brittle because accurate masks require unusually precise visual guidance. Third, spatially grounded prompts remain difficult: both models frequently miss the cave that could be explored, the passenger area of the airship, or other context-dependent regions whose interpretation depends on stronger spatial reasoning.

### 11.2 Future Directions

- Developing methods to understand and explain *how* an LLM's reasoning process leads to a particular segmentation output (Qian et al., 2024), increasing trustworthiness.
- Continuing the trend towards unified models (Zhang et al., 2024b; Fei et al., 2024) that can perform segmentation alongside other vision-language tasks, leading to more versatile and capable AI systems.

## 12 Conclusion

In conclusion, this survey offers a comprehensive overview of the emerging field of LLM-assisted segmentation for images and videos. We have categorized diverse methods, analyzed key technical innovations, and reviewed performance on standard benchmarks. Through discussing challenges and future directions, we hope this work can serve as a significant resource which might inspire further innovation at the intersection of language and vision for pixel-level understanding and prediction.

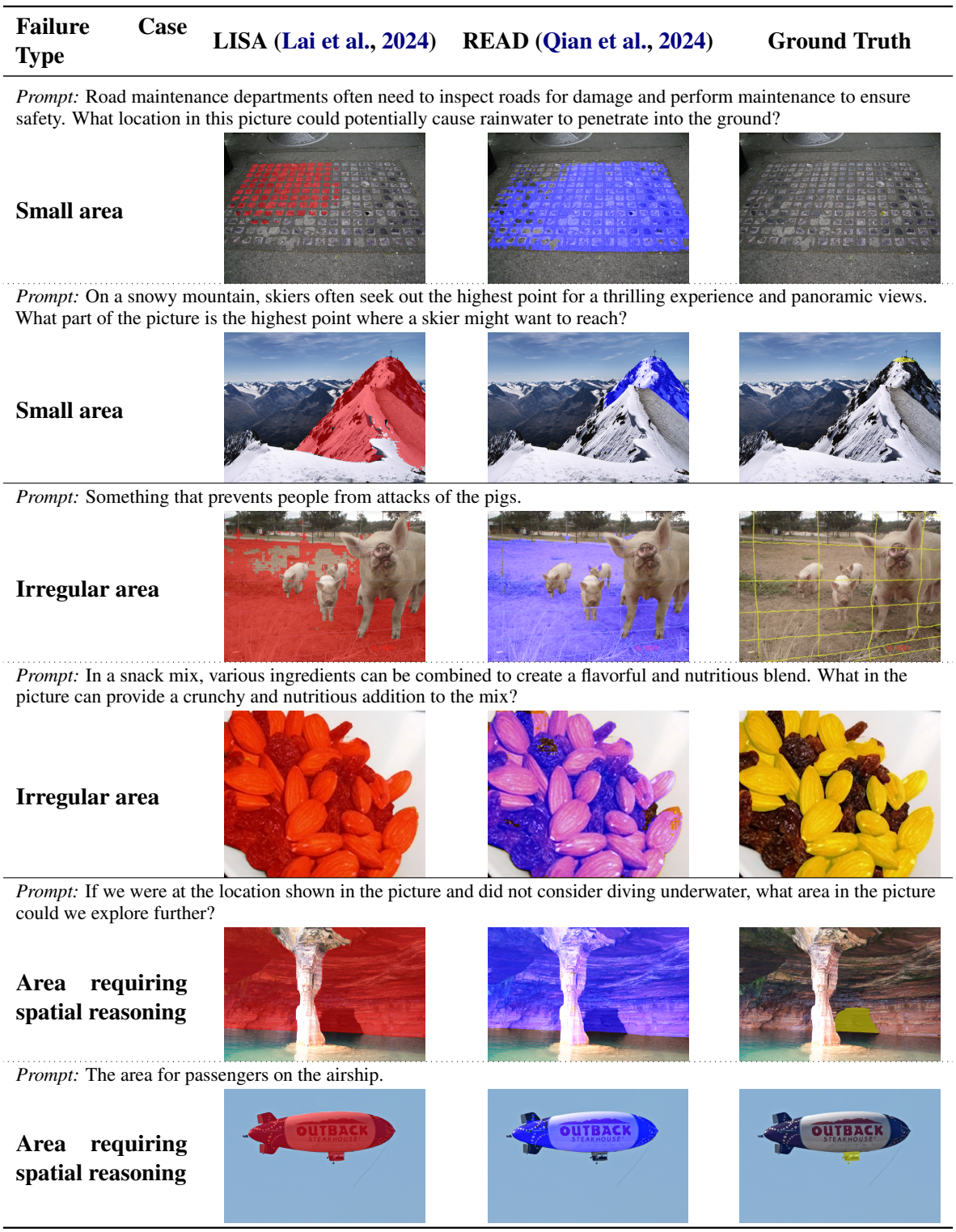


Figure 4: Qualitative failure cases in reasoning segmentation on ReasonSeg (Lai et al., 2024) test set. Each scenario includes a prompt, and blended mask results from LISA-13B-llama2-v1 (Lai et al., 2024), READ-13B (Qian et al., 2024), alongside the ground truth. The blending colors are red for LISA, blue for READ, and yellow for the GT. Examples are grouped by failure type, with same-type cases placed in adjacent rows.

## Limitations

One potential limitation of the paper is that: due to the rapid pace of development in LLM-driven visual segmentation, new methods emerge continuously, making a completely up-to-date overview challenging to maintain. Nevertheless, we believe this survey will be a useful resource for the community for a long time, and we will update it regularly to keep it up-to-date.

## Ethics Statement

We confirm that this survey strictly follows ethical research standards. All of the literature papers mentioned in the main paper are publicly available, and no human participants or personally identifiable information have been included. The aim of the survey is to promote academic understanding and support the responsible advancement of LLM-based visual segmentation research area. All previous related works have been cited appropriately, with due recognition given to their original contributions.

## References


Ali Athar, Xueqing Deng, and Liang-Chieh Chen. 2025. Vicas: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19023–19035.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. 2024. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859.

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

Hritam Basak and Zhaozheng Yin. 2025. Semidavil: Semi-supervised domain adaptation with vision-language guidance for semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9816–9828.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. 2024. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and 1 others. 2020. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338.

Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. *arXiv preprint arXiv:2412.19806*.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. 2025. The devil is in temporal token: High quality video reasoning segmentation. *arXiv preprint arXiv:2501.08549*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Ziling Huang and Shin'ichi Satoh. 2023. Referring image segmentation via joint mask contextual embedding learning and progressive alignment network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7753–7762.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Amin Karimi and Charalambos Poullis. 2025. Dsv-lfs: Unifying llm-driven semantic cues with visual features for robust few-shot segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4584–4594.

Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. 2023. Extending clip's image-text alignment to referring image segmentation. *arXiv preprint arXiv:2306.08498*.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. 2019. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.

Jiahao Li, Yang Lu, Yuan Xie, and Yanyun Qu. 2024. Relationship prompt learning is enough for open-vocabulary semantic segmentation. *Advances in Neural Information Processing Systems*, 37:74298–74324.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070.

Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. 2025. Glus: Global-local reasoning unified into a single large language model for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014a. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014b. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Chang Liu, Henghui Ding, and Xudong Jiang. 2023a. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. 2025b. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Gen Luo, Yiyi Zhou, Yunchao Ji, Chunhua Cao, Zhen Li, Haoqiang Fan, and Shiming Xiang. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10034–10043.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. 2024. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. *arXiv preprint arXiv:2411.04923*.

Rui Qian, Xin Yin, and Dejing Dou. 2024. Reasoning to attend: Try to understand how< seg> token works. *arXiv preprint arXiv:2412.17741*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021a. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Junchi Wang and Lei Ke. 2024. Llm-seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1765–1774.

Zhaoqing Wang, Chang Liu, Henghui Ding, and Xudong Jiang. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695.

Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. 2024a. Lasagna: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*.

Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. 2024b. Hyperseg: Towards universal visual segmentation with large language model. *arXiv preprint arXiv:2411.17606*.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225.

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

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. 2018. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. 2024. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115. Springer.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. 2023. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*.

Zhao Yang, Henghui Ding, Jing Lin, and Xudong Jiang. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

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

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. 2024b. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in Neural Information Processing Systems*, 37:71737–71767.

Xin Zhang and Robby T Tan. 2025. Mamba as a bridge: Where vision foundation models meet vision language models for domain-generalized semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14527–14537.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. 2024c. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14227–14238.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. 2024d. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, (2).

Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. 2024. Villa: Video reasoning segmentation with large language model. *arXiv preprint arXiv:2407.14500*.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. 2024a. Llafs: When large language models meet few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3065–3075.

Lanyun Zhu, Tianrun Chen, Qianxiong Xu, Xuanyi Liu, Deyi Ji, Haiyang Wu, De Wen Soh, and Jun Liu. 2025a. Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30231–30240.

Minjie Zhu, Yichen Zhu, Xin Liu, Ning Liu, Zhiyuan Xu, Chaomin Shen, Yaxin Peng, Zhicai Ou, Feifei Feng, and Jian Tang. 2024b. Mipha: A comprehensive overhaul of multimodal assistant with small language models. *arXiv preprint arXiv:2403.06199*.

Muzhi Zhu, Hengtao Li, Hao Chen, Chengxiang Fan, Weian Mao, Chenchen Jing, Yifan Liu, and Chunhua Shen. 2023. Segprompt: Boosting open-world segmentation via category-level prompt learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 999–1008.

Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. 2025b. Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. *arXiv preprint arXiv:2503.08625*.

Xueyan Zou, Xinlong Cheng, Guanglu Song, Yue Su, Yu Zhang, Junnan Li, Jifeng Dai, and Yu Qiao. 2023a. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems*.

Xueyan Zou, Guanglu Song, Yu Zhang, Jifeng Dai, and Yu Qiao. 2023b. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15135–15145.

# From Words to Pixels: A Comprehensive Survey on Large Language Models in Visual Segmentatin

Supplementary Material

## A Societal Impact and Responsible AI

The growing capabilities of LLM-driven segmentation models raise important considerations regarding their societal impact and the need for responsible AI development. The potential for both significant benefits and harms necessitates a careful examination of their application in the real world, particularly concerning safety, fairness, and the potential for misuse.

**Misinterpretation and Misuse** The ability to segment based on nuanced language creates risks of misinterpretation and malicious misuse. A model might fail to grasp the negative constraints or safety-critical context in an instruction, such as misinterpreting "segment everything that is not a bicycle lane" in a way that endangers a cyclist. Beyond misinterpretation, the technology itself can be repurposed for harmful applications. For example, it could be used for automated surveillance by identifying and tracking individuals based on vague textual descriptions in public video feeds, posing a severe threat to privacy. It could also facilitate the creation of sophisticated manipulated media ("deepfakes") by enabling precise, language-based object removal or replacement in images and videos. Developing models with more robust contextual understanding and establishing clear ethical guidelines and technical safeguards are crucial steps to mitigate these risks and ensure the responsible deployment of this technology.

The information-flow schematic (Fig. 1) highlights where interaction capacity increases (Parallel/Hierarchical) at the cost of training complexity, while the Sequential pattern offers interpretable chains but risks an information bottleneck. The taxonomy-to-technical map (Fig. 2) makes explicit how categories in Section 3 align with integration patterns in Section 4, helping readers quickly locate representative methods by both goal and mechanism.

## B Limitations of IoU and Complementary Indicators

While IoU (gIoU/cIoU) remains the primary mask-quality metric, it does not fully capture correctness for reasoning-heavy tasks where intermediate steps and constraints matter. We therefore recommend reporting, alongside IoU, light-weight complementary indicators:

- **Step-consistency accuracy**: agreement between intermediate reasoning statements (e.g., size/order/spatial relations inferred by the model) and measurable mask attributes.
- **Answer-guided IoU**: IoU conditioned on correct intermediate answers when tasks include textual sub-answers; separates language reasoning errors from mask decoding errors.
- **Constraint satisfaction rate**: fraction of prompts where boolean constraints (e.g., "left of", "largest", "not touching") are satisfied by the produced mask.

These indicators can be computed from existing annotations or simple derived attributes and provide a more faithful view of reasoning correctness.

## C Traditional Baselines for RES and Grounding

Table 7 reports pre-LLM referring expression segmentation results in the same format as the main RES table.

Compared to LLM-based results in Tab. 4, pre-LLM methods trail by a consistent margin across splits: (i) RefCOCO testA: best pre-LLM 76.5 vs best LLM 85.7 (+9.2); testB: 70.2 vs 83.4 (+13.2). (ii) RefCOCO+ testA: 71.0 vs 83.5 (+12.5); testB: 57.7 vs 75.2 (+17.5). (iii) RefCOCOg test: 66.2 vs 78.9 (+12.7). The gap widens on RefCOCO+ testB and RefCOCOg, where expressions are longer and more compositional; this aligns with Section 4, where dense cross-modal alignment and reasoning/chain interfaces in LLM-based systems better handle fine-grained attributes and relational cues.

| Pre-LLM Method | Backbone | RefCOCO | | | RefCOCO+ | | | RefCOCOg | |
|---|---|---|---|---|---|---|---|---|---|
| | | val | testA | testB | val | testA | testB | val | test |
| MCN (2020) | DarkNet53 | 62.4 | 64.2 | 59.7 | 50.6 | 55.0 | 44.7 | 49.2 | 49.4 |
| VLT (2021) | Swin-B | 73.0 | 76.0 | 69.6 | 63.5 | 68.4 | 56.9 | 63.5 | 66.2 |
| LAVT (2022) | BERT | 72.7 | 75.8 | 68.8 | 62.1 | 68.4 | 55.1 | 61.2 | 62.1 |
| CRIS (2022) | ResNet101 | 70.5 | 73.2 | 66.1 | 62.3 | 68.1 | 53.7 | 59.9 | 60.4 |
| ReLA (2023a) | Swin-B+BERT | 73.8 | 76.5 | 70.2 | 66.0 | 71.0 | 57.7 | 65.0 | 66.0 |
| X-Decoder (2023b) | Focal-T + DaViT-B/L +Florence | - | - | - | - | - | - | 64.6 | - |
| SEEM (2023a) | Focal-T + DaViT-B/L +Florence | - | - | - | - | - | - | 65.7 | - |

Table 7: Pre-LLM referring expression segmentation results (reported cIoU) on RefCOCO/RefCOCO+/RefCOCOg. Values are adapted from original papers.