PIXELTHINK: Towards Efficient Chain-of-Pixel Reasoning

Song Wang[1,2] **Gongfan Fang**[2] **Lingdong Kong**[2] **Xiangtai Li**[3] **Jianyun Xu**[4*]
Sheng Yang[4] **Qiang Li**[4] **Jianke Zhu**[1†] **Xinchao Wang**[2†]

[1]Zhejiang University [2]National University of Singapore
[3]Nanyang Technological University [4]AD Lab, CaiNiao Inc., Alibaba Group

Project Page: PixelThink.github.io

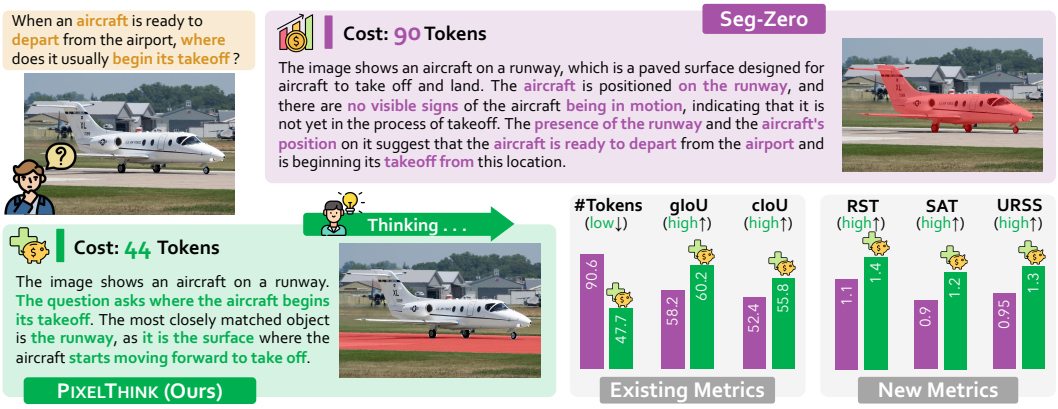

Figure 1: **Motivation of Efficient Chain-of-Pixel Reasoning (PIXELTHINK).** We propose a novel scheme for reasoning segmentation that effectively regulates reasoning length based on task *difficulty* and *uncertainty*. Our method improves segmentation quality while significantly reducing token usage. A suite of metrics is introduced for holistic evaluations of reasoning quality, segmentation accuracy, and computational efficiency.

Abstract

Existing reasoning segmentation approaches typically fine-tune multimodal large language models (MLLMs) using image-text pairs and corresponding mask labels. However, they exhibit limited generalization to out-of-distribution scenarios without an explicit reasoning process. Although recent efforts leverage reinforcement learning through group-relative policy optimization (GRPO) to enhance reasoning ability, they often suffer from overthinking – producing uniformly verbose reasoning chains irrespective of task complexity. This results in elevated computational costs and limited control over reasoning quality. To address this problem, we propose PIXELTHINK, a simple yet effective scheme that integrates externally estimated task difficulty and internally measured model uncertainty to regulate reasoning generation within a reinforcement learning paradigm. The model learns to compress reasoning length in accordance with scene complexity and predictive confidence. To support comprehensive evaluation, we introduce ReasonSeg-DIFF, an extended benchmark with annotated reasoning references and difficulty scores, along with a suite of metrics designed to assess segmentation accuracy, reasoning quality, and efficiency jointly. Experimental results demonstrate that the proposed approach improves both reasoning efficiency and overall segmentation perfor-

*Project leader.
†Corresponding authors (jkzhu@zju.edu.cn, xinchao@nus.edu.sg).

mance. Our work contributes novel perspectives towards efficient and interpretable multimodal understanding. The code and model will be publicly available.

1 Introduction

Reasoning segmentation [1, 2, 3] is an emerging vision-language task that requires predicting pixel-level masks in response to complex natural language queries. In contrast to traditional semantic or instance segmentation [4, 5, 6], which depends on predefined class labels, reasoning segmentation involves grounding fine-grained referring expressions that encode attributes, spatial relations, or contextual information. This capability is critical for embodied tasks such as interactive robotics [7, 8] and autonomous driving [9, 10]. Advances in multimodal large language models (MLLMs) [11, 12, 13, 14] have facilitated the development of reasoning segmentation via supervised fine-tuning (SFT).

Representative approaches [2, 15, 16, 17], such as the pioneering LISA [2], integrate pre-trained MLLMs with segmentation modules through additional vision-language supervision to enable language-guided segmentation. Despite achieving strong performance on in-domain tasks, these methods often face limitations in generalizing to out-of-distribution (OOD) scenarios, especially when presented with complex or ambiguous queries [18, 19]. Moreover, the absence of explicit reasoning chains reduces interpretability and hinders effective error analysis.

To overcome the limitations of SFT-based methods, recent progress in LLM research [20, 21] has motivated the adoption of reinforcement learning (RL) strategies to enhance reasoning capabilities and generalization. In particular, group-relative policy optimization (GRPO) has shown strong performance in language domains such as mathematical and code reasoning without requiring additional supervision [22, 23]. Building on this, several works have extended GRPO to visual perception tasks [24, 19, 18, 25], achieving improved out-of-distribution generalization and generating explicit, interpretable reasoning paths. Nevertheless, these methods often suffer from overthinking – producing unnecessarily verbose reasoning chains in simple cases – which leads to increased computational cost and reduced efficiency [26, 27]. As illustrated in Figure 1, Seg-Zero [18] yields incorrect segmentation results, despite utilizing a redundant reasoning process involving *twice the number of tokens*. Moreover, the lack of standardized evaluation metrics for reasoning quality hinders a thorough assessment of the benefits brought by explicit reasoning in segmentation.

In this paper, we propose PIXELTHINK under the GRPO framework for reasoning segmentation, which regulates reasoning length based on externally estimated task difficulty and internally measured model uncertainty. Each input is assigned a token budget based on its estimated difficulty and uncertainty, and GRPO is guided by soft length-aware rewards that gently penalize excessive reasoning, promoting conciseness when appropriate. To facilitate systematic evaluation, we introduce ReasonSeg-DIFF, an extended version of ReasonSeg [2], enriched with task difficulty annotations and dual-mode reasoning references (*short* and *long*). Our evaluation protocol jointly assesses segmentation accuracy (`gIoU` and `cIoU`) and reasoning score (`RScore`) using LLM-based ratings against reference reasoning chains. Additionally, we propose three efficiency-aware metrics to quantify the trade-off between segmentation performance and reasoning token usage. Specifically, `RST` and `SAT` assess the efficiency of reasoning and segmentation, respectively, while `URSS` provides a unified measure that comprehensively captures overall effectiveness across both aspects.

We conduct extensive experiments to benchmark PIXELTHINK against state-of-the-art reasoning segmentation and efficiency methods [18, 28, 2, 29]. The results demonstrate that PIXELTHINK effectively regulates reasoning length in accordance with task difficulty, while maintaining reasoning quality and further enhancing segmentation accuracy. Additionally, we present exploratory analyses on the role of reasoning in segmentation, including ablations under no-thinking conditions. These analyses confirm that necessary and concise reasoning improves segmentation performance, whereas excessive redundancy provides no additional benefit.

Our main contributions can be summarized as follows:

- We propose a novel scheme PIXELTHINK that enables efficient reasoning segmentation by leveraging external task difficulty and internal model uncertainty to guide the reward process in reinforcement fine-tuning.

- We build ReasonSeg-DIFF, a new benchmark with annotated reasoning references and difficulty scores, and establish a comprehensive evaluation protocol that covers reasoning quality, segmentation accuracy, and efficiency.
- Extensive experiments validate the effectiveness of our approach in reducing reasoning length and enhancing segmentation performance. In-depth analyses under various reasoning strategies are also conducted to inform future research.

2 Related Work

Reasoning Segmentation. Referring expression segmentation [30, 1, 31, 32, 33, 34, 35] extends traditional segmentation approaches [36, 5, 37] to open-vocabulary settings by localizing objects described in natural language. Recent works leverage multimodal large language models (MLLMs) [11, 12, 13, 14] to integrate visual and linguistic reasoning, enabling flexible segmentation from free-form queries. LISA [2] introduces step-wise alignment between textual reasoning and object grounding. A series of subsequent works [15, 17, 3, 38, 39] explore fine-tuning MLLMs with segmentation heads, leveraging token-level instructions for fine-grained prediction. Seg-Zero [18] further improves generalizability through reinforcement fine-tuning [23, 22], generating reasoning chains and reference tokens that guide segmentation modules. Despite these advances, current approaches still struggle with reasoning efficiency and adaptability to varying task difficulty levels. This work builds upon prior research and introduces an efficiency-aware reasoning scheme.

Large Reasoning Models. Large language models (LLMs) have demonstrated remarkable capabilities in multi-step reasoning through Chain-of-Thought (CoT) prompting [40, 41, 42]. Beyond prompting, recent efforts focus on optimizing the reasoning process via process reward models [43, 44, 45], reinforcement fine-tuning [46, 47, 22], and other test time scaling methods [48, 49, 50]. DeepSeek-R1 [23] employs group relative policy optimization (GRPO)[22] to elicit LLMs' latent reasoning capacity and achieves significant advances. Building on this paradigm, recent works have extended GRPO and LLM-based reasoning to the visual domain [51, 24, 25, 52, 19], enabling multimodal reasoning via vision-language models [13, 14]. However, current visual reasoning research lacks evaluation of both the reasoning process and perceptual outcomes. We bridge this gap by introducing a holistic evaluation protocol that jointly assesses reasoning quality and segmentation performance.

Efficient Inference Methods. Recent surveys [26, 27, 53, 54] have highlighted the inefficiencies of current reasoning models, including excessive token usage and redundant reasoning steps. To mitigate these issues, a variety of strategies have been explored. TALE [55] proposes allocating token budgets adaptively, while CoT-Valve [56] employs model merging to train reasoning chains of different lengths via supervised fine-tuning. Reinforcement learning-based methods such as L1 [29] and O1-Pruner [57] impose direct constraints on reasoning length during training. Further improvements include draft-based generation [58], skip mechanisms [59], and pruning strategies [60]. Self-training approaches [61] also contribute to efficiency by leveraging iterative pseudo-supervision. In the context of efficient MLLMs, efforts have primarily focused on dynamic and adaptive input compression [62, 63, 64, 65, 66], while output-level reasoning optimization remains underexplored. Our work addresses this problem by introducing a reward-driven mechanism that regulates reasoning length based on task difficulty and model uncertainty.

3 Methodology

3.1 Overview

Problem Definition. Reasoning segmentation [2, 18] aims to generate accurate segmentation masks given an image $\mathcal{I}$ and a referring expression $\mathcal{E}$. In contrast to conventional referring segmentation, which directly maps inputs to segmentation outputs, our formulation additionally requires the model to produce an explicit intermediate reasoning process $\mathcal{R}$ to improve interpretability and generalization. The task thus involves generating both the reasoning chain $\mathcal{R}$ and the segmentation mask $\mathcal{M}$.

Baseline. We follow the standard setup [18] as illustrated in Figure 2(a), where the overall framework consists of a reasoning model and a segmentation model. A multimodal large language model (MLLM), specifically Qwen2.5-VL [14], is adopted as the reasoning backbone (Reason). With an image $\mathcal{I}$ and a referring expression $\mathcal{E}$ as inputs, the model generates two distinct outputs: $\mathcal{R}, \mathcal{S} =$

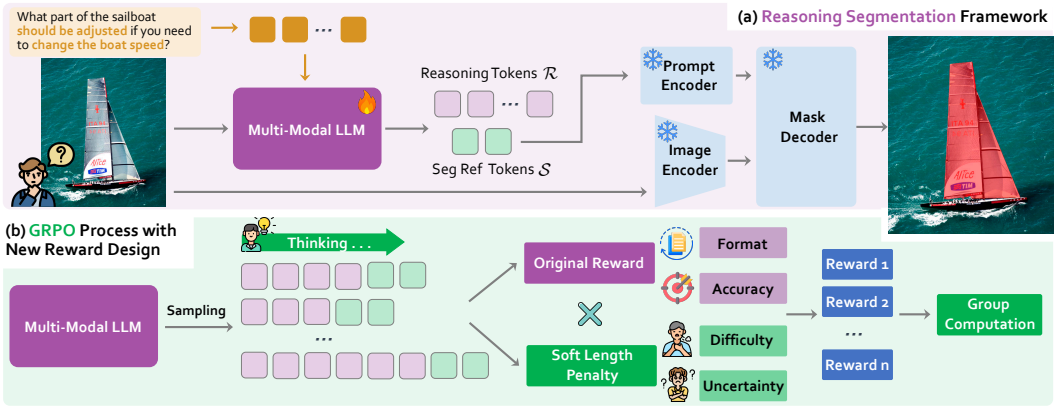

Figure 2: **Overview of PIXELTHINK.** (a) Workflow of the reasoning segmentation framework. Given an input image and query, the model generates a reasoning chain and segmentation reference that guides the segmentation outcome. (b) The group-relative policy optimization (GRPO) procedure employed during reinforcement fine-tuning. Our new reward design incorporates both task difficulty and model uncertainty, enabling the model to learn efficient reasoning strategies.

$\text{Reason}(\mathcal{I}, \mathcal{E})$, where the reasoning chain $\mathcal{R}$ is a multi-step textual explanation that reflects the model's visual understanding and reasoning process. Segmentation reference tokens $\mathcal{S}$ are spatial priors including a bounding box and two points, which serve as inputs to the segmentation model. We utilize SAM series models [37, 67] as the segmentation module, which takes the segmentation reference token $\mathcal{S}$ predicted by the reasoning model as input and produces the final binary mask $\mathcal{M}$. This modular design enables a clear decoupling of reasoning and fine-grained segmentation.

Reinforcement Fine-Tuning. We adopt reinforcement fine-tuning (RFT) to explicitly regulate output characteristics by optimizing non-differentiable objectives through task-specific reward signals. As shown in Figure 2(b), our reward formulation integrates task difficulty and model uncertainty, promoting a balanced trade-off between reasoning quality and computational efficiency.

3.2 Difficulty and Uncertainty Estimation

Task Difficulty. To achieve efficient reasoning during training, we estimate an instance-level difficulty score $\mathcal{D} \in [1, 10]$ for each sample. Following the same process in benchmark construction (Section 4.1), we prompt a large MLLM to assess difficulty across three aspects—*scene complexity*, *segmentation challenge*, and *linguistic ambiguity*—and compute the final score as their average. These difficulty priors are then used to modulate token budget allocation in reinforcement fine-tuning.

Model Uncertainty. We also quantify model-internal uncertainty based on token-level confidence in the generated reasoning sequence. For each token, we compute the gap between the highest and second-highest predicted probabilities [68, 69], using this margin to estimate certainty. The overall uncertainty score $\mathcal{U}$ is defined as:

$$\mathcal{U} = 1 - \frac{1}{T}\sum_{t=1}^{T}(p_t^{(1)} - p_t^{(2)}), \tag{1}$$

where $p_t^{(1)}$ and $p_t^{(2)}$ are the top-2 probabilities at timestep t, and T is the total number of tokens. A smaller margin indicates greater uncertainty, and the transformation ensures $\mathcal{U} \in [0, 1]$, with higher values corresponding to lower confidence. This internal self-assessment complements external task difficulty, jointly informing the adjustment of reasoning length.

3.3 Reward Design

Original Reward. We adopt the original reward design in Seg-Zero [18], which captures both reasoning validity and segmentation accuracy. The reward function comprises the following components:

$$\mathrm{R}_{\text{original}} = \mathrm{R}_{\text{format}}^{\text{reason}} + \mathrm{R}_{\text{format}}^{\text{seg}} + \mathrm{R}_{\text{accuracy}}^{\text{seg}}, \tag{2}$$

which assesses reasoning format, segmentation format, and segmentation accuracy, respectively. $\mathrm{R}^{\mathrm{seg}}_{\mathrm{accuracy}}$ comprises the evaluation of mask IoU, point-level, and bounding box-level L1 distance. The original reward $\mathrm{R}_{\mathrm{original}}$ serves as the basis for subsequent reasoning length modulation.

Soft Length Penalty. To enable controllable reasoning length across tasks of varying complexity, we introduce a soft budget penalty that adaptively modulates the reward using both external task *difficulty* and internal model *uncertainty*. In contrast to prior approaches [29, 55], this mechanism encourages concise reasoning in simple scenarios while permitting elaboration in complex or ambiguous cases, thereby mitigating overthinking and balancing reasoning adequacy with efficiency. Given the difficulty score $\mathcal{D}$ and the uncertainty score $\mathcal{U}$, we define the expected reasoning token budget L_{budget} as:

$$L_{\mathrm{budget}} = \begin{cases} L_{\mathrm{base}} + \alpha \cdot \mathcal{U}, & \text{if } \mathcal{D} \geq \tau_1 \\ L_{\mathrm{low}}, & \text{if } \mathcal{D} < \tau_2 \\ \mathrm{None}, & \text{otherwise} \end{cases} \tag{3}$$

where thresholds τ_1 and τ_2 divide tasks into *hard*, *medium*, and *easy* levels. L_{base}, α, and L_{low} are constants denoting the base budget for difficult tasks, the gain from uncertainty, and the minimal budget for easy tasks. We deliberately leave moderately difficult tasks unconstrained to allow learning flexibility in ambiguous regimes. The soft penalty is computed as:

$$s(L_{\mathrm{used}}, L_{\mathrm{budget}}) = \begin{cases} 1 - \beta \cdot (L_{\mathrm{used}} - L_{\mathrm{budget}}), & \text{if } L_{\mathrm{used}} > L_{\mathrm{budget}} \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

where L_{used} is the actual token count, and β is a small penalty factor that ensures smooth reward decay without harsh clipping. This design maintains model stability during training and avoids discouraging minor budget exceedance that may improve output quality. The final reward is computed as:

$$\mathrm{R}_{\mathrm{final}} = \mathrm{R}_{\mathrm{original}} \cdot s(L_{\mathrm{used}}, L_{\mathrm{budget}}). \tag{5}$$

This formulation achieves fine-grained adjustment over reasoning length, ensuring that token usage aligns with *task complexity* and *model confidence*. Moreover, it avoids the pitfalls of hard constraints, which could prematurely truncate informative reasoning during early training. Our empirical findings confirm that the chosen upper bounds are *sufficiently permissive* to allow learning while providing enough structure to suppress redundant output.

3.4 Reinforcement Fine-tuning Process

Training with GRPO. Following recent advances [22, 23, 18], we adopt group-relative policy optimization (GRPO) for reinforcement fine-tuning. The model is optimized to maximize the task-specific reward $\mathrm{R}_{\mathrm{final}}$ introduced in Section 3.3, which integrates reasoning quality, segmentation accuracy, and token efficiency into a unified objective. GRPO enhances training stability by comparing rewards within mini-batches at a group level, facilitating more consistent gradient updates and improving convergence with reduced variance.

Inference. At inference time, the model operates under the same architecture and prompting schema as the baseline model. It generates both the reasoning chain $\mathcal{R}$ and segmentation mask $\mathcal{M}$ with image-text pairs as inputs. *No additional labels* or *reward feedback* are required during testing. This simple yet effective scheme facilitates seamless deployment and supports token-efficient, interpretable segmentation across diverse inputs.

4 Benchmark Construction & Evaluation Protocol

While existing reasoning segmentation benchmarks primarily focus on evaluating the final segmentation mask, they often overlook the quality of the reasoning process and the efficiency of token usage. To address this gap and facilitate comprehensive evaluation, we construct ReasonSeg-DIFF, an extension of the ReasonSeg dataset [2] that includes task difficulty annotations and reference reasoning chains. This benchmark enables fine-grained assessment under varying levels of complexity.

4.1 The Construction of ReasonSeg-DIFF

As illustrated in Figure 3, the construction process consists of three parts, and we provide more details and examples in the ReasonSeg-DIFF Details part of the supplementary materials.

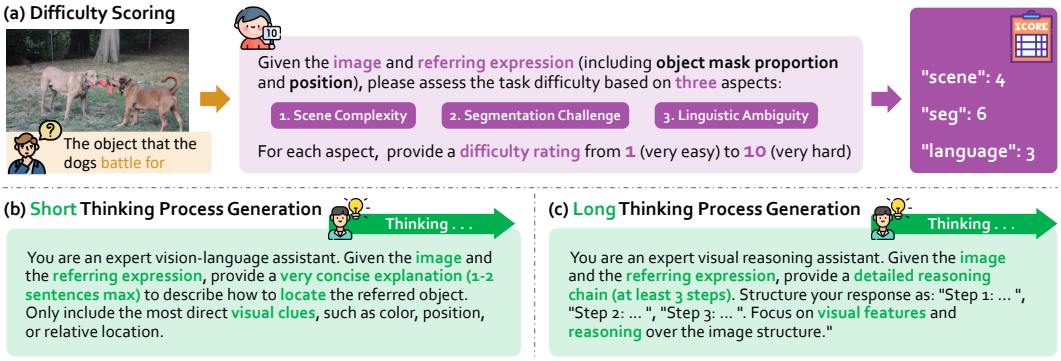

Figure 3: **The construction of ReasonSeg-DIFF.** (a) Design on the *Difficulty Scoring* scheme. (b) Generation of the *Short Thinking* process. (c) Generation of *Long Thinking* process.

Difficulty Scoring. To quantify the intrinsic reasoning challenge of each sample, we propose a structured scoring framework based on three interpretable factors: (1) *Scene Complexity*, measuring the number and similarity of distractor objects; (2) *Segmentation Challenge*, capturing the spatial size, position, occlusion, and whether the target is a whole or a part; and (3) *Linguistic Ambiguity*, evaluating how explicit the referring expression is versus the need for visual inference.

As shown in Figure 3(a), we encode visual and textual priors into a unified prompt and query a large vision-language model (Qwen2.5-VL-72B [14]) to independently score each aspect on a scale of [1–10], accompanied by a natural language explanation. The final difficulty score is computed as the average of three dimensions, yielding a holistic and interpretable measure of instance-level complexity. To ensure reliability, this score is further cross-validated against human annotations. We also categorize difficulty into three levels (*easy*, *medium*, and *hard*) using thresholds τ_1 and τ_2.

Short and Long Thinking Process Generation. To support evaluation under varying reasoning budgets, we construct two types of reference chains for each sample. The *short chain* conveys essential visual cues and confident identification in 1–2 sentences, while the *long chain* follows a structured multi-step format (*e.g.*, "Step 1... Step 2..") to emulate more comprehensive visual reasoning. Both chains are generated by prompting Qwen2.5-VL-72B with task-specific instructions and paired vision-text inputs as illustrated in Figure 3(b)(c). These only serve as evaluation references for assessing the informativeness and efficiency of model-generated reasoning.

4.2 Evaluation Protocol

With the constructed ReasonSeg-DIFF, we further introduce a comprehensive evaluation protocol for reasoning segmentation that jointly measures *segmentation accuracy*, *reasoning quality*, and *computational efficiency*. The protocol consists of three complementary components detailed below.

Segmentation Evaluation. Following established settings [1, 30], we adopt two standard metrics: `gIoU`, the mean Intersection-over-Union (IoU) across all test samples, and `cIoU`, the cumulative IoU computed as the total intersection divided by the total union over the dataset. To incorporate efficiency into segmentation assessment, we propose **Segmentation Accuracy per Token** (`SAT`), formulated as: $\texttt{SAT} = \frac{100\times\texttt{gIoU}}{P\times\sqrt{T_{\text{num}}+1}}$, where P denotes the number of model parameters (in billions), and T_{num} is the average token length of the generated reasoning chains. The *square root* in the denominator provides a soft penalty for longer outputs, discouraging unnecessarily verbose reasoning without overly punishing small increases in token length. `SAT` favors models that achieve high segmentation accuracy with minimal reasoning overhead and compact model size.

Reasoning Quality Evaluation. To assess the quality of generated reasoning chains, we employ a large language model (LLM) [70] to score predictions against the reference annotations introduced in Section 4.1. The evaluation covers three key dimensions: (1) *Completeness* — whether all necessary steps and information are included; (2) *Object Grounding* – the degree of alignment with the referred visual object; and (3) *Fluency and Clarity* – coherence, grammaticality, and readability.

Each aspect is rated on a 1–10 scale, and their average constitutes the overall reasoning score, denoted as `RScore`. To reflect reasoning efficiency, we further introduce **Reasoning Score per Token** (`RST`)

Table 1: Overall evaluation results on the proposed ReasonSeg-DIFF benchmark.

Method	Reasoning Quality			Segmentation Performance			Overall
	#Token ↓	RScore↑	RST↑	gIoU(%)↑	cIoU(%)↑	SAT↑	URSS↑
Source of ReasonSeg Data: Validation Set							
○ Seg-Zero [18]	90.79	7.67	1.14	61.63	52.56	0.92	0.99
○ Seg-Zero [18] + Prompt	57.62	7.31	1.36	62.50	59.22	1.17	1.23
○ Seg-Zero [18] + L1-Exact [29]	65.66	5.73	1.00	40.97	42.47	0.72	0.80
○ Seg-Zero [18] + L1-Max [29]	61.21	4.37	0.79	61.75	57.39	1.12	1.02
• **PIXELTHINK** (Ours)	**46.98**	6.92	**1.43**	**63.81**	**62.69**	**1.32**	**1.35**
Source of ReasonSeg Data: Test Set							
○ Seg-Zero [18]	90.58	7.67	1.14	58.20	52.37	0.87	0.95
○ Seg-Zero [18] + Prompt	57.84	7.33	1.37	58.15	53.45	1.08	1.17
○ Seg-Zero [18] + L1-Exact [29]	65.24	5.71	1.00	39.63	34.42	0.70	0.79
○ Seg-Zero [18] + L1-Max [29]	61.61	4.31	0.78	58.20	47.44	1.05	0.97
• **PIXELTHINK** (Ours)	**47.66**	6.92	**1.42**	**60.17**	**55.77**	**1.23**	**1.29**

Table 2: Difficulty-aware evaluation on the ReasonSeg-DIFF test set.

Method	Reasoning Quality			Segmentation Performance			Overall
	#Token ↓	RScore↑	RST↑	gIoU(%)↑	cIoU(%)↑	SAT↑	URSS↑
Difficulty Level: Easy							
○ Seg-Zero [18]	84.97	8.07	1.24	68.65	65.85	1.06	1.11
○ Seg-Zero [18] + Prompt	55.35	7.80	1.48	67.93	63.94	1.29	1.35
○ Seg-Zero [18] + L1-Exact [29]	68.03	6.42	1.10	51.95	43.77	0.89	0.96
○ Seg-Zero [18] + L1-Max [29]	60.15	4.73	0.86	68.40	62.15	1.25	1.13
• **PIXELTHINK** (Ours)	**44.73**	7.56	**1.60**	**70.25**	**67.49**	**1.48**	**1.52**
Difficulty Level: Medium							
○ Seg-Zero [18]	90.73	7.68	1.15	60.15	55.53	0.90	0.97
○ Seg-Zero [18] + Prompt	58.37	7.30	1.35	59.89	56.15	1.11	1.18
○ Seg-Zero [18] + L1-Exact [29]	66.03	5.60	0.98	39.97	35.01	0.70	0.78
○ Seg-Zero [18] + L1-Max [29]	61.45	4.18	0.76	59.46	47.79	1.07	0.98
• **PIXELTHINK** (Ours)	**47.00**	6.84	**1.41**	**62.05**	**58.16**	**1.28**	**1.32**
Difficulty Level: Hard							
○ Seg-Zero [18]	95.37	7.26	1.06	44.37	28.93	0.65	0.77
○ Seg-Zero [18] + Prompt	58.97	6.97	**1.29**	45.38	31.35	0.84	0.97
○ Seg-Zero [18] + L1-Exact [29]	60.94	5.29	0.96	27.61	18.94	0.50	0.64
○ Seg-Zero [18] + L1-Max [29]	63.28	4.22	0.75	46.09	30.36	0.82	0.80
• **PIXELTHINK** (Ours)	**51.79**	6.50	1.28	**46.80**	**35.05**	**0.92**	**1.03**

as: $\texttt{RST} = \frac{10 \times \texttt{RScore}}{P \times \sqrt{T_{\text{num}}+1}}$. Following our difficulty-aware setting, RScore is computed using the *short* chain for *easy* and *medium* samples, and the *long* chain for *hard* ones. This evaluation protocol favors models that generate concise yet semantically rich reasoning, especially with limited tokens.

Unified Metric. To holistically evaluate reasoning segmentation across accuracy, reasoning quality, and computational efficiency, we propose the **Unified Reasoning Segmentation Score** (URSS): $\texttt{URSS} = (1-\gamma) \times \texttt{RST} + \gamma \times \texttt{SAT}$, where $\gamma \in [0, 1]$ governs the relative emphasis on segmentation accuracy (SAT) and reasoning quality (RST). We set $\gamma = 0.7$ by default to reflect the *greater importance* of segmentation performance in practical applications.

5 Experiments

Datasets. We train exclusively on $9,000$ samples from RefCOCOg [1] *without any reasoning data*, following the same split as Seg-Zero [18] for fair comparison. Evaluation is primarily conducted on ReasonSeg-DIFF derived from ReasonSeg [2], enabling *zero-shot* assessment across varying task complexities. ReasonSeg-DIFF includes 199 validation samples and 769 test samples, stratified by difficulty into 51/102/45 and 171/411/187 for easy, medium, and hard levels, respectively. We further report results on RefCOCO, RefCOCO+, and RefCOCOg to validate the common performance.

Table 3: Performance comparison on existing benchmarks. Symbol † denotes scores reported in the paper while * denotes our reproduction with official code and model checkpoint. Our method achieves consistent improvements across the majority of benchmarks. The 7B model shows limited gains on RefCOCO due to the dataset's inherent simplicity and performance saturation.

(a) Reasoning segmentation on ReasonSeg [2].

Method	val		test	
	gIoU	cIoU	gIoU	cIoU
OVSeg [72]	28.5	18.6	26.1	20.8
ReLA [73]	22.4	19.9	21.3	22.0
Grounded-SAM [74]	26.0	14.5	21.3	16.4
LISA-7B-LLaVA1.5 [2]	53.6	52.3	48.7	48.8
LISA-13B-LLaVA1.5 [2]	57.7	60.3	53.8	50.8
SAM4MLLM [28]	46.7	48.1	-	-
Qwen2.5VL-3B [14] + SAM2 [67]	53.8	44.1	47.6	37.4
Seg-Zero-3B† [18]	62.6	58.5	56.1	48.6
Seg-Zero-7B† [18]	62.6	62.0	57.5	52.0
Seg-Zero-3B* [18]	59.1	48.8	52.5	43.4
Seg-Zero-7B* [18]	61.6	52.6	58.2	52.4
PIXELTHINK-3B (ours)	62.3	58.5	58.8	52.1
PIXELTHINK-7B (ours)	**63.8**	**62.7**	**60.2**	**55.8**

(b) Referring expression segmentation (cIoU) on RefCOCO (+/g) [1].

Method	RefCOCO	RefCOCO+	RefCOCOg
	testA	testA	test
LAVT [75]	75.8	68.4	62.1
ReLA [73]	76.5	71.0	66.0
LISA-7B [2]	76.5	67.4	68.5
PixelLM-7B [15]	76.5	71.7	70.5
MagNet [33]	78.3	73.6	69.3
PerceptionGPT-7B [76]	78.6	73.9	71.7
Seg-Zero-3B† [18]	79.3	73.7	71.5
Seg-Zero-7B† [18]	80.3	76.2	72.6
Seg-Zero-3B* [18]	76.0	70.6	68.8
Seg-Zero-7B* [18]	79.4	73.7	73.2
PIXELTHINK-3B (ours)	78.7	72.9	72.2
PIXELTHINK-7B (ours)	79.3	**74.8**	**73.9**

Implementation Details. We mainly initialize the reasoning model with Qwen2.5-VL-7B [14] and adopt SAM2-Large [67] as the segmentation backbone. Reinforcement fine-tuning is performed using the GRPO [22], with a KL divergence coefficient of 1×10^{-3} and 8 samples per update. The initial learning rate is set to 1×10^{-6}. The soft length penalty parameters are set as follows: $L_{\text{base}} = 256$, $\alpha = 25$, $L_{\text{low}} = 96$, and the penalty factor $\beta = 2 \times 10^{-3}$. The thresholds τ_1 and τ_2 are set to 5.0 and 3.5, respectively. All experiments are conducted on 8 NVIDIA A100 GPUs with DeepSpeed [71].

5.1 Main Results

Quantitative Results on ReasonSeg-DIFF. We evaluate our method on our ReasonSeg-DIFF benchmark, which incorporates fine-grained difficulty annotations and reference reasoning chains to facilitate comprehensive assessment across reasoning quality, segmentation accuracy, and efficiency. As Seg-Zero [18] is the only existing approach that jointly provides explicit reasoning chains and segmentation masks, we adopt it as the primary baseline for comparison. To examine efficient reasoning generation, we additionally compare with three length-aware baselines: (1) **Prompt**, which imposes a token limit through prompt-level constraints; (2) **L1-Exact**, and (3) **L1-Max**, two reinforcement fine-tuning strategies from [29] that incorporate different reward formulations to regulate output length. For fair comparison with Seg-Zero, all methods are fine-tuned exclusively on the RefCOCOg [1] training split without any additional annotations.

Table 1 presents the overall results on the validation and test sets of ReasonSeg-DIFF. Our method achieves substantial reductions in reasoning token usage while simultaneously enhancing segmentation accuracy. Unlike prompt-based and L1-style baselines that enforce rigid length constraints, our approach preserves reasoning quality, resulting in a *more favorable balance* between performance and efficiency. Additionally, Table 2 reports difficulty-level breakdowns on the test set, where our method consistently surpasses all baselines across easy, medium, and hard subsets.

Quantitative Results on Existing Benchmark. We further evaluate the generalization capability of our method on four widely-used benchmarks: ReasonSeg [2] and the standard referring expression segmentation datasets RefCOCO, RefCOCO+, and RefCOCOg [1]. On ReasonSeg, we compare against state-of-the-art approaches including OVSeg [72], ReLA [73], LISA [2], Grounded-SAM [74], and SAM4MLLM [28]. As reported in Table 3(a), our method achieves the highest segmentation accuracy while using significantly fewer reasoning tokens. Notably, our method with Qwen2.5-VL-3B model performs on par with prior 7B counterparts, highlighting the efficiency of our approach.

We also report results on RefCOCO, RefCOCO+, and RefCOCOg using their standard test splits. As shown in Table 3(b), our method delivers competitive performance across all three datasets, confirming its robustness and strong generalization in both in-domain and out-of-domain settings.

Qualitative Results. Figure 4 presents visual comparisons between our method and Seg-Zero across various scenes. PIXELTHINK consistently predicts more precise segmentation masks while generat-

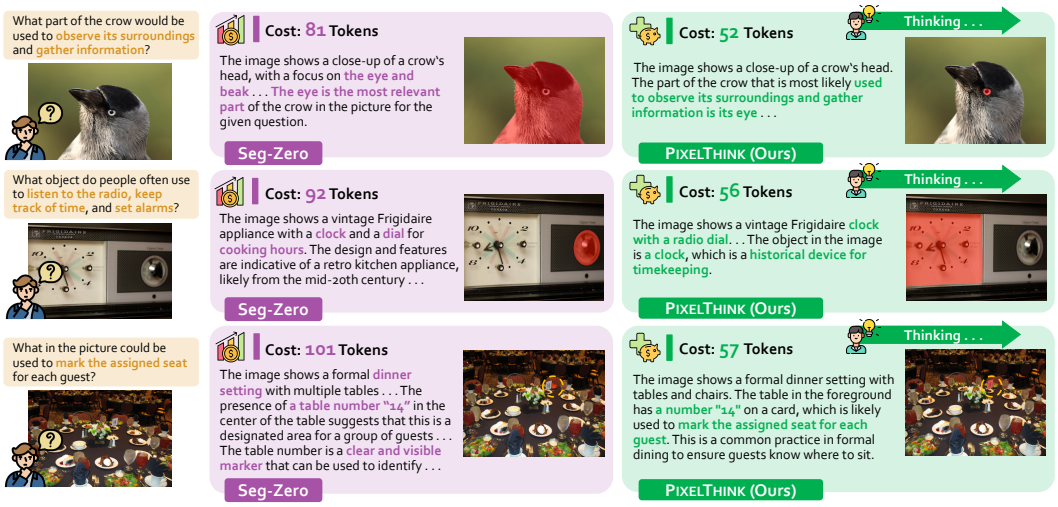

Figure 4: **Qualitative comparisons** between Seg-Zero [18] and the proposed PIXELTHINK. Representative samples across different difficulty levels are selected to highlight differences in the reasoning process and segmentation performance.

Table 5: Ablation of the difficulty splits in PIXELTHINK.

Difficulty	#Tok↓	gIoU↑	cIoU↑
-	86.92	59.65	50.92
2	61.78	**64.38**	61.80
3	**46.98**	63.81	**62.69**
4	76.89	62.28	55.78

Table 6: Ablation on the token budget allocation details.

Budget	#Tok↓	gIoU↑	cIoU↑
-	86.92	59.65	50.92
(64, 256)	**24.71**	62.12	59.57
(96, 256)	46.98	**63.81**	**62.69**
(96, 384)	60.18	60.18	53.29

Table 7: Ablation on the different no-thinking mode analyses.

Method	#Tok↓	gIoU↑	cIoU↑
No-thinking-RL	0.00	60.19	49.49
No-thinking-Prompt	0.00	60.37	51.47
Seg-Zero	90.79	61.63	52.56
PIXELTHINK(ours)	46.98	**63.81**	**62.69**

ing substantially shorter reasoning chains. In contrast, Seg-Zero frequently exhibits *overthinking*, producing lengthy and redundant explanations that fail to improve segmentation quality.

5.2 Diagnostic Experiments

We conduct diagnostic experiments on the validation set of ReasonSeg-DIFF for further exploration. Additional results and implementation details are provided in the supplementary materials.

Ablation on PIXELTHINK Scheme. We ablate the two central components of PIXELTHINK– *task difficulty* and *model uncertainty* – to analyze their individual and joint contributions. For uncertainty-guided control only, we also divide the uncertainty scores into three levels and assign token budgets accordingly. As shown in Table 4, incorporating either difficulty or uncertainty alone reduces reasoning length and yields improvements in segmentation accuracy. Their combination achieves the optimal performance, confirming the complementary nature of the two signals for effective and efficient reasoning.

Table 4: Ablation of our PIXELTHINK scheme.

Difficulty	Uncertainty	#Token ↓	gIoU(%)↑	cIoU(%)↑
		86.92	59.65	50.92
✓		**39.95**	62.07	58.35
	✓	42.41	59.54	52.37
✓	✓	46.98	**63.81**	**62.69**

Ablation on Difficulty Splits. We further investigate the impact of difficulty granularity by varying the number of difficulty levels used during training. In the 2-level setting, medium and hard samples are merged and assigned a uniformly larger token budget. In the 4-level setting, the medium category is further subdivided, with finer-grained length constraints applied. As illustrated in Table 5, the 3-level split offers the best balance between segmentation accuracy and reasoning efficiency.

Ablation on Token Budget Allocation. We next ablate the token budget configuration used for the soft length penalty in reward. Specifically, we set the base upper bounds to $L_{\text{base}} = 256$ and $L_{\text{low}} = 96$, and assess alternative parameter configurations accordingly. As shown in Table 6, our approach consistently achieves substantial reductions in reasoning token usage while improving segmentation performance, demonstrating the robustness of the proposed budget design.

No-thinking Mode Analyses. Inspired by recent investigations into no-thinking paradigms for reasoning [77, 78], we extend this line of analysis to the pixel-level segmentation task. We implement two variants within the reinforcement learning framework: No-thinking-RL and No-thinking-Prompt. As demonstrated in Table 7, incorporating appropriate reasoning steps consistently improves segmentation accuracy, highlighting the benefit of efficient reasoning over naive or omitted inference.

6 Conclusion

In this paper, we propose PIXELTHINK, an efficiency-aware reasoning scheme for segmentation that explicitly regulates reasoning length based on task difficulty and model uncertainty. By introducing a soft length penalty and reward modulation, our method enables efficient chain-of-Pixel reasoning and improving segmentation accuracy. To achieve comprehensive evaluation, we construct a difficulty-aware benchmark ReasonSeg-DIFF, and design holistic metrics that jointly assess reasoning quality, segmentation precision, and efficiency. Extensive experiments demonstrate that PIXELTHINK produces concise yet informative reasoning chains and consistently outperforms baselines across varying difficulty levels. Further discussions are presented in Section E of the Appendix.

Appendix

A Additional Implementation Details

In this section, we provide additional implementation details for the baseline models and the No-thinking methods discussed in the main paper.

A.1 Implementation Details on Baseline

Seg-Zero Re-implementation. The evaluation of Seg-Zero [18]'s 7B model is conducted using the official model checkpoint available in the public repository. Due to the unavailability of the 3B checkpoint, we re-train the model using the official codebase under the same experimental settings to ensure fair comparison. The prompts used for both models strictly adhere to the official implementation, as shown below. For consistency, the same prompt format is also adopted in our proposed PIXELTHINK.

Original Prompt from Seg-Zero

Please find "{Question}" with bbox and points.
Compare the differences between objects and find the most closely matched one.
Output the *thinking process* in <think> </think> and *final answer* in <answer> </answer> tags.
Output the *one bbox* and *points* of two largest inscribed circles inside the interested object in JSON format.
i.e., <think> *thinking process here* </think>
<answer>"Bbox": $[10, 100, 200, 210]$, "Point 1": $[30, 110]$, "Point 2": $[35, 180]$</answer>

Seg-Zero with Prompt for Short-Thinking. For the prompt-based baseline, we utilize the official model checkpoint and explicitly *incorporate a token budget constraint into the prompt*, setting the upper limit to 64 tokens during inference. The specific prompt used is provided below:

Short-thinking Prompt

Please find "{Question}" with bbox and points.
Compare the difference between objects and find the most closely matched one.
Think step-by-step and explain your reasoning process in less than 64 tokens.
Output the *reasoning* in <think> </think> and the *final answer* in <answer> </answer> tags.
Output the *one bbox* and *points* of two largest inscribed circles inside the interested object in JSON format.
i.e., <think> *short reasoning here* </think>
<answer>"Bbox": $[10, 100, 200, 210]$, "Point 1": $[30, 110]$, "Point 2": $[35, 180]$</answer>

Adapt L1 for Reasoning Segmentation. Since L1 [29] is designed to control the reasoning length of large language models (LLMs) [23, 70], its original implementation requires explicitly specifying the desired token count in the prompt. Besides, it is initially trained on the fixed-length DeepScaleR dataset [79] using L1-Exact, followed by continued fine-tuning with L1-Max. To distinguish L1 from prompt-based baseline and ensure a fair comparison with Seg-Zero without relying on *additional reasoning data*, we adopt the same prompt format used in Seg-Zero for re-implementing L1 in the reasoning segmentation task. For both variants, we independently integrate the corresponding length control functions into the reward formulation. Specifically, L1-Exact uses a 64-token upper limit, while L1-Max allows up to 128 tokens, enabling comparable final reasoning lengths.

A.2 Implementation Details on No-thinking Mode

For the implementation of No-thinking-Prompt, we directly use the official model checkpoint from Seg-Zero and apply a prompt that explicitly instructs the model not to produce any reasoning steps as below, thereby enforcing a "no-thinking" behavior. For No-thinking-RL, we follow the CLS-RL [77] framework by adopting a similar prompt format and removing the reasoning-format reward term from the GRPO reward function. We then re-train Seg-Zero using this modified reward formulation to obtain the final No-thinking-RL results.

No-thinking Prompt

Please find "{Question}" with bbox and points.
Compare the differences between objects and find the most closely matched one.
Output the final answer in the `<answer> </answer>` tag only.
The answer should include *one bbox* and the *points* of the two largest inscribed circles inside the interested object in JSON format, *i.e.*,
`<answer>`"Bbox": [10, 100, 200, 210], "Point 1": [30, 110], "Point 2": [35, 180]`</answer>`

B The ReasonSeg-DIFF Dataset

In this section, we present the dataset construction details, including the scoring prompts and representative examples from ReasonSeg-DIFF.

B.1 Prompts for Scoring and Statistics

Difficulty Scoring. For both the RefCOCOg [1] training set and the validation/test splits constructed from ReasonSeg [2] to form ReasonSeg-DIFF, we need to assign a difficulty score for each sample. To achieve this, we generate *visual descriptions* based on mask properties (*e.g.*, size and position) and *textual descriptions* derived from the referring expressions (*e.g.*, expression length and the number of spatial terms). These descriptions are incorporated into the prompt. We then instruct Qwen2.5-VL-72B [14] to rate each sample from three perspectives: (1) *Scene Complexity*, (2) *Segmentation Challenge*, and (3) *Linguistic Ambiguity*. The final difficulty score is computed as the average of these three ratings. The prompt used for this scoring process is provided below.

Difficulty Scoring Prompt

You are an expert in reasoning segmentation evaluation.
Given the image and the referring expression: "{Question}", please assess the task difficulty based on the following three aspects:
1. *Scene Complexity*:- How many objects are visible in the scene?- How many of them are potentially related or visually similar to the target?
2. *Segmentation Challenge*:
- What is the size and position of the target object?
- Are there occlusions, overlaps, or visually similar objects nearby?
- Is the mask describing the whole object or just a part? "{Visual Description}"
3. *Language Complexity*:
- Does the referring expression explicitly point to the target object?
- Or does it require additional reasoning to infer which object is referred to? "{Textual Description}"
For each aspect, please provide a difficulty rating from 1 (very easy) to 10 (very hard), and summarize in the following Python dictionary format.
i.e., {"scene": 4, "segmentation": 6, "language": 3}

Lable Statistics. We further analyze the distribution of difficulty levels, namely *easy*, *medium*, and *hard*, as determined by our scoring framework for both RefCOCOg [1] and ReasonSeg [2]. As summarized in Table A, RefCOCOg contains a significantly higher proportion of easy samples compared to ReasonSeg, which is consistent with commonly held expectations regarding the relative complexity of the two datasets. Notably, despite differences in data distribution between RefCOCOg and ReasonSeg, our method achieves superior *zero-shot* performance on ReasonSeg when trained solely on RefCOCOg.

Table A: Label statistics on difficulty distribution in training, validation and test set.

Dataset	Easy	Medium	Hard
Training Set (RefCoCOg)	3220	4810	970
Validation Set (ReasonSeg)	51	102	45
Test Set (ReasonSeg)	171	411	187

Reasoning Process Scoring. For reasoning process evaluation, we adopt the following prompt and use reasoning chains from ReasonSeg-DIFF as reference annotations. The model-generated reasoning

Table B: Ablation on the uncertainty weight.

Weight (α)	#Token ↓	gIoU(%)↑	cIoU(%)↑
0	**39.95**	62.07	58.35
25	46.98	**63.81**	62.69
35	71.10	62.66	**63.13**

Table C: Ablation on the length constrain for *medium* samples during training.

with constrain	#Token ↓	gIoU(%)↑	cIoU(%)↑
✓	**42.10**	60.92	53.84
✗	46.98	**63.81**	**62.69**

is evaluated by Qwen2.5-72B [70] across three dimensions: *Completeness*, *Object grounding*, and *Fluency & Clarity*, providing a comprehensive assessment of reasoning quality.

Reasoning Scoring Prompt

You are an expert in evaluating reasoning quality for reasoning segmentation tasks.
Given the following predicted reasoning and reference reasoning, please score the prediction in three aspects from 1 to 10:
1. *Completeness*: Does it include all necessary steps and important information?
2. *Object Grounding*: Is it aligned with the referred object in the question?
3. *Fluency & Clarity*: Is the reasoning coherent, fluent, and grammatically correct?
The question is: "{Question}"
Reference Reasoning: "{Reference Text}"
Predicted Reasoning: "{Thinking Text}"
Return a Python dictionary with keys "completeness", "grounding", and "fluency".
i.e., {"completeness": 8, "grounding": 7, "fluency": 9}

B.2 Examples from ReasonSeg-DIFF

We present several representative examples from the constructed ReasonSeg-DIFF in Figure A, covering samples categorized as *easy*, *medium*, and *hard*. Each example includes its assigned difficulty scores along with the corresponding reference reasoning chains. For *easy* and *medium* cases, we recommend the use of *short* reasoning chains, whereas *longer* chains are preferable for *hard* samples. The annotation files are available in the ReasonSeg-DIFF directory as a *supplementary attachment* for further reference.

B.3 License

The ReasonSeg-DIFF dataset is released under the Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)[3] license.

C Additional Experimental Results

In this section, we conduct more ablation experiments and provide additional qualitative results including failure cases in reasoning segmentation.

C.1 Additional Ablation Results

All experiments in this section follow the main ablation setting, using the 7B model and evaluating on the validation split.

Ablation on Uncertainty Weight. We conduct an ablation study on the uncertainty ($\mathcal{U}$) weighting factor α for hard samples to evaluate its effect on reasoning length control and segmentation performance. As shown in Table B, selecting an appropriate uncertainty weight is crucial for achieving optimal performance in both reasoning efficiency and segmentation accuracy.

Ablation on Length Constrain for Medium Samples. We also investigate the impact of applying length constrain to medium-difficulty samples. In the controlled variant, the reasoning length is limited to a maximum of 176 tokens. As shown in Table C, allowing medium cases to remain unconstrained provides greater flexibility in reasoning length, resulting in improved overall performance.

[3] https://creativecommons.org/licenses/by-sa/4.0/legalcode.

C.2 Additional Qualitative Results

In Figure B, we provide additional qualitative comparisons between our PIXELTHINK and Seg-Zero, including both segmentation masks and reasoning chains. Across a range of scenarios, PIXELTHINK consistently yields more accurate segmentation results while generating significantly shorter reasoning chains, highlighting its superior *efficiency* and *effectiveness*.

C.3 Failure Cases and Analyses

In Figure C, we present several failure cases to illustrate limitations of the current approach. In the first example, both Seg-Zero and PIXELTHINK produce incomplete segmentation results, as multiple objects in the scene satisfy the referring conditions. In the second example, Seg-Zero fails during the reasoning process by misidentifying the target object, yet still generates a partially correct mask. In contrast, our method correctly identifies the target but yields an imperfect segmentation mask. These cases reveal a key limitation of the current decoupled architecture, where the reasoning outputs and final masks are not always well aligned. In the future, we will explore tighter integration and joint optimization to enhance consistency and overall performance in reasoning segmentation.

D Additional Observations and Analyses

In this section, we provide complementary observations to further interpret the experimental results presented in the main paper. Specifically, we examine several empirical phenomena observed during training and evaluation: (1) the discrepancy between the token budget and the actual reasoning length, (2) the convergence of reasoning lengths across varying difficulty levels, and (3) the trade-off between concise reasoning and completeness in comparison to Seg-Zero.

D.1 Discrepancy between Token Budget and Reasoning Length

In our experiments, we observe that the model trained with the proposed reward framework frequently generates reasoning chains that are significantly shorter than the predefined token budget. This behavior can be attributed to several factors:

Soft Length Penalty Encourages Conservative Generation. Our reward function incorporates a *soft length penalty* that linearly penalizes the use of tokens exceeding the expected budget. Unlike hard truncation, this approach allows for flexibility while implicitly encouraging the model to stay within budget. Therefore, the model learns to *avoid unnecessary token usage* unless it contributes to improved task performance.

Accuracy-dominant Reward Prevents Token Inflation. The final reward integrates segmentation accuracy with alignment to the expected reasoning length. Since $\mathcal{R}_{\text{original}}$ primarily governs the reward dynamics, longer reasoning chains that do not lead to performance gains are *implicitly penalized*. This design encourages the generation of concise yet informative reasoning, where token usage is closely aligned with task utility.

D.2 Convergence of Reasoning Length across Difficulty Levels

Although our training framework assigns distinct token budgets according to difficulty levels, we notice that the final reasoning lengths across all categories tend to converge within a relatively narrow range. Several determinants account for this phenomenon:

Shared Decoder and Autoregressive Generation Bias. The reasoning model employs a *unified decoder* with *same prompt* to generate reasoning chains for all samples. Since this decoder is optimized across tasks with varying levels of difficulty, it learns an averaged generation pattern and tends to favor a stable reasoning length distribution. This behavior is further reinforced by the autoregressive nature of decoding, through which the model implicitly learns a preferred stopping condition based on distributional patterns observed during training.

Conservative Token Budget Design. The token budget upper bounds for each difficulty group are set conservatively high to avoid premature truncation. The soft penalty is applied only when the reasoning length exceeds the budget and does not actively encourage the model to approach the upper limit. This design allows the model to naturally converge to a reasoning length below the threshold.

D.3 RScore Performance Compared to Seg-Zero

While our method surpasses Seg-Zero in both segmentation accuracy and inference efficiency, we observe slightly lower values in the RScore, which evaluates the quality of the generated reasoning chains. This can be traced to limitations in the design of the RScore metric:

RScore Emphasizes Completeness without Length Awareness. RScore is computed based on three criteria: *completeness*, *grounding*, and *fluency*. Notably, the metric does not consider the brevity or efficiency of the generated reasoning. Thus, longer reasoning chains often receive higher completeness scores, even when parts of the explanation may be redundant.

PIXELTHINK Prioritizes Efficiency and Accuracy. PIXELTHINK is designed to generate concise yet informative reasoning under length-aware constraints. While our reasoning is more efficient, it may omit minor details which can lead to slightly lower completeness scores. However, these omissions *do not necessarily affect segmentation accuracy*, which remains higher in our approach.

RST Facilitates a More Equitable Evaluation. To address this limitation, we propose **Reasoning Score per Token** (RST), which normalizes RScore by both model size and the number of generated tokens. This metric offers a more holistic assessment of reasoning quality relative to computational cost, enabling fairer comparisons between models with varying reasoning lengths.

E Further Discussions

In this section, we further discuss the limitations of our work, highlight directions for future research and consider the potential societal impact.

E.1 Limitation and Future Work

As the first attempt to enable efficient reasoning in reasoning segmentation, our method emphasizes simplicity and practicality, focusing on token-level control guided by task difficulty and uncertainty. However, our design still relies on coarse-grained difficulty scores and manually defined budget rules, which may limit adaptiveness in more complex scenarios. Additionally, the reasoning and segmentation stages are loosely coupled, and the applicability of our framework to broader multimodal tasks remains to be fully validated.

In the future, we will explore more precise and robust difficulty estimation by leveraging self-supervised signals in combination with human feedback, as well as developing finer-grained, learnable token allocation strategies that adapt to task-specific demands. Furthermore, integrating reasoning and segmentation into a joint optimization framework improves consistency and overall performance. Extending the proposed paradigm to other vision-language reasoning tasks such as visual question answering (VQA), visual commonsense reasoning, and video understanding further demonstrates its generalizability and practical value.

E.2 Potential Societal Impact

In this work, a reinforcement learning-based fine-tuning scheme is proposed for efficient reasoning in segmentation tasks, with potential applications in domains such as autonomous driving, robotics, and medical imaging. By enabling more efficient and interpretable visual reasoning, our method supports safer and more transparent decision-making in high-stakes scenarios. However, as with many vision-language models, our approach depends on large-scale pretrained models, which may carry biases from their training data. Moreover, automated reasoning systems could be misapplied in contexts such as surveillance or critical decision-making without adequate human oversight. We advocate for responsible deployment and encourage further research on fairness, robustness, and transparency to ensure beneficial societal impact.

Figure A: **Qualitative examples from ReasonSeg-DIFF.** Representative samples are shown with annotated difficulty scores and corresponding reference reasoning chains. For easy and medium cases, short reasoning chains are preferred to capture concise visual cues, whereas hard cases benefit from longer chains that reflect more elaborate reasoning over ambiguous or complex scenes. These examples illustrate the diverse reasoning requirements across difficulty levels and support more fine-grained evaluation.

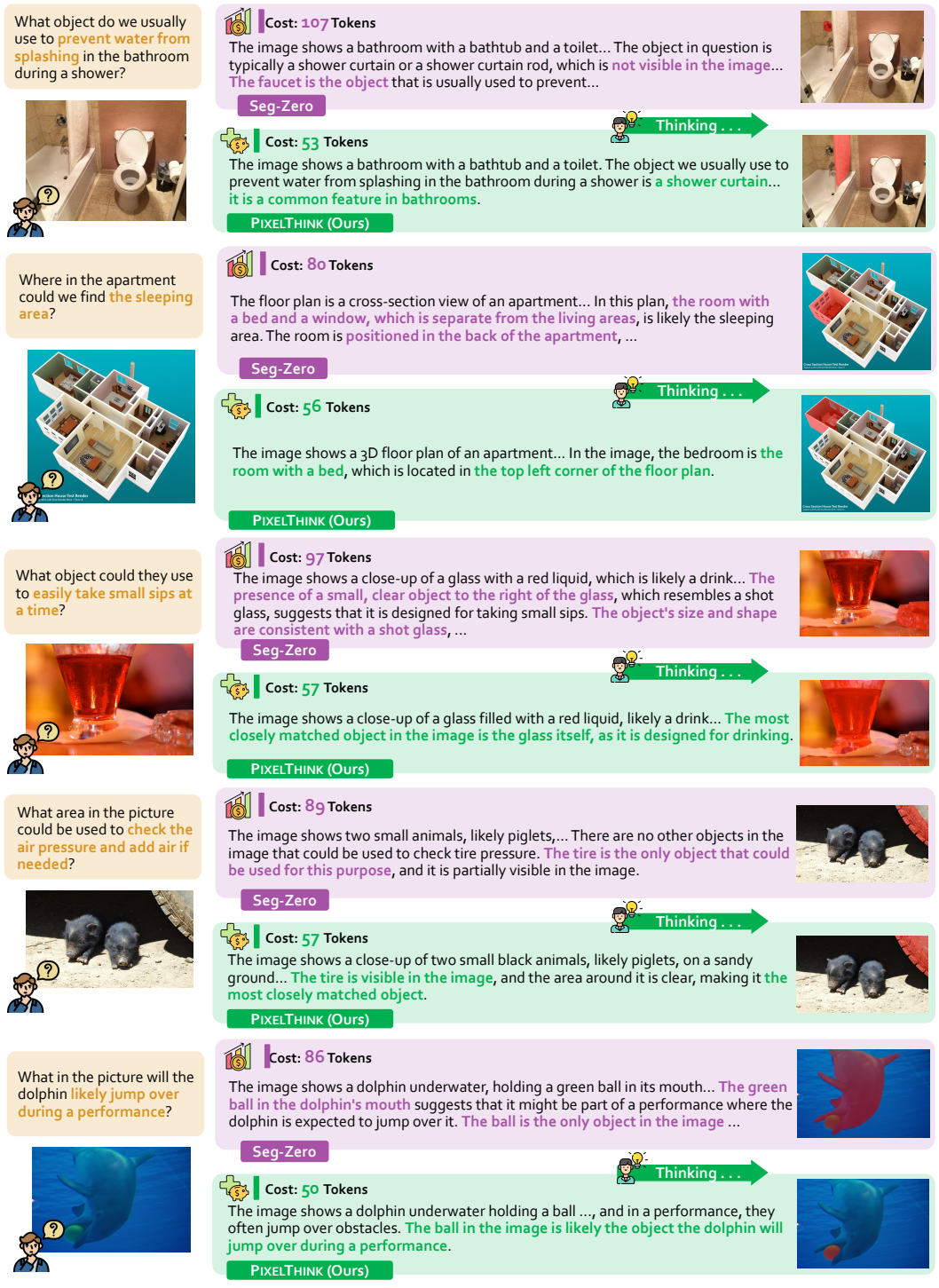

Figure B: Additional **qualitative comparisons** between Seg-Zero [18] and the proposed PIXELTHINK. Our method demonstrates consistent improvements in segmentation accuracy across diverse scenarios, accompanied by substantially shorter reasoning chains. These examples highlight the effectiveness of our efficient reasoning framework in mitigating *overthinking* while maintaining or improving segmentation quality.

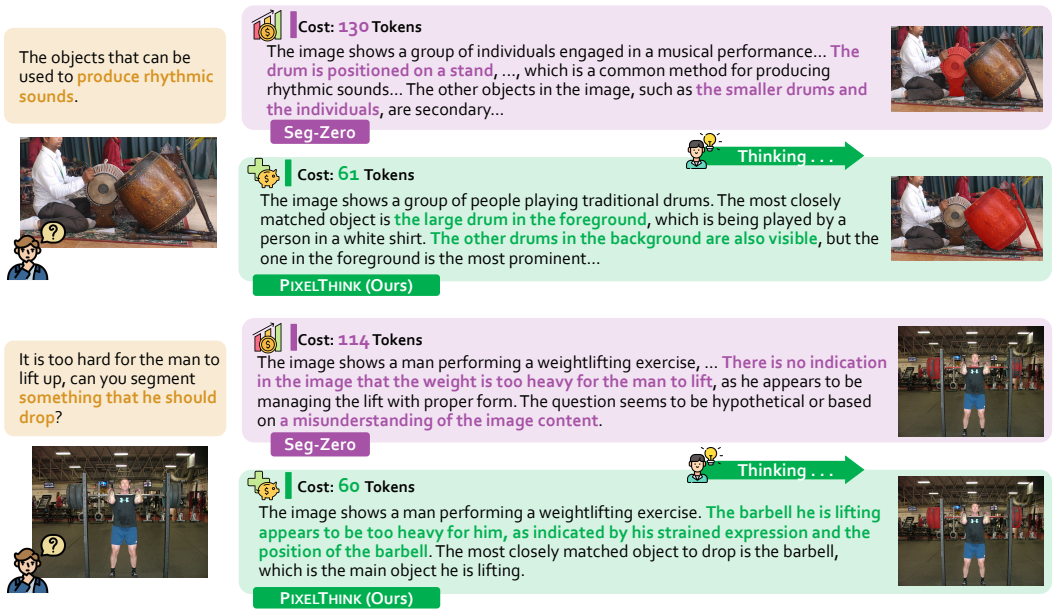

Figure C: **Failure cases** of Seg-Zero [18] and the proposed PIXELTHINK. The first example illustrates ambiguity in the referring expression, resulting in incomplete segmentation masks from both methods. The second example reveals a mismatch between reasoning and segmentation: Seg-Zero fails to identify the correct object in its reasoning, while PIXELTHINK correctly interprets the instruction but produces a partially inaccurate mask.

F License and Consent with Public Resources

In this section, we outline the details of responsible release and acknowledge the use of public resources that supported this work.

F.1 Responsible Release

Our work focuses on enhancing efficiency and controllability in reasoning segmentation using publicly available models and datasets. The proposed benchmark, ReasonSeg-DIFF, is constructed based on ReasonSeg [2], which includes human-annotated referring expressions and segmentation masks from natural scenes. We ensure that no private, sensitive, or copyrighted content is introduced during data processing. All models used are open-sourced under appropriate licenses, and we do not release any newly trained large-scale models that could pose potential misuse risks. Upon releasing our benchmark and codebase, we will include a clear license, data usage policy, and guidelines to discourage applications involving sensitive attributes, surveillance, or unauthorized identification.

F.2 Public Datasets

All experiments and the construction of our benchmark are conducted using the following publicly available datasets:

- ReasonSeg [2][4] ..... Apache License 2.0
- RefCOCO (+/g) [1][5] ..... Apache License 2.0
- MS COCO [80][6] ..... Other (specified in description)

F.3 Public Models and Implementation

We compare and validate the effectiveness of the proposed method using the following publicly available models and source codes:

- Qwen2.5-VL [14][7] ..... Apache License 2.0
- SAM2 [67][8] ..... Apache License 2.0
- Seg-Zero [18][9] ..... Apache License 2.0
- verl [81][10] ..... Apache License 2.0
- L1 [14][11] ..... MIT License

[4] https://github.com/dvlab-research/LISA.
[5] https://github.com/lichengunc/refer.
[6] https://cocodataset.org.
[7] https://github.com/QwenLM/Qwen2.5-VL.
[8] https://github.com/facebookresearch/sam2.
[9] https://github.com/dvlab-research/Seg-Zero.
[10] https://github.com/volcengine/verl.
[11] https://github.com/cmu-l3/l1.

References

[1] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

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

[3] Lanyun Zhu, Tianrun Chen, Qianxiong Xu, Xuanyi Liu, Deyi Ji, Haiyang Wu, De Wen Soh, and Jun Liu. Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation. *arXiv preprint arXiv:2504.00640*, 2025.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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, 2017.

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[7] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023.

[8] Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. *arXiv preprint arXiv:2503.00513*, 2025.

[9] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

[10] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*, 2025.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

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

[13] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

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

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

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

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

[19] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[20] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

[21] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025.

[22] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[23] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[24] R1-V Team. R1-V. https://github.com/Deep-Agent/R1-V?tab=readme-ov-file, 2025.

[25] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[26] Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *arXiv preprint arXiv:2504.10903*, 2025.

[27] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Hanjie Chen, Xia Hu, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

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

[29] Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

[30] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Eethods in Natural Language Processing*, pages 787–798, 2014.

[31] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *IEEE Transactions on Image Processing*, 2024.

[32] Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. In *European Conference on Computer Vision*, pages 108–126. Springer, 2024.

[33] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26573–26583, 2024.

[34] Shijia Huang, Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, and Liwei Wang. A mutual supervision framework for referring expression segmentation and generation. *International Journal of Computer Vision*, pages 1–16, 2025.

[35] Tao Zhang, Xiangtai Li, Zilong Huang, Yanwei Li, Weixian Lei, Xueqing Deng, Shihao Chen, Shunping Ji, , and Jiashi Feng. Pixel-sail: Single transformer for pixel-grounded understanding. *arXiv*, 2025.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

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

[39] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv*, 2025.

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[41] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

[42] Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint arXiv:2310.04959*, 2023.

[43] Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023.

[44] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.

[45] Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*, 2025.

[46] OpenAI. OpenAI o1. https://openai.com/o1/, 2024.

[47] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[48] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[49] Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling. *arXiv preprint arXiv:2502.06703*, 2025.

[50] Yuxin Zuo, Kaiyan Zhang, Shang Qu, Li Sheng, Xuekai Zhu, Biqing Qi, Youbang Sun, Ganqu Cui, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025.

[51] EvolvingLMMs Lab. Open R1 Multimodal. https://github.com/EvolvingLMMs-Lab/open-r1-multimodal, 2025.

[52] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.

[53] Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, and Bryan Hooi. Efficient inference for large reasoning models: A survey. *arXiv preprint arXiv:2503.23077*, 2025.

[54] Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*, 2025.

[55] Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*, 2024.

[56] Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. Cot-valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*, 2025.

[57] Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025.

[58] Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025.

[59] Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

[60] Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning. *arXiv preprint arXiv:2504.01296*, 2025.

[61] Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. Self-training elicits concise reasoning in large language models. *arXiv preprint arXiv:2502.20122*, 2025.

[62] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

[63] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024.

[64] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 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*, pages 19–35. Springer, 2024.

[65] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024.

[66] Zhuoyan Xu, Khoi Duc Nguyen, Preeti Mukherjee, Saurabh Bagchi, Somali Chaterji, Yingyu Liang, and Yin Li. Learning to inference adaptively for multimodal large language models. *arXiv preprint arXiv:2503.10905*, 2025.

[67] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[68] Heinrich Jiang and Maya Gupta. Minimum-margin active learning. *arXiv preprint arXiv:1906.00025*, 2019.

[69] Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*, 2024.

[70] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[71] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3505–3506, 2020.

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

[73] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023.

[74] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[75] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[76] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27124–27133, 2024.

[77] Ming Li, Shitian Zhao, Jike Zhong, Yuxiang Lai, and Kaipeng Zhang. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.16188*, 2025.

[78] Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.

[79] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, et al. Deepscaler: Surpassing o1-preview with a 1.5 b model by scaling rl. *Notion Blog*, 2025.

[80] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.

[81] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.