Decoupled Clip and Dynamic Sampling Policy Optimization for Food Reasoning Segmentation

Hikaru Tanabe
Department of Informatics
The University of Electro-Communications
Tokyo, Japan
tanabe-h@mm.inf.uec.ac.jp

Keiji Yanai
Department of Informatics
The University of Electro-Communications
Tokyo, Japan
yanai@cs.uec.ac.jp

Abstract

Food reasoning segmentation involves interpreting complex textual queries to identify corresponding regions in food images, which is a task that remains challenging for conventional supervised learning approaches. In this paper, we propose a pioneering reinforcement learning framework inspired by R1-Zero, tailored for food reasoning segmentation by integrating Multimodal Large Language Models (MLLMs) with a segmentation module. To address challenges such as entropy collapse and response inefficiency observed in GRPO-based training, we design a reward function that enforces both format compliance and segmentation accuracy, and explore the application of Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) to reasoning segmentation. Experiments on the FoodReasonSeg-Single benchmark show that our 3B RL-based model outperforms existing 7B SFT-based baselines, achieving state-of-the-art performance. These results underscore the promise of reinforcement learning in developing compact yet powerful models for visual reasoning in the food domain.

CCS Concepts

• **Information systems** → **Multimedia information systems**.

Keywords

Multimodal Large Language Models, Reasoning Segmentation, Reinforcement Learning, Food Image Recognition

ACM Reference Format:
Hikaru Tanabe and Keiji Yanai. 2025. Decoupled Clip and Dynamic Sampling Policy Optimization for Food Reasoning Segmentation. In *Proceedings of the 1st International Workshop on Multi-modal Food Computing (MMFood '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746264.3760490

1 Introduction

Reasoning segmentation [11] is the task of identifying image regions corresponding to complex textual queries that describe parts of the image. This task requires advanced visual reasoning grounded in broad world knowledge, and has recently been addressed leveraging Multimodal Large Language Models (MLLMs) by integrating their strong reasoning capabilities. In particular, solving this task in the food domain has great social impact for applications such as dietary monitoring assistants and embodied cooking robots, making the development of high-quality and lightweight models highly desirable.

MLLMs serve as a foundation for solving diverse visual tasks by integrating LLMs with extensive knowledge and reasoning capabilities, including those in the food domain, and vision encoders that offer multi-granular visual recognition [3]. Notably, milestone studies have explored methods to elicit complex reasoning abilities from MLLMs through supervised fine-tuning (SFT). Based on these strengths, there has been a growing interest in combining MLLMs with mask decoders to perform reasoning segmentation.

Meanwhile, recent research such as DeepSeek-R1 [8] has demonstrated that applying reinforcement learning strategies such as GRPO [22] to LLMs can significantly improve reasoning performance not only in-domain but also in out-of-domain (OOD) settings. In the context of MLLMs, similar improvements in vision-centric tasks have been reported by incorporating R1-style reinforcement learning strategies [10, 14, 16, 23, 27].

Reasoning segmentation also benefits from this trend [14, 15, 26]. However, RL-based training for this task faces challenges such as instability caused by entropy collapse during training, and increased computational cost and inference time due to longer responses. Moreover, no prior work has applied RL-based strategies to MLLMs in the food domain, and it is strongly anticipated that introducing such strategies can lead to high-quality solutions for the domain.

In this study, we introduce an R1-Zero [8] style RL-based training strategy for food reasoning segmentation and validate its effectiveness. To address the common issues in GRPO-based MLLM training such as entropy collapse and long responses, we design a reward function that includes response length constraints, and adopt DAPO [32], a promising optimization method that incorporates higher clipping and dynamic sampling strategy.

Experiments on a food reasoning segmentation dataset show that our 3B models trained with GRPO, with guidance from our reward design, significantly outperform 7B models trained with conventional SFT methods, which demonstrates the effectiveness of RL-based training strategies. Furthermore, we report that DAPO-based reinforcement learning achieves state-of-the-art performance while improving training efficiency. Our results suggest that RL-based food reasoning approaches represent a promising direction for future research. To the best of our knowledge, this is the first R1-Zero style reinforcement learning framework for food reasoning segmentation.

MMFood '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2046-8/2025/10
https://doi.org/10.1145/3746264.3760490

2 Related Work

2.1 Reasoning Segmentation

Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities across a wide range of vision-language tasks. Early models such as Flamingo [1], LLaVA [13], and InstructBLIP [6] established a foundation for integrating visual inputs into pretrained language models. Subsequent works have further improved vision-language alignment and instruction-following capabilities [3, 12, 17]. Other research has also explored the use of MLLMs for vision-centric tasks such as object detection [4, 31].

Reasoning segmentation is the task of segmenting regions corresponding to complex text queries provided by the user, requiring models to perform reasoning based on implicit knowledge [11]. Existing approaches, such as LISA [11], GLaMM [19], HyperSeg [28], and SAM4MLLM [5], integrate MLLMs with external mask decoder modules (e.g., SAM2 [20]) to leverage the advanced reasoning capabilities of MLLMs for understanding the implicit content of text queries and performing segmentation. However, their training strategies are limited to supervised fine-tuning (SFT), resulting in restricted performance, particularly in out-of-domain settings.

2.2 Food Image Recognition with MLLMs

In the food domain, FoodLMM [30] addresses the unique challenges of food recognition, including nutritional inference and food reasoning segmentation. By combining food-specific datasets, task-specific heads, and reasoning-guided fine-tuning, FoodLMM improves performance on tasks such as nutrient estimation and reasoning segmentation. Nevertheless, its reliance on SFT still limits its reasoning flexibility, especially for complex or novel queries.

Several studies have explored the use of MLLMs for nutrient estimation from meals [9, 24, 25, 29]. Accurate and flexible estimation of meal regions plays a crucial role in enhancing nutrient estimation [2, 25], highlighting the importance of high-quality reasoning segmentation for advancing downstream food-related tasks.

2.3 RL-based Multimodal Reasoning

Reinforcement learning (RL) has recently emerged as a promising tool to improve reasoning capabilities in LLMs and MLLMs. Typical methods such as PPO [21], RLHF [18], and ReST [7] laid the groundwork for RL in language models. More advanced techniques, including GRPO [22] and DAPO [32], have enabled efficient and stable policy optimization for structured reasoning.

GRPO introduced group-based advantage normalization together with PPO-style clipping and KL penalty policy updates, facilitating stable RL training for long-chain reasoning. However, it tends to produce overly verbose outputs, which can reduce efficiency and potentially degrade accuracy. DAPO improves upon GRPO by introducing decoupled clip, which relaxes constraints on probability increases for low-likelihood tokens, and dynamic sampling, which oversamples and filters out uniformly rewarded groups to preserve effective gradient signals. These techniques, along with other enhancements, enable more expressive yet controlled policy updates.

In the context of reasoning segmentation, Seg-Zero [14] introduced an RL-based approach where MLLM generates chain-of-thought outputs that guide the segmentation model via bounding boxes and point prompts. This decoupled architecture, optimized via reinforcement learning with verifiable rewards, achieves superior performance in zero-shot settings. In the context of multimodal reasoning, VisionReasoner [15] proposed a unified reward framework to train a single model on detection, segmentation, and counting tasks, achieving comprehensive performance in multiple tasks.

However, these studies do not sufficiently explore optimization methods beyond GRPO, and further evaluation in the food domain should be conducted. Building on these insights, we explore the application of DAPO to reasoning segmentation in the food domain. By applying DAPO together with a reward function designed to enforce format compliance and segmentation accuracy, our approach enhances training stability and efficiency while encouraging concise yet accurate reasoning in the food domain.

3 Method

3.1 Model Architecture

We adopt a decoupled architecture that pairs MLLM with frozen segmentation module, following Seg-Zero [14], as illustrated in Figure 1. Let $x \in \mathbb{R}^{H\times W\times 3}$ be an input image and q a textual query. The MLLM defines an auto-regressive policy over output tokens

$$\pi_\theta(S \mid x, q) = \prod_{t=1}^{T} \pi_\theta(s_t \mid x, q, s_{<t}), \tag{1}$$

where $S = [s_1, \ldots, s_T]$ is the generated sequence. We structure S into two semantically distinct parts enclosed by format tags: a reasoning chain r within `<think>...</think>` and a compact answer g within `<answer>...</answer>`. The answer g serializes up to $K \le K_{\max}$ segmentation references as a set of object prompts

$$\hat{\mathcal{P}} = \{(\hat{b}_i, \hat{p}_i)\}_{i=1}^{K} \tag{2}$$

The segmentation module $\mathcal{S}$ is an external mask decoder with a prompt encoder that accepts boxes and points. Given (b_i, p_i) and the image x, it returns a binary mask $m_i = \mathcal{S}(x; b_i, p_i) \in \{0,1\}^{H\times W}$.

This design yields clear responsibilities: the MLLM handles multimodal reasoning and structured geometric specification, while the mask decoder ensures accurate delineation under fixed visual priors.

In our training recipe, only the MLLM parameters θ are updated while $\mathcal{S}$ remains frozen. The MLLM thus learns to produce reasoning chains that are concise yet sufficient, and to emit geometrically well-formed (b_i, p_i) that are compatible with the frozen decoder's prompting space. The decoupling keeps the high-capacity perception of $\mathcal{S}$ intact and shifts credit assignment to the language policy, which we optimize with RL objectives described in Section 3.3.

3.2 Reward Design

Following VisionReasoner [15], the total reward is defined as the sum of four components: thinking-format, answer-format, non-repeat, and segmentation-accuracy rewards. Let $\mathcal{G}_{\text{bbox}} = \{b_j\}_{j=1}^{N}$ be the set of N ground-truth bounding boxes and $\mathcal{G}_{\text{pt}} = \{p_j\}_{j=1}^{N}$ the set of N ground-truth representative points.

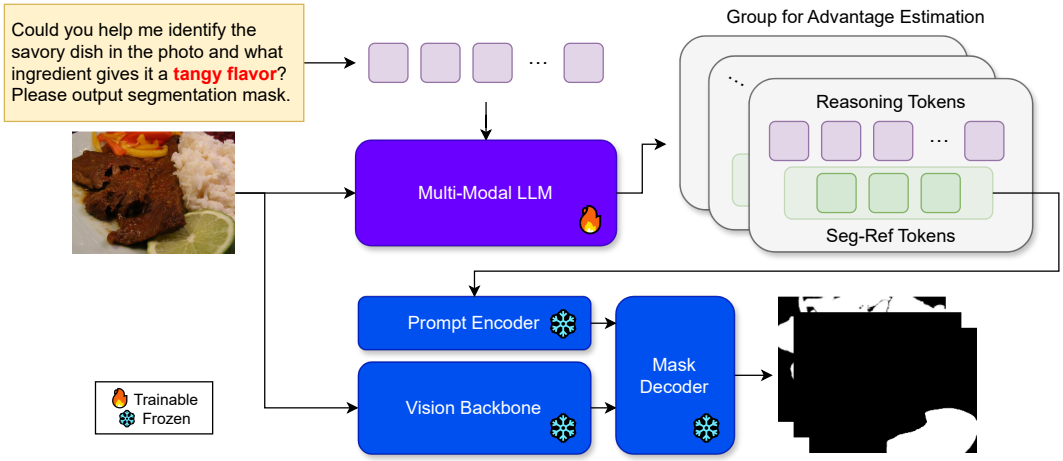

Figure 1: Overview of our model architecture. A trainable MLLM produces a reasoning chain and serialized segmentation references, which are consumed by a frozen segmentation module to output the final mask.

The thinking-format reward R_{tf} enforces that the output follows a strict reasoning–answer separation format, where the reasoning process is enclosed in `<think>...</think>` and the final answer in `<answer>...</answer>`.

The answer-format reward R_{af} ensures that all predicted bounding boxes and representative points have the correct dimensionality. Let $\{(\hat{b}_i, \hat{p}_i)\}_{i=1}^{K}$ be the K predicted object references, where $\hat{b}_i \in \mathbb{R}^4$ denotes a bounding box and $\hat{p}_i \in \mathbb{R}^2$ denotes a representative point:

$$R_{\text{af}} = \frac{1}{K} \sum_{i=1}^{K} \left(\mathbb{I}[\hat{b}_i \in \mathbb{R}^4] + \mathbb{I}[\hat{p}_i \in \mathbb{R}^2] \right), \tag{3}$$

where $\mathbb{I}[\cdot]$ returns 1 if the dimensionality condition is satisfied and 0 otherwise.

The non-repeat reward R_{nr} penalizes overly verbose reasoning by checking for repeated sentences in the reasoning part of the output and rewarding outputs in which the number of repetitions is below a predefined threshold D_{dup}.

The segmentation-accuracy reward R_{seg} evaluates spatial alignment between predicted and ground-truth references, combining region overlap and positional precision. The Hungarian algorithm is applied to match predictions to ground truth, yielding $\mathcal{M} = \{(\hat{b}_i, b_j)\}$ and $\mathcal{M}' = \{(\hat{p}_i, p_j)\}$:

$$R_{\text{IoU}}^{\text{bbox}} = \frac{1}{\max(N,K)} \sum_{(\hat{b}_i, b_j) \in \mathcal{M}} \mathbb{I}[\text{IoU}(\hat{b}_i, b_j) > \tau_{\text{IoU}}], \tag{4}$$

$$R_{\text{L1}}^{\text{bbox}} = \frac{1}{\max(N,K)} \sum_{(\hat{b}_i, b_j) \in \mathcal{M}} \mathbb{I}[\|\hat{b}_i - b_j\|_1 < \epsilon_{\text{bbox}}], \tag{5}$$

$$R_{\text{L2}}^{\text{pt}} = \frac{1}{\max(N,K)} \sum_{(\hat{p}_i, p_j) \in \mathcal{M}'} \mathbb{I}[\|\hat{p}_i - p_j\|_2 < \epsilon_{\text{pt}} \ \wedge \ \hat{p}_i \in \hat{b}_i]. \tag{6}$$

The segmentation-accuracy reward is then formulated as

$$R_{\text{seg}} = R_{\text{IoU}}^{\text{bbox}} + R_{\text{L1}}^{\text{bbox}} + R_{\text{L2}}^{\text{pt}}. \tag{7}$$

Finally, the total reward is

$$R_{\text{total}} = R_{\text{tf}} + R_{\text{af}} + R_{\text{nr}} + R_{\text{seg}}. \tag{8}$$

3.3 Policy Optimization Strategy

3.3.1 Group Relative Policy Optimization. Group Relative Policy Optimization (GRPO) [22] is introduced to simplify policy training by eliminating the need for a value function. Instead, it estimates the advantage of each sample based on its relative position within a group of responses. Specifically, for a given question–answer pair (q, a) from the data distribution $\mathcal{D}$, the behavior policy $\pi_{\theta_{\text{old}}}$ generates a set of G responses $\{o_i\}_{i=1}^{G}$. The advantage of each response is computed by standardizing its total reward against the group statistics as follows:

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}, \tag{9}$$

where R_i is the sequence-level reward for the i-th response o_i.

GRPO adopts a clipped surrogate loss with an explicit KL divergence penalty to ensure stable updates:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\ \{o_i\}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\big(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\big)\hat{A}_{i,t} \Big) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Bigg], \tag{10}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}.$$

3.3.2 Decoupled Clip and Dynamic Sampling Policy Optimization. To address the limitations of GRPO, we adopt Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) [32]. This method

enhances sample efficiency and optimization stability by modifying the clipping mechanism and filtering strategy.

DAPO decouples the clipping range into asymmetric bounds, where lower-probability tokens receive a wider allowance to increase, thereby encouraging exploration. Furthermore, it introduces a dynamic sampling constraint, which selects only partially correct samples (i.e., those whose reward is strictly between 0 and 1) to compute the gradient. These improvements enable more expressive updates while suppressing trivial or degenerate policies.

Let $T = \sum_{i=1}^{G} |o_i|$ denote the total number of tokens in the sampled group. The training objective of DAPO is defined as follows:

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\ \{o_i\}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{T} \sum_{i=1}^{G} \sum_{t=1}^{|o_i|} \min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\big(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\big)\hat{A}_{i,t} \right) \right]$$
$$\text{s.t.} \quad 0 < \left| \{ o_i \mid \texttt{is\_equivalent}(a, o_i) \} \right| < G, \tag{11}$$

where $\varepsilon_{\text{low}}, \varepsilon_{\text{high}} > 0$ define the asymmetric clipping range.

DAPO facilitates faster convergence and more diverse responses without compromising output quality. In large-scale experiments, it has shown superior performance in reasoning tasks with fewer training steps, making it a potentially effective strategy for reinforcement learning with MLLMs.

4 Experiments

4.1 Implementation Details

We employed Qwen2.5-VL 3B [3] as the MLLM, and SAM2 [20] as the segmentation module. During training, only the MLLM was made trainable, while the segmentation module was kept frozen. We employed gradient clipping with a maximum norm of 1.0 to stabilize training.

The threshold values for the reward function were set to $\tau_{\text{IoU}} = 0.5$, $\epsilon_{\text{bbox}} = 10$, and $\epsilon_{\text{pt}} = 30$. The clipping bound for GRPO was set to $\varepsilon = 2.0$, while the lower and upper clipping bounds for DAPO were set to $\varepsilon_{\text{low}} = 2.0$ and $\varepsilon_{\text{high}} = 2.8$, respectively. For DAPO, we utilized token-level policy gradient loss with an overlong penalty of 0.1 to discourage excessively long responses that might be truncated. During the rollout phase, we generated $G = 8$ responses per question using temperature sampling with $\tau = 1.0$. We report the results after training GRPO and DAPO for 100 and 50 steps, respectively. All experiments were conducted with 4 × A6000 GPUs.

4.2 Dataset

To simplify RL-based training and evaluation, we constructed a modified dataset, FoodReasonSeg-Single, based on FoodReasonSeg dataset [30].

For training, we utilized the multi-turn question–response mask annotations associated with FoodReasonSeg. For each question, we merged all response masks across turns into a single composite mask to serve as the ground truth. This design choice was informed by our experimental results, which indicated that merging multiple responses leads to higher performance compared to training on a single-turn mask. By adopting this merged-mask setup, the model is encouraged to explore a broader range of regions while the reward function and training strategy serve to limit unnecessary outputs.

However, this approach tends to cause the model to predict multiple food regions as the number of steps increases. To mitigate this issue, we incorporate early stopping during training. Future work should explore approaches that balance broad exploration with long-term training.

During evaluation, we assess the quality of mask prediction by comparing the predicted mask to a single corresponding response mask for each question, treating it as the ground truth.

Following prior studies [11, 14, 30], we evaluate segmentation quality using both gIoU and cIoU. The gIoU is defined as the average of per-image IoUs, while cIoU measures the ratio of cumulative intersection over cumulative union across the dataset.

Table 1: Evaluation results on the FoodReasonSeg-Single benchmark. † indicates results reproduced using our implemented evaluation script using the weights and prompts released in the repository of FoodLMM. * indicates that response samples where no mask was output were excluded from the evaluation metrics

Method	gIoU	cIoU
Qwen2.5-VL 3B + SAM2	0.07	0.08
LISA 7B [11] †	0.19	0.19
FoodLMM-Chat 7B [30] †*	0.25	0.20
Ours (3B, GRPO)	0.46	0.41
Ours (3B, DAPO)	**0.51**	**0.45**

4.3 Results

We evaluate our method on the FoodReasonSeg-Single benchmark.

Table 1 shows the quantitative results in terms of gIoU and cIoU. Our 3B model with GRPO significantly outperforms existing SFT-based baselines, including LISA 7B [11] and FoodLMM-Chat 7B [30], with a 0.21 gain in both gIoU and cIoU compared to FoodLMM. Furthermore, applying DAPO yields additional improvements, achieving 0.51 gIoU and 0.45 cIoU, demonstrating the effectiveness of the policy optimization strategy.

Figure 2 presents the response examples of reasoning segmentation, covering prior work as well as our models.

FoodLMM directly conditions the mask decoder on LLM features at the time of seg-token generation, which differs from our approach that uses segmentation references. Inspecting the FoodLMM response text shows that it has an object hallucination of a pumpkin in the image. As a result, the mask decoder is conditioned on a nonexistent region, leading to extract no region masks.

For the baseline, Qwen2.5-VL fails to output a parseable response format. As a result, the mask decoder cannot be conditioned by segmentation references, and the model fails to produce a mask.

By contrast, our models trained with GRPO and DAPO emit valid segmentation references and successfully output masks. Examining the respective thinking texts reveals that an object hallucination of sweet potatoes causes the segmentation reference to focus on the vegetables under GRPO. Consequently, the predicted mask also focuses on the vegetables, failing to target the steak, which

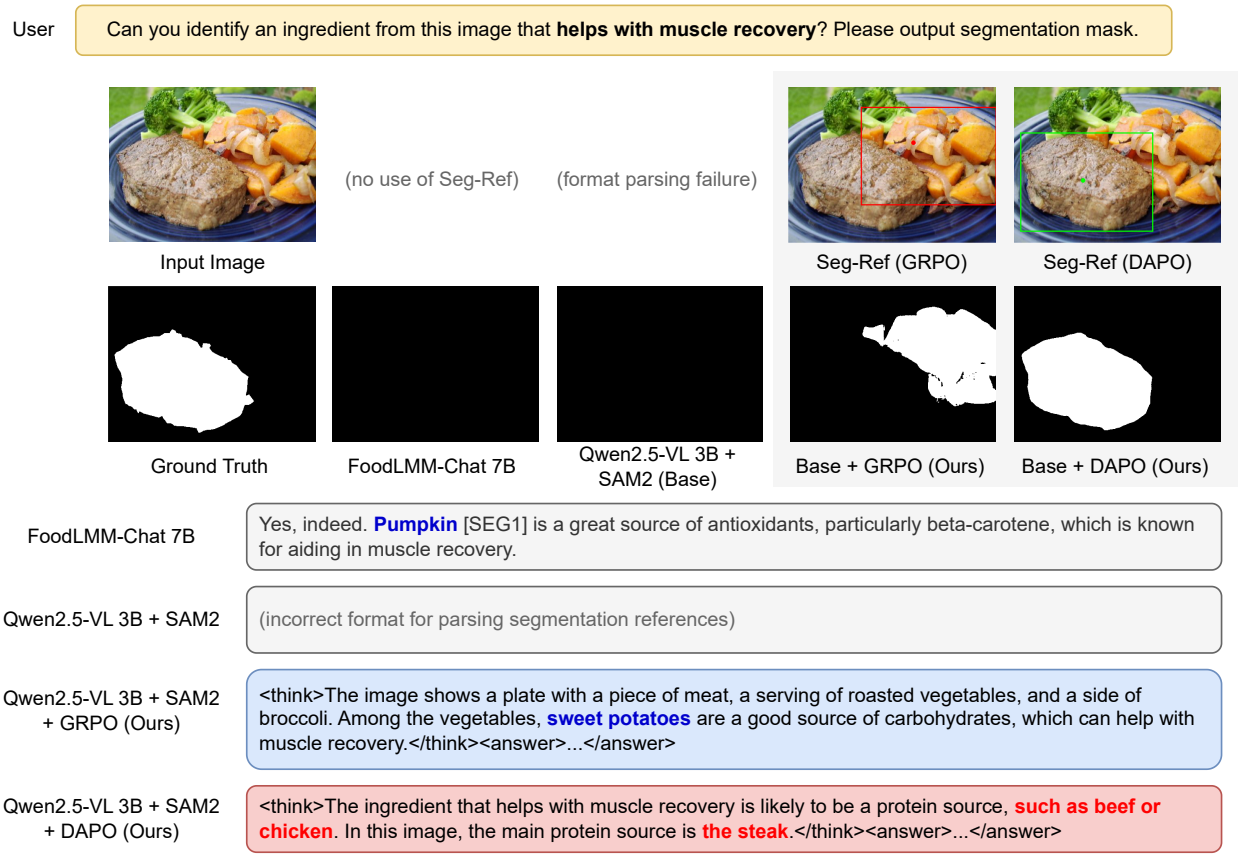

Figure 2: Reasoning segmentation results. Our models trained with GRPO and DAPO generate segmentation references, with DAPO focusing on the steak region close to the ground truth.

corresponds to more appropriate region. In contrast, the model's thinking process correctly focuses on the steak, producing a mask close to the ground truth.

These results demonstrate that our framework substantially mitigates object hallucinations and yields a more effective reasoning process compared with existing methods. Moreover, our architecture is modular and readily adaptable to diverse tasks, and the responses present the model's explicit reasoning process in a clearer and more inspectable form.

5 Discussion and Future Work

Future work includes exploring training strategies that balance broad exploration with stable long-term optimization. Our experiments revealed that training with multi-object masks sometimes led the RL model to extract multiple regions when the number of training steps increased. Conversely, using single-object ground truth limited the upper bound of achievable scores. For this reason, we adopted multi-object masks in this study. A promising future direction is to investigate trade-offs between broad exploration—such as training with multi-object masks to boost cold-start performance—and stable optimization that avoids over-segmentation.

In addition, incorporating staged training from the SFT phase could better leverage the strengths of R1-style training. Verifiable rewards for region mask quality are also promising fields for further improving segmentation accuracy. Additionally, more comprehensive evaluations on out-of-distribution data are required to validate the robustness of RL-based food reasoning segmentation.

6 Conclusion

We introduced an R1-Zero–style reinforcement learning framework for food reasoning segmentation. By designing the reward function and applying RL-based training strategy such as GRPO and DAPO, our method achieves superior performance on food reasoning segmentation. Notably, our 3B model outperforms existing 7B SFT-based baselines, highlighting the effectiveness of our RL-based training for multimodal reasoning. These findings highlight reinforcement learning as a promising foundation for developing high-quality models in the food reasoning domain.

Acknowledgments

This work was supported by JSPS KAKENHI 22H00548 and JST CRONOS JPMJCS24K4.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198* (2022).

[2] Yoshikazu Ando, Takumi Ege, Jaehyeong Cho, and Keiji Yanai. 2019. DepthCalorieCam: A Mobile Application for Volume-Based FoodCalorie Estimation Using Depth Cameras. In *Proc. of the 5th International Workshop on Multimedia Assisted Dietary Management*. 76–81.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).

[4] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024. LION: Empowering Multimodal Large Language Model with Dual-Level Visual Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. 2024. SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36. 49250–49267.

[7] Caglar Gülcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced Self-Training (ReST) for Language Modeling. *arXiv preprint arXiv:2308.08998* (2023).

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

[9] Pengkun Jiao, Xinlan Wu, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. 2024. RoDE: Linear Rectified Mixture of Diverse Experts for Food Large Multi-Modal Models. *arXiv preprint arXiv:2407.12730* (2024).

[10] Jung Uk Kim, Sungjune Park, and Yong Man Ro. 2025. MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models via Reinforcement Learning. *arXiv preprint arXiv:2502.19634* (2025).

[11] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326* (2024).

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[14] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. 2025. Seg-Zero: Reasoning-Chain Guided Segmentation via Cognitive Reinforcement. *arXiv preprint arXiv:2503.06520* (2025).

[15] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. 2025. VisionReasoner: Unified Visual Perception and Reasoning via Reinforcement Learning. *arXiv preprint arXiv:2505.12081* (2025).

[16] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-RFT: Visual Reinforcement Fine-Tuning. *arXiv preprint arXiv:2503.01785* (2025).

[17] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. 2024. Mm1: methods, analysis and insights from multimodal llm pre-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 304–323.

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 27730–27744.

[19] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2024. GLaMM: Pixel Grounding Large Multimodal Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[22] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Yuqian Li, Yifan Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300* (2024).

[23] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. *arXiv preprint arXiv:2504.07615* (2025).

[24] Hikaru Tanabe and Keiji Yanai. 2025. CalorieLLaVA: Image-Based Calorie Estimation with Multimodal Large Language Models. In *Pattern Recognition and Image Analysis (ICPR International Workshops, MADiMa 2024, Revised Selected Papers) (Lecture Notes in Computer Science, Vol. 14435)*. Springer, 62–74.

[25] Hikaru Tanabe and Keiji Yanai. 2025. CalorieVoL: Integrating Volumetric Context into Multimodal Large Language Models for Image-based Calorie Estimation. In *Proceedings of the 30th International Conference on Multimedia Modeling (MMM)*. 416–421.

[26] Song Wang, Gongfan Fang, Lingdong Kong, Xiangtai Li, Jianyun Xu, Sheng Yang, Qiang Li, Jianke Zhu, and Xinchao Wang. 2025. PixelThink: Towards Efficient Chain-of-Pixel Reasoning. *arXiv preprint arXiv:2505.23727* (2025).

[27] Zhiqiang Wang, Pengbin Feng, Yanbin Lin, Shuzhang Cai, Zongao Bian, Jinghua Yan, and Xingquan Zhu. 2025. CrowdVLM-R1: Expanding R1 Ability to Vision Language Model for Crowd Counting using Fuzzy Group Relative Policy Reward. *arXiv preprint arXiv:2504.03724* (2025).

[28] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. 2024. HyperSeg: Towards Universal Visual Segmentation with Large Language Model. *arXiv preprint arXiv:2411.17606* (2024).

[29] Dongyu Yao, Keling Yao, Junhong Zhou, and Yinghao Zhang. 2024. CaLoRAify: Calorie Estimation with Visual-Text Pairing and LoRA-Driven Visual Language Models. *arXiv preprint arXiv:2412.09936* (2024).

[30] Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. 2024. FoodLMM: A Versatile Food Assistant using Large Multi-modal Model. *arXiv preprint arXiv:2312.14991* (2024).

[31] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. *arXiv preprint arXiv:2310.07704* (2023).

[32] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv preprint arXiv:2503.14476* (2025).