Exploring Efficient Reasoning Segmentation with Small Language Models

Changsong Wen, Zelin Peng, Yu Huang, Xiaokang Yang, and Wei Shen✉

MoE Key Lab of Artificial Intelligence, School of Computer Science,
Shanghai Jiao Tong University
{changsong, zelin.peng, yellowfish, xkyang, wei.shen}@sjtu.edu.cn

Abstract. Reasoning segmentation is a challenging vision–language task that performs pixel-level segmentation guided by language reasoning over implicit textual descriptions. Existing methods typically employ large language models (LLMs) to achieve such reasoning capability, yet their substantial computational and memory demands limit practical deployment. To overcome this, we present LReSeg, the first attempt to tackle reasoning segmentation with a small language model (SLM), achieving strong performance with substantially reduced model size and computational cost. To address the challenges of scaling reasoning segmentation to SLMs, we propose two key designs: First, unlike LLMs, SLMs have limited capacity to provide sufficient spatial instruction cues for mask decoding. We introduce register tokens that aggregate text-conditioned spatial features via the SLM's self-attention and inject them into the visual token stream to enrich the mask decoder's inputs. Second, we adopt a unified encoder architecture to eliminate the redundant visual backbone of conventional dual-encoder designs, naturally ensuring feature consistency between reasoning and mask decoding. LReSeg surpasses models of similar scale with superior efficiency, while being nearly 10× smaller (800M vs. ∼8B) and 3.3× faster than conventional 7B-scale LLM-based methods. Code is available at https://github.com/downdric/LReSeg.

Keywords: Multi-modal learning · Reasoning segmentation

1 Introduction

As a fundamental task in computer vision, image segmentation assigns a label to each pixel in an image, enabling a fine-grained understanding of the scene. Segmentation serves as the foundation for a wide range of downstream applications, such as scene understanding [18], autonomous driving [22], and embodied intelligence [41]. Traditional segmentation paradigms are often limited to a fixed set of predefined categories [7] or rely on simple prompts to specify open-vocabulary target regions [33]. Recently, referring segmentation allows users to indicate objects through natural language expressions [25], but such queries are typically explicit and straightforward.

✉Corresponding author.

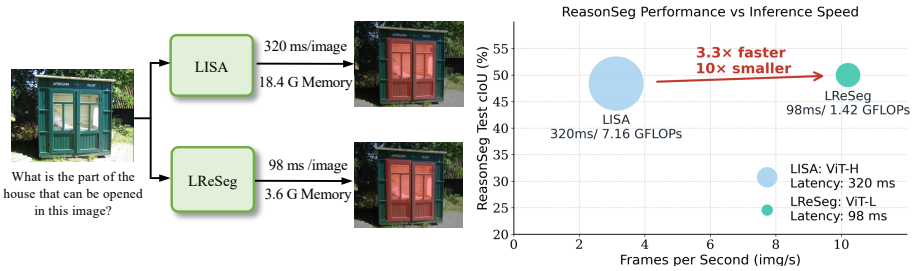


Fig. 1: Performance and efficiency of our LReSeg on ReasonSeg. Compared to LISA (LLaVA-7B), LReSeg runs 3.3× faster with over 10× fewer parameters and only 3.6 GB memory, while maintaining competitive accuracy.

LISA [19] first extends multimodal large language models (MLLMs) [26] to the more challenging reasoning segmentation task by introducing a special `<SEG>` token to represent the segmentation target. Given an image and a complex instruction, reasoning segmentation aims to produce a binary segmentation mask that localizes the objects or regions implied by the instruction. Unlike traditional referring segmentation, which relies on explicit and concise phrases, reasoning segmentation must interpret longer or abstract queries that demand contextual understanding, commonsense reasoning, and external knowledge. For instance, rather than directly mentioning *"the microwave,"* the query may take the form *"which appliance is commonly used to reheat food quickly"* in reasoning segmentation.

Existing reasoning segmentation approaches predominantly rely on large-scale language models with billions of parameters [19,38,45] to interpret complex textual instructions. Although these models exhibit strong reasoning capabilities, their substantial computational and memory costs make them impractical for real-world deployment in resource-constrained settings. In this paper, we propose the first lightweight reasoning segmentation model, termed LReSeg, which is built upon a small language model (SLM) [43]. As illustrated in Figure 1, LReSeg achieves a promising performance, while reducing inference latency from 320 ms to 98 ms per image (3.3× faster) and memory consumption from 18.4 GB to 3.6 GB compared to LISA-7B.

However, scaling reasoning segmentation down to SLMs comes with unique challenges: First, SLMs struggle to provide sufficient semantically grounded spatial cues for generating pixel-level masks. Specifically, the `<SEG>` token is essentially a textual vocabulary token embedded in a language response (*e.g.*, *"It is* `<SEG>`*"*), and thus inherently lacks the spatial instruction cues for dense visual prediction, especially given the constrained model capacity of SLMs. This motivates us to further leverage the SLM's visual tokens for the mask decoder, yet the causal attention mask prevents them from attending to subsequent reasoning text, limiting their awareness of instruction texts. Therefore, we introduce a small set of register tokens appended to the SLM's input sequence. Through the

SLM's self-attention mechanism, these tokens selectively aggregate salient visual features and bring text-conditioned dense spatial cues, which are then fused with the SLM's visual tokens to enhance segmentation.

Second, conventional methods [19,34,36] employ dual-encoder architectures (e.g., utilizing CLIP for MLLM reasoning and SAM for mask decoding), introducing substantial computational redundancy. LReSeg adopts a unified encoder architecture, naturally ensuring the visual features fed to the decoder remain consistent with the language-conditioned representations produced by the SLM (e.g., register tokens), while remaining computationally efficient.

We conduct extensive experiments on multiple vision-language segmentation datasets to demonstrate the effectiveness of our method. On the challenging reasoning segmentation benchmarks, our approach significantly outperforms other models of similar parameter scales while being more efficient in both inference latency and memory consumption. These results highlight the strong segmentation capability and efficiency of our lightweight framework.

2 Related Work

2.1 Multimodal Large Language Model

Recent progress in multimodal large language models (MLLMs) has substantially enhanced the capacity of vision-language tasks [11,12,14]. Proprietary models like GPT-4o [2] and Gemini [42] represent leading advances in unified vision-language generation. They are typically trained on large-scale, diverse multimodal corpora to support strong multimodal understanding and generation abilities. In the open-source community, models such as Qwen-VL [54], DeepSeek-VL [51,29], and Idefics3 [20] introduce advances such as multi-stage visual encoders, vision–language alignment strategies, and extended multimodal context support. These models have demonstrated competitive performance on a wide range of multimodal tasks, including visual question answering [58], detailed image captioning [9], and video-language understanding [4].

Despite their strong capabilities, large-scale MLLMs typically incur substantial computational and memory overhead, which restricts their deployment in resource-constrained settings. To address this limitation, recent research [62,16,57] has investigated multimodal small language models that seek to preserve essential reasoning and generation capabilities within stringent resource constraints. For example, Phi-3 Vision [1] employs a combination of synthetic data and filtered public web data during pretraining to achieve strong generalization with approximately 4.2B parameters. TinyLLaVA [62], LLaVA-OneVision [21], and SmolVLM [31] further demonstrate that MLLMs with as few as one billion parameters can be effectively constructed through efficient architecture design and lightweight vision–language projection modules.

2.2 Reasoning Segmentation

Recent advances in multimodal large language models (MLLMs) have opened new perspectives for reasoning segmentation, leveraging the powerful language-

driven reasoning capabilities of LLMs to facilitate pixel-level scene understanding. Recent studies have extended LLMs to segmentation tasks by introducing architectures that couple high-level linguistic reasoning with low-level visual prediction tasks [13,23,52,6,50,49].

LISA [19] performs reasoning segmentation by using MLLMs to process text and vision jointly, with SAM [17] decoding the output tokens into segmentation masks. Following this pioneering work, recent methods including GSVA [52], MMR [15], GLaMM [36], PixelLM [38], and POPEN [65] have extended segmentation capabilities further, particularly focusing on multi-object segmentation and robustness to queries without target objects. READ [34] leverages the highly activated points from `<SEG>` tokens as prompts for SAM decoder, leading to improved segmentation performance. PSALM [61] and HyperSeg [48] extend MLLMs with unified prompting methods, enabling a single model to handle diverse image and video segmentation tasks. For interactive and conversational segmentation, SegLLM [45] uses dialogue memory to support context-aware object segmentation in multi-turn settings. Recently, works like Seg-Zero [27] and VisionReasoner [28] have explored reinforcement learning strategies to enhance reasoning capabilities in segmentation tasks. LLM-based reasoning has also been expanded to video segmentation [53,10], extending multimodal reasoning capabilities to dynamic vision tasks.

Despite their impressive reasoning capabilities, current reasoning segmentation models often come with substantial computational overhead. This motivates us to explore efficient architectures capable of retaining segmentation capabilities while reducing model complexity and computational costs.

3 Methodology

3.1 Overall Framework of Reasoning Segmentation

Reasoning segmentation [19] is a vision-language task that requires identifying target regions in an image based on indirect textual descriptions. Our reasoning segmentation model consists of four components: a visual encoder f_{vis}, a modality projection layer f_{proj}, an SLM backbone f_{SLM}, and a segmentation decoder f_{seg}. The visual encoder first extracts the visual features $\mathbf{v}$ from the image:

$$\mathbf{v} = f_{\text{vis}}(\mathbf{I}), \tag{1}$$

where $\mathbf{v} \in \mathbb{R}^{N_v \times d_v}$, N_v denotes the number of visual tokens, and d_v is the feature dimension of each token.

To bridge the gap between the visual encoder and the small language model, we design a lightweight modality projection module f_{proj} that efficiently compresses high-resolution visual features $\mathbf{v}$ and aligns SLM representations. The projector applies adaptive average pooling to obtain a compact set of visual tokens, and a lightweight multi-layer perceptron (MLP) maps the pooled features into the SLM embedding space:

$$\mathbf{t}_{\text{v}} = \text{MLP}(\text{AvgPool}(\mathbf{v})) \in \mathbb{R}^{\frac{N_v}{s^2} \times d_{\text{SLM}}}, \tag{2}$$

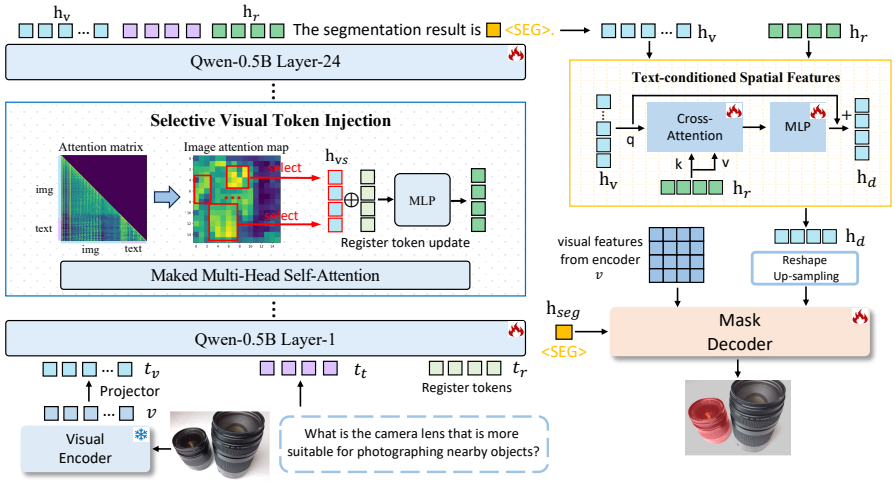


Fig. 2: Pipeline of our method LReSeg. Intermediate text-to-image attention maps are extracted from the SLM to identify salient visual tokens, which are injected into the register tokens and propagated to subsequent SLM self-attention layers to strengthen text–visual interaction. The enhanced text-conditioned spatial features, together with the <SEG> token, are finally fed into the mask decoder to generate the segmentation mask.

where s denotes the spatial downsampling stride of the average pooling operation, d_{SLM} is the embedding dimension of the SLM. This operation reduces the number of tokens, thereby lowering the computational cost for the subsequent processing [5,60].

The SLM takes visual tokens, reasoning text, special <SEG> tokens (expanded from the SLM's vocabulary), and the register tokens as inputs to perform multimodal reasoning. The decoder f_{seg} subsequently processes the <SEG> tokens, the text-conditioned spatial features (obtained by fusing the register tokens with the SLM's visual tokens), and visual features from visual encoder to generate the final mask $\hat{\mathbf{M}}$.

3.2 Unified Encoder for SLM and Segmentation

Previous reasoning segmentation approaches [19,34] typically adopt a dual-encoder architecture, where a vision-language encoder (e.g., SigLIP) is used for semantic understanding in MLLM, and the other large segmentation encoder (e.g., SAM-H with 0.6B parameters) is employed for pixel-level prediction. Such a design introduces redundant parameters and increases computational overhead. A recent study [44] has demonstrated that DINO features exhibit strong potential for reasoning in MLLMs. Moreover, DINO preserves fine-grained spatial structures, making it a versatile unified encoder that bridges reasoning and pixel-level segmentation. Therefore, we train a new MSLM that adopts DINOv3 [40] as its

unified visual encoder. The model is instruction-tuned on standard LLaVA multimodal data [26,19], enabling semantic reasoning and spatially structured visual understanding to operate on a shared visual representation.

3.3 Enriching Spatial Representations with Register Tokens

To enrich the spatial instruction cues provided by SLMs for mask decoding, we first analyze how spatial information flows within the SLM. While the visual tokens inherently encode spatial information, the causal attention mask prevents them from attending to subsequent reasoning text; consequently, the post-SLM visual stream is not text-conditioned, thus insufficient to support mask prediction. To address this, we introduce a set of register tokens at the end of the input sequence to capture text-conditioned spatial cues. These tokens aggregate spatial information from visual tokens and interact with textual tokens through the SLM's self-attention, forming text-conditioned spatial features that effectively support the decoder in mask generation.

Initially, the register tokens $\mathbf{t}_\mathrm{r} \in \mathbb{R}^{m \times d}$ are defined as a set of learnable parameters and optimized jointly with the network during training. The input sequence for the SLM is constructed by concatenating the projected visual tokens $\mathbf{t}_\mathrm{v}$, reasoning text tokens $\mathbf{t}_\mathrm{t}$, and register tokens $\mathbf{t}_\mathrm{r}$ into SLM:

$$\mathbf{h}_0 = T\left(\mathbf{t}_\mathrm{v},\ \mathbf{t}_\mathrm{t},\ \mathbf{t}_\mathrm{r}\right), \tag{3}$$

where $\mathbf{h}_0$ denotes the initial hidden state input to the SLM, T is the instruction template for reasoning segmentation.

To guide the register tokens attending to meaningful visual content, and inspired by the observation in previous work [59] that MLLMs inherently attend to semantically relevant regions, we exploit the intermediate attention of the SLM to strengthen the spatial grounding of the register tokens and enhance their sensitivity to text-aligned visual regions. We select the l-th layer of the SLM and utilize its self-attention map to compute attention scores from reasoning text tokens to visual tokens, based on which the most attended visual tokens are selected. We average the attention weights across all heads and extract the matrix from the hidden states of reasoning text tokens $\mathbf{h}_\mathrm{t}$ to visual tokens $\mathbf{h}_\mathrm{v}$:

$$\bar{\mathbf{A}}^{(l)}_{\mathrm{t}\to\mathrm{v}} = \frac{1}{H}\sum_{n=1}^{H} \mathbf{A}^{(l)}[\mathbf{h}_\mathrm{t},\ \mathbf{h}_\mathrm{v}], \tag{4}$$

where $\mathbf{A}^{(l)}$ denotes the self-attention matrix of the l-th layer, H is the total number of attention heads. For each visual token, we compute the average attention it receives from all text tokens, which measures how strongly the reasoning text attends to that visual token:

$$\alpha_j^{(l)} = \frac{1}{N_t}\sum_{i=1}^{N_t} \bar{\mathbf{A}}^{(l)}_{\mathrm{t}\to\mathrm{v}}(i,j), \tag{5}$$

where N_t is the number of text tokens. We then select the top-k visual tokens with the highest $\alpha_j^{(l)}$ values as $\mathbf{h}_{\text{vs}}^{(l)}$, and fuse them with the register tokens through the MLP update:

$$\mathbf{h}_{\text{r}}^{(l)} = \mathbf{MLP}(\tilde{\mathbf{h}}_{\text{r}}^{(l)} \| \mathbf{h}_{\text{vs}}^{(l)}), \tag{6}$$

where $\tilde{\mathbf{h}}_{\text{r}}^{(l)}$ and $\mathbf{h}_{\text{r}}^{(l)}$ are the register tokens before and after the update, $\|$ is the concatenation operation. This updated representation is then propagated through the subsequent SLM self-attention layers, enabling the register tokens to progressively acquire text-aligned spatial cues that better support segmentation.

3.4 Dual-Path Decoding for Pixel Reasoning

We adopt a dual-path decoding strategy that leverages `<SEG>` token to localize the target and text-conditioned spatial features (i.e., register tokens) to provide dense spatial cues for mask generation. We first fuse the LLM-processed register tokens $\mathbf{h}_r \in \mathbb{R}^{m \times d}$ with the visual tokens $\mathbf{h}_v \in \mathbb{R}^{N_v \times d}$ to obtain text-conditioned spatial features. Specifically, we take the visual tokens as the query and inject register tokens into the visual tokens:

$$\mathbf{h}_d = \mathbf{h}_v + \text{Softmax}\left(\frac{(\mathbf{h}_v \mathbf{W}_q)(\mathbf{h}_r \mathbf{W}_k)^\top}{\sqrt{d}}\right)(\mathbf{h}_r \mathbf{W}_v), \tag{7}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are learnable matrices and $\mathbf{h}_d$ is the text-conditioned spatial features.

Concretely, we regard $\mathbf{h}_{\text{seg}} \in \mathbb{R}^d$ of `<SEG>` token as sparse prompts indicating the semantic target, and the spatially upsampled $\mathbf{h}_{\text{d}} \in \mathbb{R}^{d \times h \times w}$ as dense spatial guidance providing pixel-aligned features. A query projected from the `<SEG>` hidden state interacts with the image feature map augmented with the text-conditioned spatial features to produce the mask logits:

$$\hat{\mathbf{M}} = \text{MLP}(\mathbf{h}_{\text{seg}})^\top \cdot \text{Upsample}(\mathbf{v} + \mathbf{h}_{\text{d}}) \tag{8}$$

We follow previous works [19,38] and simultaneously train both the segmentation and LLaVA instruction tuning tasks. The final loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{txt}} + \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{Dice}}, \tag{9}$$

where $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{Dice}}$ are combination of the binary cross-entropy loss and Dice loss for mask supervision, and $\mathcal{L}_{\text{txt}}$ denotes the language supervision loss. λ_1 and λ_2 are weighting coefficients that balance the contributions of each loss.

4 Experiments

In this section, we evaluate our method on vision-language segmentation tasks. We first introduce the datasets employed in our experiments, along with the implementation details. Then, we conduct a comprehensive comparison between our LReSeg and state-of-the-art methods. Finally, we analyze the model's efficiency and perform ablation studies to validate the contribution of each key component.

4.1 Experiments Setup

Datasets. Our training data is composed of four parts following LISA [19]:
Semantic Segmentation. We utilize ADE20K [63], COCO-Stuff [3], and LVIS-PACO [35] datasets to equip the model with generalizable dense prediction capability across diverse visual scenes. For each image, several object categories are randomly sampled and transformed into instruction-style prompts like: *"USER: [IMAGE] Can you segment the {class name} in this image? ASSISTANT: It is <SEG>."* To enhance diversity, multiple prompt templates are employed.
Referring Segmentation. We adopt the RefCOCO series [32,56] for referring segmentation datasets, including refCOCO, refCOCO+, refCOCOg, and refCLEF, to develop the model's capability to perform segmentation conditioned on natural language expressions. They are naturally converted into instruction-style prompts such as: *"USER: [IMAGE] Can you segment {description} in this image? ASSISTANT: Sure, it is <SEG>."*
Reasoning Segmentation. We adopt the ReasonSeg dataset [19], which provides image–text pairs involving visual reasoning and contextual understanding. Each sample requires the model to infer the target region based on implicit cues described in the text. We follow the instruction format used in LISA [19].
Visual Question Answering. To retain the model's general visual-language reasoning ability, we include VQA data during training. Specifically, we use the LLaVA instruct tuning dataset [26,19], which contains a wide range of instruction-following samples.

4.2 Implementation Details

MSLM training details. Our MSLM architecture employs DINOv3 [40] for visual encoding and Qwen2.5-0.5B [54] as the SLM. The vision-language projector employs an adaptive average pooling architecture with a two-layer MLP projection, generating 144 visual tokens that are subsequently fed into the SLM. We adopt a two-stage training strategy for the proposed MSLM, following LLaVA [26]. In the alignment stage, only the vision-language projector is trained on the subset of CC3M dataset [39] for 1 epoch with a batch size of 256 and a learning rate of 5e-5, using warm-up for the first 5% of total steps followed by cosine decay. The AdamW optimizer is used with a weight decay of 0.01. In the second stage, the full model is fine-tuned for 1 epoch on the LLaVA instruct tuning dataset [26] with a batch size of 128, a learning rate of 2e-5, a warm-up ratio of 3%, and using AdamW with a weight decay of 0.1. Input images are resized to 224×224 for MSLM training to ensure efficiency.
Segmentation training details. Since DINOv3 supports dynamic input resolution, the segmentation images are resized to 1024×1024, and increases the number of visual tokens fed into the MSLM to 256. The segmentation decoder adopts a SAM-like design [17] that jointly integrates text-conditioned dense spatial features and `<SEG>` tokens for mask generation. We train our model for 5 epochs on 8 RTX 3090 GPUs with a batch size of 128. We set λ_1 and λ_2 in the

Table 1: Comparison of reasoning segmentation performance. We compare our method with light-weight counterparts, including segmentation specialists, multimodal small language model–based approaches, and reasoning segmentation methods. We include existing MSLM models that use the <SEG> token as the standard implementation for reasoning segmentation. LISA* and READ* are variants configured with the same SLM as ours, e.g., Qwen2.5-0.5B.

Method	Enc.	Params	Val		Test					
			overall		short query		long query		overall	
			gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
GRES [25]	Swin-Base	226M	22.4	19.9	17.6	15.0	22.6	23.8	21.3	22.0
SEEM [67]	SAM ViT-L	440M	25.5	21.2	20.1	11.5	25.6	20.8	24.3	18.7
OVSeg [24]	Swin-Base	228M	28.5	18.6	18.0	15.5	28.7	22.5	26.1	20.8
X-Decoder [66]	FocalNet-L	341M	22.6	17.9	20.4	11.6	22.2	17.5	21.7	16.3
Grounded-SAM [37]	SAM ViT-H	774M	26.0	14.5	17.8	10.8	22.4	18.6	21.3	16.4
SmolVLM-S [31]	SigLip-Base	259M	41.2	42.6	30.0	31.4	37.5	38.5	35.4	36.7
SmolVLM-B [31]	SigLip-Base	501M	46.6	46.9	32.2	30.6	42.0	45.9	39.5	42.2
InternVL3 [64]	InternViT-300M	1.0B	48.0	48.6	37.8	34.5	51.8	48.4	47.8	45.1
LLaVA-OneVision [21]	SigLip-SO400M	897M	47.8	48.3	36.6	35.2	50.2	49.4	46.8	46.0
LISA* [19]	ViT-L & SAM-H	1.5B	48.5	48.3	37.2	33.7	49.9	48.7	46.5	45.2
READ* [34]	ViT-L & SAM-H	1.5B	49.2	50.3	39.6	33.3	51.7	51.8	48.5	47.1
LReSeg	ViT-L	805M	**52.0**	**52.8**	**41.0**	**37.5**	**54.2**	**54.7**	**50.8**	**50.0**

loss function to 2.0 and 1.5, respectively. The register tokens are randomly initialized, and the number of tokens is set to 16. The intermediate layer l used for attention-guided injection is selected as 6. During training, we fine-tune both the SLM and the segmentation decoder for the segmentation task and the remaining settings are consistent with the second training stage.

4.3 Main Results

Comparison methods. (1) **Segmentation specialists.** We first compare LReSeg with segmentation specialists of comparable scale, such as GRES [25], SEEM [67] , OVSeg [24], X-Decoder [66], and Grounded-SAM [37], which rely on visual or vision-language models. (2) **MSLM methods.** We further compare with existing MSLMs, including SmolVLM-256M (S) [31] , SmolVLM-500M (B) [31], as well as InternVL3 [64] and LLaVA-OneVision [21], both of which are based on 0.5B-scale Qwen models. Following the standard reasoning segmentation setup, we append an extended <SEG> token to the language vocabulary and feed both the <SEG> token and the visual features from the encoder into the SAM decoder for mask generation. (3) **Reasoning segmentation methods.** We compare with existing reasoning segmentation frameworks, including the representative method LISA [19] and the state-of-the-art method READ [34]. For a fair comparison, we replace their original 7B-scale LLMs with the same SLM

Table 2: Performance and efficiency comparison on the ReasonSeg dataset. We evaluate our method against the SOTA MSLM (LLaVA-OneVision) and lightweight reasoning segmentation baselines. Latency and memory are measured in ms and GB, respectively.

Method	gIoU	cIoU	Latency	TFLOPs	Memory
LISA-7B [19]	47.3	48.4	320	7.16	18.4
LLaVA-OneVision [21]	46.8	46.0	264	6.11	11.9
LISA* [19]	46.5	45.2	233	3.08	6.2
READ* [34]	48.5	47.1	238	3.09	6.5
LReSeg	**50.8**	**50.0**	**98**	**1.42**	**3.6**

Table 3: Ablation study on the ReasonSeg validation and test sets (cIoU). We ablate three key components: register token design (A1), selective visual token injection strategy (A2), and unified encoder design (A3). Metrics are reported in cIoU.

Variant	Val	Test
A1. Register Tokens		
`<SEG>` token only (baseline)	47.8	46.6
w/o `<SEG>` (register & SLM visual tokens)	48.2	46.9
w/o regitser tokens (`<SEG>` & SLM visual tokens)	48.5	47.7
Fixed (non-learnable) register tokens	48.7	47.4
A2. Selective Visual Token Injection		
Random selection of visual tokens	50.8	48.1
Uniform fusion (mean pooling)	50.6	48.4
Vision fusion at every layer	52.1	49.0
Remove causal attention mask	49.5	48.2
Remove vision-to-text attention mask	52.2	49.3
A3. Unified Encoder Design		
SigLIP backbone	50.5	47.4
CLIP backbone	49.4	47.2
DINO & SAM	51.2	48.7
LReSeg (ours)	**52.8**	**50.0**

configuration as ours, while preserving their other designs (e.g., dual-encoder structure and training strategy) unchanged.

The reasoning segmentation results on the ReasonSeg dataset [19] are shown in Table 1. Our proposed LReSeg achieves the best overall performance among all compared methods, reaching 52.8 and 50.0 cIoU on the validation and test set. Compared with segmentation specialists, LReSeg exhibits a significant improvement in both gIoU and cIoU. Although SOTA reasoning segmentation models (e.g., READ*) outperform multimodal small language models (e.g., InternVL3 and LLaVA-OneVision), LReSeg demonstrates a clear advantage. The results highlight the effectiveness of LReSeg in achieving strong reasoning segmentation performance while maintaining a lightweight and efficient architecture.

Table 4: Comparison of LReSeg and methods of similar scale on Referring Expression Segmentation benchmarks. Results are reported in cIoU.

Method	refCOCO			refCOCO+			refCOCOg	
	val	testA	testB	val	testA	testB	val	test
MCN [30]	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
VLT [8]	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
CRIS [47]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4
LAVT [55]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
X-Decoder [66]	–	–	–	–	–	–	64.6	–
ReLA [25]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
SEEM [67]	–	–	–	–	–	–	65.7	–
HieA2G [46]	73.3	75.9	69.0	64.8	69.7	56.1	62.9	63.5
LISA* [19]	71.2	72.7	68.1	59.5	66.4	53.2	65.8	67.7
READ* [34]	72.3	76.1	67.8	60.4	68.7	54.3	67.7	68.0
LReSeg	75.1	78.2	70.4	63.1	70.2	56.3	69.6	71.2

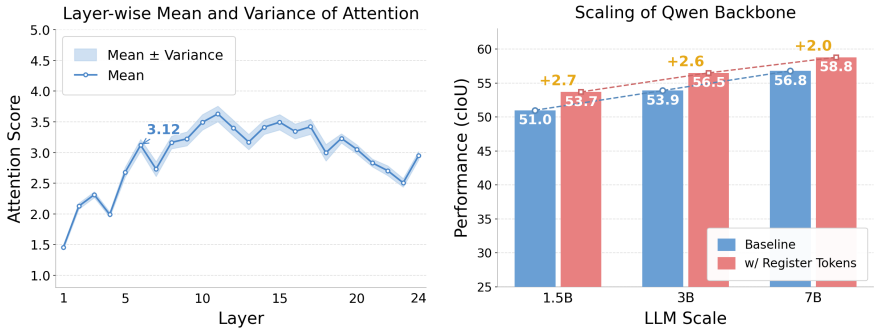
4.4 Efficiency of LReSeg

We evaluate the computational efficiency and inference latency of LReSeg in comparison to the representative segmentation models. The results are shown in Table 2. Our LReSeg achieves a favorable balance between efficiency and segmentation accuracy among models of comparable scale. Compared with reasoning segmentation variants (LISA* and READ*), LReSeg attains consistently higher validation and test performance (52.8, 50.0 cIoU) with notably reduced computation (1.42 TFLOPs), inference latency (98 ms) and memory cost (3.6 GB). Their dual-encoder design requires forwarding images twice, and the use of SAM-H for high-resolution processing introduces substantial computational overhead. LLaVA-OneVision adopts an AnyRes strategy, producing a large number of visual tokens fed into the language model, resulting in substantial encoding and attention cost.

4.5 Ablation Studies

We conduct a comprehensive ablation study to validate the effectiveness of each proposed component, as summarized in Table 3. We first ablate the **Register Tokens (A1)** to evaluate its role in improving segmentation performance. Using only the `<SEG>` token notably reduces performance, confirming that SLM struggles to capture detailed spatial structures within the `<SEG>` token. Incorporating spatial features (w/o `<SEG>`) alone or removing register tokens both yield limited improvements, indicating that register tokens and `<SEG>` tokens work complementarily. Replacing learnable register tokens with fixed ones also leads to a performance degradation.

Next, we investigate the **Selective Visual Token Injection (A2)**. Random selection or uniform visual fusion brings limited gains because the register tokens are far fewer than visual tokens, and excessive injection leads to substantial



(a) Layer-wise mean and variance of the text-to-visual attention ratio between tokens from ground-truth regions and background.

(b) Performance comparison of baseline and adding register token across different Qwen backbone scales (1.5B, 3B, 7B).

Fig. 3: (Left) Text-to-visual attention analysis across transformer layers. (Right) Effect of register tokens across different Qwen backbone scales, demonstrating consistent performance gains of LReSeg over the baseline.

background noise and text-irrelevant cues. Fusing visual information at every layer yields a favorable segmentation performance but lowers the inference speed. Removing the causal attention mask leads to a noticeable drop, as the SLM still relies on causal masking to maintain its language modeling capability due to the inclusion of LLaVA data. Removing the vision-to-text attention mask also leads to inferior performance, as it disrupts the attention patterns established during the alignment pre-training stage and introduces unnecessary computation.

Finally, we analyze the **Unified Encoder Design (A3)**. Replacing DINO with SigLIP or CLIP encoders causes a clear performance drop. While combining DINO with SAM yields slightly better results, it introduces feature inconsistency and higher computational cost. Using DINO as a unified encoder naturally ensures feature consistency between reasoning and decoding, achieving the best performance with a lightweight design. Overall, the full model achieves the highest accuracy on both validation and test splits, reaching 52.8 cIoU and 50.0 cIoU respectively, which confirms the effectiveness of our design.

4.6 Results on RefCOCO

We further evaluate LReSeg on three referring expression segmentation benchmarks, including RefCOCO, RefCOCO+, and RefCOCOg. As shown in Table 4, LReSeg achieves strong generalization performance across all benchmarks, reaching 75.1, 78.2, 70.4 cIoU on RefCOCO, 70.2, 56.3 cIoU on RefCOCO+, and 69.6, 71.2 cIoU on RefCOCOg, surpassing prior segmentation methods like VLT, CRIS, and LAVT. Despite being designed primarily for reasoning segmentation, LReSeg transfers well to the standard referring segmentation task. Moreover, LReSeg still surpasses SOTA reasoning methods, such as READ*, providing

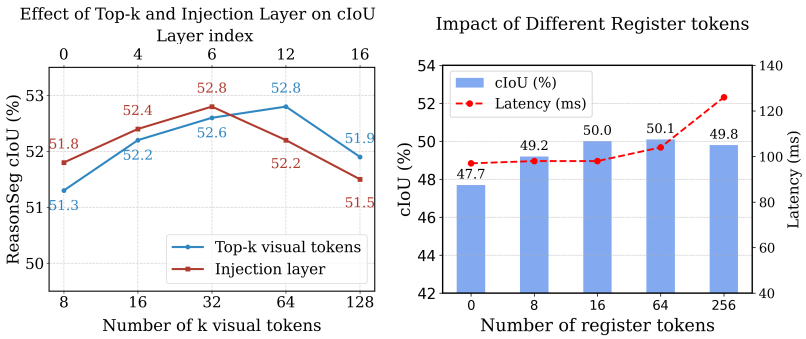


Fig. 4: (Left): impact of the injection layer and the number of selected visual tokens on model performance (ReasonSeg Val). (Right): analysis of the impact of different numbers of register tokens on performance and efficiency.

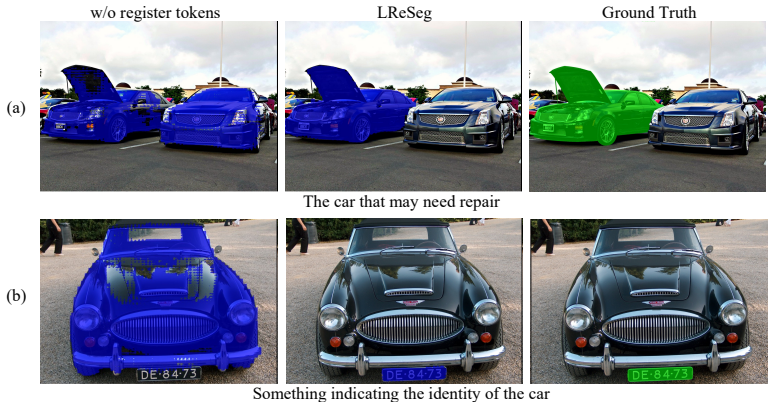


Fig. 5: Qualitative results of our method on reasoning segmentation tasks. (a) shows the baseline that performs segmentation using only the `<SEG>` token, (b) presents our proposed LReSeg, (c) denotes the ground-truth masks.

a favorable balance between efficiency and segmentation accuracy under constrained model capacity.

4.7 Analysis and Discussion

Which layer to apply visual token injection. To analyze at which layer to apply visual token injection, we extract the text-to-visual attention matrix from each transformer layer of the SLM, average the attention weights over all heads, and compute the mean attention ratio between the ground-truth region and other background tokens. A ratio greater than 1 indicates that the model attends more strongly to the correct spatial area. We then plot this attention-ratio curve with variance across all layers in the ReasonSeg dataset, as shown in Fig. 3 (left). We observe that starting from the 6-th layer, the model begins to

clearly focus on semantically relevant regions, suggesting that the $l = 6$ layer is appropriate to perform visual token injection.

Effect of register tokens across backbone scales. To validate the effectiveness of register tokens across different model sizes, we evaluate LReSeg with and without register tokens on Qwen backbones of 1.5B, 3B, and 7B parameters. As shown in Fig. 3 (right), register tokens consistently improve performance across all scales, demonstrating that the proposed design is robust to model size.

Variants of visual token number and injection layer. Fig. 4 left presents the impact of the visual token number and injection layer on the performance of ReasonSeg. Increasing the number of visual tokens improves cIoU up to $k = 64$, after which excessive spatial information introduces redundancy and noise. Meanwhile, applying selective token interaction at the 6-th layer achieves the best performance. Injecting at later layers yields slightly lower accuracy, as the feature fusion occurs too late in the reasoning process, leaving fewer opportunities for subsequent refinement within the SLM.

Efficiency of adding register tokens. We analyze the impact of the number of proposed register tokens on efficiency and accuracy. The added computational overhead is modest when the number of tokens is small. Let the original sequence length be s, appending m tokens increases the self-attention cost from $\mathcal{O}(s^2)$ to $\mathcal{O}((s+m)^2)$, which is approximately linear in m when $m \ll s$. As shown in Fig. 4 right, using $m = 16$ obtains most of the performance gains with relatively low inference latency. When m becomes large, however, the quadratic m^2 term in self-attention starts to increase, yielding limited additional accuracy but noticeably higher runtime. Therefore, we adopt a conservative register tokens budget, which augments the SLM's input sequence with a moderate number while preserving the lightweight nature of our framework.

4.8 Visualization

Fig. 5 presents qualitative results on reasoning segmentation tasks. Without dense spatial features, the baseline relying solely on the `<SEG>` token tends to produce incomplete masks with holes and mistakenly segment irrelevant object. In contrast, LReSeg effectively alleviates these issues by introducing spatial features with register tokens, leading to more coherent and precise segmentation masks that align well with textual descriptions.

5 Conclusion

In this work, we present LReSeg, a lightweight framework for reasoning segmentation that achieves strong performance with significantly reduced model size and computational cost. Through a unified visual encoder framework and enriching spatial representations by injecting text-attended visual cues into register tokens, LReSeg effectively compensates for the limited dense spatial modeling capacity of small models. Despite its compact size, LReSeg delivers the SOTA segmentation performance among models of similar scale, while offering substantial gains in latency and memory efficiency.

Acknowledgment: This work was supported by the NSFC under Grant 62322604 and 62576207.

References

1. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
4. Chandrasegaran, K., Gupta, A., Hadzic, L.M., Kota, T., He, J., Eyzaguirre, C., Durante, Z., Li, M., Wu, J., Li, F.F.: Hourvideo: 1-hour video-language understanding. Advances in Neural Information Processing Systems **37**, 53168–53197 (2024)
5. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision. pp. 19–35 (2024)
6. 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: European Conference on Computer Vision. pp. 323–340 (2024)
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
8. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16321–16330 (2021)
9. Dong, H., Li, J., Wu, B., Wang, J., Zhang, Y., Guo, H.: Benchmarking and improving detail image caption. arXiv preprint arXiv:2405.19092 (2024)
10. Gong, S., Zhuge, Y., Zhang, L., Yang, Z., Zhang, P., Lu, H.: The devil is in temporal token: High quality video reasoning segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29183–29192 (2025)
11. Guan, T., Wang, Z., Fu, P., Guo, Z., Shen, W., Zhou, K., Yue, T., Duan, C., Sun, H., Jiang, Q., et al.: A token-level text image foundation model for document understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23210–23220 (2025)
12. Guan, T., Yang, Z., Wan, J., Yang, M., Guo, Z., Hu, Z., Luo, R., Chen, R., Jiang, S., Wang, P., et al.: Codepercept: Code-grounded visual stem perception for mllms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 33542–33552 (2026)
13. He, J., Wang, Y., Wang, L., Lu, H., He, J.Y., Lan, J.P., Luo, B., Xie, X.: Multi-modal instruction tuned llms with fine-grained visual perception. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 13980–13990 (2024)
14. Hu, B., Zhang, Y.: How ai and humans express comfort differently: A corpus-based appraisal analysis. Corpus Pragmatics **10**(1), 16 (2026)
15. Jang, D., Cho, Y., Lee, S., Kim, T., Kim, D.S.: Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. In: Proceedings of the International Conference on Learning Representations (2025)

16. Jia, J., Hu, Y., Weng, X., Shi, Y., Li, M., Zhang, X., Zhou, B., Liu, Z., Luo, J., Huang, L., et al.: Tinyllava factory: A modularized codebase for small-scale large multimodal models. arXiv preprint arXiv:2405.11788 (2024)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
18. Kong, L., Liu, Y., Ng, L.X., Cottereau, B.R., Ooi, W.T.: Openess: Event-based semantic scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15686–15698 (2024)
19. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)
20. Laurençon, H., Marafioti, A., Sanh, V., Tronchon, L.: Building and better understanding vision-language models: insights and future directions. arXiv preprint arXiv:2408.12637 (2024)
21. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
22. Li, J., Dai, H., Han, H., Ding, Y.: Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21694–21704 (2023)
23. Li, X., Yuan, H., Li, W., Ding, H., Wu, S., Zhang, W., Li, Y., Chen, K., Loy, C.C.: Omg-seg: Is one model good enough for all segmentation? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 27948–27959 (2024)
24. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7061–7070 (2023)
25. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23592–23601 (2023)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
27. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
28. Liu, Y., Qu, T., Zhong, Z., Peng, B., Liu, S., Yu, B., Jia, J.: Visionreasoner: Unified visual perception and reasoning via reinforcement learning. arXiv preprint arXiv:2505.12081 (2025)
29. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
30. Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., Ji, R.: Multi-task collaborative network for joint referring expression comprehension and segmentation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 10034–10043 (2020)
31. Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Allal, L.B., Lozhkov, A., Tazi, N., et al.: Smolvlm: Redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299 (2025)

32. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: Proceedings of the European Conference on Computer Vision. pp. 792–807 (2016)
33. Peng, Z., Xu, Z., Zeng, Z., Wen, C., Huang, Y., Yang, M., Tang, F., Shen, W.: Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4562–4572 (2025)
34. Qian, R., Yin, X., Dou, D.: Reasoning to attend: Try to understand how¡ seg¿ token works. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24722–24731 (2025)
35. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., et al.: Paco: Parts and attributes of common objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7141–7151 (2023)
36. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2024)
37. 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. arXiv preprint arXiv:2401.14159 (2024)
38. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26374–26383 (2024)
39. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
40. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
41. Su, Y., Wang, Y., Hu, Q., Yang, C., Chau, L.P.: Annexe: Unified analyzing, answering, and pixel grounding for egocentric interaction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9027–9038 (2025)
42. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
43. Team, Q., et al.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 **2**(3) (2024)
44. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
45. Wang, X., Zhang, S., Li, S., Li, K., Kallidromitis, K., Kato, Y., Kozuka, K., Darrell, T.: Segllm: Multi-round reasoning segmentation with large language models. In: International Conference on Learning Representations (2025)
46. Wang, Y., Ding, H., He, S., Jiang, X., Wei, B., Liu, J.: Hierarchical alignment-enhanced adaptive grounding network for generalized referring expression comprehension. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8042–8050 (2025)

47. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
48. Wei, C., Zhong, Y., Tan, H., Liu, Y., Zhao, Z., Hu, J., Yang, Y.: Hyperseg: Towards universal visual segmentation with large language model. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
49. Wei, C., Zhong, Y., Tan, H., Zeng, Y., Liu, Y., Wang, H., Yang, Y.: Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20193–20203 (2025)
50. Wen, C., Peng, Z., Huang, Y., Shen, W.: Efficient segmentation with multimodal large language model via token routing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 10593–10602 (2026)
51. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
52. Xia, Z., Han, D., Han, Y., Pan, X., Song, S., Huang, G.: Gsva: Generalized segmentation via multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3858–3869 (2024)
53. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. In: European Conference on Computer Vision. pp. 98–115 (2024)
54. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024)
55. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18155–18165 (2022)
56. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European conference on computer vision. pp. 69–85. Springer (2016)
57. Yuan, Z., Li, Z., Huang, W., Ye, Y., Sun, L.: Tinygpt-v: Efficient multimodal large language model via small backbones. arXiv preprint arXiv:2312.16862 (2023)
58. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
59. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Mllms know where to look: Training-free perception of small visual details with multimodal llms. In: International Conference on Learning Representations (2025)
60. Zhang, S., Fang, Q., Yang, Z., Feng, Y.: Llava-mini: Efficient image and video large multimodal models with one vision token. In: International Conference on Learning Representations (2025)
61. Zhang, Z., Ma, Y., Zhang, E., Bai, X.: Psalm: Pixelwise segmentation with large multi-modal model. In: European Conference on Computer Vision. pp. 74–91 (2024)
62. Zhou, B., Hu, Y., Weng, X., Jia, J., Luo, J., Liu, X., Wu, J., Huang, L.: Tinyllava: A framework of small-scale large multimodal models. arXiv preprint arXiv:2402.14289 (2024)

63. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
64. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
65. 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 Computer Vision and Pattern Recognition Conference. pp. 30231–30240 (2025)
66. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15116–15127 (2023)
67. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in neural information processing systems **36**, 19769–19782 (2023)